

A New View of Seasonal Forecast Skill: Bounding Boxes from the DEMETER Ensemble Forecasts

By Antje Weisheimer^{1,2*}, Leonard A. Smith^{2,3} and Kevin Judd^{4,2}

¹*Meteorological Institute, Free University Berlin, C.-H.-Becker-Weg 6-10, D-12165 Berlin, Germany;*

²*Centre for the Analysis of Time Series, London School of Economics and Political Sciences (LSE), United Kingdom;*

³*Oxford Centre for Industrial and Applied Mathematics, Oxford University, United Kingdom;*

⁴*Centre for Applied Dynamics and Optimization, University of Western Australia, Perth*

(Manuscript received 31 March 2004)

ABSTRACT

Insight into the likely weather several months in advance would be of great economic and societal value. The DEMETER project has made coordinated multi-model, multi-initial-condition simulations of the global weather as observed over the last 40 years; transforming these model simulations into forecasts is non-trivial. One approach is to extract merely a single forecast (e.g., best-first-guess) designed to minimize some measure of forecast error. A second approach would be to construct a full probability forecast. This paper explores a third option, namely to see how often this collection of simulations can be said to *capture* the target value, in the sense that the target lies within the bounding box of the forecasts. The DEMETER forecast system is shown to often capture the 2m temperature target in this sense over continental areas at lead times up to six months. The target is captured over 95% of the time at over a third of the grid points and maintain a bounding box range less than that of the local climatology. Such information is of immediate value from a user's perspective. Implications for the minimum ensemble size as well as open foundational issues in translating a set of multi-model multi-initial-condition simulations into a forecast are discussed, in particular those involving "bias correction" are considered.

1 Introduction

Predictions of weather and climate using sophisticated physical models of the atmosphere and ocean are intrinsically uncertain. The underlying fluid dynamics is fundamentally nonlinear and the models are likely to be chaotic. For any given model, uncertainty in initial conditions severely limits the forecast skill in terms of a single best-first-guess forecast. For every collection of models, model inadequacy due to imperfections in each model formulation restricts the ability to forecast the evolution of a probability distribution. Ensemble forecasting using different model formulations, together with different consistent initial conditions provides a pragmatic approach to sampling, if imperfectly, these uncertainties. This may provide a clearly superior alternative to the traditional single best-first-guess forecast under a more complex model of greater, but remotely comparable, skill (Palmer, 2000).

This study demonstrates a new approach to assess the skill of ensemble simulations. Verifying probabilistic forecasts made with imperfect models is a highly ambitious and complex task; we address somewhat simpler questions: Are the distributions obtained from ensemble forecasts

likely to capture the target? and What properties define a good/useful ensemble? The bounding boxes of an ensemble will be used to define a region of model-state space within which the future is likely to fall; the probability of capture can be determined empirically. What are reliable spatial and temporal scales at which the ensemble is able to capture the target with a high probability?

How does examining the bounding box of an ensemble compare to using one of the plethora of other skill measures (e.g., Jolliffe and Stephenson, 2003) already employed in the evaluation of ensemble forecasts? Traditional measures tend to fall into two broad types, the first (like the root-mean-square error, mean-absolute-error, anomaly correlations and so on) evaluate a point measure from the forecast such as the ensemble mean, while the second (like rank histograms, probabilistic Brier scores, ignorance, and so on) interpret the ensemble as a probability forecast. The bounding box approaches take an intermediate view which differs importantly from each of these extremes. The first type of score ignores all of the distributional information in the ensemble, whereas the bounding box retains some information from the distribution while being insensitive to any mean value taken over the distribution. The second type of score aims to evaluate whether or not the verification is another drawn from the same distribution as the ensemble members, and thus will penalize an ensemble forecast that includes excep-

* Corresponding author.

E-mail: Antje.Weisheimer@met.fu-berlin.de

tionally good member forecasts depending on the relative frequency of other members of the ensemble. The bounding box asks for less than an accurate probability forecast. As will become clear below, the bounding box approaches are likely to have power in situations where the forecast skill may appear low as judged both types of traditional measures, yet the forecast still has useful information in terms of decision making.

There is currently no coherent interpretation of the multi-model multi-initial-condition ensemble as some reasonable sample of a relevant unknown probability density distribution. Rather than consider ensemble members as random draws from this hypothetical perfect ensemble distribution and then interpret the forecasts as draws from the future distribution from which the target is also drawn, we will consider the target as merely a target. We can then ask the question of whether or not we can expect the bounding box to capture the target within the ensemble, given certain statistical aspects of our ensemble forecast. Given an operational ensemble system, we can determine the frequency of capture empirically and provide this information to the users of the forecast. Alternatively, we can compute the expected rates of capture analytically under various assumptions regarding model error. Contrasting these calculations with the empirical observations then provides information for both modellers and forecasters. It is important, of course, to distinguish clearly between discussions of observed capture frequency of actual observations and those of the capture probability when properties both of the model and of the true distribution are assumed known. Furthermore, the bounding box approach gives new insight into the controversial questions of how large an ensemble need be and how model simulations might best be “post-processed” into forecasts. Whenever one attempts to interpret model simulations as forecasts, issues of “post-processing” arise. Given a very large number of single-model single-initial-condition forecasts, issues like “bias correction” for each grid point are straightforward; in the ensemble scenario this is no longer the case.

A multi-model ensemble-based system for seasonal-to-inter-annual prediction with a broad range of applications across disciplines has been developed within the joint European project DEMETER. The DEMETER system (see Palmer et al., 2004, and references therein) provides an ideal framework (i) for testing the bounding box methodology with state-of-the-art coupled forecast models, and (ii) for assessing the bounding box skill of the ensemble forecasts in the above mentioned sense. In this paper we discuss some of the results of this application. The bounding box idea is introduced in Section 2, and the DEMETER seasonal multi-model ensemble forecasts are examined in this light in Section 3. Section 4 discusses broader implications of these results with an eye towards future forecast systems, while our results are briefly summarised in the Section 5.

2 Bounding Boxes and the constraint for a meaningful ensemble

We consider the Bounding Box (BB) (Smith, 2000) of an ensemble as a prediction interval (Chatfield, 2001), i.e. a forecast range in which the verifying observation (or some

given target value) is expected to fall with some probability. Observing that the ensemble captures the target with a high probability quantifies a new distinct quality in the ensemble forecast system. We suggest that a BB which includes the target state with a high probability and a range¹ smaller than climatology is an indication for a reasonably good ensemble.

The BB of an ensemble is given by the minimum and maximum value of each ensemble component. Thus an interval is defined for each variable for each lead time at each model grid point. The criterion for an ensemble to capture the target state is simply that the target value lies between the minimum and maximum values of the corresponding ensemble component. This straight forward concept has the advantage that the BB is easy to compute and defined for any ensemble size and dimension: a two member ensemble will almost certainly define a nontrivial bounding box in a 10^7 dimensional space. Alternative measures of volume, like a convex hull, require prohibitively large ensembles, while the ensemble sizes needed for estimates of joint probabilities are truly astronomical.

Ensemble forecasts are often interpreted in terms of probability distributions (e.g., Robert and Casella, 1999; Leith, 1974; Anderson, 1996; Ehrendorfer, 1997; Palmer, 2000; Judd and Smith, 2001) and within the perfect model scenario this makes perfect sense. In that case, one would ideally draw ensemble members from what is effectively the same distribution that the target is drawn from. Given only a collection of imperfect models, it is not clear how to conceive of such a target probability distribution (Smith, 2000; Judd and Smith, 2004). In this paper, we take a different approach. The target is truly a target; we make no mention of a distribution from which it is drawn. Rather we consider the probability that forecasts fall above or below this target value.

Let E be the ensemble and assume that the ensemble members are independent and drawn from some distribution. We take the target x^* to be a point in the same space as the ensemble members, but make no assumptions about the relationship, for example, dependence or correlation, between target and ensemble members. We denote the probability that the component $y \in E$ is smaller than x^* as

$$p = \Pr(y < x^* \mid y \in E) \quad (1)$$

Then the probability that the component y is greater or equal x^* is, of course, $1 - p$.

What is the probability that the 1-dimension BB of an ensemble E includes the target state x^* ? The only manner by which the BB might *not* include the target is if the ensemble distribution was *entirely* left or right of x^* . Given an ensemble of n members, the probability that all ensemble members are smaller than x^* is p^n . Similarly, the probability for *all* ensemble members being larger or equal x^* is $(1-p)^n$. The total probability that a 1-dimensional ensemble does capture the target, P_{BB} , is then simply:

$$P_{BB}(n, d = 1) = 1 - p^n - (1 - p)^n. \quad (2)$$

This argument is easily generalised to a higher dimen-

¹ In this context, “range” denotes a measure of the volume of the bounding box. We return to this point in Section 3.4.

sional system, where each of the d coordinates is independent. Let p_i be p for the i^{th} coordinate. If

$$p_i = P(y_i < x_i^* \mid y \in E), \quad (3)$$

then the probability that a d -dimensional BB includes the target is

$$P_{BB} = \prod_{i=1}^d (1 - p_i^n - (1 - p_i)^n). \quad (4)$$

In case that the coordinates are not independent from each other, P_{BB} for a given value of n becomes even higher, since the case of independent dimensions represents a lower limit for the general conditions of dependency.

Hence, P_{BB} is a function of several variables: the one-sided probability p_i , the dimension d and the ensemble size n . Before looking in detail at the bounding boxes from the DEMETER ensemble forecasts in Section 3, we will first discuss and interpret implications of equation (4). After considering the interpretation of a single ensemble forecast, we discuss as examples the following three types of ensemble distributions: (i) the general case of an arbitrary skewed ensemble; (ii) the idealized case of a Gaussian forecast distribution; and (iii) the special case where the target is exactly at the median of the distribution.

2.1 The case of a single ensemble forecast

To clarify the problem, first consider a single day and specific lead time in the one dimensional case ($d = 1$). An ensemble prediction system and target will correspond to a single value of p_i . If p_i were known, then plugging p_i into equation (4) for a variety of values of n would provide the capture probability P_{BB} as a function of n . And thus a value of n could be found to meet any desired threshold of P_{BB} (assuming p_i is not equal to zero or to one).

In practice, of course, p_i will vary from day to day, and it is the capture rate of the ensemble system that is of interest. Inferring the capture rate as a function of n requires knowledge of the distribution of p_i . The next three subsections consider various simple illustrative examples in order to provide some insight into the sort of ensemble sizes that might be required. They are found to be operationally accessible.

2.2 Arbitrary distribution

A schematic plot of the probability density distribution (pdf) in terms of the histogram of an arbitrary 1-dimensional ensemble is given in Fig. 1a. We consider a 63-member ensemble, as this is the size of the DEMETER ensemble system discussed later. The area under the histogram curve left to the target quantifies the one-sided probability p_i . Figure 1b shows the corresponding cumulative density distribution (cdf).

Ideally, an ensemble system can be deployed with the minimum size needed to ensure the likely capture of the targets of interest. The framework above enables some insight into how the resources might be distributed. Assuming a known distribution for p_i exists that is a generic characteristic of the underlying ensemble distribution, equation (4) can be used to extrapolate the minimum ensemble size required

for any specified probability of capture. Such information on the minimal size of the ensemble would be useful in planning resource distribution within future ensemble forecast systems, as discussed in Section 4.1.

Suppose we aim to capture a target with a given probability P_{BB} using n ensemble members, equation (4) provides an estimate of the corresponding p_i -value range. For $P_{BB} \geq 0.95$ and $n = 63$, $d = 1$, this results in $0.0465 < p_i < 0.9535$. The nonlinear relationship between ensemble size, the one-sided probability and the probability to capture is displayed graphically for the 1-dimensional case in Fig. 1c. For instance, in order to capture the target given a one-sided probability $p_i = 0.01$, an ensemble of 300 members would be needed to capture with a probability of 95%.

In practice, however, p_i will vary from day to day. Assuming realistic values of n and d (specifically 63 and 10^7), and requiring a capture rate of 0.95, equation (4) implies that any value of p_i in the range 0.262 to 0.738 would have the desired capture rate. That implies that any combination of day to day variations of p_i in this range would also have a capture rate at least this large. Decreasing the n to 30 would narrow down this range of p_i to 0.472...0.528.

2.3 Gaussian distribution

Consider a Gaussian distribution $N(x_i^* - \mu_i, \sigma_i)$, as shown schematically in Fig. 2a and Fig. 2b, again for 63 ensemble members. Here μ_i and σ_i denote the mean and the standard deviation of the distribution for the i -th dimension. Then the normalised offset z_i is defined by $z_i = (x_i^* - \mu_i)/\sigma_i$. In this case the one-sided probability p_i can be replaced by the corresponding cdf(z_i), the standard normal distribution $\Phi(z_i)$, see Fig. 2b. Equation (4) reads:

$$P_{BB}(\text{Gaussian}) = \prod_{i=1}^d (1 - \Phi(z_i)^n - \Phi(-z_i)^n). \quad (5)$$

The range of values of p_i such that P_{BB} is at least 0.95 (again based on 63 ensemble members) thus transforms to $|z_i| < 1.68$. This means that, provided the forecast distribution is Gaussian, any 63-member ensemble will theoretically be able to capture the target with at least 95% probability if the absolute offset (that is, x_i) is not larger than 1.68 standard deviations.

The probability P_{BB} as a function of z and n is shown in Fig. 2c for 1-dimensional Gaussian ensembles. It demonstrates the enormous effect an offset has on the minimum number of ensemble members needed to capture the target. For instance, a normalized offset of 2.5 standard deviations ($z = 2.5$) in one dimension increases the number of n , for which P_{BB} is larger than 95%, by almost a factor of 100 compared with the zero offset situation. This suggests that issues of ‘‘bias correction’’ in translating model simulations to forecasts may play a significant role in the probability of bounding. As with any skill statistic, it is critical to distinguish between (i) evaluating an ensemble forecast with BB statistics and (ii) optimising the BB statistics of an ensemble forecast.

2.4 Centred distribution

Let us now assume that the distribution is centred on the target, specifically that the target x_i^* divides the (not necessarily symmetric) distribution into two equal parts with respect to p_i , that is $p_i = 1 - p_i = 0.5$. This is equivalent to x_i^* being the median of the ensemble distribution in the i -th dimension and $p_i = 0.5$, see Fig. 3. In this case equation (4) reads as

$$P_{BB} = \prod_{i=1}^d \left(1 - \left(\frac{1}{2}\right)^n - \left(\frac{1}{2}\right)^n\right) \quad (6)$$

$$= \prod_{i=1}^d \left(1 - \frac{2}{2^n}\right) \quad (7)$$

$$= \left(1 - \frac{1}{2^{n-1}}\right)^d \quad (8)$$

Again, the relation between P_{BB} and n can be used to explicitly estimate the minimum ensemble size needed to ensure a certain capturing probability. Suppose we claim to include the target in a 1-dimensional centred bounding box with a probability of 0.95. This immediately leads to $n = \log_2 40 = 5.3$. Thus one would need at least 6 ensemble members to capture the target with a 95% probability.

The probability P_{BB} of centred distributions is shown in Fig. 3c for different dimensions and ensemble sizes. Even for very high dimensional systems, like general circulation models, only a couple dozen of ensemble members are required to capture the target with a very high probability. For instance, in 10^7 dimensions, as few as 30 ensemble members are sufficient to almost always bound the target (compare with the general case in Section 2.2). This is not surprising because when the ensemble is drawn from a centred distribution, then one expects nearly equal numbers of ensemble members either side of the target and therefore it is unlikely that *all* ensemble members will be to the *same* side of the target. Note (i) the narrow transition zone between very high and very low probabilities in Fig. 3c, and, (ii) that P_{BB} does *not* depend on the variance of the ensemble distribution.

3 Application to a Dynamical Seasonal Multi-Model Ensemble Forecast System

3.1 The DEMETER Project

The DEMETER (Development of a European Multi-Model Ensemble Prediction System for Seasonal to Inter-annual Predictions, see also other contributions in this volume) project was designed to create a multi-model ensemble system for studying seasonal forecasts. A special focus is put on applications across disciplines ranging from down-scaling crop-yield modelling to Malaria predictions. As a result of the project an extensive database of hindcasts from 1958 to 2001 with common archiving and common diagnostic software has been made available to the public (see <http://www.ecmwf.int/research/demeter/data>).

The system comprises seven state-of-the-art global coupled ocean-atmosphere circulation models whose atmosphere and ocean components have been developed quasi-independently at different European research institutes. For all but one model the atmospheric and land-surface ini-

tial conditions were taken from the ECMWF 40-year Re-Analysis (ERA-40) data set, while the oceanic initial conditions were obtained from ocean model runs forced by ERA-40 fluxes. One of the models used a coupled initialisation method instead. Each single-model ensemble was generated by running an ensemble of nine different ocean initial conditions, which have been obtained by perturbing the imposed wind stress and SSTs. The experiments involve using a 7x9 member multi-model ensemble (seven models, nine initial conditions) to produce global seasonal hindcasts. Ensembles are started 4 times per year (1 February, 1 May, 1 August, and 1 November) and run for 6 months. Details of the models and configurations used are given in Palmer et al. (2004) and Hagedorn et al. (2005).

The DEMETER system will be used to illustrate the above introduced BB ideas in the context of realistic ensemble forecasts made with state-of-the-art atmosphere-ocean circulation models. The following analysis is based on 6-hourly hindcasts of 2m temperature of the 63 member multi-model ensemble from 1989 to 1998 on a $2.5^\circ \times 2.5^\circ$ grid and uses the ECMWF 40-year Re-Analysis (ERA-40, see <http://www.ecmwf.int/research/era>) as the target (or verification) set.

3.2 Capturing the Re-Analysis

A central question addressed within this study focuses on the relative frequency with which the DEMETER ensemble captures the ERA-40 target, and hence its potential forecast value in the BB context. The forecast value will, of course, depend not only on the application but also upon how the model simulations are translated into forecasts; we return to this point below. As a starting point, consider the raw model simulations as forecasts: no accounting is made for any potential systematic model error. Figure 4 shows how often the ensemble *fails* to capture the verifying analysis for all available 2m-temperature (T2m) 6-hourly forecast data from 1989 to 1998. The evaluation is based on the 1-dimensional bounding box for each grid point. This allows to estimate reliable spatial forecast scales in the above mentioned sense.

The ensemble is almost always able to bound ERA-40 in continental as well as in some tropical ocean areas. Regions where the target is very often (30% and more) outside the ensemble's BB include the extratropical oceans, the western tropical Atlantic, some coastal areas of South America, the ocean west of Namibia and north of Madagascar, and in the Indonesian Archipelago. The areas west of Africa and South America are characterised by upwelling of cold deep water. The resulting cold surface currents present well known, nontrivial difficulties for ocean circulation models and near surface atmospheric temperatures.

Similar statistics for different lead times are shown in Fig. 5. In the first month of the forecasts, ERA-40 falls outside the BB in most areas of the globe in approximately 5-10% of all cases. As the forecast continues, regions of consistent failure concentrate mainly on the areas noted above (that is, near eastern boundary cold ocean surface currents, the Caribbean and Indonesia and some oceanic areas in the midlatitudes). It is observed that beside the initial spin-up period of the integrations the geographical regions where the target is outside the ensemble's BB are not particularly sen-

sitive to the forecast lead time. The ability of the DEMETER ensemble to bound ERA-40 is in general better for boreal summer (JJA) than for boreal winter (DJF). The ensemble often fails to include the DJF re-analysis, especially over the Southern Ocean.

3.3 Systematic Model Error

On seasonal time-scales systematic errors in dynamical model simulations of the mean state are often at least as large as the anomalies which one is trying to predict. While removing a systematic bias based on a very large number of single-model single-initial-condition forecasts is relatively straightforward, this is not the case in the ensemble scenario. Since bounding boxes are based upon the ensemble distribution(s), the relevant “post-processing” approach is more closely related to those used for probability forecasts, rather than those for point forecasts. This subsection includes a discussion of bias correction methods for forecast distributions and their relevance to the bounding box approach. (For a discussion of MOS techniques, see Eckel, 2003, and references therein).

Inasmuch as each model is imperfect, the question how to turn model simulations into forecasts (Wilks, 1995; Smith, 2003; 2000) arises. Technically this translation corresponds to a projection operator that maps a point in the model state space into the observable physical quantity we are interested in. While often noted (Smith, 1997; 2000; Judd and Smith, 2004) this operator is more often either taken to be the identity (that is, ignored) or dealt with by *ad hoc* adjustment of the first moments of the forecast distribution. Given only model simulations and observed targets, it is not clear how to separate the role of this operator from complications due to model inadequacy (Kennedy and O’Hagan, 2001).

If we resort to merely a best-first-guess point forecast, specifically a forecast state with small expected root-mean-square error, then linear regression on the simulations (see Krishnamurti et al., 1999, and references thereof) will yield such a state; this state will almost certainly be “unphysical”. For ensemble forecasts, just as in best-first-guess forecasts, systematic differences between a model’s simulation and a given target, once quantified, will not contribute to forecast error². Jewson (2004), among others, has argued one step beyond computing a mean state, suggesting that operational ensemble forecasts provide little if any information³ beyond mean and variance.

If we interpret the ensemble members as scenarios, then kernel dressing of each ensemble member allows the construction of a non-Gaussian forecast probability function (Roulston and Smith, 2003; Raftery et al., 2003), which has proven useful in practice (Roulston et al., 2003). In the multi-model ensemble, arguably none of these interpretations are internally coherent: one should condition the forecast on the joint distribution of the all the simulations at

hand. Operational procedures to do just this are under development.

The relevance to the bounding box approach is obvious: if the raw ensemble distribution is systematically displaced (or even malformed), then any identified systematic component should be removed before the bounding box is computed. Similarly, the internal consistency of the “post-processing” methods in the previous two paragraphs can be evaluated by examining their effect on the bounding box capture rates on real data. An alternative to evaluating other methods of post-processing would be to adjust the simulation values of each model with the explicit aim of improving the bounding box (just as traditional bias removal aims to improve root-mean-square error skill scores). For example, the minimal shift required to maximize the fraction of the target values within the bounding box could account for *some forms* of systematic error without introducing the biases implied by using the ensemble mean. Or rather than merely matching the first moment or two of the target distribution, one could alternatively map the cumulative distribution function of forecasts into that of the target variables. We consider only the identity operator and the ensemble-mean-bias operator below.

In practice, the standard approach to the (currently) inevitable climate drift of coupled models followed in DEMETER acts on an essentially linear assumption. The mean drift across an ensemble of forecasts is estimated⁴ and subtracted as an *a posteriori* correction to give the bias corrected forecast anomalies (Stockdale, 1997). While this may correct the first moment of the ensemble mean as a best-first-guess forecast distribution in-sample, its general effect on probability forecasts is unclear. In fact, the coherence of “correcting” each model-variable independently is unclear, given the space-time interdependence of these variables. Ignoring the physical implications, one could, of course, also manipulate other moments of the ensemble distribution to match moments of the error distribution of the ensemble means. If these manipulations are claimed as model-space to target-space adjustments, then presumably they (or their inverse) should also be considered in the data assimilation step. If they are accounting for anything other than the simplest scalar “model drift” then their impact on the probability distribution must be carefully argued.

We note again, that the bounding box methods do consider the distribution that arises from the ensemble, they do not require that the ensemble distributions be interpreted as probability forecasts. They can merely supply useful forecast information bounding the likely range of values and accompanied by an empirical measure of how often they in fact bound.

For the remainder of this section, we consider two projection operators: the identity (taking the model values as forecasts without alteration, as done in the results above)

² As noted by a referee, this identification must, of course, be done out-of-sample.

³ Jewson (2004) provides some empirical results from a specific case to support his conclusion; we suggest that more complicated procedures be tested against this benchmark.

⁴ As will become clear in the paragraphs that follow, there are a number of unresolved difficulties with this approach. Two issues not discussed further here are (i) accounting for the estimation error in the “bias correction” which is to be applied and (ii) the fact that in this case the target is itself a model state not an observation, and hence may also be biased, especially in regions with few observations which vary in number. Both methods are included here to allow a comparison of the results.

and the standard “ensemble-mean-bias removal” applied to each individual forecast model. It is not obvious *a priori* that the ensemble sample-mean should have zero bias, even in a perfect probability forecast, given that the forecast distribution changes from day to day and that the properties of each day’s target are unknown prior to verification (assuming it has properties other than the realised value). In order to examine the properties of each method in practice, the results for both are contrasted below.

The ensemble-mean-bias removal method estimates each single-model bias relative to its mean seasonal cycle. This is done by computing the average of all the hindcasts available for each 6-hourly data of the simulation and consider this as the “climate” or mean seasonal cycle of the model, following Palmer et al. (2004). After applying a 5-day low-pass filter, hindcast anomalies are obtained by subtracting the mean model seasonal cycle to each grid point, each initial month and each lead time for each ensemble member. A similar algorithm is used for ERA-40 to produce the verification anomalies. All anomalies have been computed in cross-validation mode (Wilks, 1995), which means that only data at other times different from the target date have been used to calculate the mean seasonal cycle.

How does the ensemble-mean-bias correction alter the results of the BB analysis discussed above? Figure 6 shows the relative frequency of the ERA-40 data being outside the multi-model ensemble’s BB for all bias corrected T2m data from 1989–1998. Obviously, the correction appropriately removes the distinct errors over the cold oceans west of North America, South America and Africa, which leads to a much improved capture rate in these regions (compare Fig. 4). For large parts of the world, especially over the tropical oceans and northern Asia, however, the ensemble-mean-bias correction results in a slightly higher outside-BB-frequency. Worse still, the ensemble fails to bound the target significantly more often in the Caribbean and Indonesian warm pool areas after this bias correction than it did before the “correction” was introduced; this underlines the limits of any first moment error treatment.

3.4 On the utility of the BB criterion

For a substantial fraction of model grid points, especially over land, the DEMETER ensemble captures ERA-40 almost perfectly. It is conceivable, however, that the ability to bound is simply the result of an unrealistically large ensemble spread. Forecast ensembles might bound simply by providing wide distribution compared to the climatology. The monthly mean spread of the ensemble relative to the spread of the climatology for forecast lead times 1-3 and 4-6 months of the bias corrected data is displayed in Fig. 7. Here ensemble spread (climatological spread) is defined as the absolute difference between the maximum and minimum ensemble member (difference between the maximum and minimum historical realisation within that month). For most continental grid points the ensemble spread is smaller than the climatological range. The spread ratio is also smaller than one over the Gulf of Mexico, the Indonesian warm pool area, and parts of the Indian Ocean. In the tropical Pacific and, to a lesser extent in the Arctic as well, there is substantial variability among the seven DEMETER models leading to a much larger ensemble spread than climatological range.

These results do not change significantly upon moving to, say, ninety per cent of the climatological spread; although it is interesting to note that bounding boxes with a range of less than 0.6 times the climatological range almost never capture the target.

A high fraction of points that bound based on an ensemble which has a smaller-than-climatology spread appears to be the most desirable forecast, given the ensemble’s ability to capture/not capture ERA-40 together with the information provided by the relative ensemble spread. In this case the ensemble can be regarded as reasonably useful in the BB sense irregardless of the actual ensemble probability distribution and its interpretation. For those cases where the BB captures the target but the ensemble spread is larger than the climatology, one might still get some useful information from the ensemble forecast, for example by analyzing the probabilities allocated to a certain event. If the ensemble-to-climatology spread ratio is smaller than 1 and the ensemble often fails to bound the target, one can think of enlarging the ensemble size, hoping to obtain a better forecast which captures more often, whereas a huge ensemble spread which never-the-less fails to capture the target suggests a relatively useless forecast.

These four categories have been summarised in Figure 8 (see also Table 1 for the colour definition). Here green indicates the most easily appreciated case of a high bounding fraction and a small ensemble spread. This is true almost everywhere in Africa, South America, Australia, large parts of Asia and for certain European and North American areas. This result alone provides evidence of the promising potential of DEMETER style forecasts for useful predictions on seasonal to inter-annual time scales in these areas. Most oceanic regions, especially in the tropical belt, as well as the northern parts of America and Asia, are classified as failing to capture the target in more than 5% of all cases while having a smaller-than-climatology ensemble spread (red). The spread underestimation in the tropics is a well-known problem in weather forecasting (Puri et al., 2001). The eastern tropical Pacific, the key region for El Niño forecasts, is able to capture more often while having a rather large ensemble spread (yellow). Many oceanic grid points and parts of Greenland and Antarctica, while having a larger-than-climatology spread, are yet unable to capture ERA-40 in more than 5% of the cases (grey).

Similar statistics for forecast lead times from 1 to 6 months suggest that grid points are very likely to stay within which ever category they fall in at the 2-months lead time for the remainder of the integration period. The first month of each integration appears to be more problematic in that the ensemble spread is often too small leading to more frequent failures in capturing ERA-40. This initial failure to bound increases for the first 5–7 days of the simulation, when roughly 10% of all grid point cannot be captured, indicates transient behaviour that extends for two weeks. The ability to bound in this initial phase is better for the February start dates than for the August start dates as illustrated in Fig. 9. This suggests that, in general, classifications obtained in the second forecast month is a good indicator of performance in the further course of the simulation, up to month 6. Looking at different seasons reveals that the ensemble performs better in JJA than in DJF, with large areas where the ensemble

spread is relatively small and still captures the target with a high probability.

What is the impact of “ensemble-mean bias correction”? If, in fact, this operator merely addressed the issue of removing systematic model errors, one would expect it to lead to significant, uniform improvement in terms of the frequency when the re-analysis is captured by the ensemble’s BB. We now illustrate that this is not the case.

Based on the raw non-corrected data, roughly 50% of all 10224 grid points do capture the analysis with a probability larger than 95% (green or yellow). After the bias correction this is true for $\sim 38\%$. Tables 2 and 3 summarise the 4-colour statistics and quantify the portion of changes in colour for land and sea grid points separately, after ensemble-mean-bias removal. Approximately two out of three land points (before and after bias correction) are able to bound the analysis. While, on average, every second sea point fails to bound before having been bias corrected, this figure rises up to 73% after the correction. The number of grid points which often bound the target tends to become somewhat larger for month 4–6 lead time than for months 1–3.

Most grid points, as a result of the bias correction, do not change colour. Over land, roughly half of the grey points turn to yellow with the bias correction, i.e. do include ERA-40 when they did not before, though still with a very large ensemble spread. Real improvements in the sense of an increasing number of green points (capturing with a rather small spread) can be noticed for $\sim 30\%$ of the former red land points. However, quite a large fraction of yellow land points change for the worst - they lose their ability to bound and are re-classified grey after having been corrected. Depending on the lead time, a substantial number (15% to 27%) of all non-corrected green land points fail to capture ERA-40 after the application of the ‘ensemble-mean bias correction’. The statistics for sea points in Table 3 underlines the qualitative impression from Fig. 8. Approximately one third of all non-bias corrected yellow sea points (or, equivalently, $\sim 14\%$ of all points) lose the capability to bound the analysis and become grey. A somewhat smaller number of grid points changes from yellow to red, meaning that, while the “correction” both decreases their ensemble spread, it also increases the failure rate.

4 Future Forecast Systems

It is important, of course, to distinguish between providing the most useful interpretation today of the currently available forecasts, and attempting to improve the forecast systems of the future. In this section we move to the second task, and risk an extrapolation of the bounding box scenario based on the results of analysing DEMETER forecasts in terms of future ensemble forecast systems. The first observation, of course, is that within the green areas in Figure 8, the current system shows skill. Proof-of-value experiments to demonstrate this skill are underway.

A rough estimate of the minimum ensemble size required to provide a BB which regularly captures the target is discussed in the next subsection; the last subsection then discusses issues of multi-model and single model ensembles.

4.1 Minimum Ensemble Size

Equation (4) offers a simple relationship between the probability to capture, P_{BB} , the considered dimension d and the size of the ensemble n for a single forecast and corresponding target. Provided that the characteristics of the ensemble in terms of its one-sided probabilities p_i from equation (3) are known, this allows one to estimate the minimum size of the ensemble which would be needed so that the ensemble’s BB captured the target with any arbitrary probability P_{BB} .

Figure 10 shows that a minimum ensemble size of ~ 100 members would, within the framework of the multi-model ensemble set-up in DEMETER, provide probabilistic forecasts that would very likely (95% capture) include the ERA-40 T2m target for most locations of the globe. This is merely an increase of about 50% in computational resources above the DEMETER system. This estimate is based on all ensemble-mean-bias corrected DEMETER T2m data from 1989–1998. Over some regions over the equatorial Pacific, Central-South America, Africa, the Arabian Peninsula and India a smaller ensemble would do as well; a much larger ensemble (of some 300 members) is required to ensure a high fraction of capture over the tropical oceans of the Caribbean and the Indonesian area. In agreement with the discussion above it is found that a bias corrected ensemble would need a larger size than the non-corrected ensemble in the Caribbean and Indonesian region, while the opposite is true for grid points where the applied bias correction helps to improve capturing, e.g., most notably near the coasts of Africa and South America.

Of course, an operational seasonal ensemble would be put to more uses than the construction of bounding boxes. The ideal size of an operational ensemble would be determined by some weighted average over its applications. Although it is unclear how to construct probabilistic forecasts from an ensemble of seasonal simulations, it is clear that an upper limit on the useful ensemble size will be approached as the additional ensemble members tell us more about the individual model than the likely future state of the atmosphere. In short, model inadequacy suggests that there is an ensemble size above which addition ensemble members are effectively redundant, unless the model and the assimilation scheme are perfect.

We expect that optimal resource allocation would demand different ensemble sizes for different models (based on their computational cost and marginal value to the mix of simulations already available). It would be interesting to see if resource allocation based on redundancy criteria in the BB context suggests the same distribution of computing power as resource allocation based on redundancy in probability forecasts, however computed. The bounding box criteria will undoubtedly be less costly to compute.

4.2 Multi-Model vs. Single-Model Ensemble

One key result of both the PROVOST (Brankovic and Palmer, 2000; Doblas-Reyes et al., 2000) and the DEMETER (Palmer et al., 2004; Hagedorn et al., 2005; Doblas-Reyes et al., 2005) projects is the enhanced reliability and skill of the multi-model ensemble over a conventional single-model ensemble approach (see also Ziehmann, 2000). The

bounding box analysis supports this finding. Each of the seven single-model ensembles based on perturbed initial conditions performs clearly worse than the full multi-model ensemble, as is shown in Fig. 11. The individual models, although corrected for their specific bias, fail to capture ERA-40 target significantly more often for certain model specific regions than the super-ensemble combining all models. The single-model ensemble spread is almost always smaller than the climatology (not shown), whereas the multi-model ensemble can provide a better and useful forecast (Fig. 11 lower right). Although there is a factor of seven difference in ensemble size, the observed scaling of probability of capture with ensemble parameters suggest that it is the multi-model nature that improves the capture rate of the DEMETER ensembles, not the change in counting statistics due to the larger ensemble size.

Assuming for a moment that the ensemble members were drawn within the context of the bounding box picture of Section 2, the multi-model ensemble is also expected to be superior to ensembles based on one single model in terms of the estimated minimum ensemble size. The temperature in most parts of the globe could only be captured when the single-model ensemble would comprise on the order of 300 ensemble members; geographical regions of strong model failures would require more than 500. DEMETER style multi-model ensembles would have similar expected performance characteristics with only a hundred members or so.

5 Conclusions and Future Work

We have explored the multi-model multi-initial-condition seasonal ensemble forecasts of the DEMETER project in the context of bounding boxes. It has been demonstrated that in many relevant regions of the globe, the forecast bounding box provides potentially useful information, placing both upper and lower limits on the expected target value with a range less than that of climatology, and capturing the target more than 95% of the time. Obvious land/sea contrasts are observed in the ability to bound the ERA-40 target; it would be interesting to better understand the extent to which this reflects the relationship between the observations and the DEMETER forecast system as opposed to that between the observations and the re-analysis used as the target.

Interpreting the bounding box of an ensemble provides information which complements interpreting the same ensemble as a probability forecast. There is no direct relationship between conventional skill scores and the BB ability to capture the target. The relative frequency with which a given ensemble BB captures the target is easily generalised to higher dimensional targets which we hope to demonstrate elsewhere, while making joint probability forecasts would require astronomical ensemble sizes.

Both the raw simulations interpreted as forecasts and “ensemble-mean-bias corrected” forecasts have been analysed. Both are seen to provide useful information, but this simple approach to removing systematic errors from ensemble forecasts is shown to systematically degrade the BB forecasts in some regions. Interestingly, this “correction” seems to shift roughly equal numbers of ensembles from bound-

ing to non-bounding and vice-versa when only regions with small spread relative to climatology are considered. In regions where the spread is large compared to the climatological range, roughly one sixth of non-bounding grid points bound after the adjustment; this represents about 6% of the surface of the globe.

What is the connection between the mathematical results in Section 2 and the analysis of the DEMETER forecast in later sections? Arguably there is no direct connection, since in Section 2 we assumed full knowledge of both the distributions corresponding to the forecast model and those corresponding to the target system. In reality, we have only a finite monte carlo sample of forecasts, not the full distribution, and inasmuch as only a single target exists, the epistemological status of the “distribution from which the target is drawn” is unclear. Even if it exists, it is certainly unknown outside the perfect model scenario.

On the other hand, we can estimate the distribution of p_i by simply seeing where in the ensemble the target falls over a series of forecasts. Under the assumption that this distribution is robust, we can then place an upper bound on the minimum ensemble size required to obtain a given capture rate as follows. First form a rank histogram of the number of ensemble members less than the verification (that is, a Talagrand diagram); next make a conservative association between each bin and a value of p_i (for example, since we are looking for an upper bound on the minimum ensemble size, assign $p_i = 0$ to the first and last bins, $p_i = 1/63$ to the second and 63rd bins, and so on.) For a given value of n , equation (4) then gives the expected capture rate P_{BB} for each bin. Weighting these rates with the relative frequency of the corresponding bins then provides the capture rate for that value of n . In this way one can construct the expected capture rate P_{BB} as a function of n , and then determine the value of n required for a given rate. Note, of course, that under the assumptions above there may be some capture rates that can never be obtained. While this calculation lies beyond the scope of the current paper, it would be interesting to see how this upper bound on the minimum ensemble size varies for different forecasts. Using the bounding box approach does not require such calculations or detailed assumptions: the forecast BB can be exploited by any user who believes the observed capture rate is sufficiently large. The empirical capture rate itself can be a useful measure of skill.

A well argued, coherent, deployable framework within which to account for systematic model errors in the elements of a multi-model multi-initial-condition ensemble is needed. Similarly, there are a number of interesting questions regarding resource allocation between models in such an ensemble system if run operationally. Ultimately, mapping the joint distribution of simulations available into a single forecast is expected to outperform any piecemeal scenario-based approach. Viewing the ensembles through their bounding box statistics may prove a useful guide both in evaluating our current models and in constructing this future forecast.

6 Acknowledgements

We would like to thank all DEMETER partner groups, especially the DEMETER team at ECMWF, for assistance

with their model data and discussions. The many comments of two referees have been very helpful. This work has been partially supported by the European Union under EVK2-CT-2001-50012, a University of Western Australia small grant, and by the Office of Naval Research DRI under N00014-99-1-0056.

REFERENCES

- Anderson, J., 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate* **9**, 1518–1530.
- Brankovic, C. and Palmer, T.N., 2000. Seasonal skill and predictability of ECMWF PROVOST ensembles. *Quart.J.R.Meteorol. Soc.* **126**, 2035–2067.
- Chatfield, C., 2001. Prediction intervals. In: *Principles of Forecasting: A Handbook for Researchers and Practitioners*, (eds. J.S. Armstrong), Kluwer Academic.
- Doblas-Reyes, F.J., Deque, M. and Piedelievre, J.-P., 2000. Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Quart.J.R.Meteorol. Soc.* **126**, 2069–2087.
- Doblas-Reyes, F.J., Hagedorn, R., and Palmer, T.P. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part II: Calibration and combination. *Tellus 57A*, this issue.
- Eckel, F.A., 2003. Effective mesoscale, short-range ensemble forecasting, Ph.D. Dissertation, Department of Atmospheric Sciences, University of Washington, Seattle, Washington. Available online at www.atmos.washington.edu/ens/pubs_n_pres.html.
- Ehrendorfer, M. 1997. Predicting the uncertainty of numerical weather forecasts: a review. *Meteorol. Z., N.F.*, **6**, 147–183.
- Hagedorn, R., Doblas-Reyes, F.J., and Palmer, T.P. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus 57A*, this issue.
- Jewson, S., 2004. Do probabilistic medium-range temperature forecasts need to allow for non-normality?, www.arxiv.org/abs/physics/0310060.
- Jolliffe, I.T. and Stephenson, D.B., 2003. *Forecast Verification*. Wiley 240 p.
- Judd, K. and Smith, L.A., 2001. Indistinguishable states I : perfect model scenario. *Physica D* **151**, 125–141.
- Judd, K. and Smith, L.A., 2004. Indistinguishable states II : imperfect model scenario. *Physica D* **196**, 224–242.
- Kennedy, M. and O'Hagan, A., 2001. Bayesian calibration of computer codes *J. Royal Statistical Soc.*, **B63**, 425–464.
- Krishnamurti, T.N., Kishtawal, C.M., LaRow, T., Bachiochi, D., Zhang, Z., Williford, E., Gadgil, S. and Surendran, S., 1999. Improved weather and seasonal climate forecasts from multimodel superensemble *Science*, **285**, 1548–1550.
- Leith, C.E. 1974. Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.
- Palmer, T.N., 2000. Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Physics* **63**, 71–116.
- Palmer, T.N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., and co-authors, 2004. Development of a European Multi-Model Ensemble System for Seasonal to Inter-Annual Prediction (DEMETER). *Bull. American Meteorol. Soc.*, **85**, 853–872.
- Puri, K., Barkmeijer, J. and Palmer, T.N., 2001. Ensemble prediction of tropical cyclones using targeted diabatic singular vectors. *Quart.J.R.Meteorol. Soc.* **127**, 709–731.
- Raftery, A.E., Balabdaoui, F., Gneiting, T. and Polakowski, M., 2003. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Technical Report no. 440, Department of Statistics, University of Washington*.
- Robert, C.P., and Casella, G., 1999. *Monte Carlo Statistical Methods*. Springer Verlag 507 p.
- Roulston, M.S., Kaplan, D.T., Hardenberg, J. and Smith, L.A., 2003. Using Medium Range Weather Forecasts to Improve the Value of Wind Energy Production. *Renewable Energy* **28(4)**, 585–602.
- Roulston, M. and Smith, L.A., 2003. Combining Dynamical and Statistical Ensembles. *Tellus* **55A**, 16–30.
- Smith, L.A., 1997. The maintenance of uncertainty. In: *Proc. Int. School of Physics "Enrico Fermi"*, Bologna, Italy, 177–246.
- Smith, L.A., 2000. Disentangling uncertainty and error: On the predictability of nonlinear systems. In: *Nonlinear Dynamics and Statistics*, (eds. A.I. Mees), Birkhauser, Boston, USA, 31–64.
- Smith, L.A., 2003. Predictability past predictability present. *ECMWF Seminar Proceedings Predictability of Weather and Climate*, ECMWF, U.K., 219–242.
- Stockdale, T.N., 1997. Coupled Ocean–Atmosphere Forecasts in the Presence of Climate Drift. *Mon. Wea. Rev.* **125**, 809–818.
- Wilks, D.S., 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press 467 p.
- Ziehmann, C., 2000. Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus* **52A**, 280–299.

Table 1. Definition of the four categories and their colour codes.

	Does the ensemble capture with at least 95% frequency?	Is the ensemble spread narrower than climatology?
GREY	No	No
RED	No	Yes
YELLOW	Yes	No
GREEN	Yes	Yes

Table 2. Fraction (in %) of all grid points over land that change their 4-colour classification (see text for details) as a result of the ensemble-mean-bias correction. The rows stand for the non-bias corrected raw data; the columns for the data after the correction. In each cell, the top line corresponds to all forecast data from 1989–1998; the middle and lower lines are for all lead times of 1–3 and 4–6 months, respectively. The sum over all points in each category is given by the numbers in the last row and column, respectively.

	GREY	RED	YELLOW	GREEN	Σ
GREY	9.6	1.7	10.2	0.6	22.1
lead 1-3	13.0	2.7	8.4	0.7	24.8
lead 4-6	6.2	1.1	11.1	0.8	19.1
RED	0	10.1	0	5.1	15.1
lead 1-3	0	14.9	0	6.2	21.0
lead 4-6	0	6.5	0	4.4	10.9
YELLOW	6.8	4.0	14.7	3.0	28.5
lead 1-3	7.3	3.9	9.9	1.7	22.8
lead 4-6	6.4	3.4	20.2	4.5	34.4
GREEN	0	6.3	0	27.9	34.2
lead 1-3	0	8.4	0	23.1	31.4
lead 4-6	0	5.5	0	30.1	35.6
Σ	16.4	22.1	24.9	36.5	100.0
	20.3	29.8	18.3	31.6	100.0
	12.6	16.4	31.2	39.7	100.0

Table 3. As in Table 2, but for sea points.

	GREY	RED	YELLOW	GREEN	Σ
GREY	16.3	20.3	4.1	0.2	40.8
lead 1-3	14.4	23.7	3.9	0.2	36.5
lead 4-6	17.2	20.2	4.6	0.1	42.1
RED	0	15.5	0	0.0	15.5
lead 1-3	0	23.7	0	0.0	23.7
lead 4-6	0	10.0	0	0.1	10.1
YELLOW	14.1	6.1	21.6	0.7	42.5
lead 1-3	13.2	5.9	18.7	0.8	38.6
lead 4-6	14.3	6.6	24.8	0.6	46.4
GREEN	0	0.9	0	0.2	1.1
lead 1-3	0	0.9	0	0.3	1.2
lead 4-6	0	1.2	0	0.2	2.4
Σ	30.4	42.7	25.7	1.1	100.0
	27.6	48.5	22.6	1.4	100.0
	31.5	38.0	29.4	1.1	100.0

Figure 1 Schematic diagrams of the bounding box concept for the *general case* of an arbitrary distribution. a) Histogram as an estimation of the probability density function (pdf) of a 63-member ensemble. The dashed line indicates the position of the target x^* . The black area under the histogram curve left of the target is an estimate of the one-sided probability p_i . b) Cumulative density function (cdf) for the distribution in panel 1a. Here p_i corresponds directly to the value of the cdf for the target x^* . c) Probability P_{BB} that the 1-dimensional bounding box defined by p_i and n ensemble members captures the target. The thick dash-dotted line denotes the 95% probability. Note that the graph is symmetric with respect to $p_i = 0.5$.

Figure 2 Schematic diagrams of the bounding box concept for the case of a *Gaussian distribution*. a) Histogram as an estimation of the pdf of a 63-member ensemble. b) Cumulative density function (cdf) for the distribution in panel 2a, indicating $p_i = \Phi(z_i)$. c) Probability P_{BB} that the 1-dimensional bounding box defined by z_i and n ensemble members captures the target. The thick dash-dotted line denotes the 95% probability. Note the logarithmic scale of the vertical axis. The graph for $z_i < 0$ would be symmetric.

Figure 3 Schematic diagrams of the bounding box concept for the case where the *target is the median of the distribution*, that is $p_i = 0.5$. a) Histogram as an estimation of the pdf of a 63-member ensemble. b) Cumulative density function (cdf) for the distribution in panel 3a. c) Probability P_{BB} that the d -dimensional bounding box defined by n ensemble members captures the target. The thick dash-dotted line denotes the 95% probability.

Figure 4 Fraction of all 6-hourly T2m forecast data from 1989–1998 for which the ERA-40 re-analysis is outside the DEMETER multi-model ensemble bounding box. Isolines at 0.05, 0.1, 0.2, and 0.3.

Figure 5 Fraction of 6-hourly T2m forecasts from 1989–1998 when ERA-40 re-analysis is outside the DEMETER multi-model ensemble bounding box for lead times 1 month to 6 months. Isolines at 0.05, 0.1, 0.2, and 0.3.

Figure 6 Fraction of all 6-hourly T2m forecast data from 1989–1998 for which the ERA-40 re-analysis is outside the DEMETER multi-model ensemble bounding box where the forecasts have been filtered with an “ensemble mean bias removal” filter. Isolines at 0.05, 0.1, 0.2, and 0.3.

Figure 7 Ratio of the bias corrected multi-model ensemble spread to the spread of all historical ERA-40 re-analysis data (climatology). The spread is defined as the difference between the largest and smallest member of all realization. The data shown cover all 6-hourly T2m forecasts from 1989–1998 for lead times 1–3 months (top) and for lead times 4–6 months (bottom). Isolines at 0.5, 0.75, 1.0 (white), 1.25, 1.5, and 1.75.

Figure 8 Four-colour plot for all bias corrected 6-hourly T2m forecasts from 1989–1998. Green: spread ratio smaller than 1 AND fraction inside the multi-model ensemble bounding box larger than 95%. Yellow: spread ratio larger than 1 AND fraction inside the multi-model ensemble bounding box larger than 95%. Red: spread ratio smaller than 1 AND fraction outside the multi-model ensemble bounding box larger than 5%. Grey: spread ratio larger than 1 AND fraction outside the multi-model ensemble bounding box larger than 5%.

Figure 9 Fraction of grid points outside the bounding box for all bias corrected 6-hourly T2m forecast lead times up to 2 months and different start dates from 1989–1998.

Figure 10 Estimation of the minimum ensemble size needed to capture with a 95% probability. Data shown are based on all bias corrected 6-hourly T2m forecasts from 1989–1998. Isolines at 63, 80, 100, 120, and 160.

Figure 11 Comparison of the performance of the single-model ensemble vs the multi-model ensemble. Fraction of all 6-hourly T2m forecast data from 1989–1998 when ERA-40 is outside the bounding box. The panels on the left and the first three panels on the right show results based on the individual bounding boxes for each of the seven single-model ensembles. For comparison, the lower right panel gives the full multi-model ensemble results. Isolines in all panels at 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8.

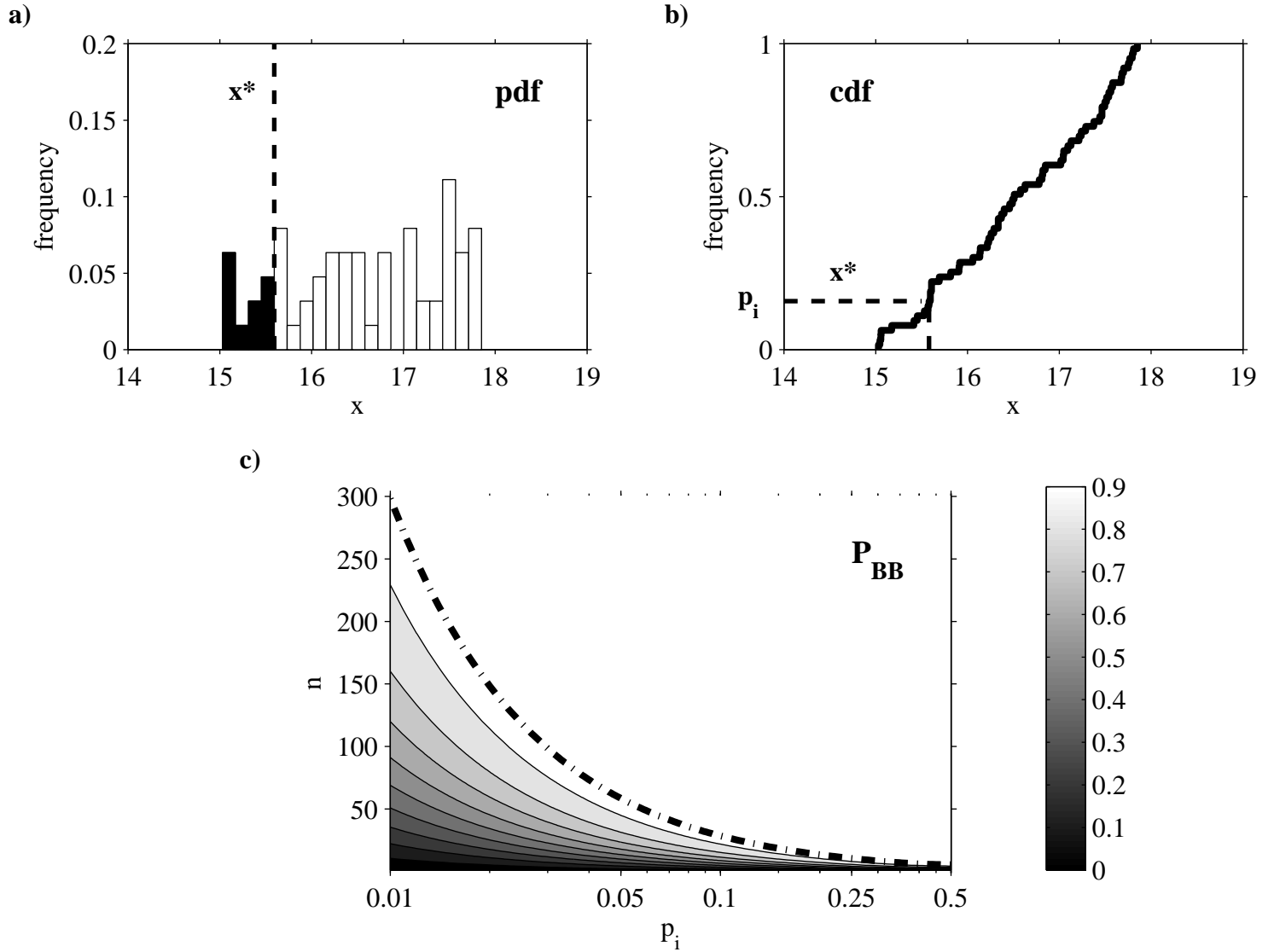


Figure 1. Schematic diagrams of the bounding box concept for the *general case* of an arbitrary distribution. a) Histogram as an estimation of the probability density function (pdf) of a 63-member ensemble. The dashed line indicates the position of the target x^* . The black area under the histogram curve left of the target is an estimate of the one-sided probability p_i . b) Cumulative density function (cdf) for the distribution in panel 1a. Here p_i corresponds directly to the value of the cdf for the target x^* . c) Probability P_{BB} that the 1-dimensional bounding box defined by p_i and n ensemble members captures the target. The thick dash-dotted line denotes the 95% probability. Note that the graph is symmetric with respect to $p_i = 0.5$.

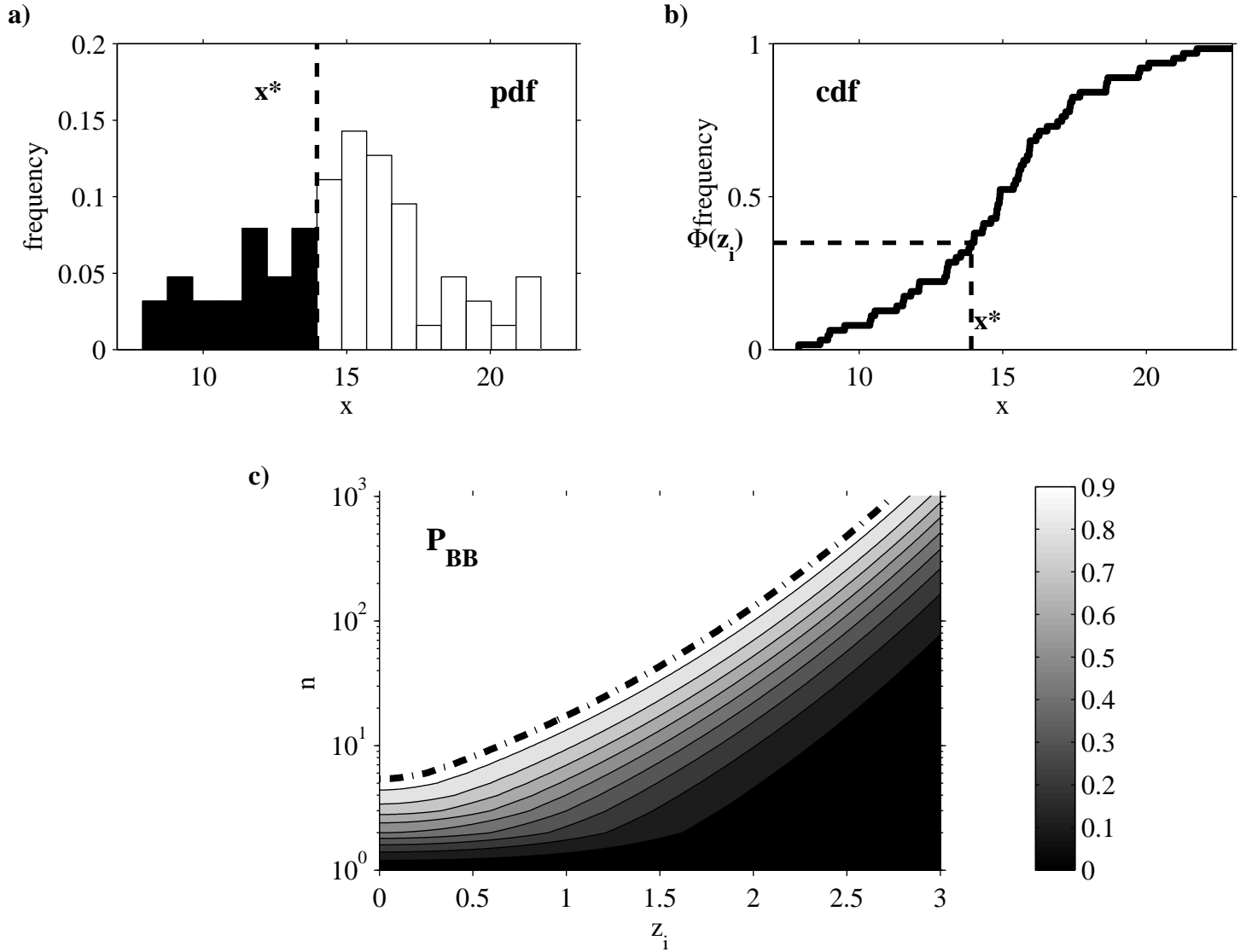


Figure 2. Schematic diagrams of the bounding box concept for the case of a *Gaussian distribution*. a) Histogram as an estimation of the pdf of a 63-member ensemble. b) Cumulative density function (cdf) for the distribution in panel 2a, indicating $p_i = \Phi(z_i)$. c) Probability P_{BB} that the 1-dimensional bounding box defined by z_i and n ensemble members captures the target. The thick dash-dotted line denotes the 95% probability. Note the logarithmic scale of the vertical axis. The graph for $z_i < 0$ would look symmetric.

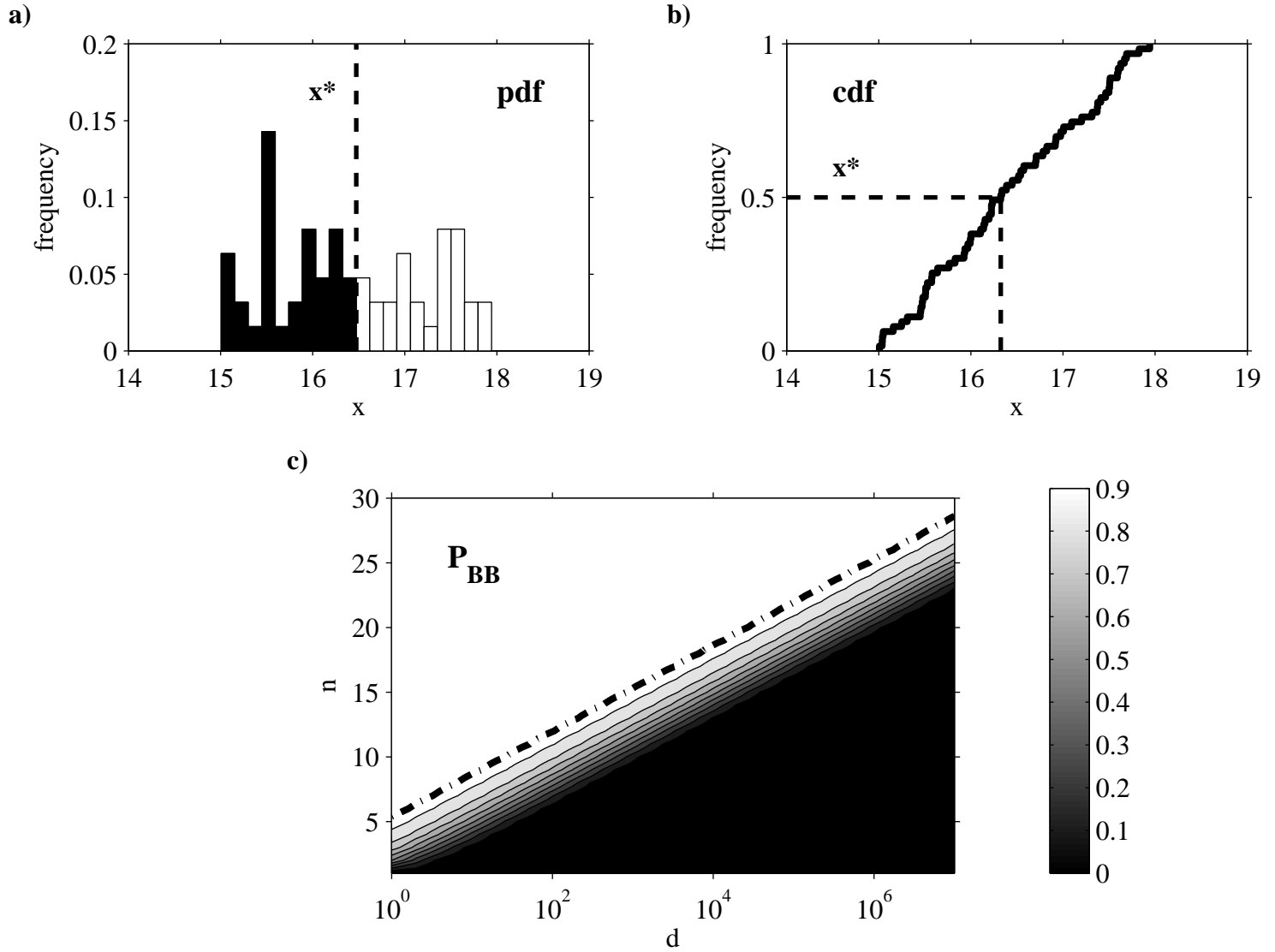


Figure 3. Schematic diagrams of the bounding box concept for the case where the *target is the median of the distribution*, that is $p_i = 0.5$. a) Histogram as an estimation of the pdf of a 63-member ensemble. b) Cumulative density function (cdf) for the distribution in panel 3a. c) Probability P_{BB} that the d -dimensional bounding box defined by n ensemble members captures the target. The thick dash-dotted line denotes the 95% probability.

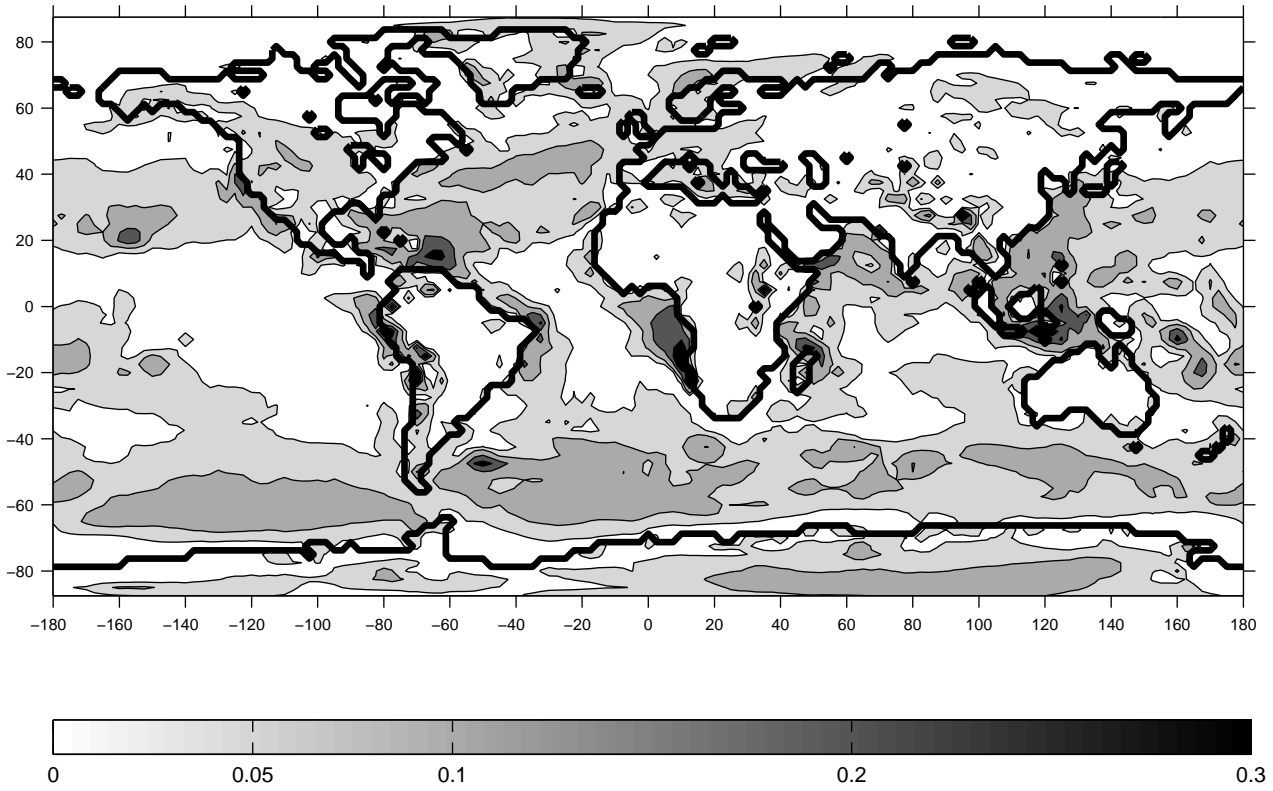


Figure 4. Fraction of all 6-hourly T2m forecast data from 1989–1998 for which the ERA-40 re-analysis is outside the DEMETER multi-model ensemble bounding box. Isolines at 0.05, 0.1, 0.2, and 0.3.

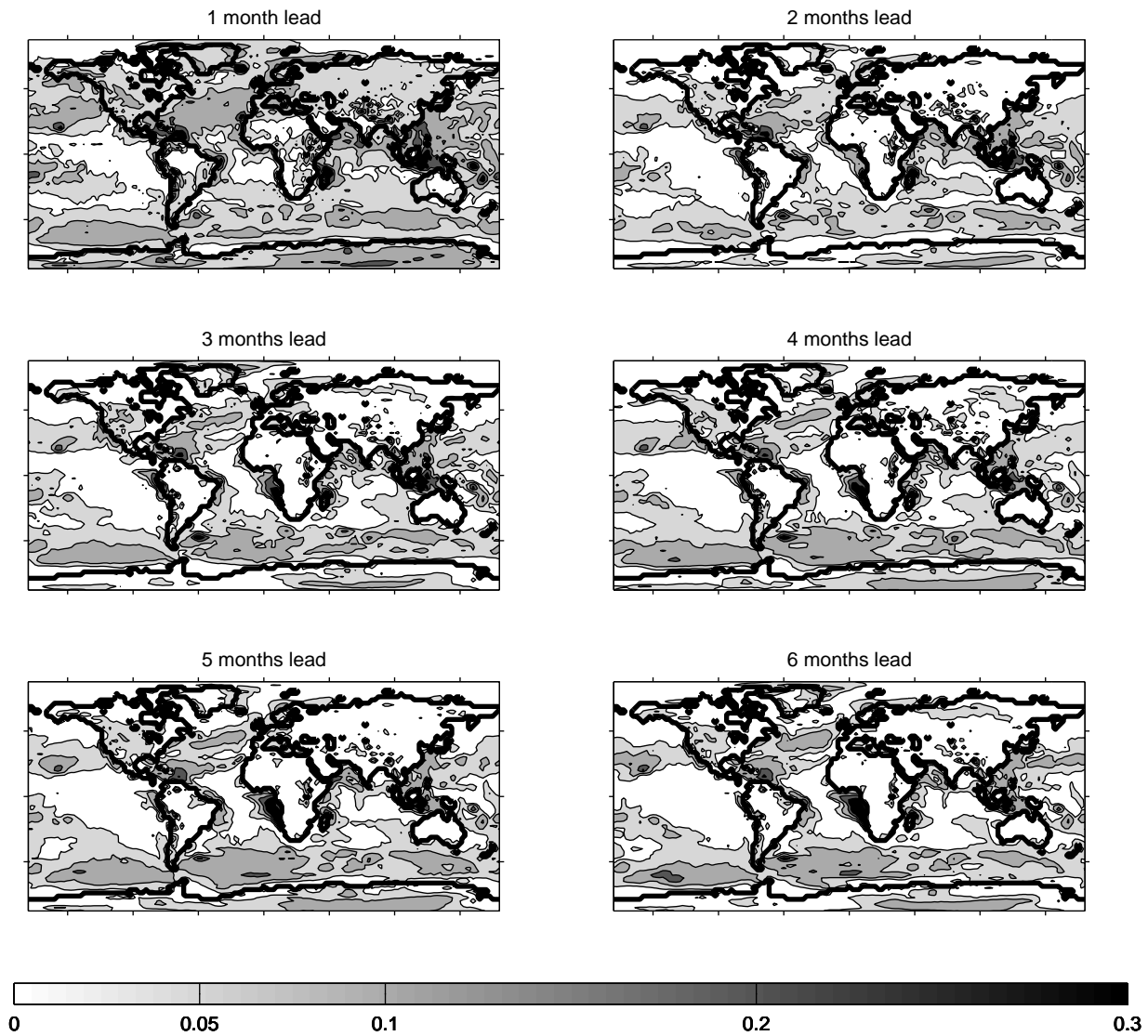


Figure 5. Fraction of 6-hourly T2m forecasts from 1989-1998 when ERA-40 re-analysis is outside the DEMETER multi-model ensemble bounding box for lead times 1 month to 6 months. Iso-lines at 0.05, 0.1, 0.2, and 0.3.

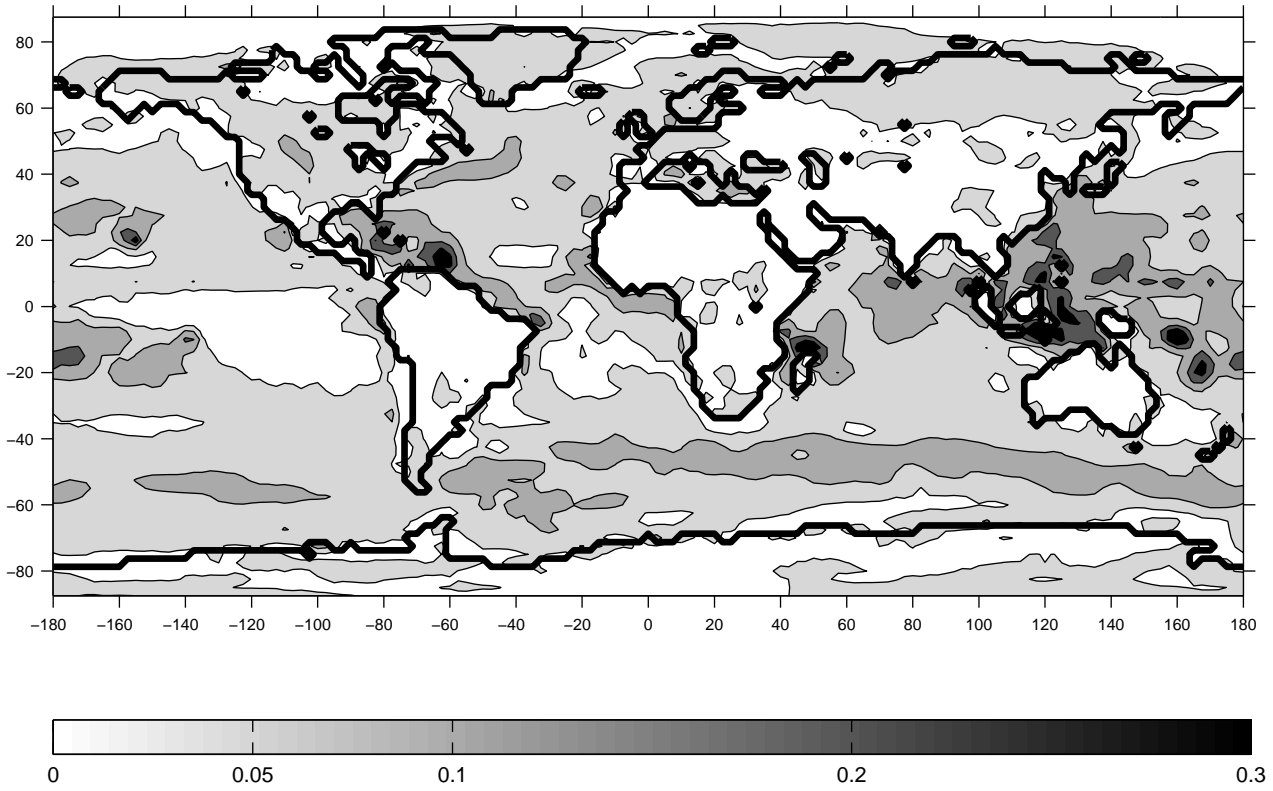


Figure 6. Fraction of all 6-hourly T2m forecast data from 1989-1998 for which the ERA-40 re-analysis is outside the DEMETER multi-model ensemble bounding box where the forecasts have been filtered with an “ensemble mean bias removal” filter. Iso-lines at 0.05, 0.1, 0.2, and 0.3.

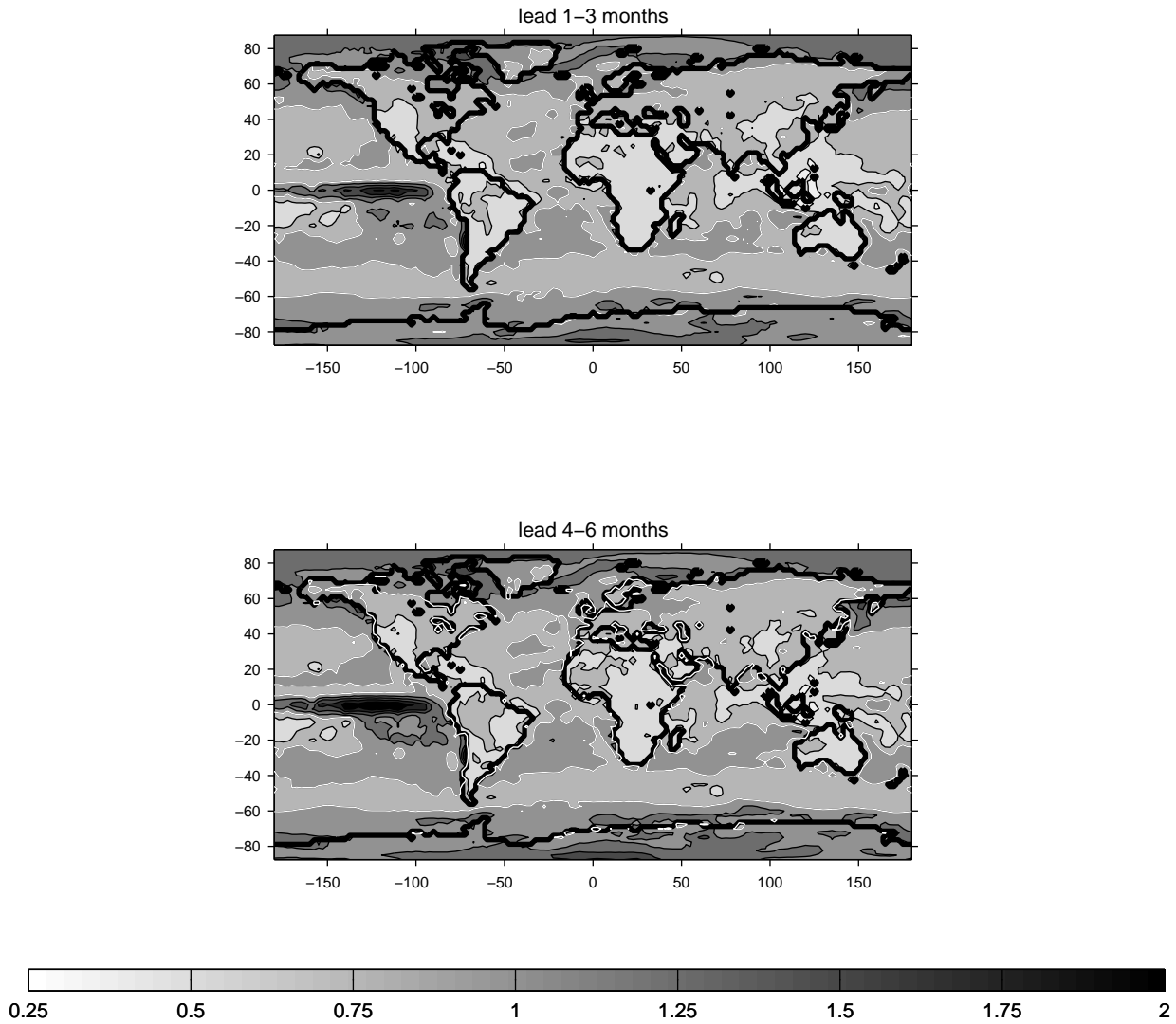


Figure 7. Ratio of the bias corrected multi-model ensemble spread to the spread of all historical ERA-40 re-analysis data (climatology). The spread is defined as the difference between the largest and smallest member of all realization. The data shown cover all 6-hourly T2m forecasts from 1989-1998 for lead times 1-3 months (top) and for lead times 4-6 months (bottom). Isolines at 0.5, 0.75, 1.0 (white), 1.25, 1.5, and 1.75.

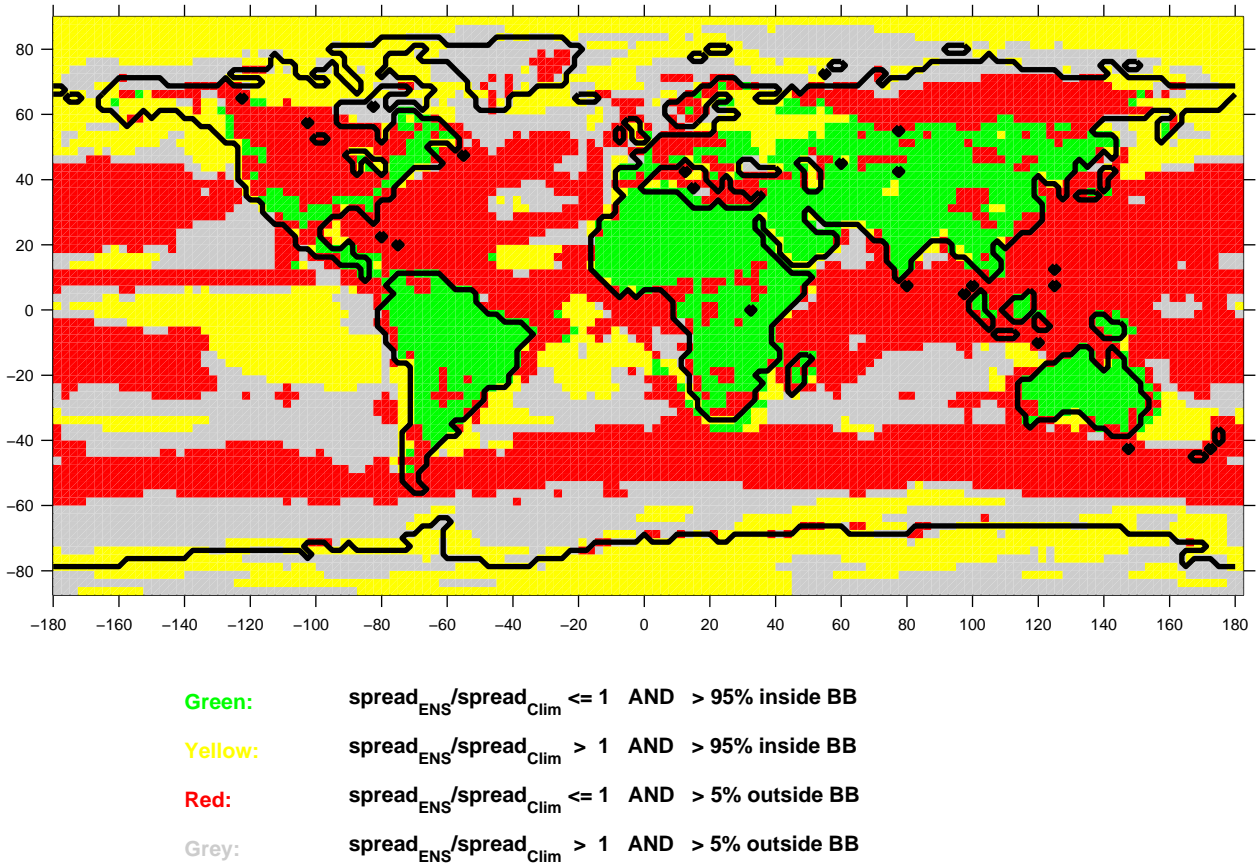


Figure 8. Four-colour plot for all bias corrected 6-hourly T2m forecasts from 1989–1998. Green: spread ratio smaller than 1 AND fraction inside the multi-model ensemble bounding box larger than 95%. Yellow: spread ratio larger than 1 AND fraction inside the multi-model ensemble bounding box larger than 95%. Red: spread ratio smaller than 1 AND fraction outside the multi-model ensemble bounding box larger than 5%. Grey: spread ratio larger than 1 AND fraction outside the multi-model ensemble bounding box larger than 5%.

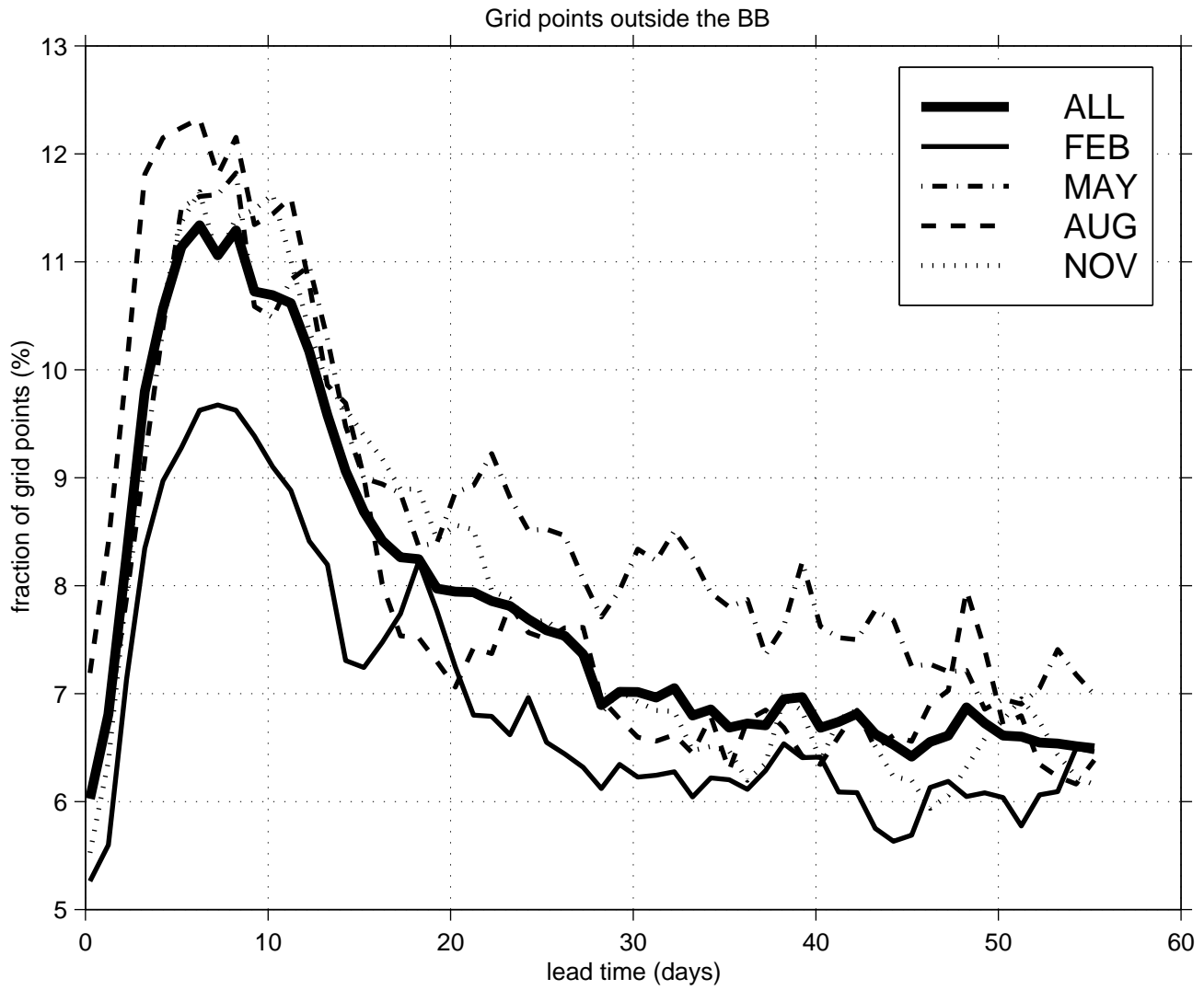


Figure 9. Fraction of grid points outside the bounding box for all bias corrected 6-hourly T2m forecast lead times up to 2 months and different start dates from 1989–1998.

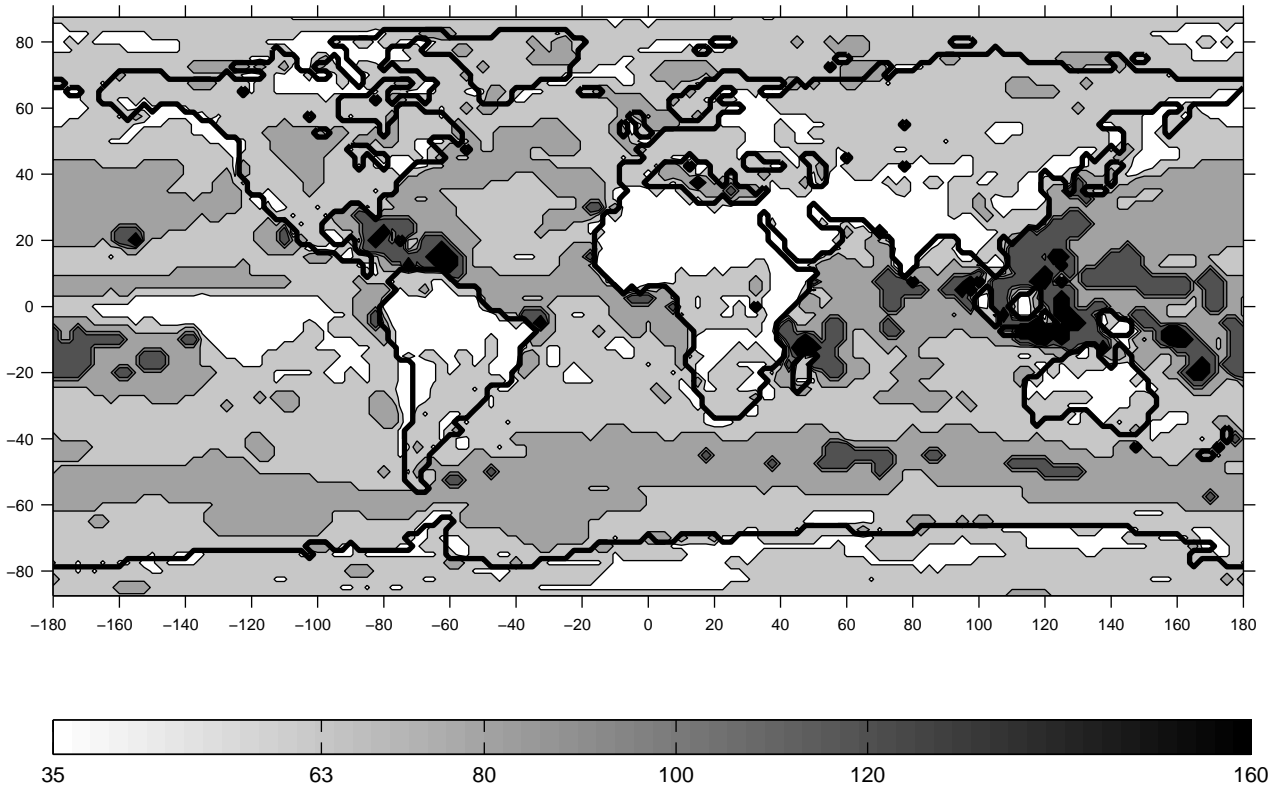


Figure 10. Estimation of the minimum ensemble size needed to capture with a 95% probability. Data shown are based on all bias corrected 6-hourly T2m forecasts from 1989–1998. Isolines at 63, 80, 100, 120, and 160.

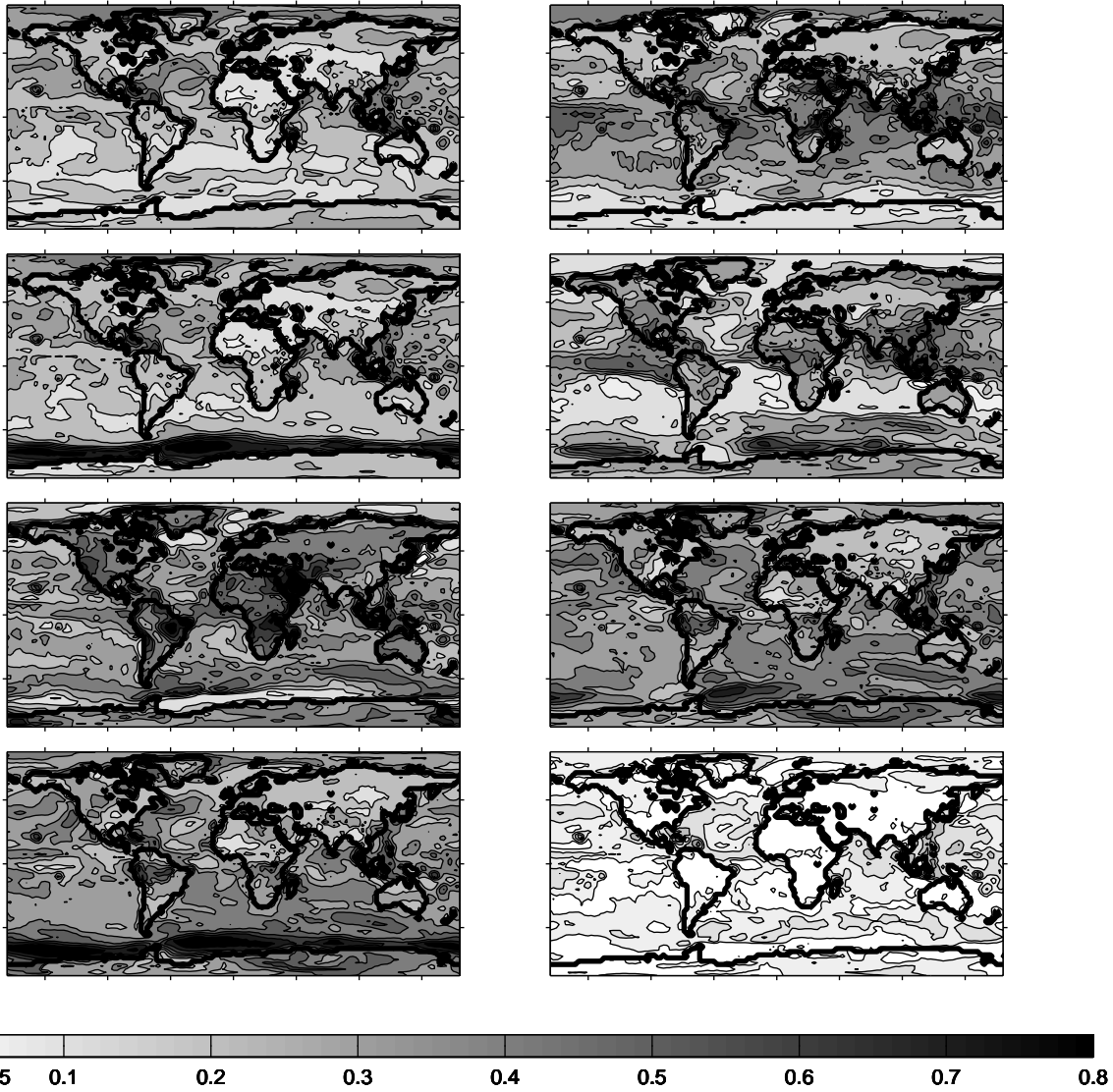


Figure 11. Comparison of the performance of the single-model ensemble vs the multi-model ensemble. Fraction of all 6-hourly T2m forecast data from 1989–1998 when ERA-40 is outside the bounding box. The panels on the left and the first three panels on the right show results based on the individual bounding boxes for each of the seven single-model ensembles. For comparison, the lower right panel gives the full multi-model ensemble results. Isolines in all panels at 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8.