Tellus (2005), 57A, 219–233
Printed in UK. All rights reserved

TELLUS

The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept

By RENATE HAGEDORN*, FRANCISCO J. DOBLAS-REYES and T. N. PALMER, *ECMWF, Shinfield Park, Reading RG2 9AX, UK*

(Manuscript received 6 April 2004; in final form 24 September 2004)

ABSTRACT

The DEMETER multi-model ensemble system is used to investigate the rationale behind the multi-model concept. A comprehensive documentation of the differences in the single and multi-model performance in the DEMETER hindcast data set is given. Both deterministic and probabilistic diagnostics are used and a variety of analyses demonstrate the improvements achieved by using multi-model instead of single-model ensembles. In order to understand the reason behind the multi-model superiority, basic scenarios describing how the multi-model approach can improve over single-model skill are discussed. It is demonstrated that multi-model superiority is caused not only by error compensation but in particular by its greater consistency and reliability.

1. Introduction

Using collective information for decision-making is common sense in both everyday life and professional business. In particular, the greater the complexity of the involved processes, the more input for our decision-making procedure can be helpful (Branzei et al., 2000). On the other hand, an overload of possibly contradictory information can lead to suboptimal decisions. It has been shown that in the real world of confusing and overwhelming information, fast and frugal heuristics (i.e. simple rules for making decisions) can be powerful tools that do surprisingly well (Gigerenzer and Todd, 1999). That is, in general decision-making theory, it is under debate whether more information leads to more success or whether 'simplicity rules the world'.

The effect, that more information does not necessarily lead to more success, has been demonstrated for the case of weather forecasting by Heideman et al. (1993). Their results suggest that 'the relation between information and skill in forecasting weather is complex' and that 'greater improvement in forecasting might be obtained by devoting resources to improving the use of information over and above those needed to increase the amount of information'. However, it is very important to note that this effect generally holds only in the case of an individual forecaster, making decisions based on different levels of information available to her or him. It must not be confused with the attempt to improve predictions by utilizing more than one decision-making system, of either subjective or objective nature. Many indications exist that such multiple decision-making systems (a group of forecasters/models) are generally superior to individual decision-making systems (a single forecaster/model).

For example, in short- and medium-range weather forecasting it has been demonstrated, in the early 1960s, that combining different forecasts from individual forecasters can be beneficial. Sanders (1963) analysed multiple-person forecasts and showed that 'the group-mean probability forecast is found to be a more skilful statement than the probability forecast of the most skilled individual'. His early findings were confirmed by later studies (Sanders, 1973; Bosart, 1975; Gyakum, 1986), and the extension of this concept from subjective multiple forecasters to objective multi-model prediction systems has also been proven successful (Clemen and Murphy, 1986; Fraedrich and Leslie, 1987). Comparisons of multi-model and single-model performance suggest that 'variations in model physics and numerics play a substantial role in generating the full spectrum of possible solutions' (Fritsch et al., 2000).

However, using more than one model addresses only one of the two main sources of error. The second source of error, uncertainties in initial conditions, can be addressed by running an ensemble of forecasts from different initial conditions. This technique, known as ensemble prediction, is used with great success at forecasting centres around the world (Tracton and Kalnay, 1993; Molteni et al., 1996). Richardson (2000) has shown that probability forecasts derived from an ensemble prediction system (EPS) are of greater benefit than a deterministic forecast produced by the same model and that, for many users, the probability forecasts have more value than a shorter-range deterministic forecast.

*Corresponding author.
e-mail: renate.hagedorn@ecmwf.int

In order to take into account both model error and uncertainties in initial conditions, the multi-model and ensemble techniques can be combined to a new approach, known as the multi-model ensemble concept (Harrison et al., 1995; Palmer and Shukla, 2000; Palmer et al., 2004). The idea of the superiority of multiple source prediction systems is based on the 'incontrovertible fact that two or more inaccurate but independent predictions of the same future events may be combined in a very specific way to yield predictions that are, on the average, more accurate than either or any of them taken individually' (Thompson, 1977). However, how 'incontrovertible' and widely accepted is this fact? Although many studies have demonstrated the success of the multi-model approach in practice, parts of the scientific community still dispute the general validity of the concept. Some of these reservations are caused by apparent misconceptions of the approach. Frequent questions in this context are the following.

(i) How can a poor model add skill?
(ii) How can the multi-model be better than the average single-model performance?
(iii) Why not use the best single model instead of the multi-model?

In this paper we attempt to clarify such misconceptions in answering the above questions and discussing in general the rationale behind the multi-model ensemble concept.

The study is based on the extensive data set of seasonal hindcasts produced in the DEMETER project (Development of a European Multi-model Ensemble System for Seasonal to Interannual Prediction). DEMETER was conceived and funded as a European Union (EU) Framework-V project in order to advance the concept of multi-model ensemble prediction by installing a number of state-of-the-art global coupled ocean–atmosphere models on a single supercomputer, and to produce a series of six-month multi-model ensemble hindcasts with common archiving and common diagnostic software. A general description of the project, the involved coupled models, the produced data set, as well as the verification, downscaling and application of the data is given in Palmer et al. (2004). Here, the DEMETER data set is used to study specifically the advantages and limitations of the multi-model ensemble approach in seasonal forecasting.

In its simplest form, a multi-model ensemble forecast is produced by simply merging the individual forecasts with equal weights. However, more complex methods of optimally combining the single-model output have been described (Krishnamurti et al., 1999; Pavan and Doblas-Reyes, 2000; Rajagopalan et al., 2002). A substantial amount of effort has been concentrated on assessing the performance of sophisticated techniques for constructing optimal multi-model ensembles. However, neither a comprehensive demonstration of the superiority of the multi-model approach for seasonal forecasting nor any substantial work on the rationale behind its success can be found in the literature. Motivated by this lack of groundwork, a comprehensive documentation of the improved performance of a multi-model ensemble system compared to single-model ensemble predictions will be presented here, and also an explanation for the multi-model superiority is proposed.

In order to illustrate the improvements found when using an 'equal-weight' multi-model ensemble forecast system separately from the likely improvements expected in optimal multi-model ensembles, the paper is split into two parts. In the first part, the basic concept of the multi-model approach is discussed along with results from the equal-weight multi-model ensemble, henceforth referred to as the simple multi-model ensemble. All issues related to advanced methods for calibrating and optimally combining models will be addressed in the second part of the paper. A careful examination of simple multi-model ensembles results is additionally motivated by the the following facts: (i) robust optimal weights are difficult to calculate given the short samples available to train the models (Kharin and Zwiers, 2002; Peng et al., 2002), i.e. often the use of the simple multi-model is the only practical way of utilizing the multi-model approach; (ii) simple multi-model ensembles may be considered as a reference method for optimal multi-model ensemble systems.

A description of the data set and diagnostic tools used can be found in Section 2. In order to document the multi-model superiority, a comprehensive comparison of simple multi-model versus single-model results is presented in Section 3. A discussion of the rationale behind the superiority of the multi-model follows in Section 4, including some theoretical considerations and practical examples. The conclusions are summarized in Section 5.

2. Data and tools

2.1. DEMETER data set

The extensive multi-model ensemble seasonal hindcast data set, produced by the DEMETER project, has been used for a comprehensive assessment of the multi-model approach. The DEMETER prediction system comprises the global coupled ocean–atmosphere models of the following institutions: the European Centre for Research and Advanced Training in Scientific Computation, France (CERFACS); Centre National de Recherche Météorologiques, France (CNRM); the European Centre for Medium-Range Weather Forecasts (ECMWF); Istituto Nazionale de Geofisica e Vulcanologia, Italy (INGV); Laboratoire d'Océanographie Dynamique et de Climatologie, France (LODYC); Max-Planck Institut für Meteorologie, Germany (MPI); UK Met Office (UKMO). In order to assess seasonal dependence on forecast skill, the DEMETER hindcasts have been started from 1 February, 1 May, 1 August and 1 November. The atmospheric and land-surface initial conditions are taken from the ECMWF Reanalysis (ERA-40) data set. The ocean initial conditions are obtained from ocean-only runs forced by ERA-40 fluxes, except in the case of MPI that used a coupled initialization method. Ocean observations have been

assimilated only in the UKMO ocean-only run after 1987. Each hindcast has been integrated for six months and comprises an ensemble of nine members. All seven models have been run for the common period of 1980–2001, although some of the models have been integrated over an even longer period (1958–2001). In order to compare single-model and multi-model results for the same period, in this study only the 22 years of the common period have been used.

2.2. *Diagnostic and evaluation tools*

The diagnosis of model results is a crucial step in assessing (and improving) model performance. Many aspects of the model performance are not independent of the chosen diagnostic. Thus, in order to answer scientific questions such as those discussed in the introduction, a broad range of diagnostic and evaluation tools has to be applied. Furthermore, both deterministic and probabilistic skill measures have to be considered. In the DEMETER project, a comprehensive verification system to evaluate the forecast quality of all DEMETER single-model and multi-model ensemble systems has been developed at the ECMWF. This system calculates a common set of verification diagnostics based on World Meteorological Organization (WMO) standards. The basic set of diagnostics can be accessed online (http://www.ecmwf.int/research/demeter/verification). This website contains the following sections.

(i) Global maps and zonal averages of the single-model bias. Hindcast anomalies are computed by removing the model climatology for each grid point, each initial month and each lead time from the original ensemble hindcasts. A similar process is used to produce the verification anomalies.

(ii) Time series of specific climate indices, e.g. related to area-averaged sea surface temperatures (SSTs), precipitation and circulation patterns.

(iii) Standard deterministic ensemble-mean scores, such as anomaly correlation coefficient (ACC), root mean square skill score (RMSSS) and mean square skill score (MSSS).

(iv) Probabilistic skill measures, such as reliability diagrams, relative operating characteristic skill score (ROCSS), Brier score, ranked probability skill score (RPSS) and potential economic value curves.

(v) Comparison of single-model and multi-model ensemble skill using scatter diagrams of area-averaged skill measures and probability density functions (PDFs) of grid-point skill scores.

All diagnostics shown in this study are based on results from this verification system.

3. Multi-model versus single-model results

Before trying to find an explanation for the claimed superiority of the multi-model, a comprehensive documentation of the differences between the single-model and multi-model performances has to be given. From a scientific point of view, it seems that a fair comparison between single-model and multi-model ensembles can only be made with same ensemble sizes. However, from an operational point of view, it also makes sense to compare existing single-model forecasts (with an ensemble size which can be afforded by a single operational centre) with a multi-model consisting of pooling together various such single models. That is, in an operational environment, the comparison of ensembles with different sizes makes sense, because it demonstrates the integrated advantages of the multi-model approach compared to using a single-model ensemble from a single operational centre. However, in order to separate differences caused by the increased ensemble size of the multi-model and differences caused by using more than one model, not only the original nine-member ensemble single models are compared to the multi-model, but also results of a single model with the same number of ensemble members like the multi-model will be studied.

3.1. *Consistency*

The scientific basis for seasonal prediction relies on the fact that the lower boundary conditions can be a major source of predictability in the atmosphere (Palmer and Anderson, 1994). Therefore, a first step in assessing the performance of seasonal forecast models is often to look at the skill of SST predictions, in particular in the tropical Pacific. A simple deterministic skill measure, such as the ACC, can give a first impression of general model performance and specific differences in single-model and multi-model skill. Comparing the single-model and multi-model ACCs for different lead times in different seasons gives a first hint of the range of possible improvements which can be achieved with the multi-model (Fig. 1). For the case here considered, seasonal SSTs in the Niño-3.4 area, in general the differences are negative, i.e. mostly the multi-model ACC is superior to the single-model ACC. The greatest relative differences can be found for the hindcasts starting in February and May, which is the period with least skill due to the well-known 'spring barrier' (e.g. Balmaseda et al., 1995). On average, the ACC of the single models is around 10% below that of the multi-model, although in particular for the three-month lead time summer hindcast (MJJ, February start date) the relative differences vary much more and reach values over -40% for the worst single model. For the August and November start dates, the overall performance of the models is already quite high, and the multi-model results are mostly not much improved compared to the single models. There are even cases in which the single-model performance is better or equal to that of the multi-model.

Furthermore, when comparing only the best single model in each season and lead time with the respective multi-model result, the difference is at maximum -8%, which does not seem to be a very significant improvement. However, two important points have to be considered when judging the gain from using a multi-model compared to a single model. First and most importantly,

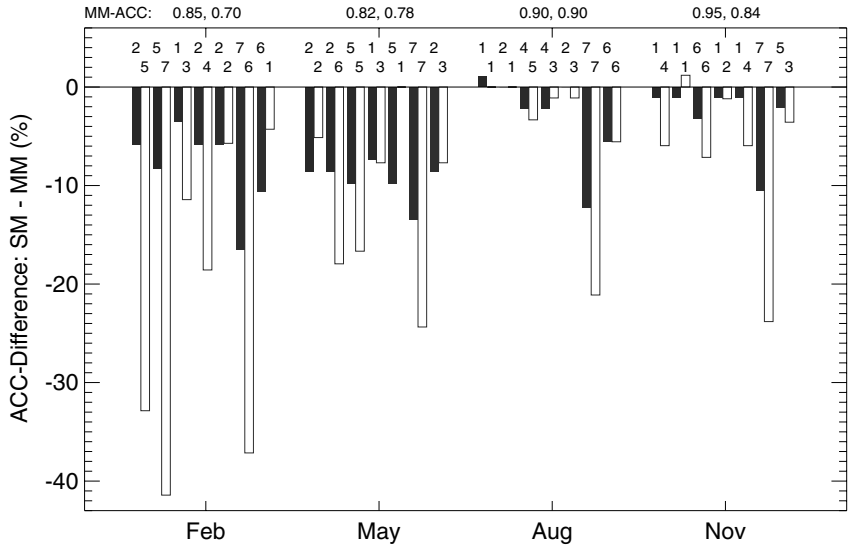

Fig 1. Relative difference between the seasonal SST ACC of the seven single models and the multi-model. Values are calculated for the Niño-3.4 area and the start dates from 1980 to 2001. Results for one-month lead time (filled bars) and three-month lead time (open bars) are shown for each season corresponding to the four start dates per year. The one-month and three-month baseline values of the multi-model are given on top of the figure. The ranking of the seven single models is added above their respective bars, with '1' marking the best model, i.e. the single model with the least difference to the multi-model value (negative differences correspond to a single-model ACC lower than the multi-model ACC).

when comparing single-model and multi-model skill in this way, we do not take into account that generally the identity of the best single model varies between different seasons, lead times, etc. For the eight cases of two lead times and four seasons considered here, all single models, except model 6, are at least once the best model. That is, it is very difficult to talk about the best single model because the best model in one case can be the worst model in another situation. For example, model 2 is the best model for the three-month lead time November and August start dates but the worst for the February start date. For a fair judgment, the single model identified as the best model across the whole range of cases should be compared to the multi-model. If, for example, model 1, which is ranked the best in three cases, is chosen as the best single model and compared with the multi-model in all eight cases, much greater gains than the above-mentioned maximum −8% can be found. That is, model 1 has a good performance in many – but not all – cases, so that, for example, for the three-month lead time February start date, a greater relative difference of −30% occurs. This points out that the main advantage of using a multi-model is not the small improvement compared to the respective best single model in individual cases, but rather the consistently better performance of the multi-model when considering all aspects of the predictions.

Other aspects, besides the season and lead time of the forecast, influencing the performance of the models and with it the identity of the best single model, are the area considered and the predicted parameter (Fig. 2). The first thing to note in Fig. 2 is the increased relative difference when considering extratropical areas. Although these improvements of up to 80% seem to be very significant on first glance, it has to be kept in mind that they are relative to a dramatically reduced overall skill. That is, it can be questioned whether variations in the ACC with a baseline around 0.3 are noteworthy or not. Apart from these partly drastically increased differences, again only little

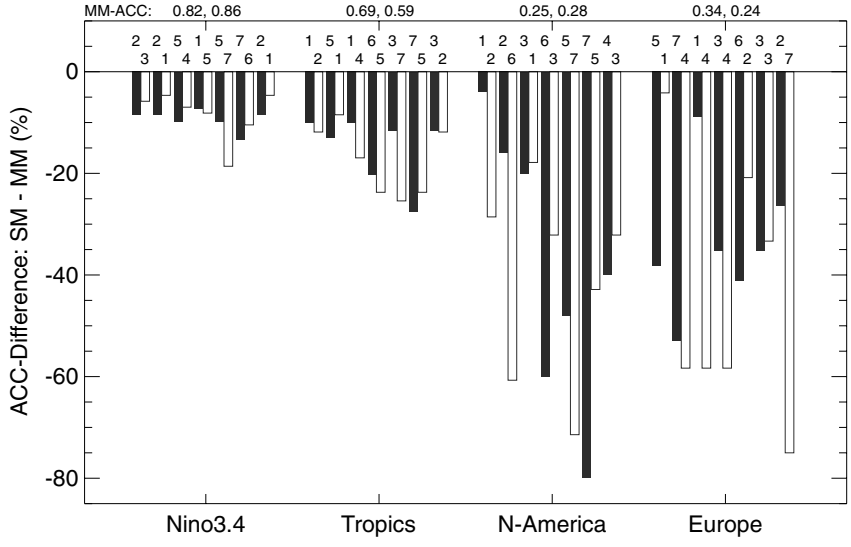

Fig 2. Relative difference between the ACC of the seven single models and the multi-model. Values are calculated for the one-month lead time seasonal mean of the May start dates from 1980 to 2001. Results for the parameter SST (filled bars) and MSLP (open bars) are shown for four different areas (Niño-3.4, Tropics, North America and Europe). The baseline values of the multi-model are given on top of the figure. The ranking of the seven single models is added above their respective bars, with '1' marking the best model, i.e. the single model with the least difference to the multi-model value (negative differences correspond to a single-model ACC lower than the multi-model ACC).

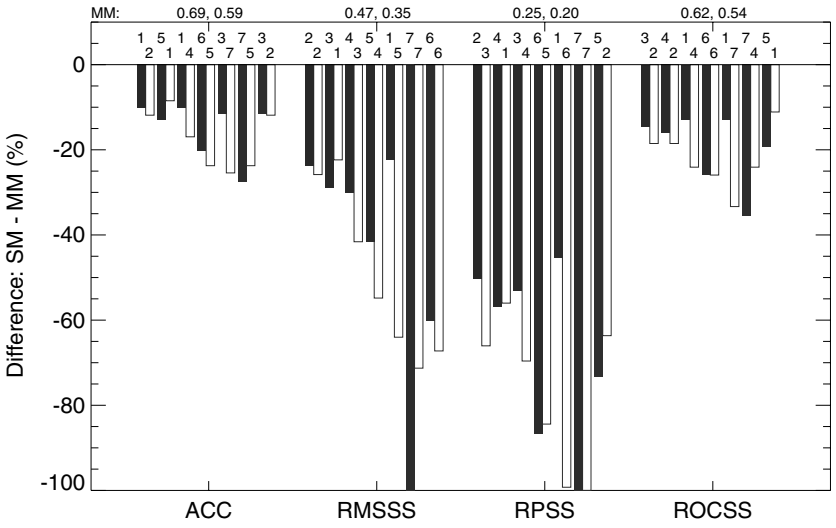

Fig 3. Relative difference between single and multi-model skill measures. Values are calculated for the one-month lead time seasonal mean of the May start dates from 1980 to 2001. The four skill measures (ACC, RMSSS, RPSS and ROCSS) are shown for the parameter SST (filled bars) and MSLP (open bars). The baseline values of the multi-model are given on top of the figure. The ranking of the seven single models is added above their respective bars, with '1' marking the best model, i.e. the single model with the least difference to the multi-model value (negative differences correspond to a single-model performance lower than the multi-model performance).

improvement is achieved when the multi-model is only compared to the respective best single model in each area and for each parameter. However, considerable variability in the identity of the best single model occurs here as well. Of particular interest is the difference in the performance of model 6 for the different parameters SST and mean sea level pressure (MSLP). Although this model has a quite low skill in predicting SST anomalies for most of the regions shown (and also for all seasons and lead times; see Fig. 1), its skill for Northern Hemisphere extratropical MSLP predictions (especially Europe) lies in the range of the most successful single models. Even if here the overall skill level is only marginal, the changes in the relative ranking of the single models point out that even an apparently poor model can add skill in other aspects of the prediction.

The second important consideration when assessing the impact of the multi-model approach is related to the choice of metric used in the diagnostic. Until now, only one particular diagnostic, the ACC, has been used to demonstrate the greater consistency of the multi-model when considering different seasons, parameters, etc., of the forecast. However, a whole range of different diagnostics exists and each of these scoring methods focuses on different aspects of the model performance. That is, the ranking of a model as best, second best, . . . , worst, depends also on the chosen score (Jolliffe and Stephenson, 2003). In order to demonstrate the degree of variability (or consistency) in the ranking of the models when applying different skill scores, Fig. 3 shows the relative differences between single model and multi-model for two deterministic measures (ACC and RMSSS) and two probabilistic skill scores (RPSS and ROCSS) [1] for both SST and MSLP. Some features of the ranking are consistent across the whole range of used skill measures. First, independent of the score applied, the differences are always negative, i.e. the multi-model results are in every case superior to all single models. Secondly, the SST performance of model 6 is ranked the worst with every diagnostic applied. However, some differences in the ranking that depend on the metric applied can be observed. For example, the ranking of model 7 varies between being the best single model in terms of ROCSS and the second worst in terms of RMSSS for the MSLP performance. The reason for the low RMSSS values for model 7 (and model 6) lies in the overactivity of these two models, which is strongly penalized by the RMSSS.

Another apparent feature in Fig. 3 is the greater relative improvement of the multi-model associated with the RMSSS and RPSS. Again, these are the skill scores with the lower absolute values compared to ACC and ROCSS. However, judging the gain of using a multi-model based only on one metric can lead to very different results. For example, relative differences of over 100% (which occur three times for RMSSS and RPSS) correspond to a reduction of the positive multi-model score to a negative value for the single model. This in turn implies an improvement from a non-skilful (worse than using climatology) single model to a skilful (better than using climatology) multi-model, which cannot be seen with the ROCSS diagnostic. Thus, for a comprehensive assessment of model performance, it is absolutely necessary to use more than one skill measure.

3.2. *Reliability and resolution*

The greater consistency of the multi-model is only one part of the explanation of the improved multi-model performance. Another, more specific aspect is the improved reliability of the predictions, with reliability having a precise technical meaning in this context. A forecast system is called reliable if the predicted probability of an event matches its frequency of occurrence when it was forecasted. That is, when considering all cases where an event is predicted to occur with a 40% probability, this event

[1] Skill scores measure the quality of a forecast system relative to a reference system (here climatology), with positive/negative scores indicating a performance better/worse than the reference. For a more detailed definition of skill scores and the RPS score and ROC skill score, in particular, see, for example, Jolliffe and Stephenson (2003) and references therein.

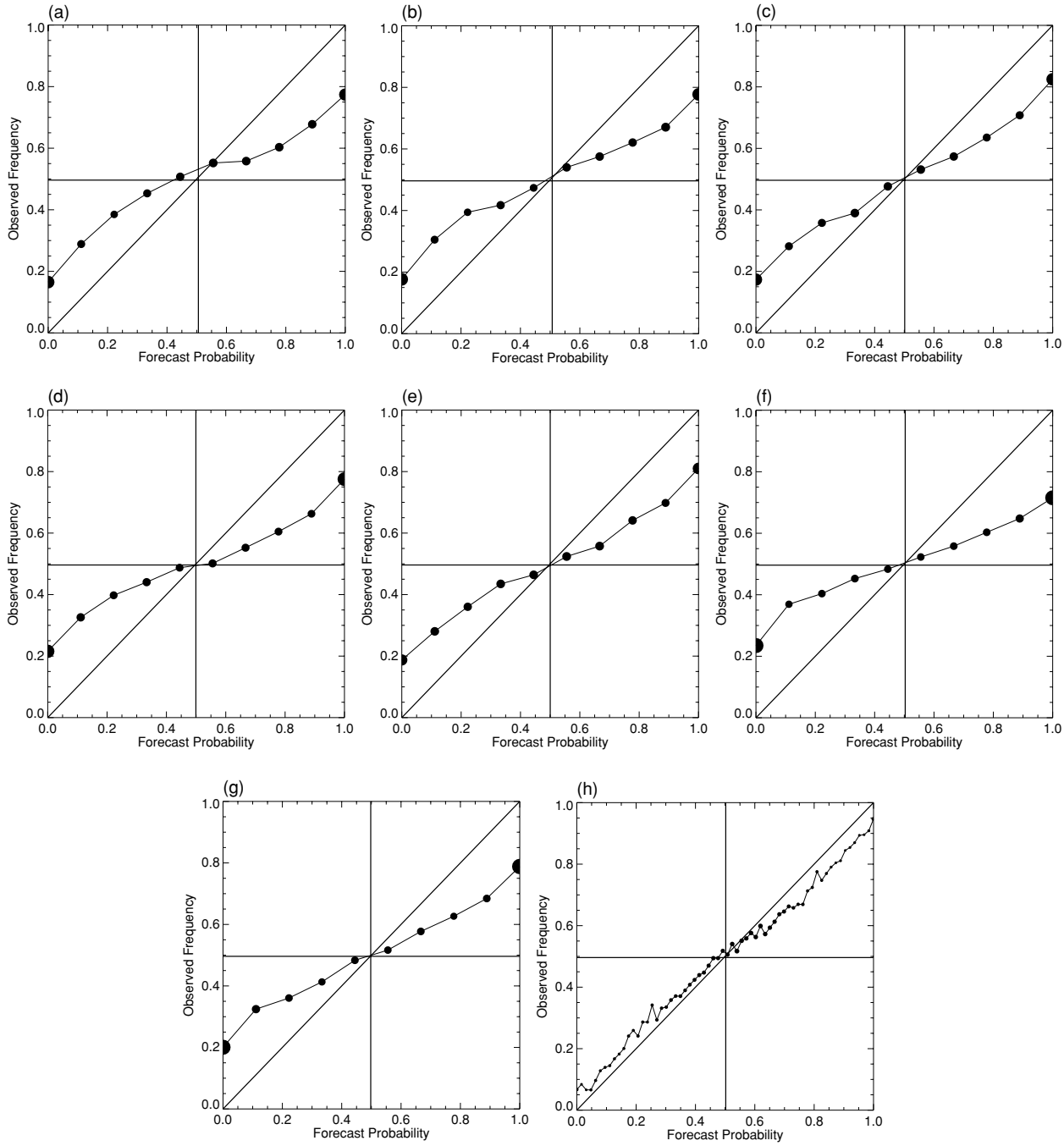

Fig 4. Reliability diagrams for the positive anomalies of seasonal averages of the 2-m temperature in summer (May start date, one-month lead time) averaged over the tropical band (±30°) for the period 1980–2001. The straight horizontal and vertical lines display the average observed frequency and forecast probability of the event. The size of the bullets represent the relative forecast frequency. The seven single model results are given in (a)–(g), the multi-model reliability diagram is shown in (h).

should verify in exactly 40% of these cases, not less and not more. As shown in the previous section, the multi-model superiority caused by greater consistency is mainly achieved by taking into account various aspects of the prediction such as season, lead time, etc. In contrast to that, major improvements in reliability can be found for each of these individual aspects of the predictions. As one example (out of many), the reliability diagrams of the seven single models as well as the multi-model are shown in Fig. 4 for the seasonal averages of the 2-m temperature in summer (May start date, one-month lead time) averaged over

the tropical band (±30°). The reliability diagram displays the accumulated proportion of forecast probabilities versus the accumulated observed frequency of the event. Every single-model ensemble proves to be overconfident, which is characterized by a too shallow slope of the line joining the points in the diagram (Figs. 4a–g). On the other hand, the reliability diagram for the multi-model ensemble fits much better the diagonal (Fig. 4h). This implies that, given a prediction with a specific probability, the multi-model will verify on average the same proportion of observed events, while the single-model ensembles will assign low (high) probabilities to cases that are observed a higher (lower) proportion of times.

Another aspect of the quality of a probabilistic forecast system is its resolution. In this context, resolution describes the ability of a forecast system to discriminate between situations that lead to different events in the future. The greater the difference between the correctly assigned forecast probability and the climatological probability of a particular event, the better is the resolution of the forecast system. Both reliability and resolution are the main components of the Brier score, which is the corresponding probabilistic score to the RMS for deterministic forecasts. In order to assess the performance of the single model and multi-model relative to the use of climatology, the Brier skill score (BSS) and its reliability and resolution components (BSS_{rel}, BSS_{res}) are given in Table 1 for the same case as in Fig. 4. It can be seen that not only the reliability component but also the resolution of the multi-model are improved compared to the single models. This results in a significant improvement of the BSS with a distinct value above zero.

In order to give a comprehensive comparison of the single-model and multi-model performances, a number of BSSs (and their reliability and resolution components) have been collected and displayed in the three scatter diagrams in Fig. 5. The diagrams contain the values of eight different regions (Northern Extratropics, Tropics, Southern Extratropics, North America, Europe, West Africa, East Africa, South Africa), all four start dates, both one-month and three-month lead times, and four different events (anomalies in the upper/lower tercile and anomalies

Table 1. BSS, the reliability component of the BSS (BSS_{rel}) and the resolution component of the BSS (BSS_{res}) for the seven single models as well as the DEMETER multi-model. Values are calculated for the seasonal averages of the 2-m temperature in summer (May start date, one-month lead time) averaged over the tropical band (±30°) for the period 1980–2001

	BSS	BSS_{rel}	BSS_{res}
Model 1	0.039	0.899	0.141
Model 2	0.039	0.899	0.140
Model 3	0.095	0.926	0.169
Model 4	−0.001	0.877	0.123
Model 5	0.065	0.918	0.147
Model 6	−0.064	0.838	0.099
Model 7	0.047	0.893	0.153
DEMETER	0.204	0.990	0.213

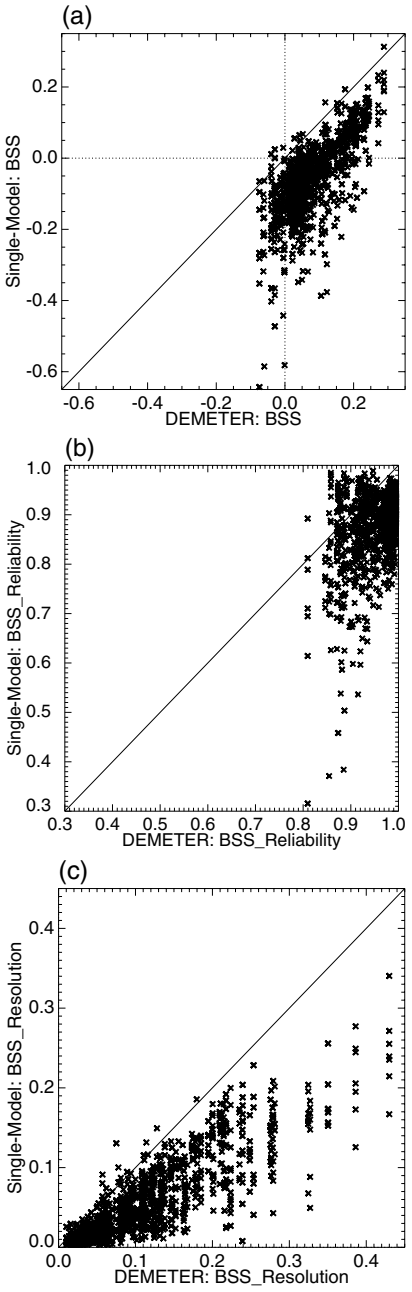

Fig 5. Scatter plots of multi-model versus single-model diagnostics of 850-hPa temperature predictions, collected over eight regions (Northern Extratropics, Tropics, Southern Extratropics, North America, Europe, West Africa, East Africa and South Africa), four start dates (February, May, August and November), two lead times (one-month and three-month), and four events (prediction of upper/lower tercile, positive/negative anomalies). (a) BSS; (b) reliability component of BSS; (c) resolution component of BSS.

above/below the mean). In order to demonstrate that the multi-model improvements are not confined to surface parameter, the results for the parameter 850-hPa temperature have been chosen. The superiority of the multi-model approach is overwhelming because most of the points (99%, 92% and 98% for BSS, BSS_{rel} and BSS_{res}, respectively) are found below the diagonal, which indicates higher scores and a better performance of the multi-model. In addition, the BSS is in many fewer cases negative for the multi-model, i.e. many more cases with skill above climatology exist for the multi-model (Fig. 5a). The magnitude of improvements shows a high variability, from moderate improvements to some cases with an extraordinary increase in skill (e.g. BSS_{rel} rises from 0.4 for a single model to 0.9 for the multi-model).

3.3. Ensemble size

In spite of the clear improvement of the multi-model ensemble performance over the single-model ensembles, an important question arises. Is the improvement in the multi-model ensemble merely due to increased ensemble size or does the additional information from different models add to the performance? In order to separate the benefits that derive from combining models of different formulation to those derived simply from the accompanying increase in ensemble size, a 54-member ensemble hindcast has been generated with the ECMWF model alone for the period 1987–1999 using the May start date. Figure 6 shows the reliability diagram for the same case as in Fig. 4 (one-month lead positive anomalies of 2-m temperature in summer over the tropical band, $\pm 30^{\circ}$), but here for the 54-member single-model ensemble and an equally sized multi-model ensemble. The multi-model ensemble for this example was constructed by randomly selecting 54 members out of the 63 available from the seven single-model ensembles. Although the increase in ensemble size in the single model results in improved reliability compared to the nine-member ensemble predictions (Fig. 4), it still does not outperform the multi-model with the same ensemble size. Both reliability and resolution, as well as BSS are still below the values of the multi-model (Table 2). This indicates that the additional information coming from the other single models adds to the improvement seen in the multi-model results.

The different rate of increase in skill related to adding more ensemble members, either from different models or the same model, can be seen in Fig. 7. As expected, and already demonstrated in Kumar et al. (2001) for the case of single models, for both single-model and multi-model hindcasts, the skill generally increases when adding more ensemble members. In the case of the 18-member/two-model ensembles, the multi-model skill varies considerably depending on the quality of the contributing single models. Combining two of the poorer models leads to a lower RPSS compared to the 18-member single-model ensembles constructed from one of the best single models. However, already in the case of the 27-member/three-model ensembles, the single-model and multi-model results are well separated. Every multi-model combination of three single models beats all single-model realizations with the same ensemble size. Furthermore, the gap between single model and multi-model increases even more when including further models into the multi-model, although it seems to stabilize with six or more models included.

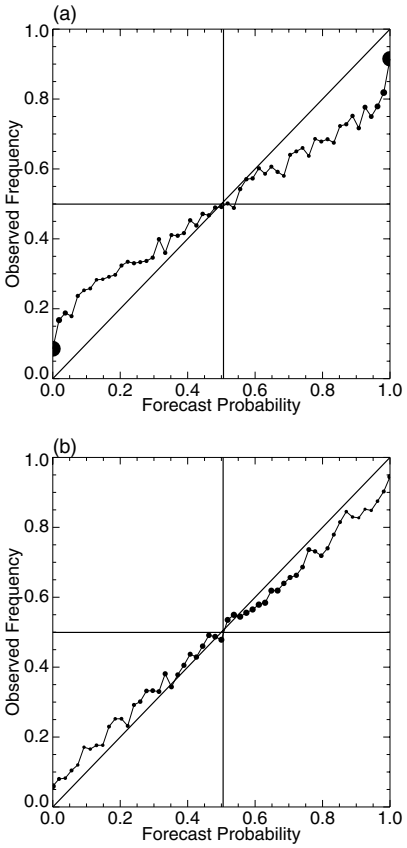

Fig 6. As Fig. 4 but for the period 1987–1999: (a) single model; (b) multi-model. Both ensembles consist of 54 members.

Table 2. As Table 1, but for the period 1987–1999 and both ensembles consisting of 54 members

	BSS	BSS_{rel}	BSS_{res}
Single model (54 members)	0.170	0.959	0.211
Multi-model (54 members)	0.222	0.994	0.227

4. Rationale behind the multi-model superiority

4.1. Conceptual background

As documented in the previous section, the multi-model approach improves both deterministic and probabilistic performances of seasonal predictions compared to single-model forecasts. This success of the multi-model approach requires clarification of how and why the multi-model works.

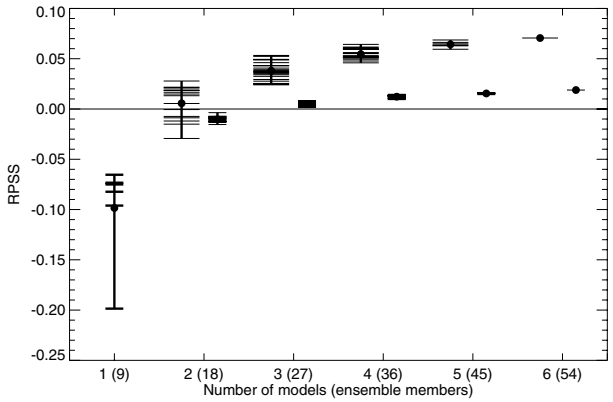

Fig 7. RPSS of one-month lead time summer precipitation hindcasts for the period 1987–1999, calculated over the tropical band (±30°). The RPSS of the single models is given in the first column, each horizontal bar representing the value of one single model with nine ensemble members. The next columns of wide horizontal bars mark the RPSS of all possible multi-model combinations composed of 2, 3, 4, 5 and 6 models. The slim horizontal bars beside the wide multi-model bars mark the RPSS of a single model with the same ensemble size as the respective multi-model (18, 27, 36, 45 and 54). For each multi-model realization, a single model was constructed by randomly choosing the same number of members as in the corresponding multi-model.

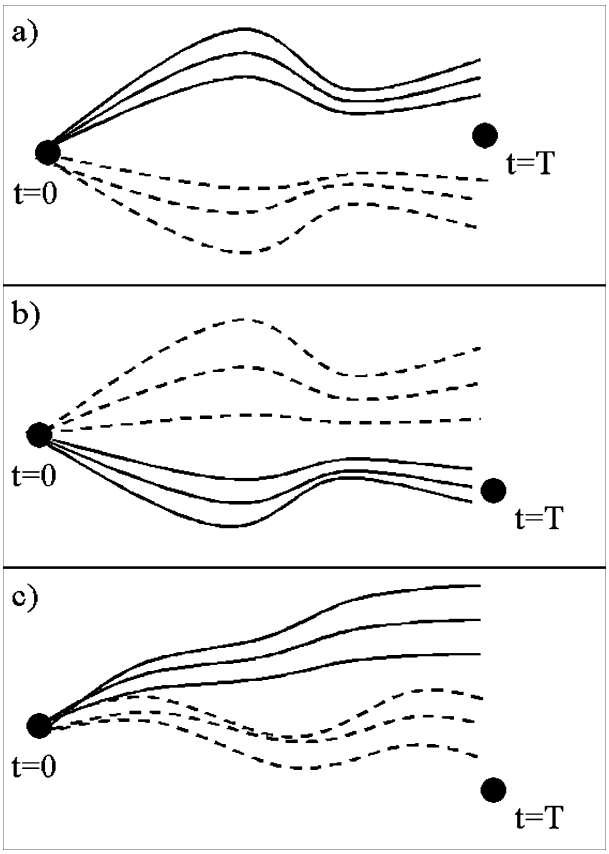

Fig 8. Idealized visualization of basic multi-model scenarios. For the sake of simplicity, only two single models and three ensemble members are included: model 1 (solid lines) and model 2 (dashed lines). (a) The multi-model provides the best prediction; (b) a single model provides the best prediction; (c) the verification lies outside the model predictions.

Every attempt to represent nature in a set of equations, resolvable on a digital computer, inevitably introduces inaccuracy. That is, although the equations for the evolution of climate are well understood at the level of partial differential equations, they have to be truncated to a finite-dimensional set of ordinary differential equations, in order to be integrated on a digital computer. The inaccuracies introduced by this process can in principle propagate upscale and infect the entire spectrum of scales being predicted by the model. The basic idea of the multi-model concept is to account for this inherent model error in using a number of independent and skilful models in the hope of a better coverage of the whole possible climate phase space. However, how, when and why does this better coverage lead to improved predictions? An idealized visualization of 'how and when' the multi-model provides better results than a single model is provided in Fig. 8. For the sake of simplicity, only two single models and three ensemble members are included in this sketch.

Comparing the performance of single model and multi-model in detail, three basic scenarios have to be considered. In the first case, the single-model ensembles lie below and above the verification, i.e. the resulting multi-model ensemble is improved because of error cancellation (Fig. 8a). This is the most obvious – but not only – reason for the multi-model superiority. Certainly, this (for the multi-model) optimal situation does not occur all the time. Thus, other reasons for the superiority must exist. The second principle scenario is that one single-model ensemble provides the best prediction (Fig. 8b) and compared to this optimal solution, the multi-model ensemble can only be worse. However, the multi-model is still improved compared to the other single model, in this case model 2. The third general possibility is that the verification lies beyond all single-model predictions, i.e. both ensembles cover an area towards only one side of the verification (Fig. 8c). In this case, the multi-model solution constitutes an improvement with regard to model 1 but a deterioration compared to model 2. However, as demonstrated in the previous section, the identity of the best model varies depending on which aspect of the forecast is considered (i.e. it is also possible that the multi-model is improved compared to model 2 but worse compared to model 1). Note that for such individual cases, it is impossible that the multi-model is worse than every single-model contribution, because starting from the worst model, every additional information always brings the prediction nearer to the verification. Despite the fact that in reality many more variations of these three basic scenarios occur (e.g. overlapping ensembles of single models), the general conclusions from considering these cases are still valid.

When assessing the difference in the performance between single model and multi-model, the diagnostic is calculated by collecting cases over a certain area, time range, etc. The resulting overall difference consists of five contributions related to the above mentioned three basic scenarios:

$$\Delta \mathrm{MS}_i = a_i + b_{\mathrm{p}i} - b_{\mathrm{n}i} + c_{\mathrm{p}i} - c_{\mathrm{n}i}. \quad (1)$$

Here, i is the index of the single model, $\Delta \mathrm{MS}_i$ is the difference in performance between multi-model and single-model i, a_i is the difference caused by scenario a cases, $b_{\mathrm{p}i}$ is the positive difference caused by scenario b cases, $b_{\mathrm{n}i}$ is the negative difference caused by scenario b cases, $c_{\mathrm{p}i}$ is the positive difference caused by scenario c cases, and $c_{\mathrm{n}i}$ is the negative difference caused by scenario c cases.

It is obvious that, as long as the cases with worse multi-model scores than the particular single model ($b_{\mathrm{n}i}$ and $c_{\mathrm{n}i}$) are balanced by the contributions from the remaining scenarios (a_i, $b_{\mathrm{p}i}$ and $c_{\mathrm{p}i}$), the overall difference in the performance will remain positive. For individual aspects of the forecast system, $\Delta \mathrm{MS}_i$ of a particular model can be dominated by $b_{\mathrm{n}i}$ and $c_{\mathrm{n}i}$ and therefore negative. However, when considering enough cases, even such supposedly better models will have failures and benefit from the then better performance of the other models. That is, the main reason 'why' the better coverage leads to improved predictions is the greater consistency of the multi-model in the long run.

4.2. *Examples*

In order to demonstrate how the above-outlined idealized scenarios are realized in the real data set, a typical example of each scenario is given in Fig. 9. The data shown are not averaged over larger areas or seasons, but represent the raw model output of monthly anomalies at grid-point level. In the first example (Fig. 9a), the ensemble members of the single models are distributed around the verification in such a way that the multi-model ensemble mean coincides exactly with the verification. In terms of the simple deterministic ranking metric shown in the diagram (difference between ensemble mean and verification), the multi-model scores the best. Also, when comparing the performance in probabilistic terms, improvements in the multi-model can be found. For example, four out of the seven single models assign zero (or negligible) probabilities to the value of the verification, whereas the multi-model assigns nearly equal probabilities for the anomaly to be above or below the verification (Fig. 9b). This case is an example of an improvement caused by error cancellation, i.e. some of the single models predict a too weak anomaly, and the other models predict a too strong anomaly, resulting in an improved multi-model prediction. Although the multi-model PDF is not as sharp as the single-model PDFs, it is much more likely to contain the verification when considering more than one case.

This can be seen also in the example of the second idealized case, in which two of the single models – but not the same as in the above example – are superior to the multi-model (Figs 9c and d). Here, models 6 and 7 give very good predictions, whereas the remaining five single models do not cover the verification at all and even predict an ensemble-mean anomaly of the opposite sign. Due to these five unsuccessful predictions, the multi-model performance – in deterministic and probabilistic terms – is worse than models 6 and 7. However, compared to the five unsuccessful single models, the multi-model performance is improved, in particular in probabilistic terms. Imagine a user, whose decision-making process is critically dependent on her knowledge of whether the temperature anomaly will be above a certain threshold (e.g. 1 K) or not. If this user were to use models 6 or 7, she would very confidently make the right decision because the models predict with 100% probability the anomaly to be above 1 K. On the other hand, if the user were to base her decision on one of the first five models, she would very confidently make the wrong decision because none of these models assigns any probability to an anomaly above 1 K. However, as shown throughout the paper, the main point is that the identity of the best single model changes depending on the aspects considered. Using models 6 or 7 results in the right decision in this particular case, but can result in the wrong decision in another case (Figs. 9a and b). The multi-model, on the other hand, assigns – independently of which single model contains the verification – at least a certain probability to the event. That is, in the long run, decision-making based on the multi-model prediction (which, for example, in this particular case gives a probability of 30% for the anomaly to be above 1 K) will be much more successful than basing the decision on only one particular model.

Even in the third case, in which the verification lies beyond all single-model predictions (Fig. 9e), similar conclusions can be drawn. None of the single-model ensemble members predicts the negative anomaly of SSTs below -1 K, and the best single-model ensemble mean is only marginally negative. In this case, the multi-model can never be better than the best single model. On the other hand, the multi-model prediction will always be better than the worse single models. In particular, the multi-model probabilities are improved compared to the single-model probabilities, with four single models having lower probabilities for negative anomalies compared to only two single models having higher probabilities and one single model with the same probability as the multi-model (Fig. 9f).

The advantages of the probabilistic multi-model forecasts become even clearer when comparing the maximum errors in the predicted probabilities between single-model and multi-model ensembles. Figure 10a shows the probability map of the occurrence of positive MSLP anomalies in the two-month lead time January 1998 multi-model prediction. The typical El Niño pattern can be seen in the high probabilities for positive MSLP anomalies over the western Pacific, and low probabilities in the

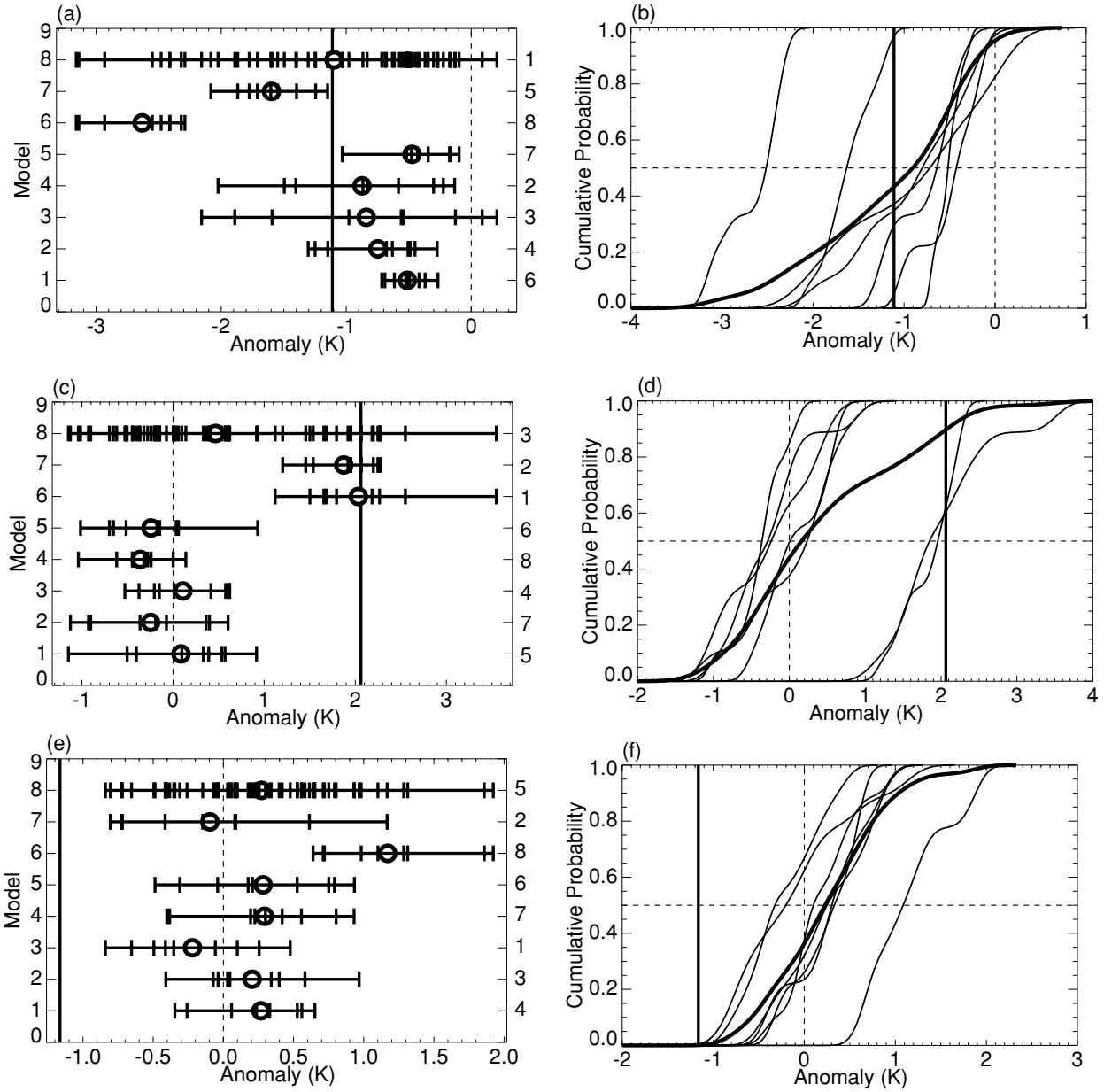

Fig 9. Practical examples corresponding to the idealized scenarios of the multi-model ensemble concept (Fig. 8). The vertical bold line in every plot represents the verification. (a) One-month lead seasonal SST hindcasts for JJA 1988 at a single grid point in the tropical Pacific. The horizontal lines represent the ensemble spread, with the vertical bars corresponding to the individual ensemble members and the open circles marking the ensemble means. The ranking of the models (in terms of the absolute error of the deterministic ensemble mean) is added at the right-hand side of the graph. (b) Corresponding cumulative PDFs to (a); the thick line marks the multi-model PDF. (c) As in (a) but for JJA 1987. (d) Corresponding cumulative PDFs to (c); the thick line marks the multi-model PDF. (e) As in (c) but for a single grid point in the North Pacific. (f) Corresponding cumulative PDFs to (e); the thick line marks the multi-model PDF.

eastern part. The main effect of pooling the single models together to the multi-model probabilities becomes obvious when looking at the global maps of the differences between predicted probabilities and the verification. In the case of the single models (Figs. 10b–h), large areas of extreme differences occur. Such extreme differences correspond to situations when most of the ensemble members of a single model predict negative anomalies in spite of positive anomalies in the verification (dark blue areas) or vice versa (dark red areas). In contrast, many fewer cases with such extreme misses occur for the multi-model probabilities

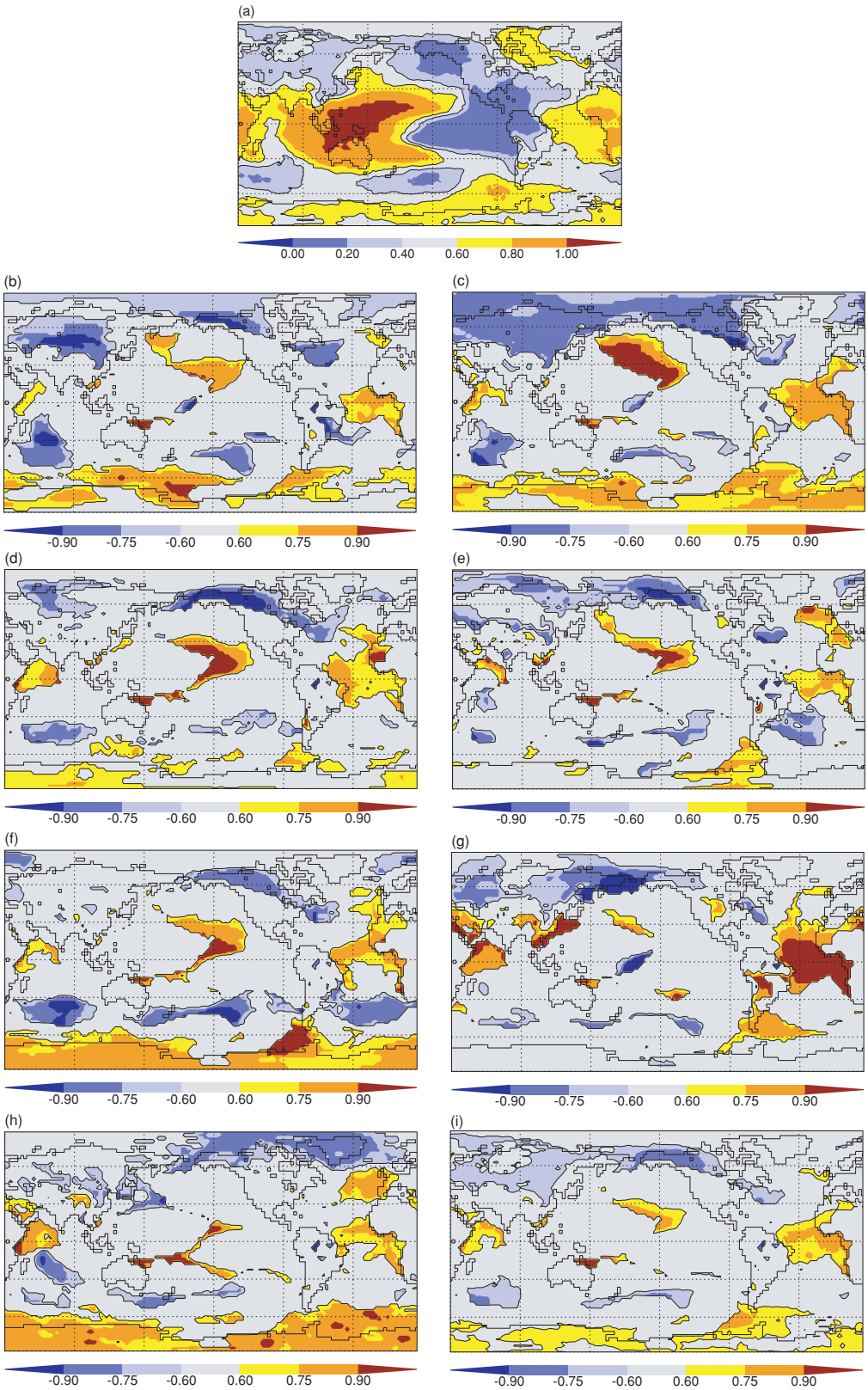

Fig 10. (a) Probability map of the occurrence of positive MSLP anomalies in January 1998 (two-month lead of integrations with start date November 1997) in the multi-model. (b)–(h) Difference between predicted probabilities and verification (0 = no occurrence; 1 = occurrence) for all single models 1–7. (i) Same as (b)–(h), but for the multi-model ensemble.

(Fig. 10i). In most areas the differences are at least below 70% or even lower, and differences above 90% are extremely rare. The reduction in the differences is caused by improving 'poor' single models with predictions from better models. That is, an error seen in a particular single model can only be reduced when other single models perform better. This is the case, for example, for the strong negative error in the central Pacific for models 1, 2 and 6, which is reduced below 50% in the multi-model. On the other hand, when all single models tend to have similar patterns of errors, as for example in the tropical Atlantic, only minor error reduction occurs. However, the comprehensive analysis of various skill scores, showing the improved performance of the multi-model, indicates that cases with error reduction prevail over cases in which no error reduction can be achieved.

These practical examples of how the idealized cases are realized and what are the effects, both in deterministic and probabilistic terms, have demonstrated 'how, when and why' the multi-model results can be superior to single-model results. However, with these idealized cases and examples it has not been proven that the multi-model concept gives improved results under all circumstances. It is of course possible to construct a scenario in which one model is superior to all other components of the multi-model system in every aspect considered. In such a hypothetical scenario, adding information of always inferior models would lead to multi-model predictions worse than the always superior single model. That is, the key to the success of the multi-model concept lies in combining independent and skilful models, each with its own strengths and weaknesses. On the other hand, one might argue that today's global models are not necessarily independent of each other. However, the differences in the error characteristics, shown by the components of the DEMETER multi-model system, support the assumption of sufficient independence in this case. Therefore, when designing a multi-model it seems to be worthwhile to test the level of independence and skilfulness beforehand, although it might be difficult to exactly define the level of skill and independence necessary for the single models to be able to create a successful multi-model. If two or more of the single models show a very similar error characteristic, this particular error characteristic might gain too much weight in a simple multi-model with equal weights. A possible solution for this problem could be the use of non-equal weights. Furthermore, if the forecast quality assessment detects a single model that is consistently worse than the other contributions, it should be excluded from the multi-model, although the definition of consistent lack of skill might depend on specific user requirements. Therefore, as long as the individual components are able to make a positive contribution to a relevant aspect of the prediction, the multi-model will benefit from this additional information. As such, the proof of the multi-model concept depends on the components of the system and can only be given in diagnosing real data sets, as has been done in the previous section.

5. Summary

This study was motivated by the question 'whether and why' the multi-model ensemble concept can improve single-model ensemble predictions. Since other studies (Doblas-Reyes et al., 2000) have already investigated the question 'whether' multi-model ensemble forecasts are superior to single-model ensembles – although not in coupled seasonal mode – the focus of this paper has been placed on the 'why'. However, before trying to find an explanation for the multi-model success, first a comprehensive documentation of its superiority was given. It was demonstrated that the judgement of whether the multi-model has a significantly improved performance depends strongly on the method chosen for this assessment. The degree of improvement that can be achieved with a multi-model depends not only on the aspect of the prediction considered, (i.e. which season, lead time, parameter, etc., is chosen for the assessment), but even more on the choice of reference. That is, when the multi-model is compared to the best single model in each individual assessment, the improvement does not seem to be very significant. However, when assessing the degree of superiority in such a way, we do not take into account that the identity of the best single model varies. For a fair comparison between single-model and multi-model performance, the multi-model should always be compared to the same single model (e.g. the best single model over the whole range of aspects). In this way, the key argument for employing a multi-model, its consistently better performance across all aspects of the predictions, becomes much clearer.

Furthermore, the consideration of some idealized cases and related practical examples enables us to answer the questions posed in the introduction of this paper.

(i) *How can a poor model add skill*? Indeed, if a model is consistently, over the whole range of aspects, worse than average, it cannot contribute to the multi-model skill (except in cases of error cancellation). However, it has been demonstrated that none of the components of the DEMETER multi-model system is a poor model in this sense. It has been argued that skill of the single-model components is a prerequisite of the multi-model concept. Under this assumption, the question is asked wrongly, because in this framework, no 'poor' model exists.

(ii) *How can the multi-model be better than the average single-model performance*? First, the relation between the average skill of the single models and the performance of the multi-model is not linear, in particular when considering probabilistic diagnostics. That is, averaging the skill of the seven single-model forecasts does not correspond to the skill of the combined seven single models, the multi-model forecast. Only when considering a linear metric, and if the verification always were to be beyond the single-model predictions (Figs. 9c and e), the multi-model performance would be similar to the average single-model performance. In practice, however, mostly the verification lies between different single-model predictions. Thus, error cancellation and non-linearity of the diagnostics are the main reasons

for the multi-model performance being superior to the average single-model performance.

(iii) *Why not use the best single-model instead of the multi-model*? Similar to the argument that none of the single models can be defined as a poor model, it is difficult to define *the* best single model. It has been shown that the identity of the best single model varies depending on the aspect of the prediction considered. In real life, a user has to decide beforehand which single model to choose for the decision-making process. This single model might have a better performance than the multi-model in some situations. However, in the long run the multi-model will give more reliable predictions.

All the above given conclusions have been made under the assumption of using the simple equal weight multi-model. The key argument for the success of the multi-model concept has been that the combination of the single models, with all its strengths and weaknesses, leads to more consistency and a more reliable forecast system. A logical question arising from this argument is, why do we have to combine strengths and weaknesses of the single models? Is it not possible to eliminate the weaknesses and keep only the strengths? In the above shown analysis of the performance of the single models, it seemed that, for example, the SST predictions of model 6 are often worse than average. If this turns out to be a robust feature, giving a lower weight to the SST forecasts of model 6 might be a way of improving the multi-model ensemble even more. However, this concept of applying different weights to the single models when combining them to the multi-model ensemble forecast is not as straightforward as it might seem at first glance. Various methods of finding optimal weights exist, and all constraints and pitfalls related to these methods will be the topic of the second part of this paper.

Finally, returning to the starting point of this contribution, whether more information leads to more success or 'simplicity rules the world', it has been demonstrated that, in the case of the multi-model ensemble system presented here, more information – even if it is sometimes the wrong information – leads to more success when considering the whole range of aspects of the forecast system. However, this question can also be handed over to the second part of the paper, where it will be investigated whether more advanced methods, based on more information about the past performance of the single models, can lead to an improved system, or whether the simplicity of the multi-model with equal weights can hardly be beaten.

Acknowledgments

This work was supported by the EU-funded DEMETER project (EVK2-1999-00024). The authors would like to thank the whole seasonal forecast group at the ECMWF for their invaluable scientific and technical support throughout the whole project.

References

Balmaseda, M. A., Davey, M. K. and Anderson, D. L. 1995. Decadal and seasonal dependence of ENSO prediction skill. *J. Climate* **8,** 2705–2715.

Bosart, L. F. 1975. Sunya experimental results in forecasting daily temperature and precipitation. *Mon. Wea. Rev.* **103,** 1013–1020.

Branzei, R., Tijs, S. and Timmer, J. 2000. Collecting information to improve decision-making. http://ideas.repec.org/p/dgr/kubcen/200026.html.

Clemen, R. T. and Murphy, A. H. 1986. Objective and subjective precipitation probability forecasts: some methods for improving forecast quality. *Wea. Forecasting* **1**. 213–218.

Doblas-Reyes, F. J., Déqué, M. and Piedelièvre, J.-P. 2000. Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Q. J. R. Meteorol. Soc.* **126,** 2069–2088.

Fraedrich, K. and Leslie, L. M. 1987. Combining predictive schemes in short-term forecasting. *Mon. Wea. Rev.* **115,** 1640–1644.

Fritsch, J. M., Hilliker, J., Ross, J. and Vislocky, R. L. 2000. Model consensus. *Wea. Forecasting* **15,** 571–582.

Gigerenzer, G. and Todd, P. M. 1999. *Simple Heuristics That Make Us Smart*. Oxford University Press, Oxford.

Gyakum, J. R. 1986. Experiments in temperature and precipitation forecasting. *Wea. Forecasting* **1,** 77–88.

Harrison, M. S. J., Palmer, T. N., Richardson, D. S., Buizza, R. and Petroliagis, T. 1995. Joint ensembles from the UKMO and ECMWF models. In: *ECMWF Seminar Procedings: Predictability*,Vol. 2, ECMWF, Reading, UK, pp 61–120.

Heideman, K. F., Stewart, T. R., Moninger, W. R. and Reagan-Cirincione, P. 1993. The weather information skill experiment (WISE): the effect of varying levels of information on forecast skill. *Wea. Forecasting* **8,** 25–36.

Jolliffe, I. T. and Stephenson, D. B. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, New York.

Kharin, V. V. and Zwiers, F. W. 2002. Climate predictions with multimodel ensembles. *J. Climate* **15,** 793–799.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z. et al. 1999. Improved weather and seasonal climate forecasts from multi-model superensemble. *Science* **285,** 1548–1550.

Kumar, A., Barnston, A. G. and Hoerling, M. P. 2001. Seasonal predictions, probability verifications, and ensemble size. *J. Climate* **14,** 1671–1676.

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

Palmer, T. N. and Anderson, D. L. T. 1994. The prospects for seasonal forecasting. *Q. J. R. Meteorol. Soc.* **120,** 755–793.

Palmer, T. N. and Shukla, J. 2000. Editorial to DSP/PROVOST special issue. *Q. J. R. Meteorol. Soc.* **126,** 1989–1990.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M. et al. 2004. Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Am. Meteorol. Soc.* **85,** 853–872.

Pavan, V. and Doblas-Reyes, F. J. 2000. Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamical features. *Climate Dyn.* **16,** 611–625.

Peng, P., Kumar, A., van den Dool, H. and Barnston, A. G. 2002. An analysis of multi-model ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.* **107,** doi:10.1029/2002JD002712.

Rajagopalan, B., Lall, U. and Zebiak, S. E. 2002. Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.* **130,** 1792–1811.

Richardson, D. S. 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **126,** 649–668.

Sanders, F. 1963. On subjective probability forecasting. *J. Appl. Meteorol.* **2,** 191–201.

Sanders, F. 1973. Skill in forecasting daily temperature and precipitation: some experimental results. *Bull. Am. Meteorol. Soc.* **54,** 1171–1179.

Thompson, P. D. 1977. How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.* **105,** 228–229.

Tracton, M. S. and Kalnay, E. 1993. Operational ensemble prediction at the National Meteorological Center: practical aspects. *Wea. Forecasting* **8,** 379–398.