TECHNICAL MEMORANDUM

560

Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts

F. J. Doblas-Reyes, A. Weisheimer, M. Déqué[1], N. Keenlyside[2], M. McVean[3], J. M. Murphy[3], P. Rogel[4], D. Smith[3] and T. N. Palmer

[1] Centre National de Recherches Météorologiques, Météo-France
42 Av. G. Coriolis, Toulouse, France
[2] Leibniz-Institut für Meereswissenschaften,
Düsternbrooker Weg 20, Kiel, Germany
[3] Hadley Centre, Met Office, Fitzroy Road, Exeter, United Kingdom
[4] CERFACS, 42 Av. G. Coriolis, Toulouse, France.

Submitted to the Quart J Roy Meteorol Soc

November 2008

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

Abstract

The relative merits of three forecast systems addressing the impact of model uncertainty on seasonal/annual forecasts are described. One system consists of a multi-model, whereas two other systems sample uncertainties by perturbing the parameterisation of reference models through perturbed parameter and stochastic physics techniques. Ensemble re-forecasts over 1991 to 2001 were performed with coupled climate models started from realistic initial conditions. Forecast quality varies between the systems due to the different strategies for sampling uncertainties, but also to differences in initialisation methods and in the reference forecast system. Although the multi-model experiment has an ensemble size larger than the other two experiments, most of the assessment was done using equally-sized ensembles. The three ensembles show similar levels of skill: significant differences in performance typically range between 5 and 20%. However, a nine-member multi-model shows better results for seasonal predictions with lead times shorter than five months, followed by the stochastic-physics and the perturbed-parameter ensembles. Conversely, for seasonal predictions with lead times longer than four months, the perturbed-parameter ensemble gives more often better results. Both the stochastic-physics and perturbed-parameter ensembles improve the reliability with respect to their reference forecast systems, but not the discrimination ability. Annual-mean predictions showed lower forecast quality than seasonal predictions, but a substantial number of cases had positive skill. Only small differences between the systems were found. The full multi-model ensemble has improved forecast quality with respect to all other systems, mainly from the larger ensemble size for lead times longer than four months and annual predictions.

1. Introduction

Seasonal time-scale climate predictions are now made routinely at a number of operational meteorological centres around the world, in many cases using comprehensive coupled dynamical models of the atmosphere, oceans and land surface. The non-linear nature of the climate system makes dynamical climate forecasts sensitive to uncertainty in both the initial state and the model used for their formulation (Palmer 2001). In other words, the main uncertainties at the source of forecast error are of two types:

1. semi-empirical relationships. Uncertainties in the initial conditions, which are accounted for by generating an ensemble from slightly different atmospheric and ocean analyses (Stockdale et al. 1998). The perturbations of the initial conditions are either of an optimal statistical nature (Tang et al. 2005) or based on insight into the dynamics of the physical system (Balmaseda et al. 2008).

2. Uncertainty in model formulation, due to the inability of dynamical models of climate to simulate every single aspect of the climate system with arbitrary detail (Palmer 2000). Climate models have limited spatial and temporal resolution, so that physical processes that are active at smaller scales (e.g. convection, orographic wave drag, cloud physics, mixing, etc) must be parameterized using

As a consequence of these uncertainties, an individual forecast is of limited value and, instead, sets of forecasts are carried out to predict the range of possible evolutions of climate. The ensemble method attempts to deal with uncertainties in the initial condition, while several methods to address model uncertainty have been proposed (Palmer 2000):

1. The multi-model method empirically samples errors that occur due to structural inadequacy in individual climate models by using models with different formulations and parameterizations (Palmer et al. 2004). This approach relies on the fact that global climate models have been developed somewhat independently at different climate institutes, using different numerical schemes to represent the dynamics and applying different parameterizations of physical processes.

2. Given that some of the most important model uncertainties are in the specification of the parameters that are used in the physical parameterizations (Murphy et al. 2004), the perturbed-parameter approach samples model uncertainty by creating ensembles of alternative variants of a single model in which multiple uncertain parameters are perturbed (Stainforth et al. 2005; Collins et al. 2006).

3. The stochastic-physics approach considers that processes taking place at unresolved scales are not adequately represented by the current parameterizations because, among other things, with the use of bulk formulae they assume that there is an ensemble of subgrid processes in quasi-equilibrium with the resolved-scale flow. Palmer (2001) suggested that sub-grid processes could be represented by simplified nonlinear stochastic-dynamic models as an alternative to the deterministic bulk-formula approach. Shutts (2005) and Shutts and Palmer (2007) showed that a cellular automaton scheme to introduce stochastic perturbations in the physical tendencies had a beneficial impact in a medium-range global forecast model, while Jin et al. (2007) employed a state-dependent stochastic multiplicative forcing to improve El Niño-Southern Oscillation (ENSO) simulations in a simplified model.

These three methods are, to a significant degree, complementary. Only the multimodel approach samples structural parameterisation uncertainties, whereas only the stochastic-physics approach samples uncertainties arising from the effects of unresolved sub-grid scale variability on the grid scale parameterisation outputs. The perturbed parameter approach samples a plausible range of sustained changes to the deterministic outputs of the parameterisations that are not accounted for in the stochastic-physics approach, and only to a limited degree in the multi-model ensemble. Note also that the use of initial-condition ensembles with either the multi-model or the perturbed-parameter approaches provides ensembles of simulations that sample both sources of uncertainty. The stochastic-physics approach, instead, samples both sources when an initial-condition ensemble is run with a single-model version.

A third source of forecast error arises from uncertainties in external forcing, including solar variability, changes in the distribution of volcanic aerosols, and changes in atmospheric composition by human activities (Doblas-Reyes et al. 2006). However, this is not specifically addressed in this paper.

A comparison of the forecast quality of three forecast systems that use the approaches to address model uncertainty described above is carried out in this paper. The results depend on the effectiveness of the above techniques for sampling modelling uncertainties, but also on the choice of reference forecast system used to implement the techniques (Section 4.3), and on the methods used to initialise the forecasts (Section 2). The relative merits of the three systems are highlighted and strategies to reduce model uncertainty are suggested. The comparison is performed using comprehensive estimates of potential and actual skill in an innovative way. A set of scores is computed for a large number of regions and several variables, taking into account that results might also change with lead time and that forecast quality has important seasonal variations. Furthermore, considering that the sample used is relatively small, the whole assessment is based on a thorough treatment of statistical inference to determine which features are statistically significant and might be interpreted as being robust. This approach to forecast quality assessment is necessary because long-range forecasts typically have low skill and because seasonal forecast systems show small differences in skill (Wang et al. 2008).

The reader will observe that the three approaches have been implemented on different forecast systems. The multi-model being an ensemble of opportunity, no reference forecast system can be defined. As for the perturbed-parameter and stochastic-physics approaches, there is no single-model system that has been able to implement them both as yet. However, this paper tries to give some insight on the benefits that one could expect from each approach.

A brief summary of the experiment used to illustrate the relative merits of each approach follows in Section 2. Section 3 describes the methods employed to estimate the different attributes of the forecast quality. The main results are in Section 4 and 5, with a discussion of the effect of the ensemble size in Section 6. The main conclusions are summarised in Section 7.

2. Experimental set-up

To assess the relative merits of the three approaches to address model uncertainty, a coordinated set of forecast experiments has been performed within the framework of the EU-funded ENSEMBLES project. Sets of ensemble seasonal and annual re-forecasts over the period 1991 to 2001 were carried out as part of the experiment known as Stream 1. The re-forecasts were started at 00 GMT on the 1st of May, running for seven months, and on the 1st of November, running for fourteen months. The contributions to each forecast system are:

- Multi-model: IFS/HOPE (ECMWF), ARPEGE/OPA (Météo-France), GloSea, DePreSys_IC (both UK Met Office) and ECHAM5/OM1 (IfM-GEOMAR Kiel) for the multi-model, each of them running nine-member initial-condition ensembles, which allows for a multi-model with 45 members.

- Perturbed parameter: Nine-member ensembles with DePreSys_PP (UK Met Office), each member sampling a different set of sustained parameter perturbations to the HadCM3 model.

- Stochastic physics: Nine-member ensembles with IFS/HOPE CASBS (ECMWF), each member started from different initial conditions and sampling a different time series of random stochastic physics perturbations.

The atmospheric initial conditions, including land surface (except for DePreSys), were derived from the ERA-40 reanalysis (Uppala et al. 2005), except in the case of ECHAM5/MPIOM, which generates the ensemble using lagged initial states from a sea surface temperature (SST) restored coupled simulation. Every system includes the interannual evolution of greenhouse and trace gases (CO_2, CH_4, N_2O and CFCs).

2.1. Multi-model

The IFS/HOPE forecast system (Anderson et al. 2007) used the atmospheric IFS cycle 29r2 with a horizontal truncation of T_L95 and 40 vertical levels. The ocean model has a horizontal resolution of 1°, with an equatorial refinement of 0.3°, and 29 levels in the vertical. The coupler OASIS2 is used to interpolate the fields exchanged once per day between the oceanic and atmospheric grids. Ocean initial conditions have been taken from an ocean re-analysis (Balmaseda et al. 2008). Atmospheric perturbations based on singular vectors have been applied in a similar way as in the operational medium-range ensemble forecasts (Rodwell and Doblas-Reyes 2006). IFS uses a climatological annual cycle of five types of aerosol (sea-salt, desert

dust, organic matter, black carbon) and interannual variations of total solar irradiance. This experiment will be referred to henceforth as IFS/HOPE control.

The ARPEGE/OPA model uses the cycle 24t2 of ARPEGE-IFS for the atmosphere, OPA8.2 as the ocean model and the GELATO sea-ice model (Salas y Melia 2002). The ocean and atmosphere are coupled with OASIS3.

The GloSea forecasting system (Graham et al. 2005) is based on HadCM3 (Gordon et al. 2000), but with an ocean equatorial refinement similar to IFS/HOPE. It includes the radiative effect of variable ozone concentration and interannual total solar irradiance. The effect of volcanic aerosols is included during the re-forecasts by damping the concentration at the start date with a time scale of one year.

Initial-condition uncertainties in these three systems (IFS/HOPE, ARPEGE/OPA and GloSea) are represented by generating the nine-member ensemble from different ocean initial conditions. This is achieved by creating different ocean analyses. A control ocean analysis is forced with momentum, heat and mass flux data from ERA-40, and perturbed ocean analyses are created in parallel by adding daily wind stress perturbations to the ERA-40 momentum fluxes, as described in Anderson et al. (2007). In addition, in order to represent the uncertainty in the initial SSTs, SST perturbations are added to or subtracted from the initial field at the start of the re-forecasts, including a vertical extrapolation for consistency. In the case of OPA, the initial conditions are restored towards the positively and negatively perturbed SSTs from one month before the re-forecast data to obtain coherently perturbed temperature profiles (Rogel et al. 2005).

The ECHAM5/OM1 model (Keenlyside et al. 2008) was run at T63 and 1° resolution for the ocean and atmospheric component, respectively. This system used atmospheric, ocean and soil initial conditions taken from a three-member ensemble coupled simulation where the model SST were restored to observed SST values (Keenlyside et al. 2005). The nine-member ensemble is generated by different combinations of ocean and atmosphere states from the SST-restored runs. This model includes the effect of anthropogenic sulphate aerosol and interannual total solar irradiance during the re-forecasts. The effect of volcanic aerosols is included as in GloSea.

2.2. DePreSys and the perturbed-parameter method

DePreSys (Smith et al. 2007) is based on the HadCM3 climate model. The version used here includes an enhanced representation of the atmospheric sulphur cycle. Flux adjustments are used to restrict the development of regional biases in sea surface temperature and salinity (using an improved method based on Collins et al. 2006). DePreSys contributes to the experiment in two different ways. DePreSys_IC uses HadCM3 with standard parameter settings following those of Gordon et al. (2000), contributing nine-member perturbed initial condition ensembles as part of the multi-model ensemble. DePreSys_PP generates a perturbed-parameter ensemble by sampling modelling uncertainties in poorly constrained multiple atmospheric and surface parameters. Eight versions[1] with simultaneous perturbations to 29 parameters were used, added to the standard model version to make a nine member ensemble. The eight perturbed versions were selected among a set of 128 combinations of model parameters previously used to simulate long term

[1] The set of parameters is described in
http://www.ecmwf.int/research/EU_projects/ENSEMBLES/table_experiments/pert_param_desc.html

climate change (Webb et al. 2006). The criteria used for the selection were based on 1) choosing the 16 members with the best simulation of present-day climate using a non-dimensional measure of the average distance between members in terms of both climate sensitivity and model parameter values, and 2) picking members out of the previous 16 by sampling a wide range of climate sensitivities (from 2.6-7.1ºC) and a wide range of ENSO amplitudes (diagnosed from the simulated standard deviation of monthly central equatorial Pacific SST anomalies, which range from 0.5-1.2ºC compared to the observed value of 0.8 ºC). Each of the perturbed versions of HadCM3 required a separate set of flux adjustment fields due to the effects of the atmospheric perturbations on simulated surface heat and water fluxes. In order to create initial conditions for the re-forecasts, each model version was run in assimilation mode from December 1989 to November 2001. During this integration, the atmosphere and ocean were relaxed towards atmospheric and ocean analyses (Smith and Murphy 2007), wherein values were assimilated as anomalies with respect to the model climate in order to minimize climate drift after the assimilation is switched off. The assimilation integration was itself started from an initial state taken from a simulation of 20th century climate. The forcings in the assimilation and the re-forecasts are the same, except that during the re-forecasts total solar irradiance was estimated by repeating the previous eleven-year solar cycle and volcanic aerosol was specified to decay exponentially from the concentration at the start date with a time scale of one year.

2.3. Stochastic physics

IFS/HOPE was also run with the CASBS (Cellular Automaton Stochastic BackScatter) stochastic parameterization (Berner et al. 2008). The stochastic parameterization is based on the idea of the backscatter of kinetic energy from unresolved spatial scales (Shutts 2005). At each time step, the level of dissipation associated with parameterization of convection, orographic wave drag and numerical dissipation (horizontal diffusion and semi-Lagrangian interpolation error) is calculated. A fraction of the dissipation is re-injected into the atmospheric model near the truncation scale to account for energy transfer out of the sub-grid scale and back to the resolved scale. The scales onto which this energy is backscattered are determined by a simple cellular automaton, which essentially plays the role of a stochastic number generator. The initial conditions and ensemble generation employed were as in the IFS/HOPE control experiment.

3. Forecast quality assessment

Various measures of forecast quality have been used to assess the relative merits of the three forecast systems. The scores include the anomaly correlation (ACC) and root mean square error (RMSE) of the ensemble mean and, for dichotomous probability forecasts, the Brier skill score (BSS) with respect to climatology and the relative operating characteristic (ROC) skill score (ROCSS) (Jolliffe and Stephenson 2003). All forecast quality measures have used ERA40 as the reference dataset, except for precipitation for which GPCP (Adler et al. 2003) was taken as the reference.

Ensemble forecasts have been widely used to issue probability forecasts (e.g. Richardson 2001), although they are not the only method available for this purpose (Stephenson *et al*. 2005). In the case of a dichotomous event, given an ensemble of simulations, a simple way of obtaining a probability forecast consists in computing the fraction of ensemble members for which the value of a given variable exceeds a threshold. More sophisticated methods of obtaining an estimate of the forecast probability distribution function (PDF) from the ensemble have been proposed (e.g. Roulston and Smith 2003; Stephenson *et al*.,

2005), but given the limited sample size of long-range forecasts a simple, frequentist, non-parametric approach has been used.

The BSS is a measure of the relative benefit of the forecasts with respect to using the naïve climatological probabilities. It is defined as $BSS=1-BS/BS_c$, where BS is the Brier score, defined as the sum over all forecasts of the quadratic distance in probability space between the forecast probability and an observational step function that takes the value one (zero) if the event does (not) verify, and BS_c is the Brier score of the climatological forecast. The BSS has also been decomposed into the sum of two components (Murphy 1986): the reliability (RELSS) term that measures the relative bias of conditional means, and the resolution (RESSS) term that measures the relative variance of the conditional means. Computing the forecast probabilities as a fraction of the ensemble members satisfying a threshold-based criterion implies that the maximum set of probabilities issued is determined by the ensemble size plus one. However, it is common to simplify the forecasts using a smaller number of probability categories (Doblas-Reyes et al. 2008), which in this paper is the same for all forecast systems regardless of their ensemble size. The effect of this simplification has been taken into account in the BS. The BSS decomposition used here includes two additional terms in the resolution component that account for the within-bin variance and covariance of the probability forecasts, as described in Stephenson et al. (2008).

The ROC is a signal detection curve for dichotomous forecasts obtained by plotting a graph of the hit rate (total number of correct forecasts, or hits, divided by the total number of events observed) and the false alarm rate (number of false alarms divided by the total number of events observed) over a range of thresholds. In the case of probability forecasts, these thresholds are the range of probabilities issued. The area under the ROC curve, A, is a measure of discrimination, or the variance of the forecasts conditioned on the observations. The ROC skill score (ROCSS) is defined as ROCSS=2A-1.

Attribute diagrams (Hsu and Murphy 1986; Jolliffe and Stephenson 2003) illustrate several of the forecast quality attributes defined above. They are made of a reliability diagram that allows the visualization of the reliability and resolution of a set of probability forecasts for a specific dichotomous event, and a histogram of the probability forecasts. The reliability diagram illustrates the conditional relative frequency of occurrence of the event as a function of the forecast probability. In the idealised case of infinite sample and ensemble sizes, the diagonal line represents the result for a set of forecasts with perfect reliability. If for those forecasts with probability p the average frequency the forecasts verify is different from p, the probability forecasts obtained from the ensemble are not trustworthy. This situation will appear in the diagram as a point away from the diagonal. If the corresponding curve is shallower than the diagonal, the forecast system is said to be overconfident, while if it is steeper the system is underconfident. The sum of the vertical square distance of all the points to the diagonal (weighted by the sample size of each probability category) is an estimate of the lack of reliability of the system as measured by the Brier score. In the same way, the sum of the vertical distance of the points to the horizontal line corresponding to the observed climatological frequency of the event measures the forecast resolution, i.e., the ability of the system to issue reliable forecasts different from the naive climatological probability. This means that if the reliability curve were to be horizontal, as in the case of a certain type of random forecasts or for a climatological forecast, the conditional frequency of occurrence would not depend on the forecast probabilities and the system would have zero resolution (and no skill with respect to a climatological forecast).

Every forecast quality measure has been computed taking into account the systematic error of the forecast systems. For the deterministic measures, ACC and RMSE, forecast anomalies are estimated by removing the mean over the period 1991-2002 of all the re-forecasts available for a given lead time and start date in cross-validation mode. The anomalies for the reference dataset are estimated for the same calendar period. In the case of the multi-model and the perturbed-parameter ensembles, the mean is estimated separately for each one of the single models or model versions and the anomalies are computed from the respective climate mean. For the probabilistic measures, the events are defined using percentiles of the distribution. The threshold that defines the event is chosen separately for the verification dataset and the set of re-forecasts, and is computed from all the available years for the same start date and lead time. The reader should note that the flux corrections applied in the perturbed-parameter case does not prevent the simulations from drifting over land nor constrain the model variability, making necessary estimates of a climatological distribution for the re-forecasts. This way of dealing with the systematic error is similar to the "bias-corrected relative frequency" used by Hamill and Whitaker (2006), but different from the one used in Smith et al. (2007). In the latter case, re-forecast values are bias corrected using the difference between a long climate integration and a suitable reference dataset for the same period. Anomalies are then computed with respect to the mean of the reference dataset over any specific period. The centred ACC obtained with these anomalies is lower than the one discussed in this paper, the difference being larger as the sample size is reduced. This implies that DePreSys, by having long runs performed with the same forecast system, offers the possibility of providing more accurate bias-corrected predictions than if only the re-forecasts were available. However, only predictions for anomalies have been considered in this paper and no attempt to compute bias-corrected predictions has been made for consistency with current operational activities.

The main objective of this paper being a comparison of the quality of several experiments, it is paramount to assess the robustness of the results. For the first time, statistical inference has been applied throughout for the comparison of seasonal/annual forecast experiments. Confidence intervals for the scores have been computed using a bootstrap method, where the re-forecast/reference pairs were resampled 1,000 times with replacement (Nicholls 2001; Lanzante 2005; Jolliffe 2007). The scores were then computed for each of the 1,000 samples, ranked and the intervals for specific confidence levels estimated. Inference tests for the differences in scores between two forecast systems have been carried out with a two-sample test based on the differences of the 1,000 bootstrap estimates of the scores.

4. Seasonal forecast quality

4.1. Prediction of tropical sea surface temperature

The performance of seasonal forecast systems has been traditionally tested on tropical SSTs. This is because the main global source of seasonal predictability comes from the interannual variability related to ENSO. As an illustration, Figure 1 shows the SST forecast anomalies over the Niño3.4 (5°N-5°S,170°W-120°W) region for the three experiments. The re-forecasts have been initialized on the 1st of May 1997, a year in which the most intense warm ENSO event of the 20th Century was recorded. The observed anomalies were close to 3 K, a value that some of the ensemble members of the perturbed-parameter and multi-model ensembles attain, suggesting that such an extreme event could have been predicted with non-negligible probability by both systems. However, although the stochastic-physics ensemble members predict the occurrence of a warm anomaly in the central Pacific, the spread of the ensemble is not large enough to encompass the observed anomalies and the probability of an extreme warm event is severely underestimated.

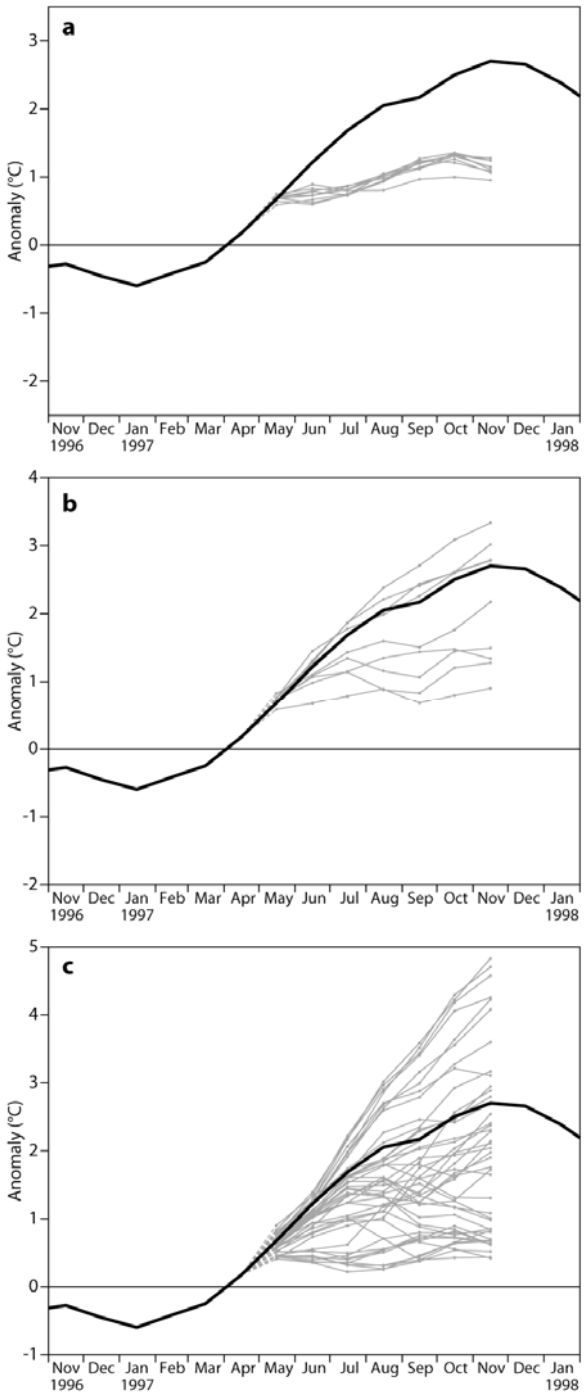

Figure 1: Monthly forecast anomalies of sea surface temperature (SST) averaged over the Niño3.4 (5°N-5°S,170°W-120°W) for the a) stochastic-physics, b) perturbed-parameter and c) multi-model ensembles. The ensembles were initialized on the 1st of May 1997. The solid black line shows the observed anomalies.

As individual events can not be used to assess forecast quality, a simple way to evaluate the performance of the three forecast systems on seasonal and annual time scales over the equatorial oceans is displayed in Figures 2 and 3. Figure 2 displays the ensemble-mean RMSE and the ensemble standard deviation around the ensemble mean for SST averaged over the Niño3.4 (5°N-5°S,170°W-120°W) and eastern tropical Indian ocean (0-10°S,90°E-110°E) regions using re-forecasts of the May and November start dates over lead time for the multi-model, perturbed-parameter and stochastic-physics ensembles. For comparison, the RMSE of a simple statistical model based on persisting the SST anomaly of the month previous to the start date is also

shown. The accuracy of the re-forecasts measured by the RMSE generally decreases with lead time, although all forecast systems show higher skill than the simple persistence model. For Niño3.4, the multi-model ensemble has the smallest RMSE, although followed closely by the other two experiments. The three experiments also show differences in the spread, as in Figure 1. They can be ranked in decreasing order of spread as multi-model, perturbed-parameter and stochastic-physics ensembles. The spread of the multi-model ensemble matches the RMSE reasonably well, which can be considered as a desirable feature in a well-calibrated prediction system. Instead, the perturbed-parameter and stochastic-physics ensembles have a larger RMSE than the ensemble spread for all lead times of the May and November start dates. Ensembles with this sort of behaviour are usually called under-dispersive, as the dispersion of the ensemble does not cater for all the error growth. For the eastern Indian ocean SST, again the multi-model ensemble shows higher skill and a better match between the RMSE and the spread than the other two ensembles, although the spread is overestimated. For this region, as for Niño3.4, it is found that beyond the second month the perturbed-parameter ensemble has a larger spread than the stochastic-physics ensembles. SSTs for other tropical regions also show skill with respect to persistence and climatology (e.g. Figure 4).

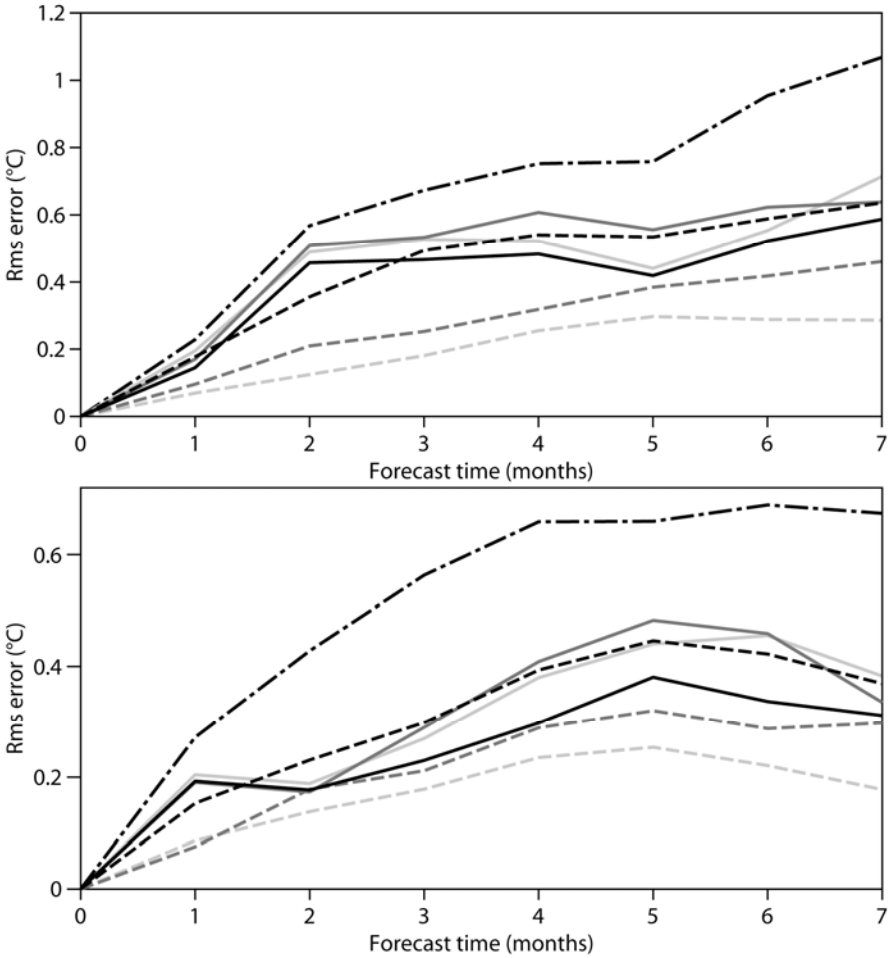

Figure 2: Ensemble-mean root mean square error (RMSE, solid) and ensemble spread (dashed) as a function of the forecast month for the anomalies of sea surface temperature (SST) averaged over the Niño3.4 (5°N-5°S,170°W-120°W, top) and east Indian ocean (0-10°S,90°E-110°E, bottom) regions using the May and November start date re-forecasts. Results are for the stochastic-physics (light grey), perturbed-parameter (medium grey) and multi-model (dark grey) ensembles, respectively. The RMSE of a simple statistical model based on persisting the SST anomaly of the month previous to the start date is shown with a dotted-dashed black line.

Figure 3 depicts the RMSE and spread of SST re-forecasts over the Niño3.4 region for the November start date up to the fourteenth month of integration. Note that with respect to Figure 2 the sample size is now reduced to a half because only re-forecasts for the November start date are available for a forecast period longer than seven months. Both the perturbed-parameter and the multi-model ensembles show a similar evolution of the RMSE, the multi-model RMSE growing above the perturbed-parameter ensemble after forecast month six. The ensemble spread matches the RMSE for the multi-model up to month six. Beyond that time, the ensemble underestimates the spread, as is also the case for the perturbed-parameter ensemble from the beginning of the integrations. The stochastic-physics ensemble has a larger RMSE and smaller spread than the other two systems. The tropical Pacific is the only region that shows some substantial skill using monthly mean data at the end of the first year of forecast, the tropical Atlantic and Indian oceans showing a faster loss of skill with time that makes them less accurate than persistence (not shown). Given the important teleconnections with the tropical Pacific, the skill over that region might be at the origin of some annual-mean forecast skill in other tropical and extra-tropical regions discussed below.

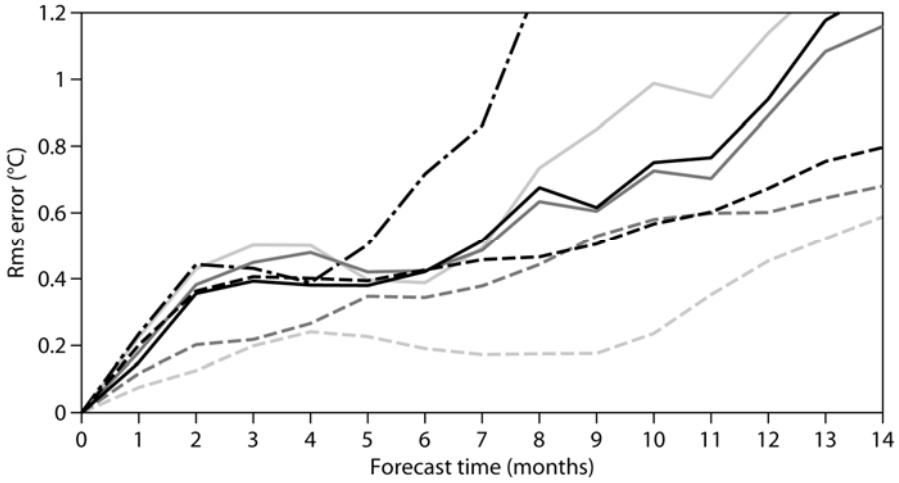

Figure 3: As in Figure 2, but for Niño3.4 SST using the November start date re-forecasts up to 14 months.

The forecast quality for the tropical SST shows that while the three experiments perform well over the tropical Pacific and other tropical regions in terms of RMSE, the multi-model shows a lower RMSE and a better fit between the RMSE and the ensemble spread than the other two forecast systems up to lead times of five or six months. Beyond that time, the performance of the multi-model and perturbed-parameter ensembles is at a similar level, in spite of the larger ensemble size of the multi-model. An overall underestimation of the spread is found for long lead times, which is much larger for the stochastic-physics ensemble than for the other two forecast systems.

4.2. Ensemble characteristics: skill and spread

Previous forecast quality assessments of seasonal forecast systems show low skill over most areas outside the tropical Pacific (e.g. Palmer et al. 2004; Wang et al. 2008). In this low-skill scenario, the most adequate way of treating climate information on seasonal and annual time scales is in the form of probability forecasts. Probability forecasts formulated from dynamical forecast systems are computed using information from the ensemble of simulations. Hence, a preliminary assessment requires an analysis of the ensemble characteristics, which here are described in terms of the ratio between the mean spread around the ensemble mean and the ensemble-mean RMSE, or spread-to-RMSE ratio. The link between the spread-to-RMSE ratio

and a deterministic measure of potential skill such as the ensemble-mean anomaly correlation coefficient (ACC) is also discussed in this section.

The multi-model has a larger ensemble size than the other two forecast systems. To determine the best forecast system, it is paramount to discard the possibility that any improvement is purely a consequence of using a larger ensemble size. In order to carry out a fair assessment of the benefits of using each forecast system independently of the benefits of increasing the ensemble size, a reduced multi-model with nine ensemble members has been considered. However, it is important to bear in mind that, by the current construction of a multi-model system, it is likely that a multi-model will tend to have a larger ensemble size than other approaches to address model uncertainty because different institutions end up pooling their resources together. The nine members were randomly selected from the larger 45-member multi-model ensemble, but taking at least one member from each of the five forecast systems contributing to the multi-model ensemble. The results described below are robust and agree well with those obtained for different samples of nine-member ensemble taken from the 45-member multi-model ensemble. A comparison of the full multi-model ensemble with the other experiments is discussed in Section 6.

The global distribution of the boreal summer (June-to-August) ensemble-mean anomaly correlation and of the spread-to-RMSE ratio for near-surface temperature in the re-forecasts started on the 1st of May with a lead time of one month is shown in Figure 4. The correlation is mostly positive and becomes statistically significant in large areas, especially over the tropical and subtropical Pacific, the North and equatorial Atlantic and some continental regions. Areas with significant skill tend to agree between the three forecast systems, which increases the robustness of the result. These areas can be considered as having significant potential skill, which might translate to actual skill if the variability of the forecasts is similar to the variability in the observations. This level of skill is typical of, if not better than, state-of-the-art seasonal forecast systems (e.g. Palmer et al. 2004; Saha et al. 2006). The ratio spread-to-RMSE for the multi-model is significantly larger than one over some sparse areas (e.g. Southern Europe), suggesting an excess of spread compared to the ensemble-mean error. In many regions, however, the ratio is close to one for the reduced multi-model ensemble. Note, however, that while a value of one indicates a desirable consistency between the ensemble spread and the forecast error, it does not guarantee that such consistency is being generated for the right reasons, i.e. that the processes giving rise to the ensemble spread are necessarily identical to those giving rise to the forecast errors. The stochastic-physics ensemble is more under-dispersive, not only over the tropics, as discussed above, but also over most ocean and continental regions. The perturbed-parameter ensemble shows an under-dispersive behaviour over the tropical regions and over-dispersion in the northern subtropics. An analysis of the spatial distribution of the spread in those areas where the spread-to-RMSE ratio is low (not shown) indicates that in the case of the stochastic-physics and perturbed-parameter ensembles, the ensemble spread tends to be too small instead of the RMSE being too large. This suggests that the ratio could be improved if those two experiments could increase the spread without degrading the RMSE.

Areas where the ensemble-mean skill for near-surface temperature is high do not in general agree with those where the spread matches the RMSE (Figure 4). This shows that consistency between adequate ensemble spread, measured by a ratio spread-to-RMSE close to one, and potential skill, as estimated by the ACC, while intrinsically desirable, does not necessarily indicate an enhanced likelihood of an accurate forecast. In other words, with the current forecast systems the spread might not be a useful predictor of the ensemble-mean skill.

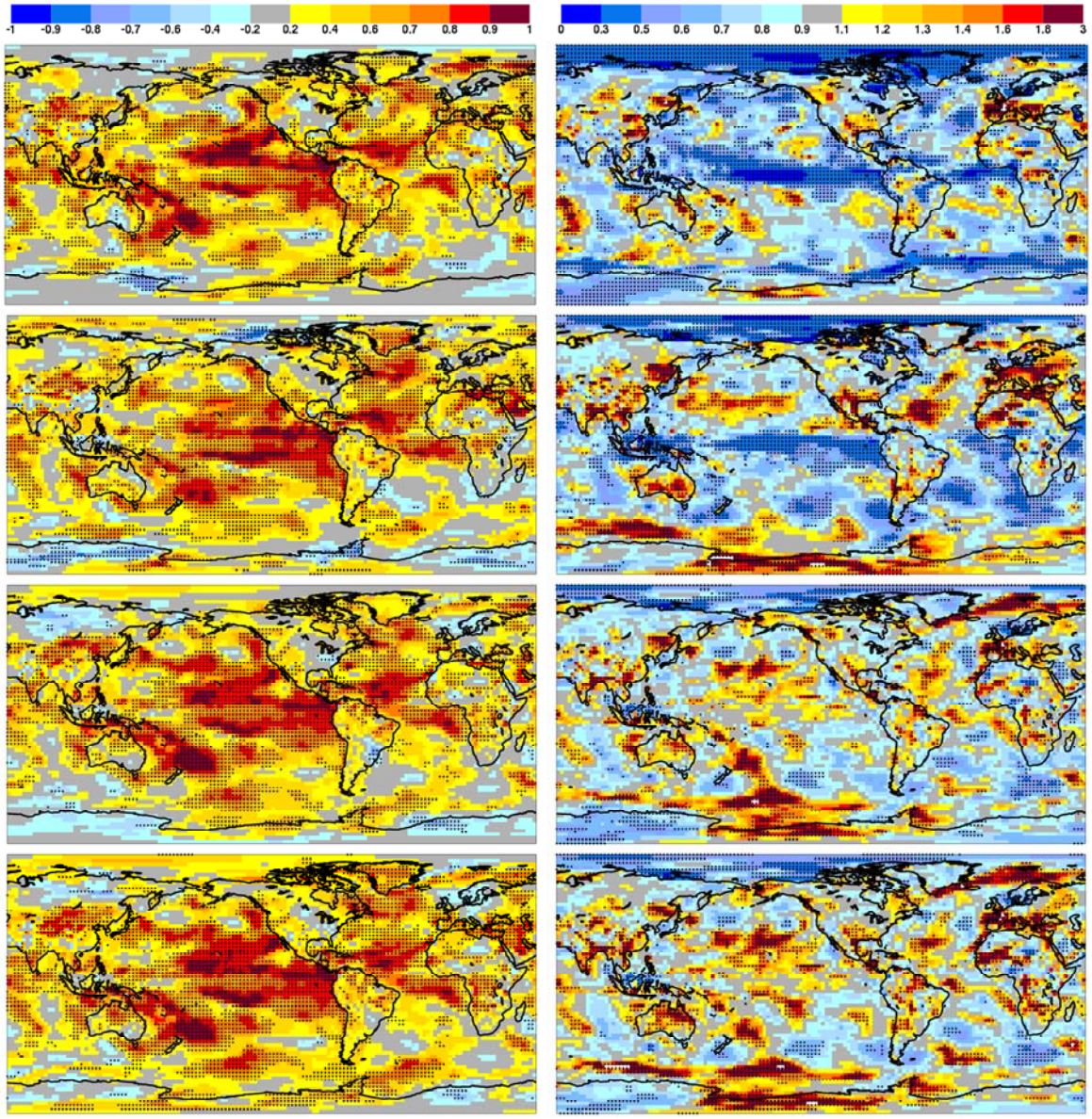

Figure 4: Ensemble-mean anomaly correlation (left column) and ratio between the ensemble spread and the ensemble-mean root mean square error (RMSE, right column) of mean near-surface temperature for June to August using the May start date re-forecasts. The first, second, third and fourth rows show results for the stochastic-physics, perturbed-parameter, reduced multi-model and 45-member multi-model ensembles, respectively. The black dots depict the grid points where the correlation (ratio spread-to-RMSE) is significantly different from zero (one) with 95% confidence.

Predictions of seasonal precipitation display a lower skill than those of temperature. A strong annual cycle of skill with lower skill during boreal summer has been found (not shown). Figure 5 shows the ensemble-mean ACC and the spread-to-RMSE ratio for one-month lead boreal winter (December-to-February) precipitation from the re-forecasts started on the 1st of November. As for near-surface temperature, skill is higher over the tropics, the three systems have a similar level and spatial distribution of the ACC and there is little agreement between regions with statistically significant correlation and a ratio spread-to-RMSE close to one. This last feature again suggests that a well calibrated ensemble in a deterministic sense does not guarantee (or preclude) a significantly high potential skill. There is a large similarity between the multi-model and the perturbed-parameter ensembles, with no obvious general under-dispersion of the latter. By contrast, the

stochastic-physics ensemble again shows an underestimation of the spread with respect to the ensemble-mean RMSE, especially in the tropics.

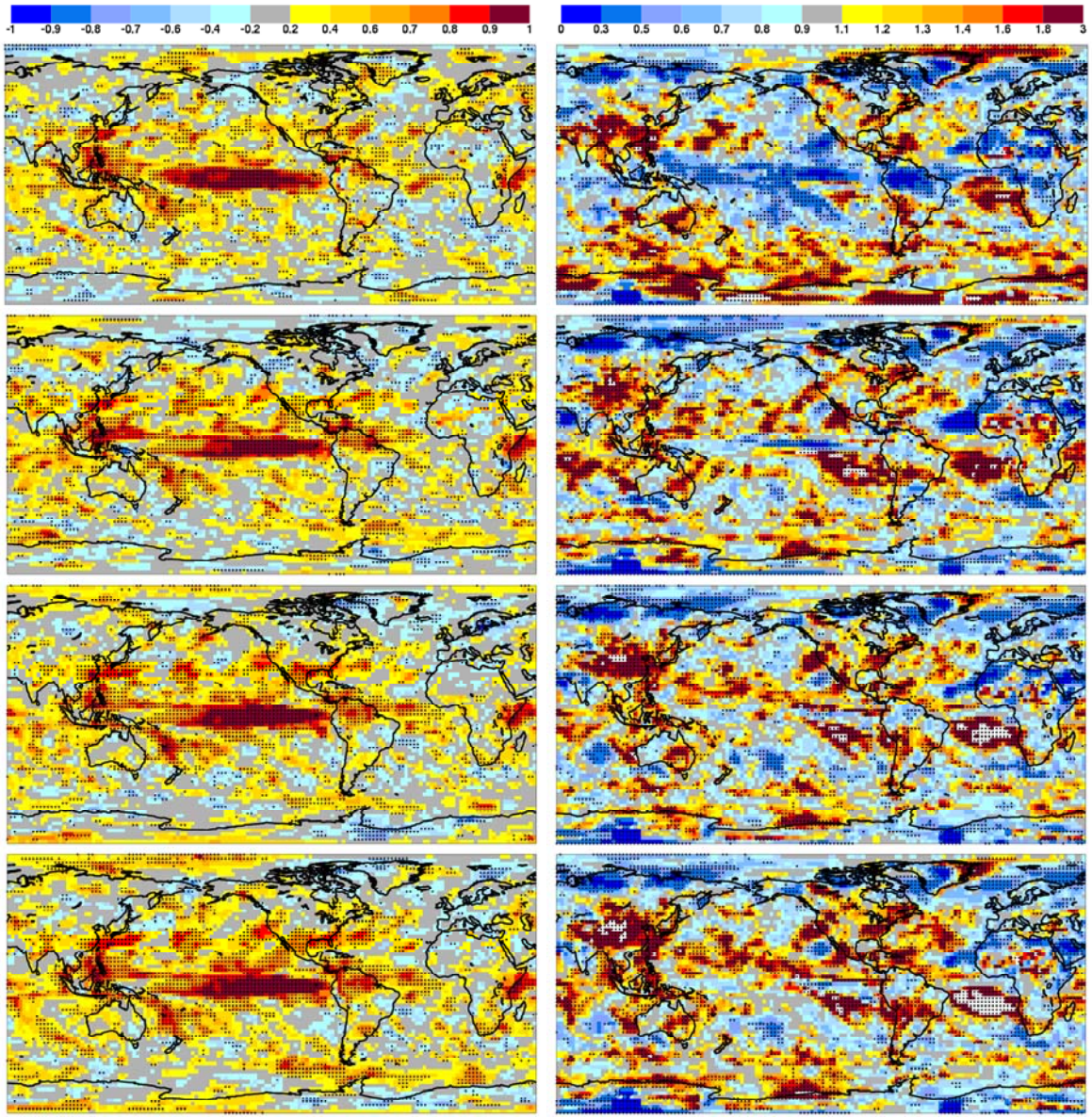

Figure 5: Ensemble-mean anomaly correlation (left column) and ratio between the ensemble spread and the ensemble-mean root mean square error (RMSE, right column) of mean precipitation for December to February using the November start date re-forecasts. The first, second, third and fourth rows show results for the stochastic-physics, perturbed-parameter, reduced multi-model and 45-member multi-model ensembles, respectively. The black dots depict the grid points where the correlation (ratio spread-to-RMSE) is significantly different from zero (one) with 95% confidence.

Figures 4 and 5 show that the characteristics of the three forecast systems in terms of ensemble-mean skill are very similar. This implies that it could be difficult to determine which approach gives the best results overall. For instance, Déqué (2007) examined the re-forecasts discussed here and found that the stochastic-physics ensemble gives better results than the other systems for 500-hPa geopotential height over the Northern Hemisphere. However, other variables and regions might give different results. Figure 6a shows the ensemble-mean ACC of near-surface temperature over the northern extratropics for both start dates and several forecast periods. The values, which in a large proportion of cases are statistically significantly

different from zero with a 95% confidence level, vary between the three forecast systems, with the best performer depending on the start date and lead time. To make the task of determining the best forecast system even more complicated, confidence intervals overlap. Figure 6b, that shows the spread-to-RMSE ratio, helps to shed a clearer picture on one of the features mentioned above, i.e. the lack of spread of the stochastic-physics ensemble. The ratio has been computed by previously averaging the square of the mean spread and the mean square error over all grid points. Both the reduced multi-model ensemble and the perturbed-parameter ensembles show a remarkable match between the spread and the RMSE for lead times longer than zero, especially in the case of the former. Instead, the stochastic-physics ensemble underestimates the spread in every instance. However, this is not necessarily an indication of lower reliability when compared to the other two systems, as will be illustrated below.

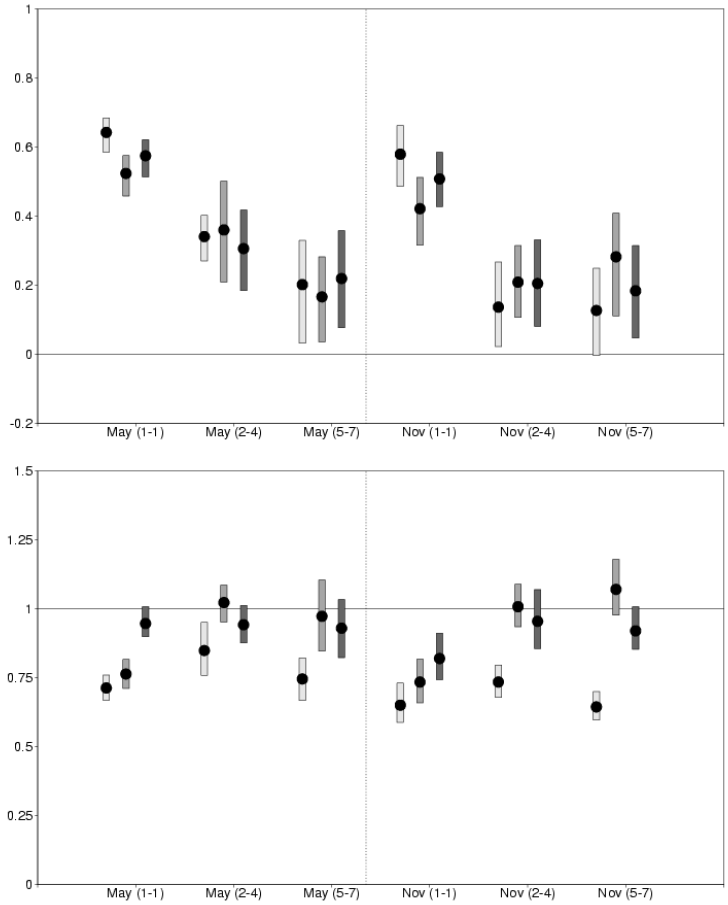

Figure 6: Ensemble-mean anomaly correlation (top) and ratio between the ensemble spread and the ensemble-mean root mean square error (RMSE, bottom) for the anomalies of near-surface temperature over the Northern Hemisphere. The horizontal axis covers the forecast period 1-1 (first month), 2-4 and 5-7 months for the two start dates May and November, results for each start date being separated by the vertical dotted line. The bars are for the stochastic-physics (light grey), perturbed-parameter (medium grey) and reduced multi-model (dark grey) ensembles, respectively. The black dots depict the sample values, obtained by averaging variances and covariances over all grid points before computing the correlation or the spread-to-RMSE ratio, and the bars show the 95% confidence intervals.

The results described above can be generalized to other variables and lead times, as Figure 7 and Table 2 illustrate. The scatter plots have been created with scores computed for seasonal predictions of several variables (500 hPa geopotential height, 850hPa temperature, precipitation, near-surface temperature and

mean sea level pressure), over a number of regions (Table 1) for both May and November start dates and for several lead times. This allows for a comprehensive comparison of pairs of experiments that goes beyond a visual inspection of specific scores for individual regions and variables and is easier to interpret than the spatial distribution of the forecast quality measures of Figures 4 and 5.

Table 1: Regions used in the computation of the forecast quality measures. The first four regions include land and ocean grid points, while only land points have been considered in the rest of the regions.

	Latitude (south, north)	Longitude (west, east)
Europe	35° to 75°	-12.5° to 42.5°
North America	30° to 70°	-130° to -60°
Northern Hemisphere	30° to 87.5°	0° to 360°
Southern Hemisphere	-87.5° to -30°	0° to 360°
Tropics	-20° to 20°	0° to 360°
Mediterranean	30° to 47.5°	-10° to 40°
Australia	-45° to -11°	110° to 155°
Amazon	-20° to 12°	-82.5° to -35°
Southern South America	-55° to -20°	-75° to -35°
Western North America	30° to 60°	-130° to -82,5°
Eastern North America	25° to 50°	-85° to -60°
Northern Europe	47.5° to 75°	-10° to 40°
West Africa	-12.5° to 17.5°	-20° to 22.5°
East Africa	-12.5° to 17.5°	22.5° to 52.5°
Southern Africa	-35° to -12.5°	-10° to 52.5°
Southeast Asia	-10° to 20°	95° to 155°
East Asia	20° to 50°	100° to 145°
Southern Asia	5° to 30°	65° to 100°
Central Asia	30° to 50°	40° to 75°
North Asia	50° to 70°	40° to 180°

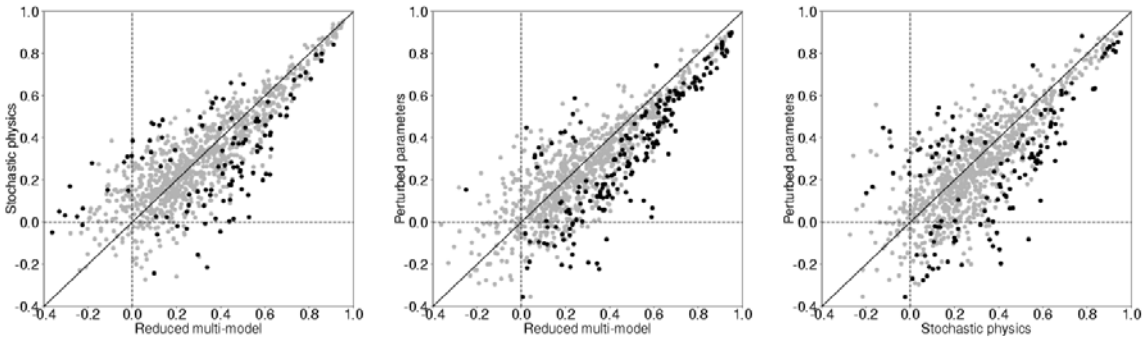

Figure 7: Scatter plots comparing the ensemble-mean correlation of different forecast systems: stochastic-physics versus reduced multi-model (left), perturbed-parameter versus reduced multi-model (centre) and perturbed-parameter versus stochastic-physics (right). Each dot shows the ensemble-mean ACC for the seasonal prediction of a specific parameter (500 hPa geopotential height, 850 hPa temperature, precipitation, near-surface temperature and mean sea level pressure), start date (May and November) and lead time (lead times from zero up to four months), over each one of the regions in Table 1 for a pair of forecast systems. Black dots are used for cases where the differences between two forecast systems are statistically significant with 95% confidence. A summary of the percentage of cases where a forecast system is significantly better than the other can be found in Table 2.

The results show that a) there is a clear relationship between the ensemble-mean skill of the different experiments because when one performs relatively well, they all tend to perform well, b) there is a large scatter in the characteristics of the skill because many cases can be found when any given system performs either better or worse than an alternative system, and c) average differences in ensemble-mean skill are relatively small compared to the scatter as in 80-90% of cases no significant difference can be found between the performance of any two systems, when equal ensemble sizes are being considered (Table 2). However, a modest number of statistically significant differences in performance do emerge, depending on the lead times considered.

Table 2: Percentage of cases in which one of the forecast systems is statistically significantly better (with 95% confidence) than another one for different forecast quality measures. The scores have been obtained for each region in Table 1 and for the variables 500 hPa geopotential height, 850 hPa temperature, precipitation, near-surface temperature and mean sea level pressure. The event anomalies above the upper tercile and the median and below the lower tercile have been considered for the probability forecasts. The comparison is carried out separately for lead times shorter than five months (forecast periods 1-3, 2-4, 3-5, 4-6 and 5-7 months), for which predictions for both start dates (May and November) have been used, and lead times longer than four months (forecast periods 6-8, 7-9, 8-10, 9-11, 10-12, 11-13 and 12-14 months), for which only predictions for the November start date were available. For instance, the pair of numbers in the first box is for the percentage of scores for which the perturbed-parameter ensemble is statistically significantly better than the stochastic-physics ensemble (3.8), while the second number (9.9) is for the percentage of cases for which the stochastic-physics ensemble is significantly better than the perturbed-parameter ensemble. In total, 1,000 (scores using the ensemble mean) and 3,000 (probabilistic scores) different cases have been computed for lead times shorter than five months, and 700 (scores using the ensemble mean) and 2,100 (probabilistic scores) for lead times longer than four months.

	Lead time (months)	ACC	Ratio spread / RMSE	BSS	RELSS	RESSS	ROCSS	BSS(∞)
Perturbed parameter vs stochastic physics	0-4	3.8/9.9	11.2/3.7	3.9/11.3	2.8/8.9	1.5/3.7	3.3/8.1	3.8/11.5
	5-11	8.4/2.4	4.0/29.6	10.9/2.0	8.9/1.5	2.3/0.7	8.1/1.7	10.8/2.0
Reduced multi-model vs perturbed parameter	0-4	17.1/1.6	4.7/2.8	19.6/1.3	14.8/0.6	6.2/1.2	12.9/1.7	20.5/1.3
	5-11	4.4/9.6	22.3/4.4	5.6/8.3	4.6/4.7	1.0/2.6	3.5/8.6	5.5/8.1
Reduced multi-model vs stochastic physics	0-4	7.7/3.7	12.7/2.2	9.6/3.9	6.6/2.3	4.3/1.7	6.6/3.8	10.2/3.9
	5-11	7.0/2.6	8.4/15.7	10.9/2.0	9.0/1.5	1.6/0.5	6.6/2.0	11.3/2.0
Multi-model vs perturbed parameter	0-4	31.2/0.9	5.8/5.4	49.2/0.1	36.2/0.0	16.4/0.3	28.4/0.5	17.2/2.2
	5-11	9.4/5.1	5.6/1.9	31.7/1.0	28.0/0.7	3.8/2.1	8.3/2.8	4.4/11.6
Multi-model vs stochastic physics	0-4	14.8/2.4	11.8/3.8	34.2/0.4	25.6/0.1	13.1/0.5	16.6/1.4	6.7/5.0
	5-11	14.0/2.3	13.0/4.9	45.6/0.4	37.7/0.1	6.8/0.7	13.5/1.2	8.2/3.7

For lead times of up to four months the reduced multi-model ensemble has more often than not a higher correlation than the stochastic-physics and perturbed-parameter ensembles, i.e. there are more symbols below the diagonal, although there are still cases where one of the other two forecast systems can be better than the reduced multi-model ensemble. Among the other two forecast systems, the stochastic-physics is more often than not better than the perturbed-parameter ensemble. Table 2 also summarizes the proportion of cases when a forecast system is significantly better (measured with respect to an ideal ratio of one) than the other two in terms of the spread-to-RMSE ratio. The reduced multi-model ensemble performs better than

both the perturbed-parameter and stochastic-physics ensembles. In contrast with the correlation, the stochastic-physics ensemble performs more often worse than the perturbed-parameter ensemble.

Results for lead times longer than four months are computed only from re-forecasts of the November start date (Table 2). Interestingly, there are many skilful predictions even at these longer lead times. The reduced multi-model still has more statistically significantly higher correlations than the stochastic-physics ensemble. However, in contrast to shorter lead times, the perturbed-parameters ensemble shows a larger proportion of cases with higher skill than either of the other two systems. The perturbed-parameter ensemble has also more often a better matching between the spread and the RMSE than the stochastic-physics ensemble, although the reduced multi-model is more often significantly better than the other two.

4.3. Forecast quality for probability predictions

Values of the spread-to-RMSE ratio close to one are a desirable feature of all ensemble systems and have been interpreted as a precondition to achieve reliability. However, it has been demonstrated in the previous section that in the seasonal forecast context it does not necessarily have a direct link with the potential skill of the ensemble mean. In addition to that, values of the spread-to-RMSE ratio close to one do not guarantee that the ensemble would translate into a set of reliable probability forecasts and, hence, a more direct measure of reliability is needed. This can be obtained by formulating probability forecasts from the ensemble, which is also a way to include in the climate information provided to an eventual user as many sources of uncertainty as possible. Besides, the ensemble-mean ACC suggests that there is potential skill in the re-forecasts, so estimates of aspects of probability forecast accuracy beyond reliability are required.

The forecast quality of probabilistic predictions is multi-faceted; this implies that different skill measures can lead to slightly different conclusions. Attributes diagrams offer a comprehensive illustration of several forecast quality properties of a set of dichotomous probability forecasts, such as reliability and resolution, a measure of forecast accuracy. As an illustrative example, Figure 8 shows the attribute diagrams for predictions of 1-month lead boreal summer (June-to-August) precipitation (May start date) above the upper tercile over the tropical band. Probability forecasts for all grid points and the eleven years of the sample are pooled together to construct the diagram. Each forecast probability bin is represented by a solid circle whose area is proportional to the bin sample size (i.e., the number of probability forecasts in the interval over all the years and grid points of the region), so that a histogram of the probability forecasts is implicitly embedded in the attributes diagram. For each interval of probability, a representative probability is defined as the weighted mean of the individual forecast probabilities included in the category. This is different from, and more precise than, defining the representative forecast probability as the centre of the probability interval of the category (Bröcker and Smith 2007). The vertical line represents the average forecast probability, while the horizontal line is for the climatological frequency of the event. The black dashed line in the diagram separates skilful from unskilful areas in the sense of the BSS: categories with forecast probabilities lower (higher) than the mean probability that fall below (above) this line, contribute positively to the BSS; otherwise they contribute negatively to the BSS.

The reliability curves for all systems are shallower than the diagonal, although the reduced multi-model ensemble is slightly closer to the diagonal that either the stochastic physics or perturbed parameters. However, also for this forecast system the confidence intervals are not wide enough across the full forecast probability range to encompass the diagonal. The 95% confidence intervals, which are represented with the

grey bars, are also far away from the diagonal, an indication of the overconfidence of these systems. On the other hand, the confidence intervals are away from the climatological frequency line for most of the probability categories showing that the forecasts have statistically significant resolution, which is measured by the vertical distance between the reliability curve and the horizontal line. The first interpretation of the diagram is that these particular probability forecasts issued with the simple counting method, although skilful, should be considered unreliable and be subject to some form of calibration before use.

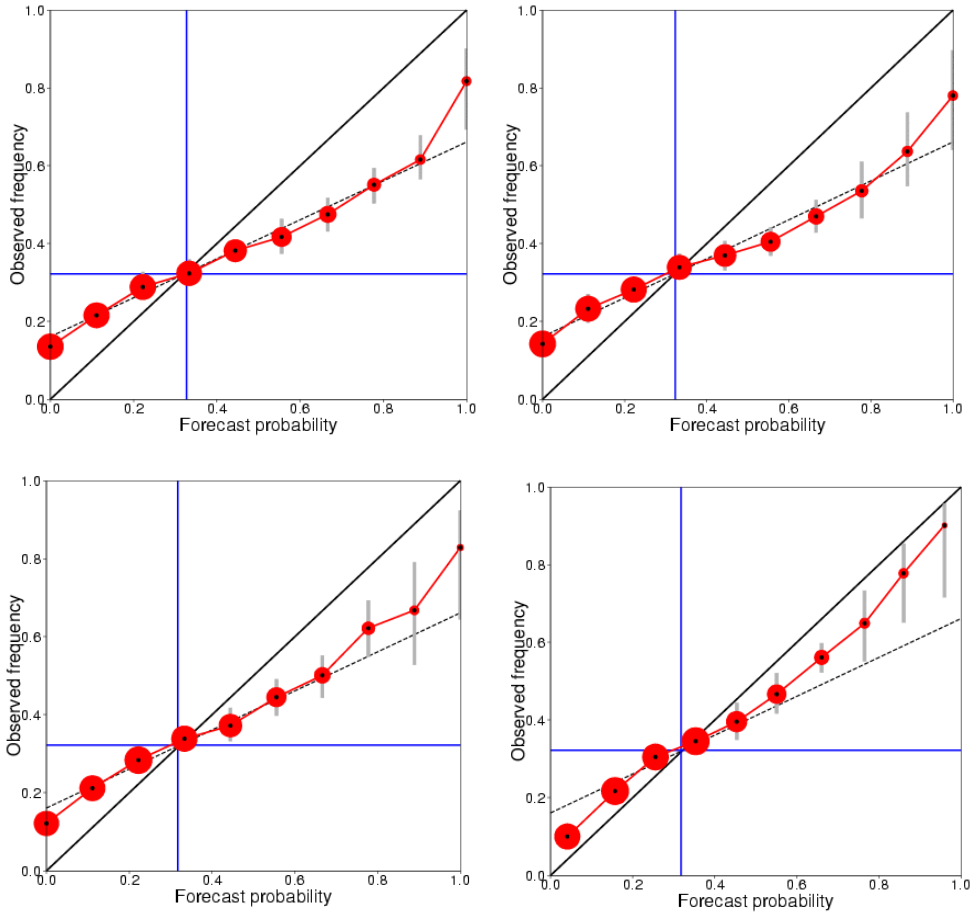

Figure 8: Attributes diagrams for predictions of boreal summer seasonal precipitation above the upper tercile over the tropical band (20°N-20°S) for the stochastic-physics (top left), the perturbed-parameter (top right), the reduced multi-model (bottom left) and the 45-member multi-model (bottom right) ensembles. The forecast period is 2-4 months and the start date the first of May. The size of the dots is proportional to the number of probability forecasts included in each of the ten probability bins. The vertical solid lines correspond to the average forecast probability and the horizontal solid line the observed climatological frequency of the event. The dashed grey lines delimit the areas where the predictions are skilful in the sense of the Brier score: predictions with points above the line to the right of the average probability and below the lines to the left of it contribute positively to the Brier skill score. Vertical grey bars depict the 95% confidence intervals for each bin of the reliability diagram.

As an example of how skilful the systems are for regions of special interest Figure 9 shows the ROCSS of near-surface temperature over the Mediterranean region and of precipitation over Northern Europe for values above the upper tercile as a function of the start date and the forecast period, starting from the first forecast month of the predictions initialized on the 1st of May and ending with the four-month lead seasonal

predictions initialized on the 1st of November. The ROCSS decreases with lead time in all instances except for the perturbed-parameter experiment in the November start date for Mediterranean temperature. The sample value of the forecast quality measures, displayed with black dots, although low, is most of the times positive in the case of temperature, but not so much for precipitation. This is typical of seasonal forecasts over extra-tropical regions (Hagedorn et al. 2005; Saha et al. 2006). The confidence intervals, as in the case of the ensemble-mean ACC, are large and straddle the no-skill zero line, which should be interpreted as a large fraction of the positive values actually being not different from zero. Besides, there is a big overlap between the confidence intervals of the three forecast systems. On this basis, it is again difficult to conclude that any specific system is consistently better than the others.

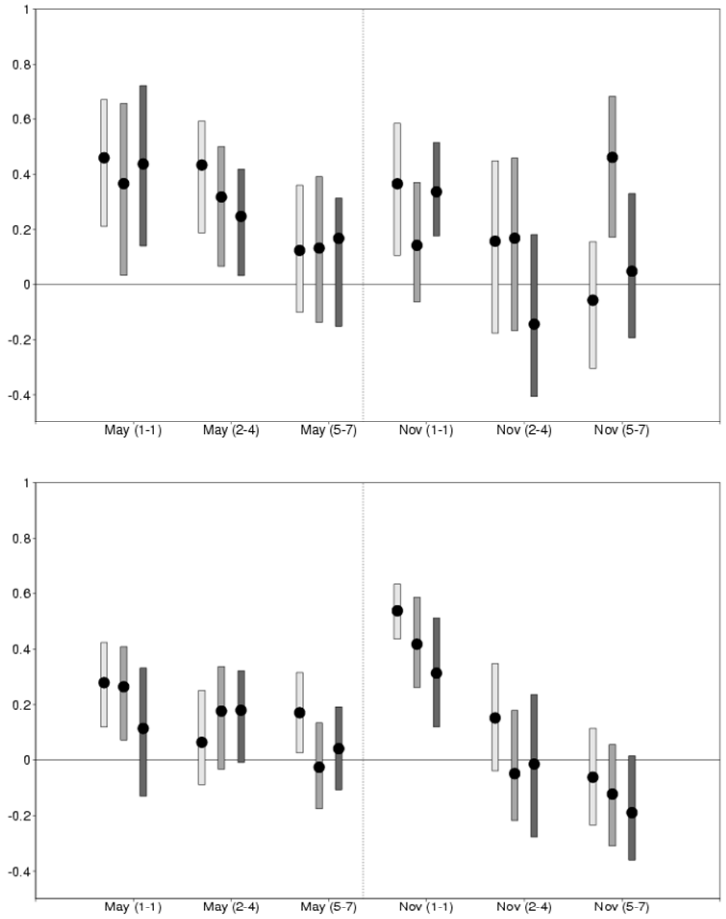

Figure 9: ROC skill score of near-surface temperature over the Mediterranean region (top) and precipitation over Northern Europe (bottom) for predictions of values above the upper tercile. The horizontal axis covers the forecast periods 1-1 (first month), 2-4 and 5-7 months for the two start dates May and November, results for each start date being separated by the vertical dotted line. The bars are for the stochastic-physics (light grey), perturbed-parameter (medium grey) and reduced multi-model (dark grey) ensembles, respectively. The black dots depict the sample values and the bars show the 95% confidence intervals. The contingency table required to compute the scores was computed by pooling together (not averaging) the re-forecasts from all the points of predefined regions over the 1991 to 2001 period.

To overcome the fact that differences in forecast quality between the three forecast systems are in many cases small and not statistically significant, as in the case of the ensemble-mean ACC, a thorough comparison has been carried out by computing the scores for all the regions listed in Table 1 in a similar way as for the ensemble-mean scores. Probability forecasts for three different events have been considered: values above the median and the upper tercile, and values below the lower tercile. Figure 10 shows the scatter plots of BSS and ROCSS for the stochastic-physics, perturbed-parameter and reduced multi-model experiments for both start dates and seasonal forecast ranges with lead times between zero and four months, making a total of 3,000 cases. For both skill scores there is a large range of values, from unskilful predictions (lower than zero) to values close to 0.5 and 1 for BSS and ROCSS, respectively. There is a large spread of the scores around the diagonal; in other words, comparing the scores of two experiments for a specific case (region, lead time, start date, variable and event) might give a completely opposite result to the scores for a different case. To better interpret Figure 10, Table 2 shows the proportion of cases whose scores (along with the reliability and resolution skill scores) are significantly better for one experiment than for another one. Once again, the conclusions depend on the lead time. For lead times up to four months, the BSS is more often significantly better for the reduced multi-model than for the other two experiments. Between the perturbed-parameter and the stochastic-physics ensembles, the latter shows a larger proportion of cases with significantly higher BSS than the former. The larger proportion of cases where reliability is significantly better for the reduced multi-model (Table 2) suggests that the better performance in terms of BSS can be largely attributed to an improved reliability, although resolution also plays a role. The improved reliability of the reduced multi-model ensemble agrees with this experiment having the best performance in terms of the ratio spread-to-RMSE. Both resolution and ROCSS are more often significantly better in the reduced multi-model than in the two other experiments. The improvement in terms of ROCSS and resolution, both measures of potential skill, is especially important because the discrimination and the resolution can only be enhanced by using additional sources of forecast information, while the reliability could be improved a posteriori using climatological information from the observations in a calibration process (Doblas-Reyes et al. 2005). Table 2 also shows that for lead times of up to four months the BSS of the stochastic-physics ensemble shows a better performance than the perturbed-parameter ensemble, which is explained by the superiority of the former in terms of reliability, resolution and ROCSS. It is important to bear in mind that the proportions of statistically significant differences are small and never higher than 20%. In other words, while a larger proportion of points in Figure 10 are below the diagonal, most of the points are grey rather than black. Furthermore, the ratios between the proportion of significantly different cases are also relatively small, between 3 and 6. This should invite the reader to interpret the results with caution and, in a simplification effort, avoid automatically discarding any of the approaches.

The ranking of the experiments for predictions with lead times shorter than five months, with the reduced multi-model showing better results than the stochastic-physics and perturbed-parameter experiments, does not hold for longer lead times. Figure 11 illustrates the ROCSS for seasonal predictions with lead times longer than four months. Predictions for these long forecast periods are skilful for a number of regions, lead times, events and variables, although the level of skill is noticeably lower than for the predictions scored in Figure 10. Table 2 shows that, by contrast with the results for the shorter lead times, the perturbed-parameter ensemble has better performance in terms of reliability, resolution and ROCSS than both the stochastic-physics and reduced multi-model ensembles. Among the reduced multi-model and stochastic-physics ensembles, the former shows better results than the latter by all measures. In fact, the stochastic-physics ensemble shows a dramatic loss of skill for the longer lead times. A comparison of the forecast quality for

the three experiments for lead times shorter than five months with the November start date re-forecasts only (i.e., discarding the May start date re-forecasts) gave basically the same results as those shown in Table 2 for both start dates (not shown). This suggests that the change in the ranking of the experiments with the lead time of the predictions is not due to the unavailability of the May start date re-forecasts for predictions with lead times longer than four months.

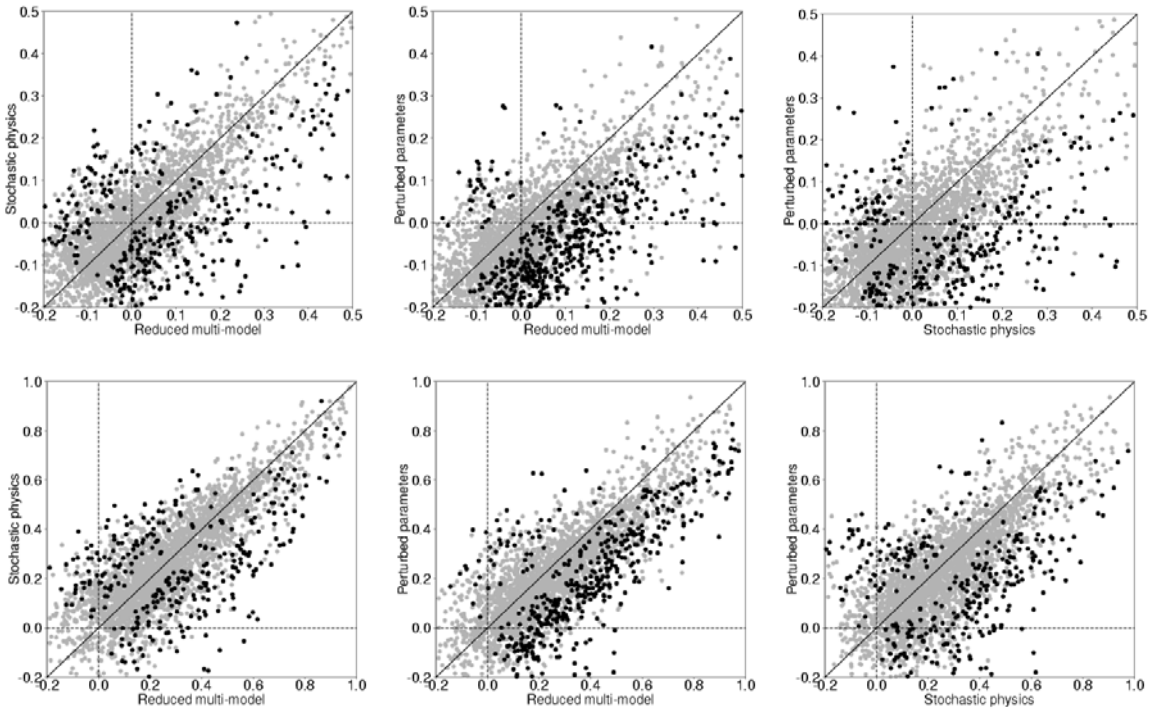

Figure 10: Scatter plots comparing the Brier skill score (top row) and the ROC skill score (bottom row) of the stochastic-physics versus reduced multi-model ensembles (left), perturbed-parameter versus reduced multi-model ensembles (centre) and perturbed-parameter versus stochastic-physics ensembles (right). Each dot represents the skill scores for seasonal predictions of a specific variable (500 hPa geopotential height, 850 hPa temperature, precipitation, near-surface temperature and mean sea level pressure), event (values above the median and the upper tercile and values below the lower tercile), start date (May and November) and lead times (lead times of up to four months), over one of the regions in Table 1 for a pair of forecast systems. Black dots are used for cases where the differences between the scores of a pair of forecast systems are statistically significant with 95% confidence. A summary of the percentage of cases a forecast system is significantly different from another can be found in Table 2.

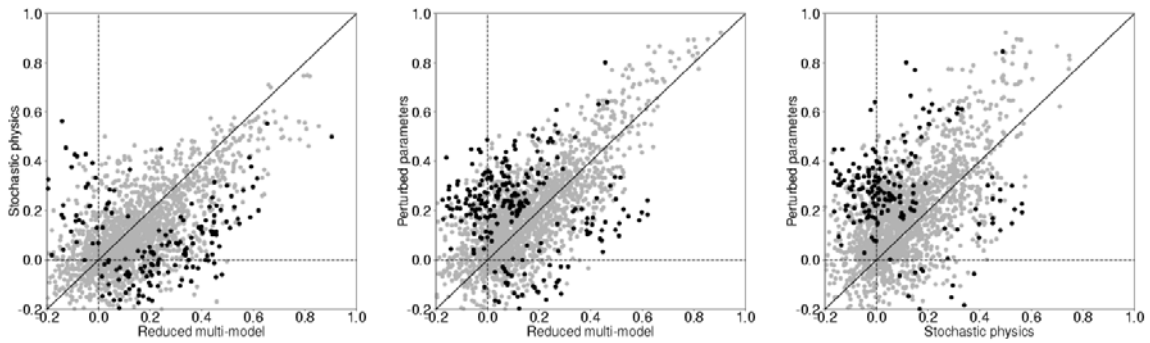

Figure 11: As Figure 10, but for the ROCSS for seasonal predictions with lead times longer than four months from the November start date re-forecasts.

The change of ranking between the stochastic-physics and the perturbed-parameter experiments with the lead time might be due to either 1) the efficiency of each approach to represent model uncertainty depending on the forecast time scale or 2) purely to one of the reference forecast systems, IFS/HOPE or DePreSys_IC, having better forecast quality than the other depending on the lead time. As there is no reference system for the reduced multi-model, the multi-model approach can not be considered in this context. Table 3 summarizes the forecast quality comparison between IFS/HOPE and DePreSys_IC. While for predictions with lead times shorter than five months IFS/HOPE has more often significantly better reliability and ROCSS, the opposite occurs for longer forecast periods: DePreSys_IC is more often a significantly better forecast system (taking into account all measures of probabilistic forecast quality) than IFS/HOPE for lead times longer than four months. This implies that, with the current experiments, it is impossible to determine which one of the two approaches, stochastic physics or perturbed parameter, is better because the differences between the reference forecast systems used to implement them are similar to the differences in forecast quality found between the predictions using the two approaches to address model uncertainty. A way to resolve this question might be to use the same reference forecast system to implement both approaches, something that is not technically possible with our systems in the short term.

Table 3: Same as Table 2, but for a comparison between probability seasonal predictions issued with different forecast systems. Note that the rows labelled 0-4 have been computed with 3,000 cases, while the second one uses 2,100 because for longer lead times only the November (instead of both May and November) start date is available.

	Lead time (months)	BSS	RELSS	RESSS	ROCSS
IFS/HOPE vs DePreSys_IC	0-4	7.6/4.9	6.0/4.3	2.9/2.5	6.1/0.3
	5-11	3.2/6.8	3.1/4.3	0.4/3.2	1.9/7.7
IFS/HOPE control vs stochastic physics	0-4	2.4/14.7	1.2/16.6	2.4/3.2	2.6/7.7
	5-11	6.1/5.0	3.9/5.1	1.2/0.5	4.8/3.5
DePreSys_IC vs DePreSys_PP	0-4	4.1/8.4	1.7/10.1	2.4/1.2	4.8/3.0
	5-11	3.5/9.5	2.4/10.4	1.4/0.6	4.7/3.4
Multi-model vs reduced multi-model	0-4	38.8/0.4	27.7/0.2	13.7/0.2	21.5/8.3
	5-11	46.8/0.1	34.0/0.1	6.7/0.8	14.5/1.3
5-member ensemble multi-model vs 5-member ensemble DePreSys_PP	0-4	14.8/1.9	12.6/1.0	4.7/1.2	9.4/2.6
	5-11	3.0/7.4	2.7/6.4	0.8/2.3	2.2/8.5

However, an aspect that can be assessed is whether any of the approaches actually improves the forecast quality of their reference system. Table 3 shows the results of the comparison between the IFS/HOPE system run with and without stochastic physics and between DePreSys_IC and DePreSys_PP. The reader is reminded that all these experiments have nine-member ensembles. The use of the stochastic-physics approach improves the forecast quality of the IFS/HOPE seasonal predictions with a lead time shorter than five months. This is reflected in a larger proportion of cases with significantly better reliability, resolution and ROCSS for the stochastic-physics ensemble, which as a result gives a large proportion of cases with better BSS. These results agree well with those presented in Berner et al. (2008). However, for lead times longer than four months only the reliability of the stochastic physics predictions improves with respect to IFS/HOPE control; the rest of the scoring measures are more often significantly better for the control than for the stochastic-physics experiment. In the comparison between DePreSys_IC and DePreSys_PP, the perturbed-parameter approach improves both the BSS and the reliability of the system, but degrades the

ability to discriminate between events and non-events by showing a larger proportion of cases for which the resolution and ROCSS of DePreSys_IC is more often significantly better than for DePreSys_PP. Therefore, the use of DePreSys_PP instead of DePreSys_IC as seasonal forecast system becomes a complicated trade off between a large gain in reliability and a certain loss of resolution. An additional factor to consider in this trade off is the increased computational cost of DePreSys_PP with respect to DePreSys_IC. Although the cost of the re-forecasts of both systems is the same, DePreSys_IC requires the estimation of a model climate and flux-correction terms for a single version of the forecast system, while the use of DePreSys_PP implies that the same process has to be undertaken with nine different model versions. In a seamless context where the same forecast system is used across different time scales the burden of estimating the model climate and the flux-correction terms might be shared with the integrations carried out to perform anthropogenic climate-change projections.

The results described above suggest that the versions of the stochastic-physics and perturbed-parameter employed in this paper have difficulties to improve all aspects of the forecast quality with respect to their reference systems, in particular the resolution and the ROCSS (which measures the ability to discriminate between events and non-events). The main and important advantage of the application of these two approaches is an increase in reliability that, as a consequence, improves the BSS provided the degradation of the resolution term is not too large. The improvement in probability forecast reliability is an indication of the success of both approaches in addressing model uncertainty. Nevertheless, the results of this comparison depend on the specific reference model where the model-uncertainty approach is implemented. The same approach implemented on a different reference forecast system might lead to slightly different conclusions.

5. Forecast quality of annual-mean predictions

Results in Section 4 give an indication of the positive forecast quality of seasonal predictions with long lead times (up to eleven months). In this section the forecast quality of annual-mean values from ensemble annual re-forecasts is assessed. The use of annual-mean predictions is not widespread yet, in part due to its relative novelty, the lack of assessments of the associated level of predictability and forecast quality, and the scarcity of applications. To assess the possibility of issuing predictions of annual means with the forecast systems considered in this paper and to avoid any useful skill being a simple consequence of expected predictability arising from the early stages of the forecasts, we have considered annual-mean predictions with a two-month lead time. The predictions are for the annual mean over a complete calendar year, i.e., from January to December, as the only re-forecasts available longer than one year have the 1st of November as the start date.

Figure 12 displays a comparison of the BSS and ROCSS for the annual-mean predictions. All forecast systems show some positive skill scores, especially for the ROCSS, which can be as high as 0.6. For certain regions and events this might be high enough to allow its use in specific applications. However, the proportion of cases with statistically significant differences is small. On the basis of these results it is difficult to select one of the experiments as superior to the other two. The scores are more or less equally spread around the diagonal and the number of statistically significant cases is rarely larger for one of them. However, as shown in the next Section that discusses the impact of the ensemble size on the forecast quality of the multi-model, the 45-member multi-model is superior to the other two forecast systems in the case of annual-mean predictions.

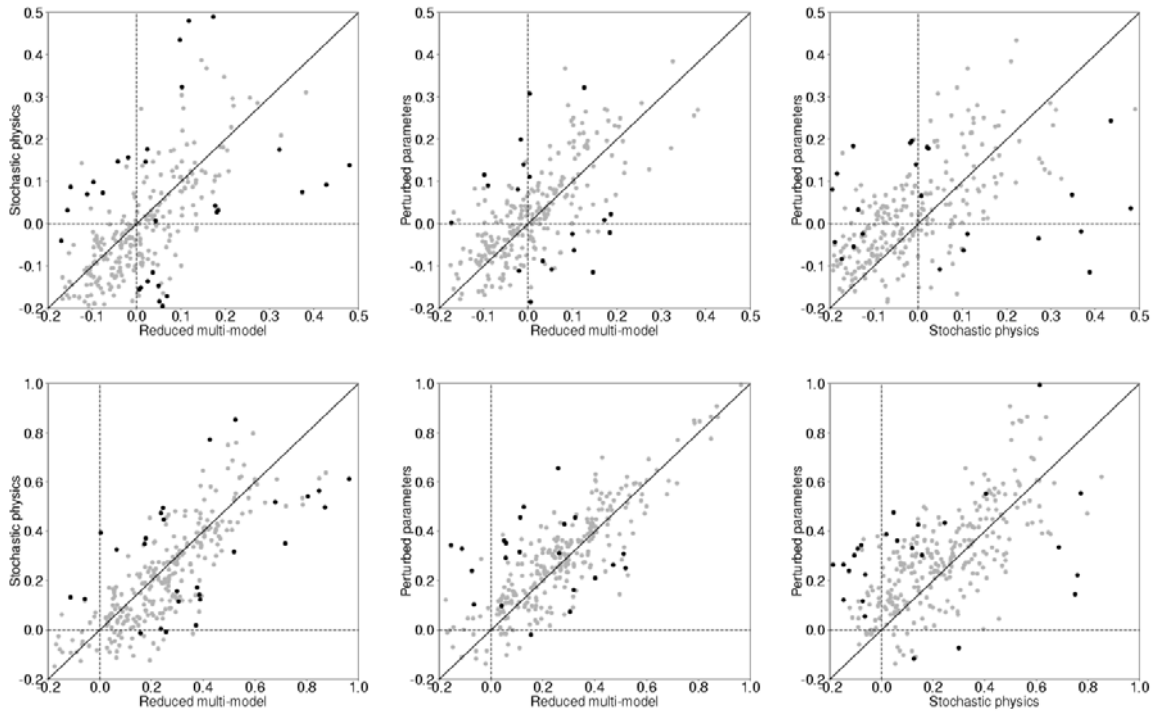

Figure 12: As Figure 10, but for annual-mean predictions from the November start date re-forecasts.

6. Effect of the ensemble size

The previous sections showed that the reduced multi-model has better performance than the other two forecast systems for predictions with lead times shorter than five months. For longer lead times the perturbed-parameter ensemble shows more often better results than the stochastic-physics and reduced multi-model experiments. Although the multi-model approach has been assessed with a reduced version that has the same ensemble size as the other two experiments, the multi-model experiment has a much larger ensemble available. Ensembles of up to 45 members can be constructed from the different forecast systems contributing to the multi-model. Palmer et al. (2004) and Weigel et al. (2007), among others, have shown that forecast quality, in particular reliability, is sensitive to the size of the ensemble used to compute the probabilities. In this section, the effect of the increased multi-model ensemble size on forecast quality is assessed and the multi-model compared to the other two experiments.

The ensemble-mean ACC and the ratio spread-to-RMSE of the multi-model ensemble are shown in Figures 4 and 5 along with the values for the three ensembles discussed in previous sections. Both, ACC and spread-to-RMSE ratio, show little change between the reduced and the 45-member multi-model ensembles. However, when probability forecasts are considered, more important differences can be found, as illustrated in Figure 8. The reliability curve of the multi-model becomes steeper than for the three ensembles discussed previously. The increased steepness of the multi-model forecasts improves both the reliability (by decreasing the distance to the diagonal) and the resolution (by increasing the distance to the horizontal line corresponding to the climatological frequency of the event). This is carried out at the expense of a reduction in the variance of the forecast probabilities, which is illustrated by a reduction in size of the symbols for the extreme forecast probabilities. This is, nevertheless, a desirable feature. Although a sharp set of probabilities, i.e. forecast probabilities with large variance, is a desirable feature of a forecast system, sharp forecasts may

be particularly harmful in the case of overconfident reliability diagrams, such as the ones obtained for the stochastic-physics, perturbed-parameter and reduced multi-model ensembles.

Table 2 shows the results of the comparison between the 45-member multi-model versus the stochastic-physics and perturbed-parameter experiments in terms of forecast quality. The multi-model has more often significantly better scores than the other two experiments for all scores and sets of lead times. The improvement is particularly clear for reliability. Figure 13 illustrates some of the results summarized in Table 2 in a comparison with the perturbed-parameter ensemble. For seasonal predictions with lead times longer than four months, and especially for annual-mean predictions, the multi-model performs more often significantly better than the perturbed-parameter ensemble. In particular, the number of seasonal prediction cases with negative BSS of the perturbed-parameter ensemble is reduced in the case of the multi-model ensemble. These results contrast with results described in the previous section, where the perturbed-parameter experiment had more often significantly better scores than the reduced multi-model for lead times longer than four months, and can be explained as an effect of the larger ensemble size of the multi-model. As the BSS depends strongly on the ensemble size for small ensembles, Ferro (2007) has developed an analytical expression to estimate the Brier score for an infinite ensemble size, BSS(∞). This estimate is a function of the sample BSS and the variance of the probability forecasts (also known as sharpness). Estimates of the BSS(∞) in a comparison of the multi-model with the perturbed-parameter and stochastic-physics ensembles give very similar results to the comparison of the BSS of the reduced multi-model with the perturbed-parameter and stochastic-physics ensembles, all having nine-member ensembles (Table 2).

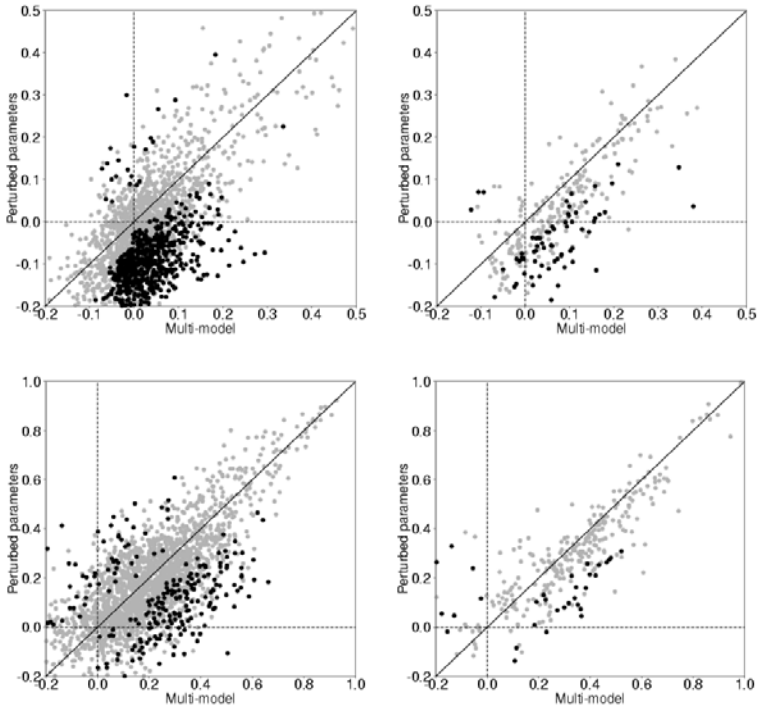

Figure 13: As Figure 10, but for the perturbed-parameter versus the 45-member multi-model using the seasonal predictions initialized on the 1st of November with lead times longer than four months (left) and the annual-mean predictions (right).

Confirming the validity of the BSS(∞) estimates, Table 2 illustrates that estimates of the BSS(∞) for equally-sized ensembles are similar to those obtained for the standard BSS. The BSS(∞) shows once more that the multi-model ensemble performs more often better than any of the two other experiments regardless of the ensemble size for lead times shorter than five months, and that for longer lead times its superiority is an effect of the larger ensemble size. To support this hypothesis, especially in the context of a comparison with the perturbed-parameter ensemble, Table 3 shows results of the comparison between the full and the reduced multi-model ensembles. The 45-member multi-model shows percentages of significant differences with respect to the reduced multi-model ensemble that is smaller than the percentages of significant differences with respect to the perturbed-parameter ensemble, while the opposite is found for longer lead times. This is in agreement with the good behaviour of the perturbed-parameter ensemble for lead times longer than four months.

Although in previous comparisons the reduced multi-model and the perturbed-parameter ensembles had the same ensemble size, the reduced multi-model experiment is a combination of initial-condition and multi-model ensemble. The reduced multi-model was constructed from five different models, while the perturbed-parameter experiment uses only one member from each of the nine model versions of DePreSys_PP. There is still the possibility that the change in ranking with lead time between the reduced multi-model and the perturbed-parameter experiments is somehow related to the different number of quasi-independent forecast systems used. To check this possibility, a comparison between a five-member multi-model ensemble (taking one member from each single-model ensemble) and a perturbed-parameter ensemble made up of five different model versions taken randomly out of the nine available was carried out, the results being summarized in Table 3. A comparison of Table 2 and Table 3 shows that the better performance of the reduced multi-model with respect to the perturbed-parameter ensemble for lead times shorter than five months is also found when using the five-member ensembles. Instead, the five-member perturbed-parameter ensemble for lead times longer than four months is superior to the five-member multi-model in a similar way as the perturbed-parameter experiment is more often significantly better than the reduced multi-model. The main difference when the five-member ensembles are used with respect to the results for the nine-member ensembles appears to be a reduction in the proportion of cases when the reliability of the multi-model is significantly better than that of the perturbed-parameter experiment.

The superiority of the 45-member multi-model can not be automatically considered as an intrinsic characteristic of the approach because when equally-sized ensembles are used, the perturbed-parameter ensemble performs better than the multi-model for lead times longer than four months. However, currently multi-models tend to have larger ensemble sizes than other forecast systems, which in practice gives a multi-model forecast system an advantage in terms of forecast quality. It should be borne in mind that the superiority of the 45-member multi-model ensemble with respect to the perturbed-parameter and stochastic-physics ensembles emphasizes the relevance of the collaborative nature of multi-model forecasting, where pooled resources allow for large forecast ensembles.

7. Summary and conclusions

Estimations of the uncertainty in dynamical climate forecasts are affected by 1) the lack of perfect knowledge of the state of the climate system (Palmer and Anderson 1994) and 2) model inadequacy (Smith 2002). The relative merits of three methods to address the impact of model uncertainty on the forecast quality of seasonal and annual predictions have been assessed in this paper using a coordinated experiment with different forecast systems. The three methods are multi-model, perturbed parameter and stochastic physics. Two of the methods, stochastic physics and perturbed parameter, have been implemented in different forecast systems, which implies that the differences in the performance of the experiments are not just due to the approach to model uncertainty, but also to the different characteristic of the reference forecast systems. The third approach, the multi-model, builds on an ad hoc collection of forecast systems, and a comparison with the experiments implementing the two other approaches somehow depends on the pool of single-model systems the multi-model is made of. In spite of that, different seasonal forecast multi-model ensembles seem to share common features (Wang et al. 2008), such as the gain in accuracy with respect to the single models and the increased reliability of the probability forecasts issued, which are also found in this study.

The comparison between the three experiments has been carried out in terms of forecast quality, using a set of forecast quality measures that estimate potential skill, ensemble spread and probability forecast reliability, resolution and discrimination. Part of the relevance of the forecast quality assessment arises from the fact that, unlike in the climate-change problem, the future skill and reliability of seasonal and annual predictions can be estimated from predictions carried out for past events and, hence, could feed back information about the drawbacks of each approach when used for climate projections at longer time scales (Palmer et al. 2008). This is particularly the case for the perturbed-parameter ensemble, which has been used in a parallel experiment to assess anthropogenic climate change. This study used ensemble re-forecasts over the period 1991 to 2001 with two start dates per year carried out in a coordinated experiment with coupled global climate models started from realistic initial conditions. This is not a long period to carry out a comprehensive comparison, but given the significant computational cost of performing even such a relatively short re-forecast period, this effort offers a unique opportunity to assess, for the first time, the skill of the different approaches in a controlled setup. To take into account the relatively small sample size, estimates of the confidence intervals of the forecast quality measures have been obtained to determine which results can be considered statistically significant. This is an innovative approach that goes beyond more traditional assessments that tend to look at the performance of climate forecasts over tropical oceans.

All the forecast systems analyzed show a general level of forecast quality similar to each other and also to state-of-the-art systems (Palmer et al. 2004; Saha et al. 2006). In particular, predictions of tropical SST, which are at the origin of a large part of seasonal time-scale predictability, are more skilful than a basic statistical persistence model. In a more general context, seasonal predictions formulated with the 45-member multi-model ensemble show improved forecast quality with respect to the predictions from the two other experiments analysed. Based on comparisons of ensembles of equal size for lead times up to four months, it could be found that this superiority is intrinsic to the multi-model system used in this comparison for lead times shorter than five months. For longer lead times, the improvements of the multi-model are mostly due to its larger ensemble size, for which different institutions have pooled resources together. A comparison of a reduced multi-model with the same ensemble size as the perturbed-parameter and stochastic-physics experiments shows that for lead times longer than four months, the perturbed-parameter experiment gives

more often significantly better results than the reduced multi-model, especially in terms of probability forecast resolution and discrimination. The perturbed-parameter ensemble also has slightly better forecast quality than the stochastic-physics ensemble for seasonal predictions at those long lead times, while the opposite occurs for lead times shorter than five months. Unfortunately, it is not possible to use this result to claim that the stochastic physics method is superior to the perturbed parameter approach for lead times shorter than five months and vice versa for lead times longer than four months because the corresponding reference forecast systems rank in a similar way. Furthermore, it is important to bear in mind that the proportion of cases with statistically significant differences in the comparison between the reduced multi-model, the perturbed-parameter and the stochastic-physics ensembles are small, typically between 5 and 10%.

Annual-mean predictions have also been analyzed. Although the forecast quality is reduced with respect to seasonal forecasts at short lead times, there are a substantial number of cases with positive skill, which might lead to the development of specific applications of this type of predictions. As in the case of predictions at the seasonal time scale, the multi-model shows more often a better forecast quality than the other two experiments, although this is mainly due to its larger ensemble size.

Interestingly, when the stochastic-physics and perturbed-parameter ensembles are compared to their reference forecast systems, both improve the reliability of the probability forecasts, but only the stochastic-physics ensemble for lead times shorter than five months improves in terms of resolution and discrimination ability. This assessment suggests the need to understand the reduction with increasing lead time in the proportion of cases with significantly better ROCSS for the seasonal predictions of the stochastic-physics experiment.

It has been shown that the stochastic-physics and multi-model experiments, for short lead times, and the perturbed-parameter and the multi-model experiments, for long lead times, give slightly better forecast quality. These results suggest that all these methods deserve attention in the future. The multi-model is an ad-hoc approach that benefits from a large ensemble size from institutional collaboration. The perturbed parameter, while being an expensive method, it allows a large degree of optimization of the ensemble properties. The other approach, stochastic physics, is directly physically based and offers the opportunity to improve forecast quality while addressing a reduction of systematic model error. It is interesting that the stochastic-physics and perturbed-parameter experiments are competitive with the multi-model, while both methods can provide the opportunity to design ensembles according to specific criteria addressing the model inadequacy problem, hence providing an additional route to improved forecast performance beyond that of steadily developing the forecast models themselves.

The results presented in this paper are a lower limit of the forecast quality a user could expect from state-of-the-art seasonal forecast systems. Recent studies for target regions have shown that a careful post-processing (Coelho et al. 2006) and combination of information from different forecast systems, including statistical-empirical schemes (Stephenson et al. 2005), and predictor's expertise (Schubert et al. 2007) can significantly increase the forecast quality of seasonal predictions. This is certainly possible in our case, since the three forecast systems described here possess complementary strengths in their abilities to sample structural, parameter and initial-condition uncertainties. Furthermore, specific post-processing of the output of the forecast systems discussed here is desirable and might give improved results from those described above, in particular in terms of probability forecast reliability.

Acknowledgements

This work was supported by the ENSEMBLES project (GOCE-CT-2003-505539). The authors acknowledge the significant contributions by David Anderson, Magdalena Balmaseda, Judith Berner, Thomas Jung, Kristian Mogensen, Franco Molteni, Jean-Philippe Piedelièvre and Tim Stockdale. All the data used in this paper are available from http://www.ecmwf.int/research/EU_projects/ENSEMBLES/data/index.html.

8. References

Adler, R.F., G.J. Huffman, A. Chang, R. Ferraro, P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, P. Arkin and E. Nelkin, 2003. The version 2 vlobal precipitation climatology project (GPCP) monthly precipitation analysis (1979-present). *J. Hydrometeor.* **4**, 1147-1167.

Anderson, D. L. T., T. Stockdale, M.A. Balmaseda., L. Ferranti, F. Vitart, F. Molteni, F.J. Doblas-Reyes, K. Mogensen and A. Vidard, 2007. *Development of the ECMWF seasonal forecast System 3*. ECMWF Technical Memorandum 503 [Available from http://www.ecmwf.int/publications/library/do/references/show?id=87744].

Balmaseda, M.A., A. Vidard and D.L.T. Anderson, 2008. The ECMWF ocean analysis system ORA-S3. *Mon. Weather Rev.*, **136**, 3018-3034.

Berner, J., F.J. Doblas-Reyes, T.N. Palmer, G. Shutts and A. Weisheimer, 2008. Impact of a cellular automaton backscatter scheme on the systematic error and seasonal prediction skill of a global climate model. *Philosophical Transactions of the Royal Society A*, **366**, 2561-2579, 10.1098/rsta.2008.0033.

Bröcker, J. and L.A. Smith, 2007. Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651-661.

Coelho, C.A.S., D.B. Stephenson, M. Balmaseda, F.J. Doblas-Reyes and G.J. van Oldenborgh, 2006. Towards an integrated seasonal forecasting system for South America. *J. Climate*, **19**, 3704-3721.

Collins, M., B.B.B. Booth, G. Harris, J.M. Murphy, D.M.H. Sexton and M.J. Webb, 2006. Towards quantifying uncertainty in transient climate change. *Climate Dyn.*, **27**, 127-147.

Déqué, M., 2007. ENSEMBLES stream-1 hindcasts: from the season to the decade with four coupled models. Proceedings of the ECMWF Workshop on Ensemble Prediction, 133-139 (available from http://www.ecmwf.int/publications/library/do/references/list/12052007).

Doblas-Reyes, F.J., R. Hagedorn, T.N. Palmer and J.-J. Morcrette, 2006. Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. *Geophys. Res. Lett.*, **33**, L07708, doi:10.1029/2005GL025061.

Doblas-Reyes, F.J., C.A.S. Coelho and D.B. Stephenson, 2008. How much does simplification of probability forecasts reduce forecast quality? *Meteorol. App.*, **15**, 155-162.

Ferro, C.A.T., 2007. Comparing probabilistic forecasting systems with the Brier score. *Wea. Forecasting*, **22**, 1076-1088.

Gordon, C., C. Cooper, C.A. Senior, H. Banks, J.M. Gregory, T.C. Johns, J.F.B. Mitchell and R.A. Wood, 2000. The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.*, **16**, 147-168.

Graham, R.J., M. Gordon, P.J. Mclean, S. Ineson, M.R. Huddleston, M.K. Davey, A. Brookshaw and R.T.H. Barnes, 2005. A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model. *Tellus A*, **57**, 320-339, doi: 10.1111/j.1600-0870.2005.00116.x.

Hagedorn, R., F.J. Doblas-Reyes and T.N. Palmer, 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus A*, **57**, 219-233.

Hamill T.M. and J.S. Whitaker, 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Weather Rev.*, **134**, 3209-3229.

Hsu, W.-R. and A.H. Murphy, 1986. The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285-293.

Jin, F.-F., L. Lin, A. Timmermann and J. Zhao, 2007. Ensemble-mean dynamics of the ENSO recharge oscillator under state-dependent stochastic forcing. *Geophys. Res. Let.*, **34**, L03807, doi:10.1029/2006GL027372.

Jolliffe, I.T., 2007. Uncertainty and inference for verification measures. *Wea. Forecasting*, **22**, 137-150

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

Keenlyside, N.S., M. Latif, M. Botzet, J. Jungclaus and U. Schulzweida, 2005. A coupled method for initializing El Niño Southern Oscillation forecasts using sea surface temperature. *Tellus A*, **57**, 340-356, doi: 10.1111/j.1600-0870.2005.00107.x.

Keenlyside, N.S., M. Latif, J. Jungclaus, L. Kornblueh and E. Roeckner, 2008. Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, **453**, 84-88, doi: 10.1038/nature06921.

Lanzante, J.R., 2005. A cautionary note on the use of error bars. *J. Climate*, **18**, 3699-3703.

Murphy, A.H., 1986. A new decomposition of the Brier score: Formulation and interpretation. *Mon. Weather Rev.*, **114**, 2671-2673.

Murphy, J.M., D.M.H. Sexton, D.N. Barnett, G.S. Jones, M.J. Webb, M. Collins and D.A. Stainforth, 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768-772.

Nicholls, N., 2001. The insignificance of significance testing. *Bull. Amer. Meteorol. Soc.*, **82**, 981-986.

Palmer, T.N., 2000. The prediction of uncertainty in weather and climate forecasting. *Rep. Prog. Phys.*, **63**, 71-116.

Palmer, T.N., 2001. A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction. *Q. J. R. Meteorol. Soc.*, **127**, 279-304.

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

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

Palmer, T.N., F.J. Doblas-Reyes, A. Weisheimer and M. Rodwell, 2008. Towards "seamless" prediction: Calibration of climate-change projections using seasonal forecasts. *Bull. Amer. Meteorol. Soc.*, **89**, 459-470.

Richardson, D.S., 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol. Soc.,* **127**, 2473-2489.

Rodwell, M. and F.J. Doblas-Reyes, 2006. Predictability and prediction of European monthly to seasonal climate anomalies. *J. Climate*, **19**, 6025-6046.

Rogel, P., A. T. Weaver, N. Daget, S. Ricci and E. Machu, 2005. Ensembles of global ocean analyses for seasonal climate prediction: impact of temperature assimilation. *Tellus A*, **57**, 375-386.

Roulston M.S. and L.S. Smith, 2003. Combining dynamical and statistical ensembles. *Tellus A*, **55**, 16-30.

Saha, S., S. Nadiga, C. Thiaw, J. Wang, W. Wang, Q. Zhang, H. M. Van den Dool, H.-L. Pan, S. Moorthi, D. Behringer, D. Stokes, M. Peña, S. Lord, G. White, W. Ebisuzaki, P. Peng and P. Xie, 2006. The NCEP Climate Forecast System. *J. Climate*, **19**, 3483-3517.

Salas y Melia, D., 2002. A global coupled sea ice-ocean model. *Ocean Modelling*, **4**, 137-172.

Schubert, S., R. Koster, M. Hoerling, R. Seager, D. Lettenmaier, A. Kumar and D. Gutzler, 2007. Predicting drought on seasonal-to-decadal time scales. *Bull. Amer. Meteorol. Soc.*, **88**, doi:10.1175/BAMS-88-10-1625.

Shutts, G. J., 2005. Kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.*, **131**, 079-3102.

Shutts, G. J. and T.N. Palmer, 2007. Convective Forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization. *J. Climate*, **20**, 187-202.

Smith, L.A., 2002. What might we learn from climate forecasts? *PNAS*, **99**, 2487-2492.

Smith D.M. and J.M. Murphy, 2007. An objective ocean temperature and salinity analysis using covariances from a global climate model. *J. Geophys. Res.*, **112**, C02022, doi:10.1029/2005JC003172.

Smith, D., S. Cusack, A.W. Colman, C.K. Folland, G.R. Harris and J.M. Murphy, 2007. Improved surface temperature prediction for the coming decade from a global climate model. *Science*, **317**, 796-799, doi: 10.1126/science.1139540.

Stephenson D.B., C.A.S. Coelho, M. Balmaseda and F.J. Doblas-Reyes, 2005. Forecast Assimilation: A unified framework for the combination of multi-model weather and climate predictions. *Tellus A*, **57**, 253-264, doi: 10.1111/j.1600-0870.2005.00110.x.

Stephenson, D.B., C.A.S. Coelho and I.T. Jolliffe, 2008. Two extra components of the Brier score decomposition. *Wea. Forecasting*, **23**, 752–757.

Stockdale, T.N., D.L.T. Anderson, J.O.S. Alves and M.A. Balmaseda, 1998. Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. *Nature*, **392**, 370-373.

Tang, Y., R. Kleeman and A.M. Moore, 2005. Reliability of ENSO dynamical predictions. *J. Atmos. Sci.*, **62**, 1770-1791.

Uppala, S. M. and 45 others, 2005. The ERA-40 reanalysis. *Q. J. R. Meteorol. Soc.*, **131**, 2961-3012.

Wang, B., J.-Y. Lee, I.-S. Kang, J. Shukla, C.-K. Park, A. Kumar, J. Schemm, S. Cocke, J.-S. Kug, J.-J. Luo, T. Zhou, X. Fu, W.-T. Yun, O. Alves, E. K. Jin, J. Kinter, B Kirtman, T. Krishnamurti, N. C. Lau, W. Lau, P. Liu, P. Pegion, T. Rosati, S. Schubert, W. Stern, M. Suarez and T. Yamagata, 2008. Advance and prospectus of seasonal prediction: Assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980-2004). *Climate Dyn.*, in press.

Webb, M.J., C. A. Senior, D. M. H. Sexton, W. J. Ingram, K. D. Williams, M. A. Ringer, B. J. McAvaney, R. Colman, B. J. Soden, R. Gudgel, T. Knutson, S. Emori, T. Ogura,Y. Tsushima, N. Andronova, B. Li, I. Musat, S. Bony and K. E. Taylor, 2006. On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles. *Climate Dyn*., **27**, 17-38.

Weigel, A.P., M. A. Liniger and C. Appenzeller, 2007. Generalization of the discrete Brier and ranked probability skill Scores for weighted multimodel ensemble forecasts. *Mon. Weather Rev.*, **135**, 2778-2785.