TECHNICAL MEMORANDUM

621

Forecast quality assessment of the ENSEMBLES seasonal-to-decadal Stream 2 hindcasts

F.J. Doblas-Reyes[1], A. Weisheimer, T.N. Palmer, J.M. Murphy[2] and D. Smith[2]

Research Department

[1] Current affiliation: IC3, Barcelona, Spain
[2] Met Office, Exeter, UK

April 2010

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

1 Introduction

The non-linear chaotic nature of the climate system makes dynamical climate model forecasts sensitive to small perturbations introduced by both the initial state of forecasts and variations in model formulation. Thus, individual forecasts with one fixed model are of limited value. Instead, ensembles of forecasts are used to assess the range of possible evolutions of future climate on a range of timescales. Four techniques to represent different types of uncertainties and generate ensembles for seasonal-to-decadal (s2d) forecasting have been explored in ENSEMBLES:

1. Uncertainties in the initial conditions of each single model are accounted for by generating an ensemble of slightly different atmospheric and ocean analyses (*Palmer*, 2000). The perturbations of the initial conditions are either of an optimal statistical nature or based on insight into the dynamics of the physical system.
2. Forecast error occurs partly because climate models cannot simulate every single aspect of the climate system with arbitrary detail. The multi-model method partially samples errors that occur due to structural inadequacy in individual climate models (e.g. different model formulations and approximations, systematic biases) by using different coupled models (*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.
3. There are uncertainties in the specification of the parameters that are used in the parameterizations in climate models. Many of the physical parameters in the models either do not have a direct equivalent in the real climate system, or their numerical values are not precisely known. By perturbing these parameters within a single model, errors in specific parameterizations can be accounted for *(Murphy et al.*, 2004*; Staniforth et al.*, 2005).
4. Due to the coarse finite spatial resolution of the typical climate models, the representation of processes on spatial scales smaller than the truncation scales and their feedback on larger scales remains subject to considerable uncertainty. The impact of unresolved scales can be approximated by stochastic perturbations of the physical tendencies in the model (*Palmer*, 2001; *Shutts*, 2005).

In s2d forecasting, the combination of the initial-condition ensemble methodology (1) with the multi-model concept (2) leads to the *multi-model ensemble approach*. Together with the *perturbed physical parameter* (3) and *stochastic parameterisation* (4) *strategies* which can also be used in ensemble mode, it forms the basis of the s2d ensemble prediction system in the ENSEMBLES project. The main goal of this report consists in assessing the relative benefit of approaches (2), (3) and (4) to represent the impact of model uncertainty on forecast error. It is important to bear in mind that these three approaches address different aspects of model inadequacy.

Several ENSEMBLES project reports (deliverables D1.8, D1.17 and D1.18, Milestone M1.2 and Major Milestone MM1.2, all of them, as well as the other ENSEMBLES documents mentioned below, available from http://ensembles-eu.metoffice.com) give a preliminary assessment of the relative merits of the three approaches to represent model uncertainty. The present report, after summarizing the experimental setup of the Stream 2 s2d simulations in Section 2, focuses on documenting the progress and updated results since the analysis of Stream 1, the first intercomparison experiment carried out by ENSEMBLES. Section 3 shows results from the seasonal and annual hindcasts, while Section 4

describes the preliminary results obtained with the innovative decadal hindcasts. A summary and outlook on future work beyond ENSEMBLES are given in Section 5.

2 Experimental setup of the Stream 2 s2d hindcast simulations

A coordinated set of forecast experiments, labelled Stream 2, has been performed using forecast systems that employ the three approaches. The following integrations were carried out, depending on the forecast time scale, over the period 1960-2005:

- Seasonal: seven-months long, every year starting on the 1st of February, May, August and November.
- Annual: 14-months long, every year starting on the 1st of November.
- Decadal: 10-year integrations, once every five years starting in November.

The ENSEMBLES multi-model ensemble (MME) for seasonal-to-annual hindcasts comprises global coupled atmosphere-ocean climate models from the UK Met Office (UKMO), Météo-France (MF), the European Centre for Medium-Range Weather Forecasts (ECMWF), the Leibniz Institute of Marine Sciences at Kiel University (IFM-GEOMAR) and the Euro-Mediterranean Centre for Climate Change (CMCC-INGV) in Bologna (Italy). The five systems contributed to the seasonal experiment, and all of them performed the annual hindcasts except for CMCC-INGV. All models include the main radiative forcings. None of the coupled models has flux adjustments. The atmosphere and ocean were initialized using realistic estimates of their observed states and each model was run from an ensemble of nine different initial conditions. Table 1 summarizes the main model components, forcings and initialization strategies.

Four (ECMWF, UKMO, CMCC-INGV and MF) of the five modelling groups used a similar approach to initialise the ocean state by adding small perturbations of the wind stress and SST fields. This strategy is illustrated in Figure 1. Ensembles of ocean analyses were generated by perturbing the wind-stress. At the start of each coupled hindcast, additional SST perturbations were added to the ensemble of ocean analyses. MF added SST perturbations directly to their nine ocean analyses. The perturbations are random samples from the differences between two quasi-independent data sets for wind stress and SSTs. These quasi-independent data sets have been generated based on different analysis schemes and using various observational data. They also vary in their lengths and averaging periods. The differences between these analysis data sets reflect the inherent uncertainties in the wind stress and SST analyses. The data sets used are CORE (*Large and Yeager*, 2004) and ERA-40 (*Uppala et al.*, 2005) for wind-stress and Reynolds 2DVAR, Reynolds OIv2, NCEP ERSSTv2 and HadISST1.1 for different historical periods of SSTs. UKMO follows the same initialization strategy of ECMWF and MF, but uses an improved anomaly assimilation approach for soil moisture.

One of the models (IFM-GEOMAR) used in the MME followed a substantially different initialisation methodology. Initial conditions were obtained from three coupled simulations in which the model SSTs were restored to observations in the band 30ºS-30ºN, following the method described by *Keenlyside et al.* (*2005*). The simulations covered the period 1950-2005, included all major radiative forcings and were initialised from three 20th-century simulations. The initial conditions for the nine ensemble members were obtained from different initial condition permutations from these three runs. Radiative forcings in the hindcast and initialisation experiments were identical.

The MME for the decadal forecasts is made of four of the forecast systems used in the Stream 2 seasonal-to-annual forecasts. CERFACS, ECMWF, IFM-GEOMAR and the UKMO, where CERFACS employed a version of ARPEGE/NEMO similar to that used by MF for seasonal-annual hindcasts, except for the sea-ice (where LIM was used instead of GELATO). ECMWF used the atmospheric IFS cycle 33r1 (*Bechtold et al., 2008*) with a horizontal truncation of T_L159 and 62 vertical levels, while IFM-GEOMAR employed the same version of the ECHAM5/MPI-OM1 as used in the seasonal-to-annual forecasts, but with differences in its initialization that are described below.

The experimental design, i.e. the initialisation, ensemble generation, radiative forcing, and model configuration, of the decadal forecasts was identical to that of the seasonal-to-annual hindcasts with the following exceptions. IFM-GEOMAR employed anomaly initialisation, where observed SST anomalies were added to the model climatology and the combined SST restored into the coupled model (see *Keenlyside et al*., 2008). CERFACS used a similar initialization to that employed by MF, but including also a simple sea-ice initialisation step based on the observed extension and the climatological thickness. This means that among the multi-model contributors, three systems are initialised in the "observation space", and another one is initialised in its own "model space".

DePreSys (*Smith et al.*, 2007) is a forecast system based on the HadCM3 climate model that contributed to both the seasonal-to-annual and the decadal experiments with a perturbed-parameter ensemble (PPE). The original DePreSys system employs a single variant of HadCM3 using standard settings for physical parameters following *Gordon et al.* (2000). For the Stream 2 hindcasts an updated system (DePreSys_PP) was created, employing an ensemble of model variants sampling modelling uncertainties through perturbations to poorly constrained atmospheric and surface parameters. These perturbations were applied to a configuration of HadCM3 updated to include a fully interactive representation of the sulphur cycle, and flux adjustments (calculated separately for each model variant) to restrict the development of regional biases in sea surface temperature and salinity (*Collins et al*., 2009). Eight model variants[1] with simultaneous perturbations to 29 parameters were used, added to the standard model variant to make a nine member ensemble. The eight perturbed variants were selected from a set of 128 combinations of model parameters previously used to simulate long term climate change (*Webb et al.*, 2006). The first stage of the selection process involved choosing 16 perturbed variants (*Collins et al*., 2009), consisting of the ensemble member giving the best simulation of present-day climate (according to the metric employed by *Murphy et al*., 2004), plus 15 members designed to maximise a non-dimensional measure of the average distance between members in terms of climate sensitivity and model parameter values. Eight model variants were then chosen from the 16 members, picked to sample 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). In order to create initial conditions for the hindcasts, each model variant was run in assimilation mode from December 1958 to November 2005, including time-varying radiative forcing derived from observed changes in well-mixed trace gases, ozone, sulphate and volcanic aerosol, and solar irradiance. During this integration, the atmosphere and ocean were relaxed towards atmospheric and ocean analyses, wherein values were assimilated as anomalies with respect to the model climate in order to minimize climate drift after the assimilation is switched off at the beginning of each prediction. Atmospheric analyses were taken from ERA-40 and ECMWF operational analyses), while analyses of ocean anomalies were created using an updated version of the *Smith and Murphy* (2007)

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

scheme, based on anomaly covariances calculated from the ensemble of model variants, with adjustments to improve the fit to observations. The anthropogenic forcing in the assimilation and hindcast integrations was the same, switching from values based on observations to values taken from the SRES A1B emission scenario after the year 2000. For well-mixed GHGs, global concentrations of CO2, CH4, N2O, CFC11, CFC12, CFC113, HCFC22 were specified. This involves a small element of cheating up to the year 2000, but in practice the difference between the observed and the SRES forcings would be quite small. For ozone, 2-D (zonal mean latitude-height) fields were specified. For sulphate aerosols, the model used emissions, converted into concentrations via a parameterisation of aerosol chemistry and transport included in HadCM3. Anthropogenic emissions of surface and elevated SO2 and land-based DMS were specified. Also natural SO2 emissions from quasi-continuously erupting volcanoes, distributions of chemical oxidants (OH, HO2, H2O2 and O3) and seawater DMS concentrations (assumed not to have any anthropogenic component) were added. The aerosol scheme models the direct and first indirect effects of sulphate radiative forcing, but not the second (or "cloud lifetime") indirect effect. Forcings from soot or biomass aerosols were not included in DePreSys. The solar activity forcing was considered by repeating an estimate of the previous observed 11-year cycle, and volcanic forcing is included by specifying volcanic aerosol concentrations in four equal area latitudinal zones (90°S-30°S, 30°S-0°, 0°-30°N, 30°N-90°N). The aerosols are distributed evenly above the tropopause with a constant mass mixing ratio. The volcanic aerosol present at initialisation is reduced exponentially with an e-folding time scale of one year, thus avoiding assuming prior knowledge of natural forcing events such as the eruption of Mount Pinatubo.

A new version of the stochastic physics approach (*Palmer et al*., 2009) has been used for a stochastic-physics ensemble (SPE) seasonal hindcast experiment with the ECMWF coupled model. The approach is based on the idea of a stochastic representation of the equations of motion at the computational level and as such focuses on uncertainty related to unresolved processes. Conventional physical parameterization schemes describe the effects of subgrid-scale processes in models of weather and climate by deterministic bulk formulae which depend on the local resolved-scale variables. However, through the upscale cascade of energy, the neglected unresolved subgrid-scale variability can have an impact on the larger scales in the model and thus contribute to model errors on different spatial and temporal scales. SPEs provide a methodology for representing model uncertainty due to variability of the unresolved scales. ECMWF has recently revised its stochastically perturbed parameterization tendency (SPPT) scheme and developed the stochastic backscatter scheme (SPBS; *Palmer et al*., 2009). SPPT applies univariate Gaussian perturbations to the wind, temperature and humidity tendencies of physical processes in the form of multiplicative noise with a smoothly varying pattern in space and time. A two-scale version of the perturbations with a shorter characteristic spatio-temporal scale on the order of 6 hours and 500 km together with a longer characteristic spatio-temporal scale of 30 days and 2500 km has been used. The SPBS scheme is based on the idea of backscatter of kinetic energy from unresolved scales. It is formulated in terms of a spectral streamfunction forcing field estimated from the numerical, convective and orographically induced dissipation rate and uses vertical phase correlations. A preliminary set of SPE seasonal hindcasts for the May and November start dates over the period 1991-2005 have been completed and results of which will be discussed in this report. ECMWF plans to extend these hindcasts to the full Stream 2 period and start dates as soon as the detailed settings of the scheme will be finalised.

All the simulations, except those run by CMCC-INGV, were carried out on ECMWF's super-computers. Model output is available either from the ECMWF mass storage system (MARS), or from a new server for public data dissemination, which has been installed for ENSEMBLES as described in Deliverable D2A4.3 and Milestone M2A.4.2. For more information and access to the server, the reader is invited to visit
http://www.ecmwf.int/research/EU_projects/ENSEMBLES/data/data_dissemination.html.

3 Forecast quality assessment of the seasonal and annual hindcasts

The basis of predictability on seasonal timescales lies in the slowly evolving components for the climate system, like the ocean, that act as boundary conditions for the atmosphere with its shorter intrinsic time scales. A prime example of a coupled atmospheric and oceanic phenomenon is the ENSO event in the tropical Pacific, which is the dominant mode of seasonal and interannual climate variability. ENSO affects the weather and climate remotely in many parts of the globe via well-known teleconnections. Because of its direct importance for seasonal predictability in the tropics as well as in the extratropics, in this chapter we will first analyse SSTs with regards to their systematic errors and forecast quality in Sections 3.1 and 3.2, before discussing air temperature and precipitation predictability in Section 3.3.

3.1 Systematic errors in SSTs

Although initialized using observations, seasonal forecast models develop, over the forecast time, systematic errors that lead the models to drift away from the observed state. Figure 2 shows the mean model drift for the SST, estimated from all ensemble members and hindcasts, in the Niño3 region (5ºS-5ºN, 150ºW-90ºW) for each of the four start months.

The annual cycle of the model SSTs is in a good agreement with observations (panel b). Considerable progress has been made since DEMETER in reducing the systematic SST errors, in particular on longer lead times. While the SST drift in DEMETER (not shown, see *Weisheimer et al.*, 2009) ranged between +2º C and -7º C for up to six months lead, the ENSEMBLES models have a much reduced drift with an overall amplitude of less than ±1.5º C. Global maps of SST biases at lead times of 2-4 and 4-6 months for the ENSEMBLES models are shown in Figure 3. Biases are generally small over the tropics. However, there are still substantial areas, e.g. over the cold upwelling regions at the eastern ocean boundaries, where systematic errors are large and have, despite all efforts, not much improved in ENSEMBLES.

3.2 Tropical sea surface temperature

The systematic errors discussed in Section 3.1 have been corrected by linearly removing the long-term mean over the hindcast period for a given start date and lead time. The corrections were applied in cross-validation mode (by leaving one out) in order to emulate real-time forecast conditions as closely as possible.

Figure 4 shows the temporal evolution of ensemble-mean root-mean square error (RMSE), ensemble spread and anomaly correlation coefficient (ACC) for the Niño3 SST hindcasts anomalies for all individual ENSEMBLES models (panels a and c) and for the multi-model ensemble MME (panels b and d). For comparison, a simple statistical persistence hindcast is also given. Note that the individual

model ensembles consist of 9 members, while the multi-model ensemble based on 5 models has 45 members in total.

The interpretation of the results described below, and in section 4, is complicated by a number of factors. Firstly, differences in skill between the systems will in general depend upon differences in the models and initialisation techniques used, as well as upon the different techniques for sampling modelling uncertainties. For example, the perturbed parameter approach (DePreSys_PP) is applied within a model typical of the class of models used for decadal-centennial climate change projections, but of lower resolution compared to the models used in the multi-model and stochastic physics approaches (Table 1). On the other hand, the DePreSys_PP experiments have the advantage (in a decadal prediction context) of having been built upon a model tested extensively in previous longer term projections, whereas this is not the case for most of the other models. A second complicating factor is that the different ensemble forecasting systems have, in general, different ensemble sizes. While the full MME consists of 45 ensemble members, the individual model ensemble, the perturbed parameter ensemble (PPE) and the stochastic physics ensemble (SPE) are based on ensembles of nine members only. The ensemble size does have an effect on the scores estimated and, in general, favours ensembles with a larger size because of reduced sampling uncertainty. Except for the probabilistic Brier skill score, we are not aware of any analytic expressions to compute the scores that take into account the effect of limited ensemble size on the scores. A more detailed discussion of this issue and possible suggestions to overcome it is presented in Section 3.3.

In a perfect ensemble, over a large number of ensemble forecasts, the RMSE of the ensemble mean would equal the ensemble spread about the ensemble mean. A general feature of all single-model ensembles is, however, that the ensemble spread is substantially smaller than the RMSE (Figure 4a), that is, each individual ensemble is strongly underdispersive, a known source of overconfidence in probability forecasts. As it has been demonstrated in numerous studies (e.g., *Palmer et al.,* 2004; *Weigel et al.,* 2008), the multi-model combination effectively reduces the RMSE while the ensemble spread is increased leading to overall improved skill. For the ENSEMBLES MME SSTs this leads to an almost perfect match between the RMSE and spread (Figure 4b). Results for anomaly correlation in Figure 4c and 4d indicate that the MME also improves the correlation skill versus the individual models with correlations above 0.7 at seven-month lead time.

The reader is reminded that the MME and the perturbed physical parameter ensemble (PPE) have completed the full set of Stream 2 seasonal-to-annual hindcasts. They consist of seven-month long seasonal hindcasts over the Stream 2 hindcast period 1960 to 2005 started in February, May, August and November of each year, and 14-month long annual hindcasts over the same period but only started in November of each year. Figure 5-8 show the RMSE, ensemble spread and anomaly correlation as a function of lead time for the MME (panels on the left) and the PPE (panels on the right) for the seasonal hindcasts. Over the tropical ocean, the MME reveals a near-perfect match between RMSE and ensemble spread, while the PPE is underdispersive by not developing enough spread. For the Niño regions in the tropical Pacific, the MME has a slightly larger spread than RMSE, whereas the PPE is underdispersive. The deterministic scores RMSE and ACC show more skill in the MME than in the PPE. It is worth noting that the MME in these analyses performs better (or equally well) than a simple statistical reference forecasting system that uses persisted SST anomalies.

A similar comparison for the annual forecast range is displayed in Figure 9-12. There is substantial skill in predicting the SSTs on these long lead times for the tropics as a whole and in particular for the Niño regions. In contrast with Figure 5-8, the climatological hindcast is the statistical reference forecasts for the annual predictions, which, by definition, has an ACC of zero at all lead times. For the MME, the anomaly correlation drops to 0.5 at around month nine and remains nearly constant thereafter. Remarkably, the above-mentioned good match between the RMSE and ensemble spread is further sustained over the extended forecast lead time with an approximately linear error and spread growth. The PPE performs roughly similar with slightly larger RMSE and lower ACC than the MME.

As described in the Introduction and in Section 2, ensemble seasonal hindcasts with a stochastic physics approach have also been performed. By the time of writing, the SPE simulations have not completed the full Stream 2 start dates. Thus, in the following a subset of seven-month long hindcasts started in May and November each year over the reduced hindcast period 1991-2005 will be used to compare results from these simulations with the MME and PPE.

Figure 13-19 summarise the performance of the three systems in terms of tropical Pacific SSTs. The MME is characterized by an overall very good agreement between RMSE and ensemble spread over all lead times. The PPE is underdispersive with larger RMSE than the MME. The stochastic physics ensemble shows a good match between forecast error and spread over the first few months of the forecast, but become overdispersive for longer lead times. The RMSE is comparable with, or smaller than, the corresponding error from the PPE.

Figure 15, Figure 17 and Figure 19 show a comparison of the three systems in terms of Brier skill score (BSS) for the lower and upper tercile SST events on seasonal mean lead times of 2-4 and 5-7 months for different start dates. These scores are computed using grid point values instead of regional averages as in, for instance, Figure 13. The reference hindcasts for computing the BSS is climatology. The BSS computation is based on the method introduced in *Ferro* (2007), which uses an analytical expression to estimate the BSS for a hypothetical ensemble with infinite size. The uncertainty ranges have been estimated using 1,000 bootstrap resamples with replacement. As it becomes evident from these plots, the skill varies with region, lead time, start date and analysed event. The mentioned overconfidence of the PPE hindcasts, that is too little spread in the ensemble compared with the mean forecast error, seems to translate into generally smaller BSS values. The MME gives the highest scores, while the SPE shows relatively high skill, although not as high as the MME.

3.3 Air temperature and precipitation

In this section a comparison of the seasonal-to-annual skill of the three ensemble forecasting systems MME, PPE and SPE is discussed in terms of near-surface air temperature and precipitation. As mentioned above, due to its nature, the three systems have different ensemble size: 45 for the MME and nine for both the PPE and the SPE. The estimation of skill measures are affected by sampling uncertainty. An ensemble with n members can, at best, resolve relative frequencies (or probabilities) of 1/(n+1). It is also well-known that some skill scores, e.g. the BSS, suffer from a negative bias for small ensemble sizes (*Müller et al.*, 2005; *Weigel et al.*, 2007). These problems make it very difficult, if not impossible, to directly compare scores obtained from ensembles of different sizes.

One way around the problem is to create ensembles of similar sizes by re-sampling the larger ensemble (MME) so that it matches the size of the smaller ensembles (PPE and SPE). With this approach one needs to ensure that the re-sampling is robust, i.e. that the basic statistical properties of the forecasts are not changed in the re-sampled ensemble of smaller size. One way to test this is to

compare the BSS that takes into account some aspects of the impact of the ensemble size (such as the debiased BSS after *Müller et al.*, 2005 and *Weigel et al.*, 2007 or the infinite ensemble size BSS after Ferro, 2007) for both the re-sampled ensemble of smaller size and the full ensemble. The agreement of the scores between the smaller size re-sampled ensemble and the full-size ensemble can be taken as an indication of the adequacy of the re-sampling. To our knowledge, there are no corrected versions of other probabilistic scores available. Hence, the full 45-member multi-model ensemble has been re-sampled to match the ensemble size of nine for the PPE and SPE. The re-sampled MME with nine ensemble members is known henceforth as "reduced MME". It has been constructed by taking the first two ensemble members from each of the five individual models except for the CMCC-INGV model, from which only the first member has been used.

Figure 20 shows the impact of the ensemble size in terms of ACC, BSS, reliability skill score (REL) and resolution skill score (RES). Each panel is a scatter plot of scores for near-surface temperature and precipitation over six large-scale regions (tropical land and ocean, NH extratropical land and ocean, SH extratropical land and ocean) with values obtained from the full MME with 45 members on the horizontal axis and the corresponding values from the reduced MME with nine members. The numbers in the corners indicate the percentage of cases where each system is superior. Values in brackets are those for statistically significant differences with 95% confidence level. Panel a shows that, for the vast majority of cases, the ACC for the full MME is larger than the corresponding values from the reduced MME. Assuming that the reduced MME is a robust subsample of the full MME, this fact points towards a negative ACC bias for small ensemble sizes. Panel b shows a scatter plot of the debiased BSS. The debiased BSS corrects for the negative BSS bias due to small ensemble sizes. As can be seen, the scores are clearly (and significantly) higher for the reduced MME with almost 94% of all cases having larger BSS than for the full MME. Most of these differences are statistically significant. Similar results were found when the infinite ensemble size BSS (*Ferro*, 2007) was used. This result suggests several things: a) the reduced MME is not a perfect subsample of the full MME as it does not preserve its statistical characteristics in terms of BSS, b) the debiased BSS might not be a perfect way to effectively remove all impact of ensemble size on the BSS, c) due to the existing bias of the reduced versus full MME, all further comparisons and conclusions based on the reduced MME have to be interpreted with caution. Points a) and b) are probably not independent from each other and the effect found can as well come from the combined influence of imperfect re-sampling and imperfect BSS estimate. However, due to the lack of any other sound methodology to account for the different sizes of the ensembles, we will proceed discussing the reduced MME and the debiased BSS. Panels c and d of Figure 20 show scatter plots for the reliability and resolution component of the BSS. Reliability is the most affected component by the ensemble size, the full MME showing the largest reliability, while the resolution also shows an increase with ensemble size in favour of larger ensembles.

A comparison of the MME and PPE performance in terms of ACC, debiased BSS, REL and RES similar to Figure 20, is presented in Figure 21 and Figure 22. The MME performs better than the PPE using both the full and reduced MME. While there is a clear impact of ensemble size on the ACC, the interpretation of the debiased BSS results is somewhat hampered by the fact that the debiased BSS is in favour of the reduced MME when compared to the full MME (panel b in Figure 20). Both reliability and resolution are also higher for the MME than for the PPE.

On longer than seasonal time scales with lead times of 6-8, 8-10, 10-12 and 12-14 months the results are slightly changed, as shown in Figure 23. The comparison of the debiased BSS of the PPE with the full MME (panel a) reveals that the PPE has higher scores than the MME in roughly 60% of the cases. However, if the comparison is done with the reduced MME (panel b), this number drops to only 27%. Again, interpretation of these results is hindered by the fact that the debiased BSS is in favour of the reduced MME when compared to the full MME (panel b in Figure 20). The main conclusions from the seasonal time scales in terms of reliability with better scores for the MME still hold for these longer forecast ranges (Figure 24). However, the resolution components in the MME and PPE are now closer at this longer lead time (Figure 24). For cases with high resolution the MME tends to perform better, whereas for cases with low resolution the PPE performs best.

A comparison of all three ensemble forecasting systems MME, PPE and SPE is given in Figure 25 and Figure 26. Here either the full MME or the reduced MME is compared with the PPE and the SPE. Whereas the MME turns out to be the most skilful overall system in terms of debiased BSS (Figure 25), the SPE performs relatively better versus the MME than the PPE does. For example, the full MME has higher scores than the PPE in approximately 60% of the cases. If compared with the SPE, the MME has higher scores only in roughly 44% of the cases. While the SPE seems slightly better than the full MME in terms of debiased BSS (panel c), this is no longer the case if the reduced MME is considered instead of the full MME (panel d). As discussed above, this apparent paradox is due to the imperfect sampling of the reduced MME and the limitations of the debiased BSS. A comparison of the reliability and resolution components of the BSS and the ROC skill score for the PPE and the SPE in Figure 26 shows that the SPE has more often better reliability and resolution and equal ROC skill scores than the PPE. Note that this comparison is not affected by ensemble size as both the PPE and the SPE have a similar ensemble size.

The above mentioned comparisons are based on near-surface temperature and precipitation over six large-scale regions around the globe. A skill assessment, also for near-surface temperature and precipitation, but for 21 standard regions over land (*Giorgi and Francisco*, 2000) is shown in Table 2. The table summarises, for each of these regions, the BSS estimated for an infinite size ensemble for the MME, PPE and SPE for lower and upper tercile temperature and precipitation events in JJA and DJF for lead times of 2-4 months. The scores vary with the region, variable, season and event. Overall the MME results in the largest number of positive scores (indicated by green cells in the table) with many of them being significantly positive (indicated by bold and underlined font).

Within the 21 regions, two are European: Southern Europe (labelled MED) and Northern Europe (NEU). All three systems have significantly positive skill in predicting summer temperatures in Southern Europe. Precipitation is, in general, less skilful although the PPE, and to some extent also the SPE, have significantly positive skill in predicting summer precipitation over Northern Europe.

4 Forecast quality assessment of the decadal hindcasts

The basis for investigating decadal prediction rests on evidence from observed low frequency climate variations, results from idealised modelling studies and evidence that forced climate change can also provide skill (see *Meehl et al.*, 2009, for a recent review). The ENSEMBLES Stream 2 decadal hindcasts provided a first opportunity to assess the benefits of combining projections from different models in a coordinated experiment, following initial studies carried out with individual climate models (*Smith et al.*, 2007; *Keenlyside et al.*, 2008; *Pohlmann et al.*, 2009). The inevitable existence of biases in the models used for the decadal hindcasts necessitates (as in seasonal prediction) the use of

strategies to account for these systematic errors when comparing forecasts against observations. This was achieved by expressing each hindcast as anomalies relative to either a long-term model climatology (if available), or to the average of other hindcasts. Some groups also adopted a strategy of initialising their hindcasts using observed anomalies added to an estimate of the model climatology in order to reduce model drift during the hindcasts, whereas other groups initialised the predictions using full observed fields in order to provide starting conditions as close as possible to the real climate system. The optimal strategy for decadal forecast production in the presence of model biases remains an open question for future work.

4.1 Air temperature and precipitation

Model inadequacy causes forecasts to drift away from the observed climate towards an imperfect model climate. In some cases model drift is usually not described in detail because the initialization of the hindcasts is carried out by assimilating observed anomalies into the model climate (*Smith et al.*, 2007; *Keenlyside et al., 2008*), a method that is expected to reduce the model drift. In other cases, the reason lies on the idea that the drift is small enough to not destroy the initial-condition information (*Pohlmann et al., 2009*). Here we consider that forecast drift is an important feature of the decadal forecasting problem worth discussing.

Figure 27 shows examples of the global-mean near-surface temperature for the multi-model components. While CERFACS, ECMWF and HadGEM2, which are initialized with realistic initial conditions, have a strong drift that is in some cases larger than 2 K, IfM does not have an apparent drift, a consequence of the initialization being carried out close to the attractor of the coupled model. The spatial distribution of the bias for the average prediction over the forecast years 2 to 5 is shown in Figure 28, where the maps show mostly a cold bias over the oceans, except for a warm bias over the eastern tropical basins, and a variable behaviour over land. In contrast with what Figure 27 would suggest, IfM also has some biases, which tend to cancel out when averaging over the globe. Figure 28 also displays an estimate of the bias in variability by showing the ratio of the predicted and observed standard deviation. The standard deviation has been computed using the four-year mean available for each ensemble member. Three of the systems (IfM, HadGEM2 and ECMWF) overestimate the ratio, which means that the variability of the hindcasts is higher than that found in the re-analysis, while CERFACS tends to underestimate it except in some regions like the West tropical Pacific. Figure 29 depicts the mean biases for precipitation, with an underestimation in the extratropics and an important overestimation over the tropical regions for all systems except for CERFACS. This system, in agreement with an excessively cold system, underestimates the mean precipitation almost everywhere. The tropical biases are suggestive of a double ITCZ and an excessive penetration of the equatorial Pacific cold tongue towards the west.

Given the sizable drift of the forecast systems, only hindcast anomalies have been considered to assess predictability, following standard practice in seasonal forecasting. Forecast anomalies have been estimated in cross-validation mode by removing the mean model climate for the specific forecast period using the hindcasts for which there are reference data available, again following common practice in seasonal forecasting. For instance, to obtain the anomalies of the average 5-10 year forecast period from the hindcasts initialized in November 1970, the model climate is estimated by averaging the data for the 5-10 year forecast period from all the hindcasts for which there is reference data, except the hindcast started in November 1970. This implies that data from the 1960, 1965, 1975, 1980, 1985, 1990 and 1995 hindcasts (seven start dates) are used, because no reference data for the period

2005-2010 and 2010-2015 (i.e., the verifying dates of the hindcasts with start dates in 2000 and 2005) were available. The anomalies for the reference dataset are estimated for the same calendar period. Figure 30 displays the anomalies of the global-mean near-surface temperature. Each system has slightly different characteristics, with IfM having the largest variability and CERFACS the lowest. The re-analysis shows an upward trend from the early 1980s. This upward trend is shown by all the systems, although they also have a different degree of fit to the interannual variations found in the re-analysis. The spread is large in most cases, the ensemble encompassing the observations. However, the reader should bear in mind that a three-member ensemble is too small to extract any conclusions about the spread behaviour.

Figure 31a depicts a set of hindcasts made with DePreSys_PP similar to those shown in Figure 27. Note that this is only a reduced set, as DePreSys_PP has one hindcasts started each year, instead of one every five years. The plot illustrates the behaviour of the nine versions used to construct the perturbed-parameter ensemble, with each version being systematically colder or warmer than the others. When anomalies with respect to the climatology obtained from the hindcast dataset are considered (Figure 31b), the hindcasts reproduce the upward trend as already found for the systems contributing to the multi-model, with DePreSys_PP overestimating the warming in the most recent hindcasts.

The ensemble-mean correlation has been used to estimate the skill of the predictions. Figure 32a shows that the correlation of the global-mean near-surface temperature for the four systems contributing to the multi-model is, in most cases, larger than 0.8. The differences in skill between the systems are quite likely not statistically significantly different with high confidence, except for the initial years of the IfM system. A remarkable increase of skill with forecast time is found for all systems, a feature attenuated in the CERFACS hindcasts. Figure 32b shows the ensemble-mean correlation for DePreSys_PP, this time using all the hindcasts available for this experiment. Although DePreSys_PP also shows variations in skill, there is not an obvious increase with forecast time. However, when DePreSys_PP is sampled taking only one hindcast every five years trying to mimic the sampling used for the multi-model, the ensemble-mean correlation (not shown) suggests a behaviour similar to that shown in Figure 32a. This suggests that at least part of the increase in skill with forecast time might be due to the small sample used in the multi-model ensemble, the sample of predictions with a shorter forecast time (one to four years) includes a disproportionately large contribution from dates covering the period before 1970 compared to the predictions with a posterior valid date. This simple effect might induce an apparent increase in skill with forecast time. To test this hypothesis, the DePreSys_PP correlation has been computed with the hindcasts initialized in the period 1960-1970, which is a period when there is not a clear trend in the re-analysis, the skill being very close to zero. Instead, the skill is similar to that shown in Figure 32b when hindcasts with a one-year interval in start dates and started later than 1985 are considered. However, the reduced sample of the multi-model might not be the only reason of the increase of skill with forecast time. Another test comparing the ensemble-mean correlation DePreSys_PP hindcasts (Figure 32b) against a parallel set of uninitialised hindcast experiments performed with DePreSys_PP (NoAssim, see Section 4.3) shows that the skill of NoAssim does not change with forecast time regardless of the sampling. This suggests that the initialization of the hindcasts could also play a role, the impact of the initialization being more apparent once the model is well into the forecast.

An illustration of the spatial distribution of the skill for near-surface temperature is shown in Figure 33. The systems have statistically significant skill over large regions, especially over the tropics and the North Atlantic, but also over large parts of the continents and even Europe. Interestingly, both the multi-model and DePreSys_PP seem to have a similar distribution of the skill, with the largest differences appearing over the tropics. Figure 33g shows the skill for DePreSys_PP when one-year interval start dates are used. A comparison with the skill of Figure 33e shows that the spatial distribution doesn't change substantially and that although the values are reduced, the areas with significantly positive skill increase.

A more thorough description of the skill appears in Figure 34. The ensemble-mean correlation of 850-hPa air temperature over the Northern and Southern Hemispheres and the tropical band are shown for all the forecast systems as a function of the forecast time. The four components of the multi-model ensemble have a similar skill and is of the same order as that of DePreSys_PP, which has been computed with nine instead of three-member ensembles but with the same sample of start dates as the multi-model. Encouragingly, the multi-model ensemble mean, which consists of the average of 12 individual projections, gives higher scores on average than the single models. However, a large overlapping of the 95% confidence intervals warns about extracting conclusions on the relative superiority of one system with respect to another. Additional analyses carried out with DePreSys_PP but with one-year start date intervals shows much reduced confidence intervals, which supports the need of carrying out larger samples of start dates when trying to assess the relative merits of different forecast systems or even of sensitivity experiments.

Figure 34 shows that the skill increases for longer forecast times, being larger for the 6-10 year average (i.e., five years ahead) than for the 2-5 year average. Several hypotheses can be considered to explain this behaviour. This is in agreement with the results shown in Figure 32. To test the same hypotheses considered above, the ensemble-mean skill was estimated for both DePreSys_PP Assim and NoAssim with one-year interval start dates. The characteristics of the DePreSys_PP NoAssim experiment will be described in more detail in Section 4.3. In both cases the skill increased with forecast time over the Northern Hemisphere and the tropics, while it did not over the Southern Hemisphere. The higher sampling frequency of the DePreSys_PP experiments discards any role (at least for regions out of the Southern Hemisphere) of the longer forecast periods take more data from later dates (when the trend is stronger) when just one start date every five years is available. As the NoAssim experiment, which has not been initialized using observational information, also shows the increase of skill with forecast time, a possible impact of the initialization linked to some sort of initial shock or to the damping of the volcanic aerosol within the hindcast can also be discarded as the main reason for the increase in skill. A possibility is that the increased skill results from the slightly larger degree of temporal averaging performed for the longer lead times in Figure 34.

4.2 Ocean variables: ocean heat content and Atlantic meridional overturning circulation

Figure 35 shows the drift of the ocean temperature averaged over the top 300 metres (a proxy for the upper ocean heat content) for the forecast period two to five years. Data were available for just three of the four contributors to the multi-model: ECMWF, CERFACS and IfM. The drift has been computed with respect to the corresponding re-analysis, i.e., the one used to initialize the hindcasts. The reader should be aware that these ocean re-analyses (except in the case of IfM, which does not use sub-surface observations) are affected by a known deficiency in the XBT data assimilated. The drift in

the ocean temperature bears some similarity with the pattern found for near-surface air temperature, although there are some differences: the tropical cooling tends to be more confined, there are warm areas close to the Equator even in those systems with overall SST cooling and complex structures appear in the extratropics. In particular, there are large errors in the western boundary currents, which are warmer in the model than in the verifying re-analysis for both CERFACS and ECMWF.

To better illustrate the predictions of upper ocean heat content, Figure 36 depicts the hindcasts averaged globally and over the tropics. Data for the three ocean re-analyses are also displayed in the figure to give an idea of the uncertainty in the estimates of the ocean state. While ECMWF and CERFACS have a strong cold drift, IfM shows a slight warming of the hindcasts before 1980 and a stronger one, when the IfM analysis shifts to a cooler mean state, after that date. Note that the relatively small drift of the IfM system is the consequence of the compensation of the cold and warm systematic errors shown in Figure 35. The corresponding anomalies show a large variability. The global-averaged hindcast anomalies of ECMWF and CERFACS have a rising trend, except for the 1960 start date, and follow the tendency reflected in the ocean re-analysis. IfM does not show the same rising trend in its analysis, although as for the two other sets of hindcasts the IfM hindcasts also follow the corresponding reference. However, some hindcasts missed completely the target, as in the case of ECMWF with the 1970 hindcast. In terms of correlation, all the systems have low skill, except for the global-average predictions from CERFACS.

Previous studies (e.g. *Collins et al.*, 2006) suggest that an accurate initialization of the Atlantic meridional overturning circulation (AMOC) could allow skilful predictions of the Atlantic multidecadal variability (AMDV) a few years in advance. However, past AMOC fluctuations have been poorly observed and a large uncertainty in the ocean analyses exists. This uncertainty is observed even when ocean re-analyses carried out using the same ocean model are considered. The uncertainty in the ocean reference implies that an assessment of the forecast quality of the AMOC predictions would necessarily give highly uncertain estimates. However, in this document we attempt to assess the ability of some of the experiments to simulate realistic AMOC anomaly variations.

Figure 37 shows the mean AMOC intensity as a function of latitude and depth for the ECMWF ORA-S3 re-analysis (*Balmaseda et al.*, 2007). There are maxima at around 1000 metres around 15°S and 35°N. As the ocean data were interpolated into a common 1° regular grid, the estimates obtained using the interpolated data are shown in Figure 37b for comparison. The impact of the interpolation on the meridional circulation can be seen as an increase of the value at the maxima and a set of discontinuities. Further comparisons allowed the assessment of the impact on the interannual variability of the AMOC and suggested that estimates from the interpolated data compared well with the original data provided that the values were averaged over a range of latitudes. Figure 38 shows the AMOC estimates obtained when the intensity was averaged over 10° of latitude around the maximum found in the corresponding ocean re-analysis. A striking result is the disagreement between the ocean re-analysis estimates in terms of mean state and variability. Two additional re-analyses (not shown in this figure) have been diagnosed and do not suggest any further agreement with those in Figure 38. As explained above, the large differences in the different references available add to the uncertainty of the predictions. As for the hindcasts, every system has a certain degree of drift in the AMOC intensity, especially the ECMWF system (Figure 38c). The anomalies (Figure 38b) depict in an even clearer way the differences between the ocean re-analyses, while for the hindcasts it is difficult to extract any conclusion about the predictability of the AMOC due to the large interannual variability and the limited size of the sample.

In the ECMWF case, the hindcasts experience a transition after the first couple of years of the simulations towards a shallower (by ~500 m) AMOC cell, weakening the northward branch and strengthening the southward one. As a consequence of the shallowing of the meridional overturning cell and the reduction of its vertical gradient, the AMOC intensity decreases to around 10 Sv. Model results (e.g. *Drijfhout et al.*, 2008) suggest that whereas interannual variability in the ocean overturning is largely driven by surface winds, variability on decadal and longer timescales (and probably the drift too) is primarily driven by buoyancy fluxes. Contributions to the buoyancy fluxes comprise fresh water forcing by precipitation, evaporation, runoff and sea ice melting or formation, and thermal forcing by turbulent fluxes (sensible and latent heat), radiative fluxes and the latent heat of fusion associated with formation or melting of sea ice. It is likely that most of these processes, in particular the missing ones such as those related to sea ice, are responsible for the drift of the AMOC in this model.

An index of the Atlantic multidecadal oscillation (AMO), calculated as the average North Atlantic SSTs north of 10ºN is shown in Figure 39. As suggested by *Dijkstra et al.* (2006), some agreement can be found between the AMOC and AMO indices from the reference datasets, especially if the upward trend in the AMO is discounted. This supports the use of the AMO as a proxy indicator for the intensity of the AMOC, although the impact of global warming should be somehow separated from the interannual and decadal variability. Predictions of the AMO index are skilful in all systems, especially because they reproduce the upward trend observed in the analysis since the mid 1970s. The ensemble-mean correlation of four-year average predictions is above 0.6 for all forecast periods. A more pessimistic picture has been found with an alternative AMO index estimated as SST averaged over a northern (40-60°N, 60-10°W) and a southern (40-60°S, 50°W-0°) Atlantic box (*Latif et al.*, 2006). In this case, the strong SST drift over the southern Atlantic Ocean (Figure 28) might adversely affect the simulations and limit the reproducibility of that specific AMO index.

4.3 Perturbed-parameter decadal predictions

The perturbed parameter (DePreSys_PP) seasonal-decadal hindcasts consisted of nine member ensembles of alternative variants of HadCM3, as described in Section 2. Results from these are described in Sections 3.1, 3.2 and 4.1. Here we briefly illustrate the impact on skill of initialising these hindcasts with observations. This is done by comparing the initialised hindcasts against a parallel set of hindcasts employing identical time-varying specifications of external forcing agents (trace gases, sulphate and volcanic aerosol, and solar irradiance). These "No_Assim" hindcasts used the same set of start dates as those used for the initialised hindcasts, but initial states were taken from transient climate change projections started from pre-industrial conditions (*Collins et al.*, 2009), rather than from recent observations. Each of the nine model variants in the No_Assim ensemble used the same set of parameter perturbations as one of the DePreSys_PP variants, and took initial conditions from a transient climate change simulation also using the same set of parameter perturbations.

Figure 40 shows an example of the DePresys_PP Assim and No_Assim results, plotting a time series of global pattern correlations for nine year average hindcasts of surface temperature throughout the Stream 2 period. While individual members of the DePreSys_PP Assim ensemble sometimes give better results than the ensemble-mean (not shown), the average skill of individual members is consistently smaller (compare dashed and solid red curves). This confirms that the beneficial impact of averaging decadal projections from different models in a multi-model ensemble, demonstrated in Section 4.2, also applies when combining different variants of a perturbed parameter ensemble. Note,

however, that the climate anomalies projected by different members of a multi-model or perturbed parameter ensemble will diverge due both to the effects of different model physics and also to the effects of different realisations of internal climate variability, so it remains for future work to isolate the specific impact of differences in model formulation. Figure 40 also shows that the skill in DePreSys_PP Assim increases for more recent hindcasts. This feature is also seen in the No_Assim hindcasts (blue curve in Figure 40), showing that this arises mainly from the strengthening influence of external forcing, particularly that due to man-made greenhouse gases. However, the average correlation skill is slightly smaller in the No_Assim hindcasts (0.25 versus 0.30), indicating that initialisation provides a modest increase in average skill.

Figure 41 shows the correlation between time series of nine-year mean hindcast and observed precipitation anomalies over land, during the Stream 2 period. Not surprisingly, the skill in DePreSys_PP is generally lower than for surface temperature; however many regions do show a modest positive correlation with observations. The pattern of spatial variations in skill is rather similar in the uninitialised No_Assim hindcasts. Nevertheless, the average skill is lower in the No_Assim hindcasts (spatial average of 0.04 versus 0.12 in the DePreSys_PP Assim hindcasts), indicating that initialisation of the DePreSys_PP hindcasts does have some positive impacts on precipitation projections, notably over the south-western North America, the north eastern region of South America and Indonesia. Analysis of the variation of skill with forecast lead time (not shown) shows that the (spatially averaged) positive skill found in the nine-year mean does not arise exclusively from short forecast lead times, however more research is needed to understand reasons for the spatial and temporal variations in predictability.

5 Summary

The ENSEMBLES system for climate prediction on seasonal, interannual and decadal time scales has been developed and tested. It includes three different methods to tackle the problem of model uncertainty: the multi-model initial condition ensemble technique (MME), the perturbation of physical parameters (PPE) and stochastic parameterizations of sub-grid processes (SPE). In order to compare the methodologically different approaches and to assess their relative merits in an s2d framework, a set of common simulations was defined, performed and analyzed.

In the analyses presented in the previous sections, MME and SPE perform similarly well for the seasonal predictions, while PPE has a similar performance to MME in longer time scales. Benefits of the stochastic physics approach on seasonal time scales are found in reducing the ensemble-mean RMSE and increasing the ensemble spread when compared to the control version of the corresponding coupled model. The major benefit of the MME lies in the closer match between spread of the ensemble and RMSE, which leads to an increase of the reliability. However, the MME also shows a larger discriminatory ability between the occurrence of the events and non-events. Most of the MME superiority is linked to its larger ensemble size. The annual integrations started in November have shown a relatively high level of skill in all systems over many areas. The skill is above the level obtained with climatological and persistence forecasts, especially for temperature.

The preliminary ENSEMBLES decadal hindcasts have shown that, apart from the existence of substantial skill in interannual predictions, the PPE and MME systems offer a similar level and spatial distribution of skill, with a spread surprisingly good for the level of experience available for this type of forecasts. In all cases, the decadal predictions suggest that the large uncertainty in the observed state

especially for the XXth Century is a major problem for the assessment of decadal prediction skill. More detailed assessments are being carried out at present and will be published soon.

As most of the data used to obtain these results are publicly available, deeper analyses and more detailed descriptions will follow soon. The authors expect this document to serve as documentation of the experiments and to guide further analysis of the datasets.

An important aspect of the experiments described here is the link between climate simulations carried out at different time scales. Within ENSEMBLES, there is the opportunity to assess the forecasting problem across different time scales: the s2d and the climate communities try to tackle the prediction of monthly, seasonal, interannual, decadal and longer-range climate fluctuations. At the moment, this is achieved by using separate forecasting systems for the different time and spatial scales (global and regional) with not much cross-linking and interaction between them. Recently, the need to explore the utility of "seamless" prediction methods across weather and climate time scales has been proposed because shorter timescales are known to be important in their feedback on the longer time scales. For example, accurate representations of El Niño (La Niña) events in climate prediction models are a key aspect of predictions of climate variability and change at the decadal and longer time scales. Motivated by these seamless prediction ideas, a new method to assess the trustworthiness probabilistic multi-model projections of climate change based on analyses of reliability diagrams of corresponding multi-model seasonal hindcast ensembles was proposed by *Palmer et al.* (2008). The suggested method provides constraints, necessary but not sufficient, for improving the reliability of probabilistic regional anthropogenic climate change projections of variables affected by atmospheric circulation. This is just but one example of the benefits to be achieved by a closer collaboration between the forecasting and climate-change assessment communities, of which the work described here is a preliminary attempt.

6 References

Alessandri, A., A. Borrelli, S. Masina, P.D. Pietro, A. Carril, A. Cherchi, S. Gualdi and A. Navarra (2010): The INGV-CMCC seasonal prediction system: Improved ocean initial conditions. *Mon Wea. Rev*., in press.

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

Bechtold, P., M. Koehler, T. Jung, F.J. Doblas-Reyes, M. Leutbecher, M.J. Rodwell, F. Vitart, and G. Balsamo, 2008: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Quart. J. Royal Meteor. Soc.*, **134**, 1337-1351.

Collins, W.J., N. Bellouin, M. Doutriaux-Boucher, N. Gedney, T. Hinton, C.D. Jones, S. Liddicoat, G. Martin, F. O'Connor, J. Rae, C. Senior, I. Totterdell and S. Woodward, 2008: *Met Office Hadley Centre Technical Note*, **74**.
available from: http://www.metoffice.gov.uk/publications/HCTN/index.html

Collins, M., B.B.B. Booth, B. Bhaskaran, G.R. Harris, J.M. Murphy, D.M.H. Sexton and M.J. Webb, 2009: Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles. *Climate Dyn*., submitted.

Daget, N., A.T. Weaver and M.A. Balmaseda, 2009: An ensemble three- dimensional variational data assimilation system for the global ocean: sensitivity to the observation- and background-error variance formulation. *Quart. J. Roy. Meteor. Soc.*, **135**, 1071-1094.

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

Giorgi, F. and R. Francisco, 2000: Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM, *Clim. Dyn.*, **16**, 169-182.

Jungclaus, J.H., M. Botzet, H. Haak, J. Marotzke, U. Mikolajewicz, E. Roeckner, N. Keenlyside, M. Latif and J.-J. Luo, 2006: Ocean circulation and tropical variability in the coupled model ECHAM5/MPI-OM. *J. Climate*, **19**, 3952-3972.

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*, **57A**, 340-356.

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.

Large, W. and S. Yeager, 2004: Diurnal to decadal global forcing for ocean and sea-ice models: the datasets and flux climatologies. *NCAR Technical Note*, NCAR/TN-460+STR, CGD division of the National Center for Atmospheric Research.

Meehl, G. A., L. Goddard, J. Murphy, R. J. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M. A. Giorgetta, A. Greene, E. Hawkins, G. Hegerl, D. Karoly, N. Keenlyside, M. Kimoto, B. Kirtman, A. Navarra, R. Pulwarty, D. Smith, D. Stammer and T. Stockdale, 2009: Decadal prediction: Can it be skillful? *Bull. Am. Meteor. Soc.,* in press.

Müller, W.A., C. Appenzeller, F. Doblas-Reyes and M. A. Liniger, 2005: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Climate*, **18**, 1513-1523.

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.

Palmer, T.N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phy.*, **63**, 71-116.

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

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. Meteor. Soc.*, **85**, 853-872.

Palmer, T.N., F.J. Doblas-Reyes, A. Weisheimer and M. Rodwell, 2008: Reliability of climate change projections of precipitation: Towards "seamless" climate predictions. *Bull. Amer. Meteor. Soc.*, **89**, 459-470.

Palmer, T.N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G.J. Shutts, M. Steinheimer and A Weisheimer, 2009: Stochastic parametrization and model uncertainty. *ECMWF Tech. Memo*. 598, 42 pp.

Pohlmann, H., J. H. Jungclaus, A. Köhl, D. Stammer and J. Marotzke, 2009: Initializing Decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate*, **22**, 3926-3938.

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

Shutts, G, 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **131**, 3079-3102.

Smith, D., 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. Colman, C. Folland, G. 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.

Stainforth, D.A., T. Aina, C. Christensen, M. Collins, N. Faull, D.J. Frame, J.A. Kettleborough, S. Knight, A. Martin, J.M. Murphy, C. Piani, D. Sexton, L.A. Smith, R. A. Spicer, A.J. Thorpe and M.R. Allen, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433**, 403-406.

Stockdale, T.N., D.L.T. Anderson, M.A. Balmaseda, F.J. Doblas-Reyes, L. Ferranti, K. Mogensen, T.N. Palmer, F. Molteni and F. Vitart, 2010: ECMWF seasonal forecast System 3 and its prediction of sea surface temperature. *Climate Dyn.*, submitted.

Uppala, S.M., et al., 2005: The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961-3012.

Weigel, A.P., M.A. Liniger and C. Appenzeller, 2007. The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **135**, 118-124.

Weigel, A.P., M.A. Liniger and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.,* **134**, 241-260.

Weisheimer, A., F.J. Doblas-Reyes, T.N. Palmer, A. Alessandri, A. Arribas, M. Deque, N. Keenlyside, M. MacVean, A. Navarra and P. Rogel, 2009: ENSEMBLES - a new multi-model ensemble for seasonal-to-annual predictions: Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.*, **36**, L21711, doi:10.1029/2009GL040896.

partner	atmospheric model; resolution	ocean model; resolution	initialization		external forcing	additional components, comments	references
			atmosphere and land	ocean			
ECMWF	IFS CY31R1; T159/L62	HOPE; 0.3°-1.4°/L29	ERA-40/oper. analysis, atmospheric singular vectors	wind stress perturbations to generate ensemble of ocean reanalyses; SST perturbations at initial time	observed global well-mixed GHGs and sulphate aerosol and A1B from 2000, observed solar activity, no volcanic aerosol nor ozone	operational seasonal forecasting system S3	Stockdale et al. (2010); Balmaseda et al. (2008)
	IFS CY33R1; T159/L62	HOPE; 0.3°-1.4°/L29	ERA-40/oper. analysis, atmospheric singular vectors	wind stress perturbations to generate ensemble of ocean reanalyses; SST perturbations at initial time	"	used for the decadal hindcasts only	Bechtold et al. (2008)
	IFS CY35R2; T159/L62	HOPE; 0.3°-1.4°/L29	ERA-40/oper. analysis, atmospheric singular vectors	wind stress perturbations to generate ensemble of ocean reanalyses; SST perturbations at initial time	"	used with the stochastic physics approach	Palmer et al. (2009)
UKMO	HadGEM2-A; N96/L38	HadGEM2-O; 0.33°-1°/L20	ERA-40/oper. analysis, anomaly assimilation for soil moisture	wind stress perturbations to generate ensemble of ocean reanalyses; SST perturbations at initial time	observed global well-mixed GHGs, ozone and sulphate aerosol emissions and A1B from 2000, persisted solar activity and volcanic aerosol	fully interactive sea ice module	Collins et al. (2008)
	HadAM3; 3.75x2.5°	HadOM	anomaly assimilation of ERA-40/oper. analysis	anomaly assimilation of an ocean reanalysis	"	perturbed-parameter ensemble	Smith et al. (2007)
MF	ARPEGE4.6; T63	OPA8.2; 2°/L31	ERA-40/oper. analysis	wind stress, SST and water flux perturbations to generate ensemble of ocean reanalyses	observed global well-mixed GHGs and sulphate aerosol and A1B from 2000, no solar activity nor volcanic aerosol, dynamical ozone	GELATO sea ice model	Daget et al. (2009); Salas y Melia (2002)
IFM-GEOMAR	ECHAM5; T63/L31	MPI-OM1; 1.5°/L40	initial condition permutations of three coupled climate simulations from 1950 to 2005 with SSTs restored to observations		observed global well-mixed GHGs, ozone and sulphate aerosol emissions and A1B from 2000, persisted solar activity and volcanic aerosol	-	Keenlyside et al. (2005); Jungclaus et al. (2006)
CMCC-INGV	ECHAM5; T63/L19	OPA8.2 ; 2°/L31	AMIP-type simulations with forced SSTs	wind stress perturbations to generate ensemble of ocean reanalyses, SST perturbations at initial time	observed global well-mixed GHGs and sulphate aerosol and A1B from 2000, no volcanic aerosol nor ozone	dynamical snow-sea ice model and land-surface model	Weisheimer et al. (2009); Alessandri et al. (2010)

Table 1: Overview of the forecast systems contributing to the ENSEMBLES s2d hindcasts

Near Surface Temperature

	multi-model				perturbed parameters				stochastic physics			
	JJA		DJF		JJA		DJF		JJA		DJF	
	cold	warm	cold	warm	cold	warm	cold	warm	cold	warm	cold	warm
Australia	**11.5**	**13.9**	3.2	6.7	-0.3	**11.0**	0.5	5.2	7.0	**17.3**	**11.8**	8.0
Amazon Basin	0.2	17.1	4.5	**23.4**	-13.7	2.8	-6.3	11.2	3.9	14.7	2.6	16.9
Southern South America	**9.2**	**9.0**	1.8	**9.9**	-2.8	7.2	2.9	**14.7**	**16.9**	8.8	4.5	**9.3**
Central America	5.9	**11.6**	-2.6	4.5	2.4	5.5	-3.9	3.3	1.2	-0.3	0.2	-3.7
Western North America	10.2	**12.2**	6.3	**12.5**	6.7	-1.2	3.3	8.9	2.8	8.0	6.4	4.7
Central North America	-0.2	**-7.3**	-3.3	10.4	-8.5	**-12.7**	7.2	13.8	**-21.4**	**-20.3**	-2.6	8.8
Eastern North America	4.1	-7.0	-4.5	10.1	-9.9	-14.7	**32.2**	8.2	-13.4	**-10.9**	-11.3	4.0
Alaska	-0.8	-0.9	-0.6	0.6	-0.4	-2.9	6.5	4.8	0.5	**12.2**	**-20.3**	-1.0
Greenland	**15.1**	**8.7**	**13.2**	**12.3**	12.7	-1.5	**17.3**	**15.1**	3.2	3.1	12.3	16.3
Mediterranean	**18.0**	**12.8**	5.8	4.3	**18.3**	**15.5**	**-17.5**	**-14.5**	**22.7**	**12.2**	6.2	2.6
Northern Europe	-3.3	0.2	4.9	0.5	1.1	4.6	-0.6	-4.0	4.6	6.3	1.5	5.2
Western Africa	7.9	7.0	7.3	**20.5**	-14.8	3.6	3.6	10.6	7.8	-2.0	10.9	**15.8**
Eastern Africa	**9.4**	**7.3**	-7.7	0.9	**-19.5**	-7.1	-3.9	-5.4	**-9.7**	-3.1	-3.7	8.2
Southern Africa	**14.0**	4.7	1.7	**10.6**	-3.2	10.2	-1.7	2.7	0.0	7.7	6.0	13.6
Sahel	**12.9**	7.2	**11.5**	**15.4**	**9.9**	**13.1**	6.6	**15.7**	**16.3**	**10.1**	**13.9**	**14.7**
South East Asia	**8.6**	**12.4**	**11.6**	13.4	-9.3	4.2	13.9	6.1	-0.6	9.6	3.8	1.6
East Asia	10.6	**10.2**	0.3	5.8	8.3	**10.5**	-4.2	10.1	6.4	**14.1**	3.1	-0.4
South Asia	**8.7**	**13.3**	**14.4**	**10.6**	4.3	9.2	0.1	9.3	12.9	**15.7**	**13.8**	**18.1**
Central Asia	**14.3**	**8.2**	-2.4	7.1	**14.1**	**11.8**	-2.0	**19.1**	**21.1**	**10.1**	-8.5	6.5
Tibet	**16.9**	**16.1**	-0.1	4.1	7.8	7.2	-10.4	3.8	8.3	**15.7**	5.6	7.6
North Asia	**7.3**	3.9	4.2	**8.5**	6.2	**8.4**	-1.5	**12.6**	4.2	1.6	-1.9	1.2

Table 2a: Seasonal skill over 21 standard land areas estimated from the multi-model, perturbed parameter and stochastic physics ensembles for near-surface temperature tercile events. The areas as labelled in the first column follow the definitions given in Giorgi and Francisco (2000). The scores shown are the Brier skill score assuming an infinite ensemble size multiplied by 100. The lead time is 2-4 months based on May and November start dates over the hindcast period 1991-2005. Green cells indicate events with a positive BSS, while red cells indicate events with a negative BSS. Those BSS that have been printed in bold font and are underlined show BSS that are statistical significantly different from zero.

Precipitation

	multi-model				perturbed parameters				stochastic physics			
	JJA		DJF		JJA		DJF		JJA		DJF	
	dry	wet	dry	wet	dry	wet	dry	wet	dry	wet	dry	wet
Australia	7.6	7.0	0.9	3.0	5.1	8.0	**12.4**	5.2	2.5	5.0	**10.5**	6.6
Amazon Basin	**10.3**	**10.3**	**16.0**	14.3	**8.8**	5.4	3.4	0.5	**12.2**	**11.4**	**16.1**	16.8
Southern South America	6.2	7.1	4.6	6.0	1.3	1.6	-4.5	-1.7	3.3	**9.0**	-4.7	0.2
Central America	**9.2**	**7.8**	**23.4**	**18.9**	**12.9**	5.2	**23.3**	**25.9**	**10.6**	7.7	**24.9**	**23.7**
Western North America	2.4	**8.1**	**7.2**	**7.8**	4.5	**7.5**	4.5	4.9	**9.1**	**8.4**	5.7	5.3
Central North America	0.6	2.2	7.7	**10.4**	-3.5	-5.7	**10.0**	**10.4**	1.7	3.0	2.1	5.5
Eastern North America	-1.9	-1.1	**8.3**	**10.6**	**-9.6**	**-11.1**	9.7	**13.2**	**-15.0**	-6.8	7.5	2.1
Alaska	-1.3	0.0	4.0	-2.2	-2.3	-1.0	**11.3**	3.7	-4.3	-0.7	0.2	-2.5
Greenland	2.6	**2.8**	**-3.7**	**-3.0**	1.4	0.2	**7.5**	-1.7	**-6.8**	-2.6	-2.2	-2.1
Mediterranean	-1.2	1.2	-1.0	-1.3	-6.1	-4.4	-3.0	0.1	-0.9	0.1	**11.5**	10.7
Northern Europe	2.3	2.1	-3.1	**-4.7**	**7.7**	**11.5**	-1.8	-1.6	**8.2**	6.0	6.6	1.6
Western Africa	-1.5	-0.1	-0.5	1.3	**-10.9**	-3.8	4.8	-1.6	-4.8	2.4	**-13.7**	-0.1
Eastern Africa	-2.8	1.8	3.9	2.5	-7.0	**-7.6**	**14.4**	**13.2**	-1.5	3.4	0.9	5.7
Southern Africa	3.5	1.0	5.7	**9.5**	**7.2**	4.7	**6.0**	**11.3**	7.8	9.2	7.7	8.9
Sahel	**-4.6**	**-3.6**	**-3.2**	-1.5	**-9.2**	**-6.7**	-2.7	-2.4	**-10.0**	-1.0	**-8.2**	-3.6
South East Asia	**14.3**	9.7	**8.8**	**8.3**	5.5	4.8	5.6	8.3	10.3	1.1	9.6	**12.5**
East Asia	0.5	-0.5	4.7	**4.6**	**5.6**	1.4	8.9	3.6	2.8	0.6	8.9	**15.7**
South Asia	0.2	0.9	**6.5**	**7.4**	0.6	-2.7	7.0	9.4	2.7	1.9	5.5	**10.2**
Central Asia	-0.8	0.2	7.4	5.7	0.8	-3.1	**10.3**	**8.4**	-1.5	0.2	2.9	1.6
Tibet	5.5	3.5	**6.5**	5.4	-1.4	-0.9	1.2	7.8	4.2	**6.4**	**10.7**	**10.0**
North Asia	**2.4**	**2.6**	3.1	0.6	3.3	2.9	2.1	-1.0	1.0	0.6	2.5	-1.9

Table 2b: As Table 2a, but for precipitation.

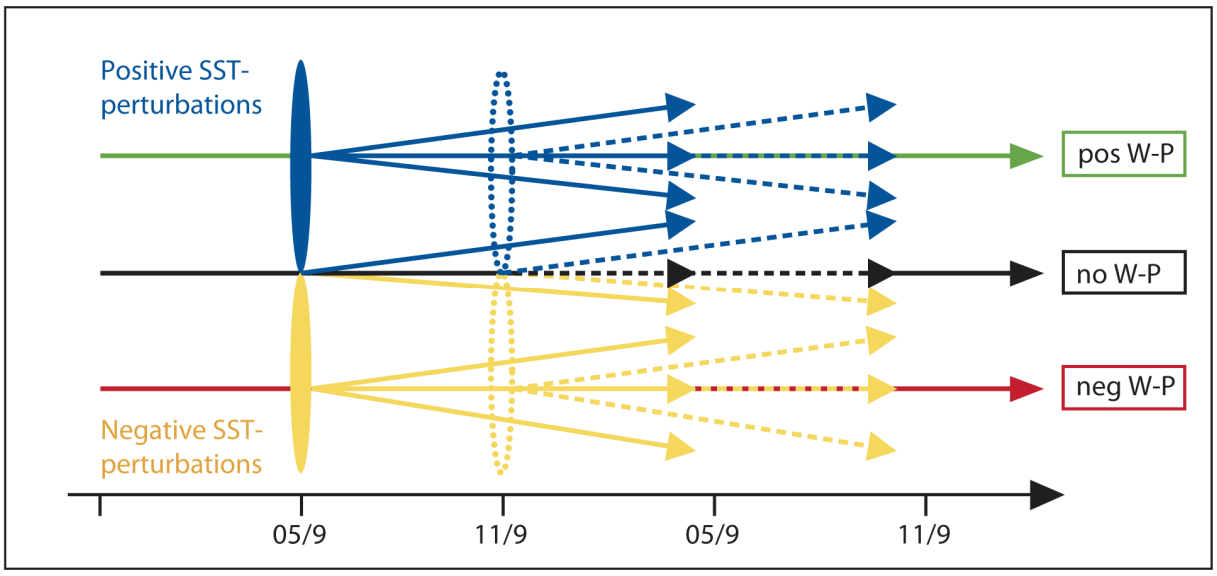

Figure 1: Schematic to illustrate the use of wind stress and SST perturbations to generate the initial condition ensemble for some ot the multi-model ENSEMBLES hindcasts. An ensemble of several ocean analyses covering the entire hindcast period (in the figure, one unperturbed and four wind-stress perturbed members) are shown by the dashed horizontal lines. The dark grey dotted line illustrates the unperturbed ocean analysis while the two red (green) analyses were generated by adding positive (negative) wind stress perturbations. At each start date of the hindcast, eight SST perturbations were added to the wind stress perturbed ocean analyses. These are indicated in yellow and act as pairs of positive and negative SST perturbations.

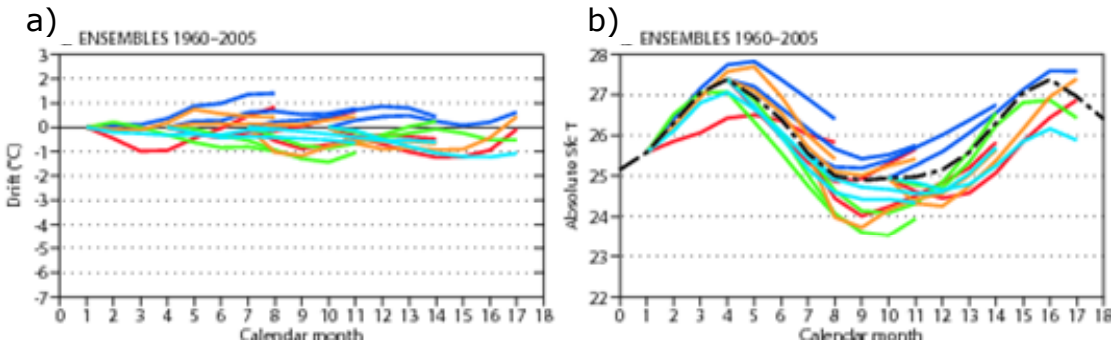

Figure 2: Systematic model error in Niño3 SST for the five individual seasonal forecast models. The plot on the left shows the drift from the verification over the seven-month forecast time, while the plot on the right shows the absolute SSTs for the individual models (colour) together with the climatological annual cycle (black). All four start dates are shown. The drift has been estimated from all available ensemble members for each start date separately over the hindcast period 1960-2005. Colour code: red MF, dark blue ECMWF, green UKMO, orange IFM-GEOMAR and light blue CMCC-INGV.

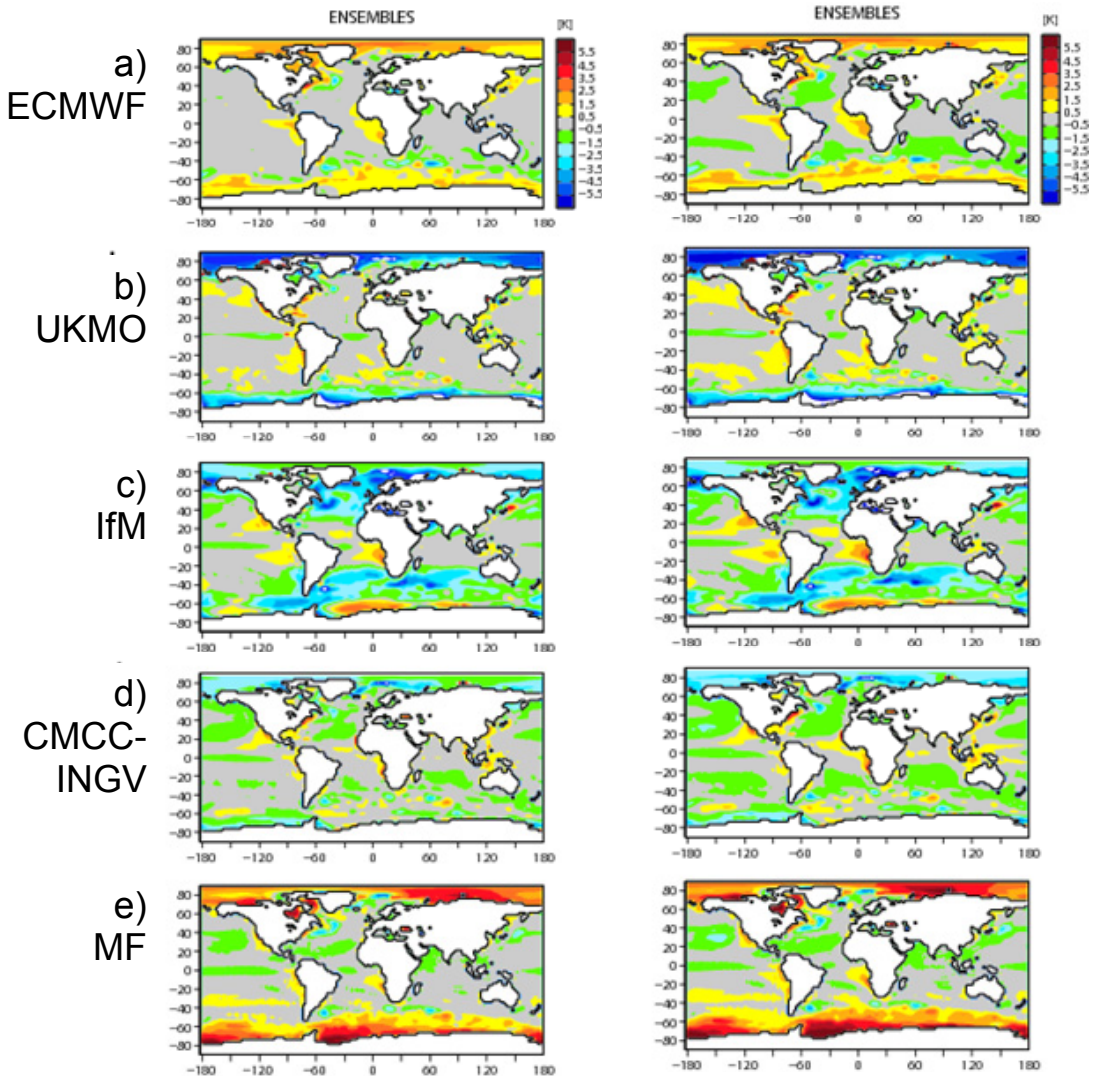

Figure 3: Mean SST bias in the five individual seasonal forecast models. Results for the average of the lead time 2-4 months are shown on the left, while the results for the lead time 4-6 months are shown on the right.

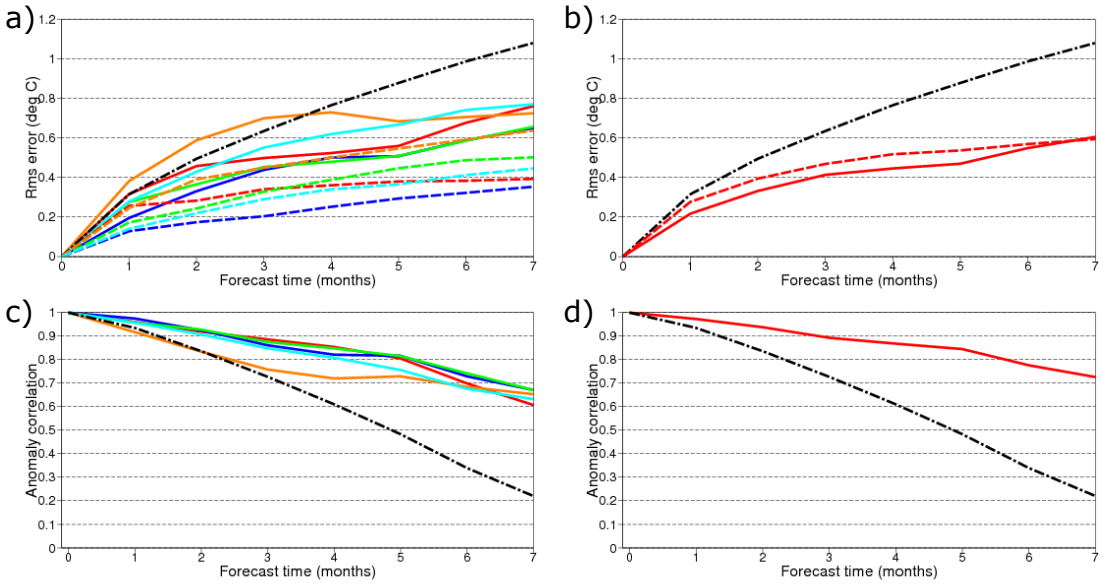

Figure 4: Seasonal skill of Niño3 SSTs from the individual models (left) and from the multi-model ensemble (right) up to 7 months ahead. The top plots show the evolution of RMSE (solid) and ensemble spread (dashed) as a function of forecast time together with the RMSE of a simple persistence forecast (black dash-dotted). The bottom plots show the anomaly correlation over forecast time (solid), again with the corresponding correlation from the persistence forecast (black dash-dotted). All scores have been estimated based on the four start dates over the 1960-2005 hindcast period. Colour code: red MF, dark blue ECMWF, green UKMO, orange IFM-GEOMAR and light blue CMCC-INGV.

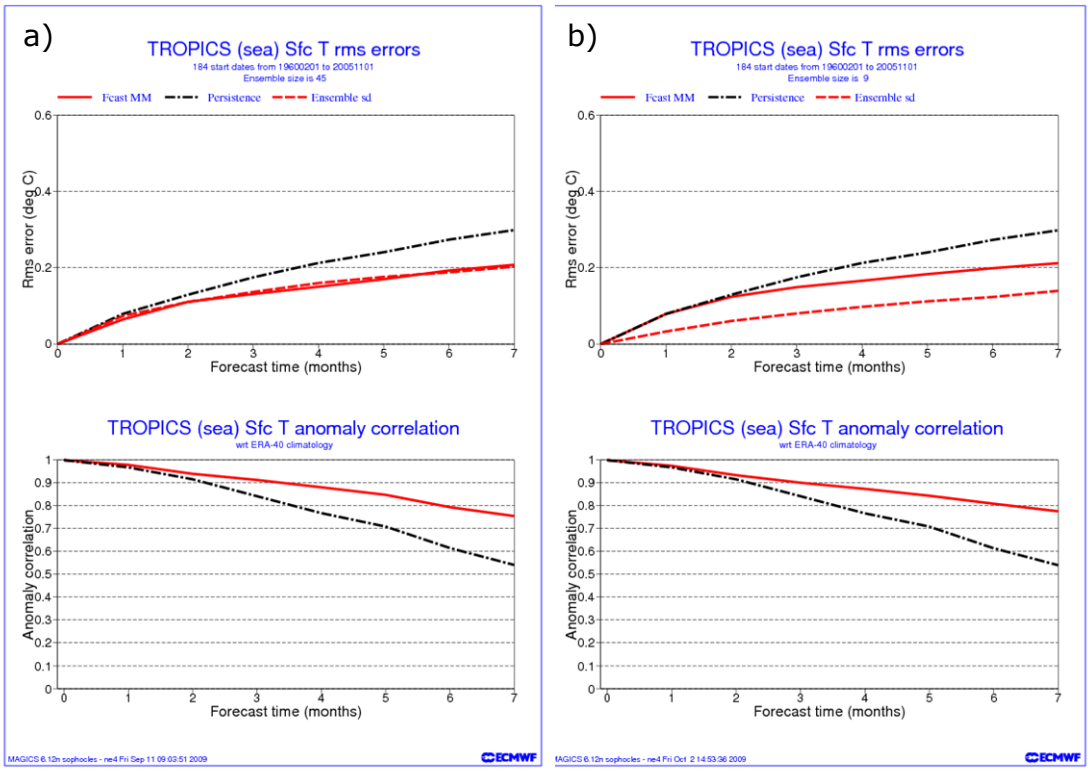

Figure 5: Seasonal skill of SSTs in the tropics from the multi-model ensemble (left) and the perturbed parameter ensemble (right) up to 7 months ahead. The top plots show the evolution of RMSE (solid red) and ensemble spread (dashed red) as a function of forecast time together with the RMSE of a simple persistence forecast (black). The bottom plots show the anomaly correlation over forecast time (red), again with the corresponding correlation from the persistence forecast (black). All scores have been estimated based on the four start dates over the 1960-2005 hindcast period.

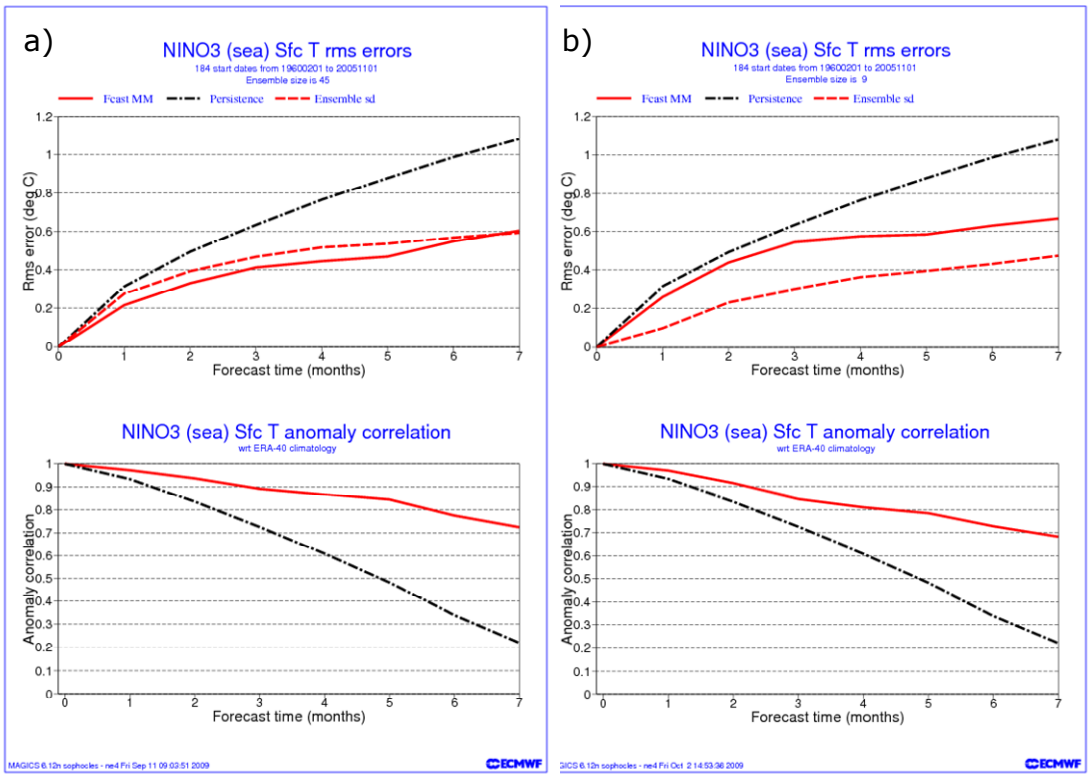

Figure 6: As in Fig. 5, but for Niño3 SSTs.

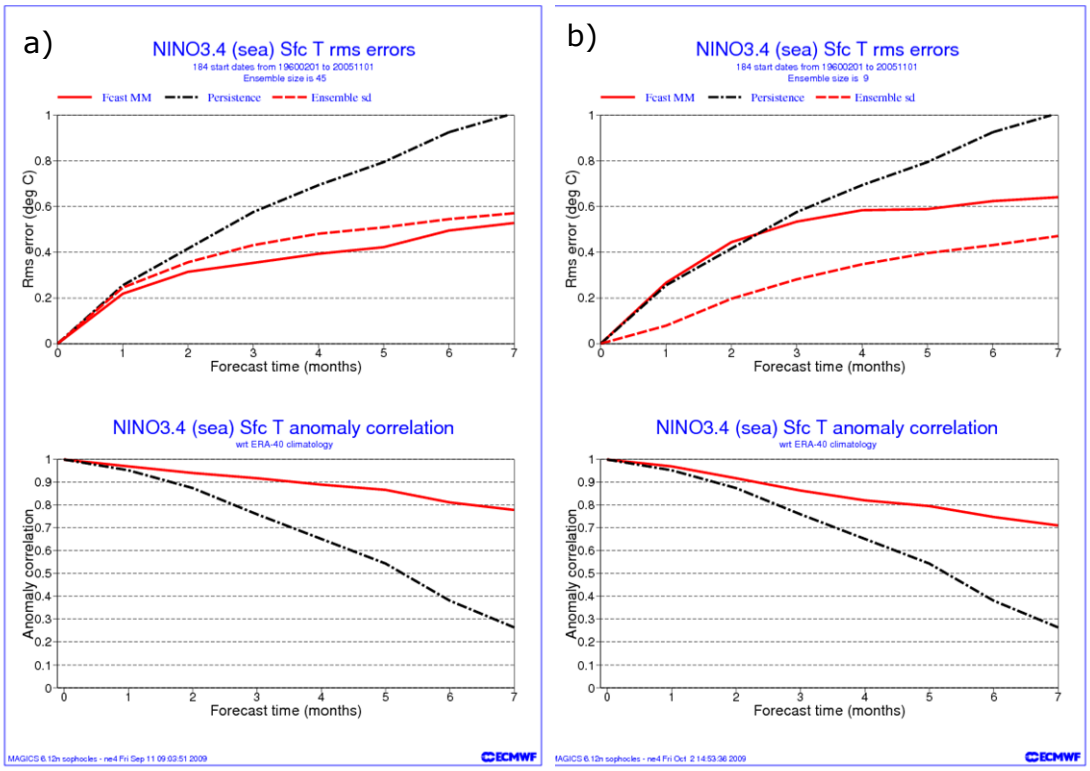

Figure 7: As Fig. 5, but for Niño3.4 SSTs.

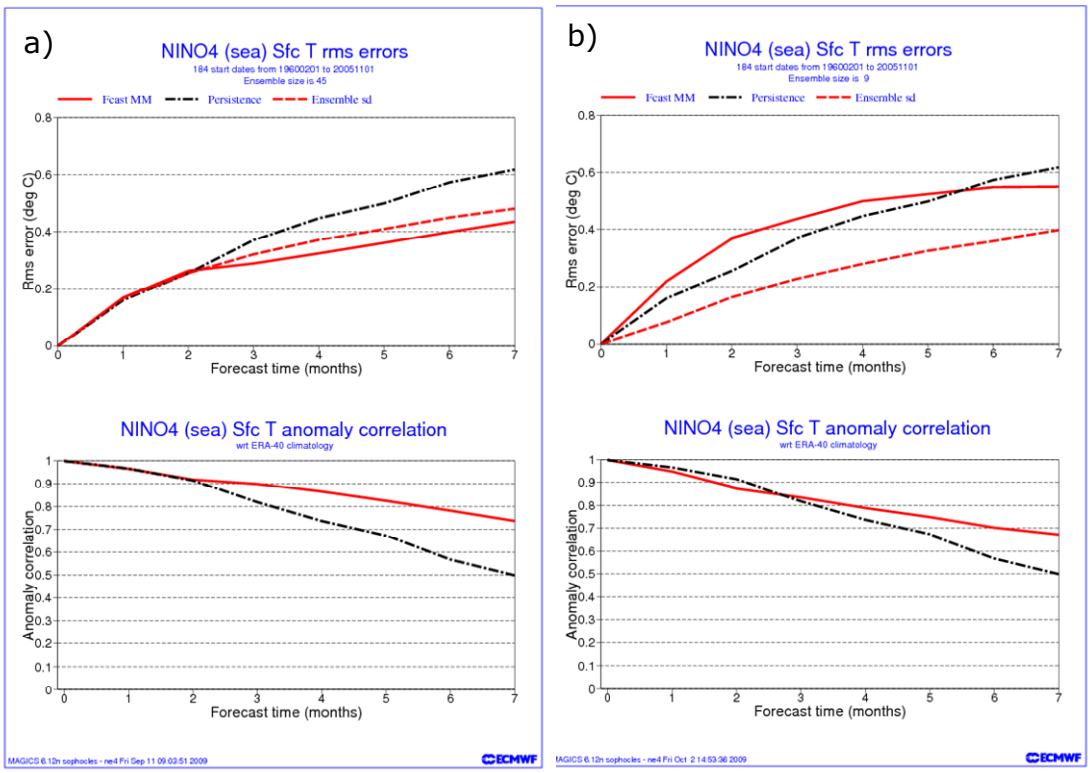

Figure 8: As Fig. 5 but for Niño4 SSTs.

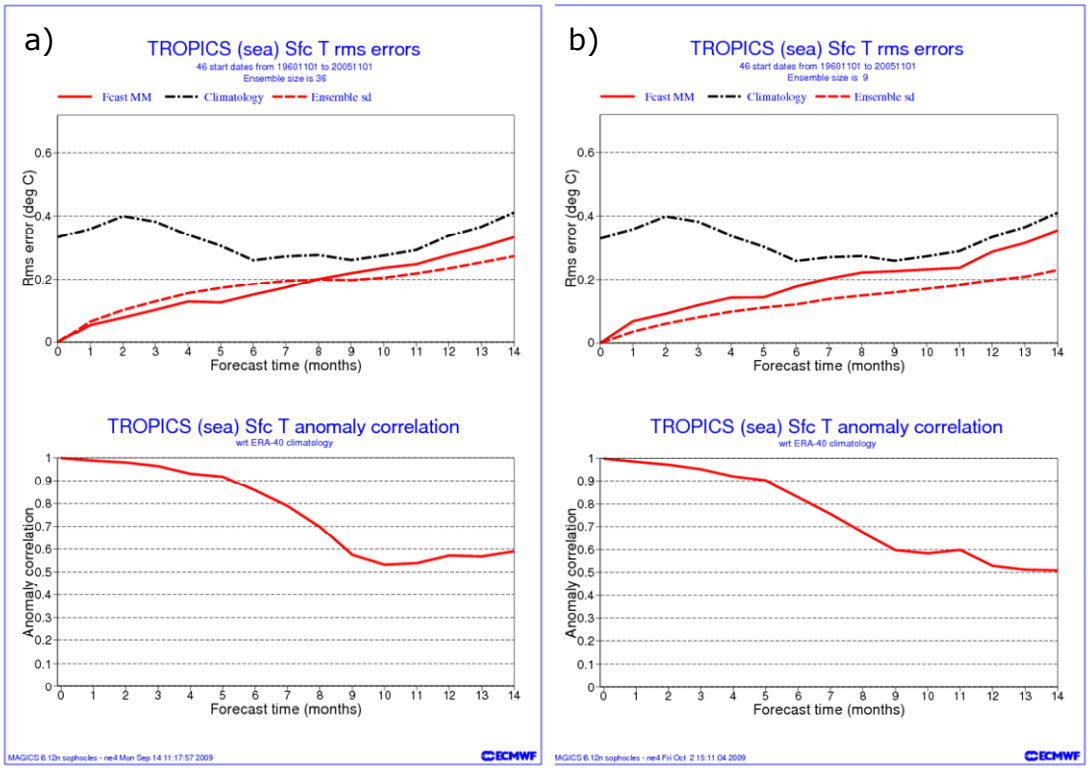

Figure 9: Annual hindcast skill of SSTs in the tropics from the multi-model ensemble (left) and the perturbed parameter ensemble (right) up to 14 months ahead. The top plots show the evolution of RMSE (solid red) and ensemble spread (dashed red) as a function of forecast time together with the RMSE of a simple climatology forecast (black). The bottom plots show the anomaly correlation over forecast time (red). All scores have been estimated based on the November start dates over the 1960-2005 hindcast period.

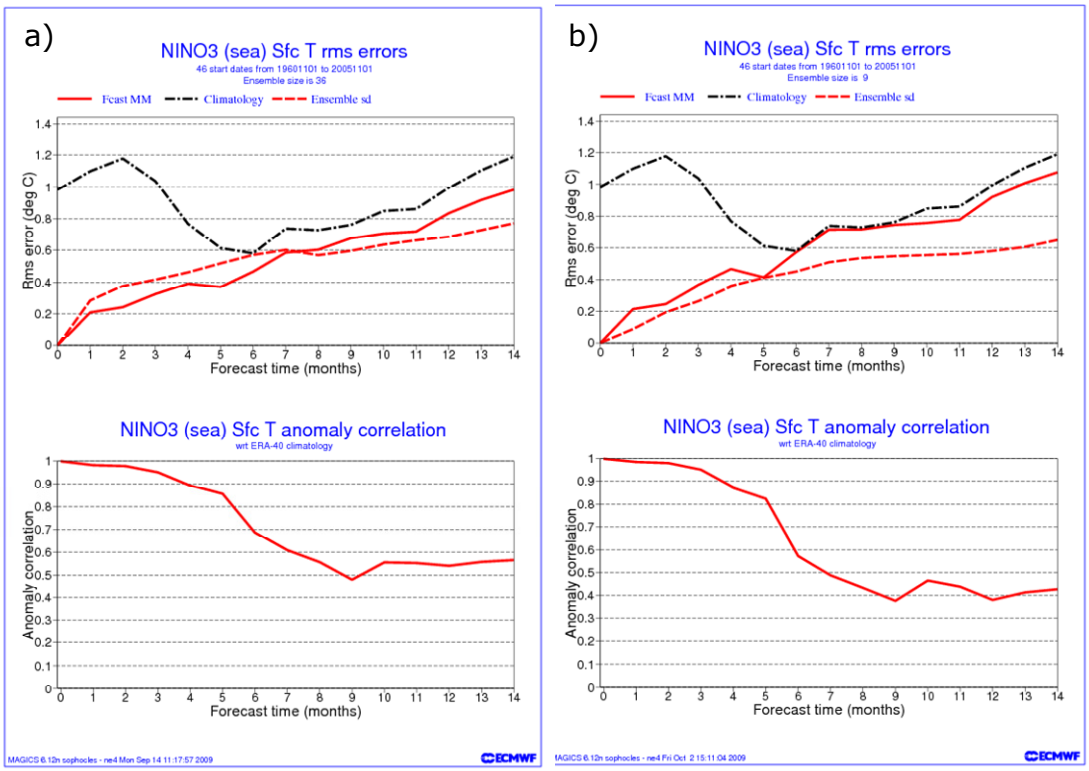

Figure 10: As in Fig. 9, but for Niño3 SSTs.

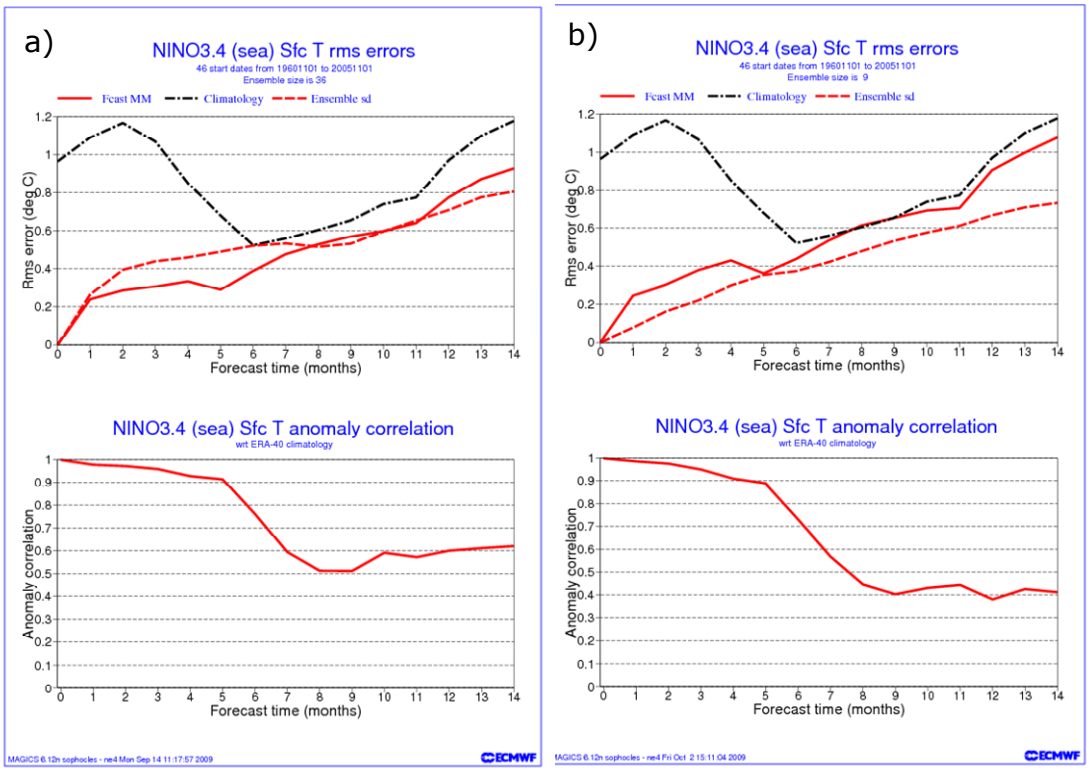

Figure 11: As in Fig. 9, but for Niño3.4 SSTs.

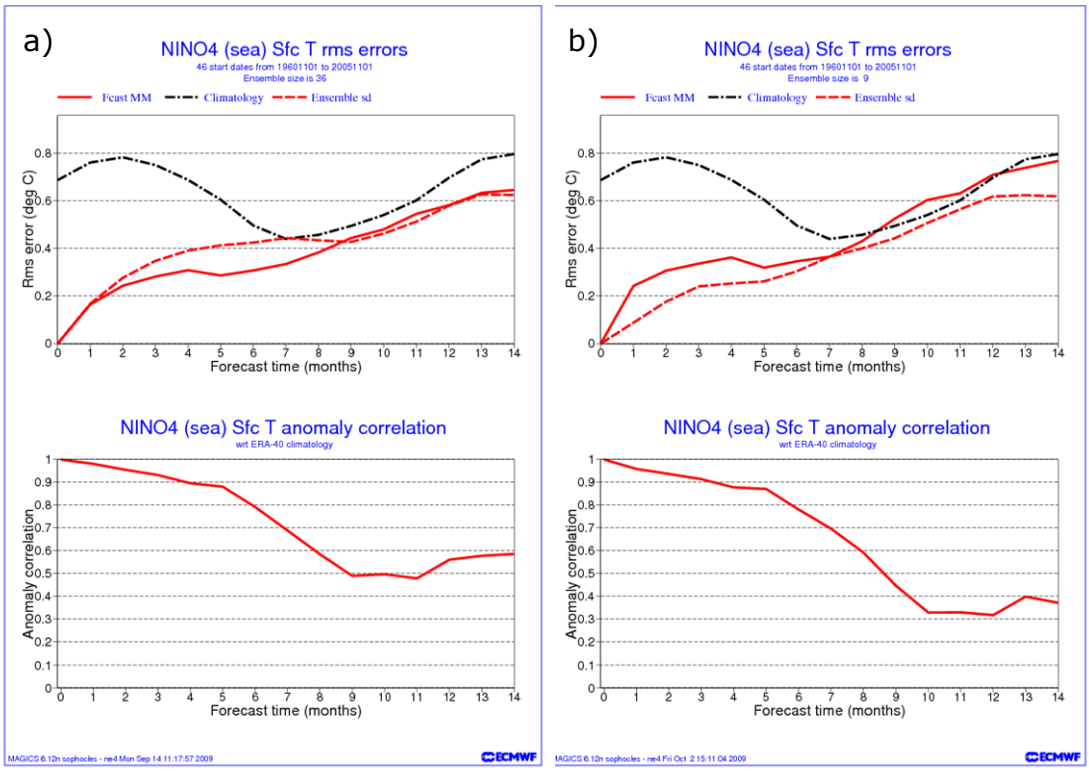

Figure 12: As in Fig. 9, but for Niño4 SSTs.

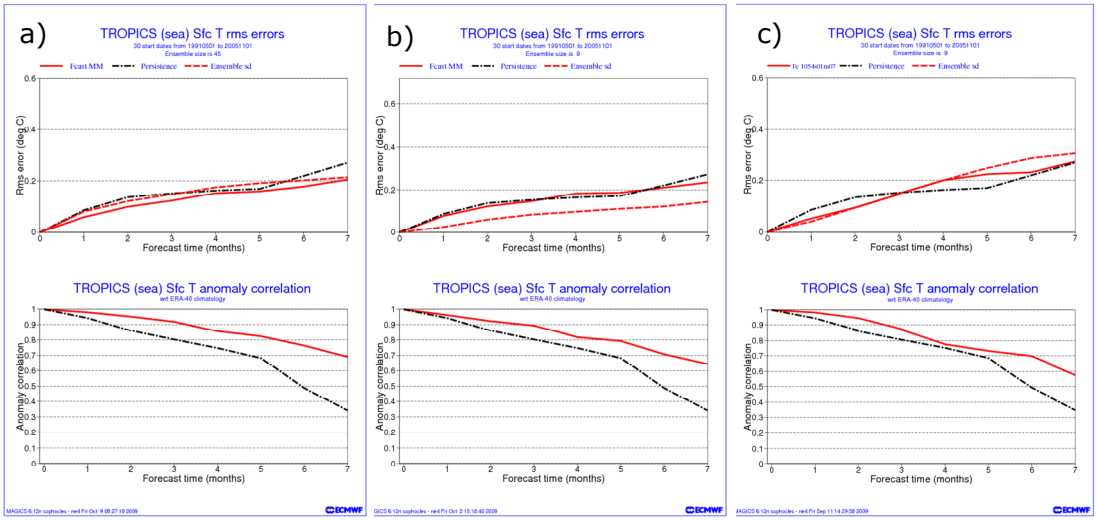

Figure 13: Seasonal skill of SSTs in the tropics from the multi-model ensemble (left), the perturbed parameter ensemble (middle) and the stochastic physics ensemble (right) up to seven months ahead. The top plots show the evolution of RMSE (solid red) and ensemble spread (dashed red) as a function of forecast time together with the RMSE of a simple persistence forecast (black). The bottom plots show the anomaly correlation over forecast time (red), again with the corresponding correlation from the persistence forecast (black). All scores have been estimated based on the May and November start dates over the 1991-2005 hindcast period.

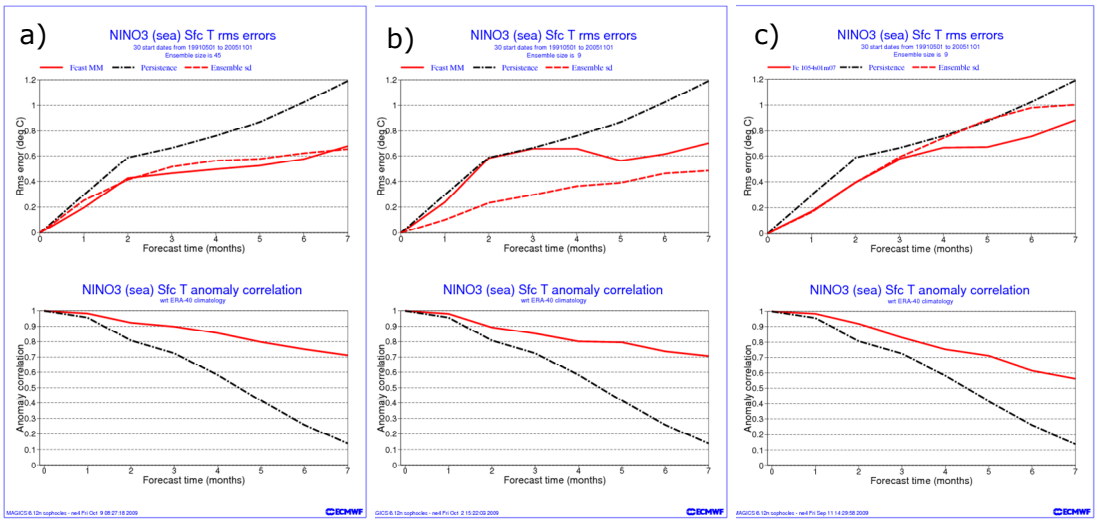

Figure 14: As in Fig. 13, but for Niño3 SSTs.

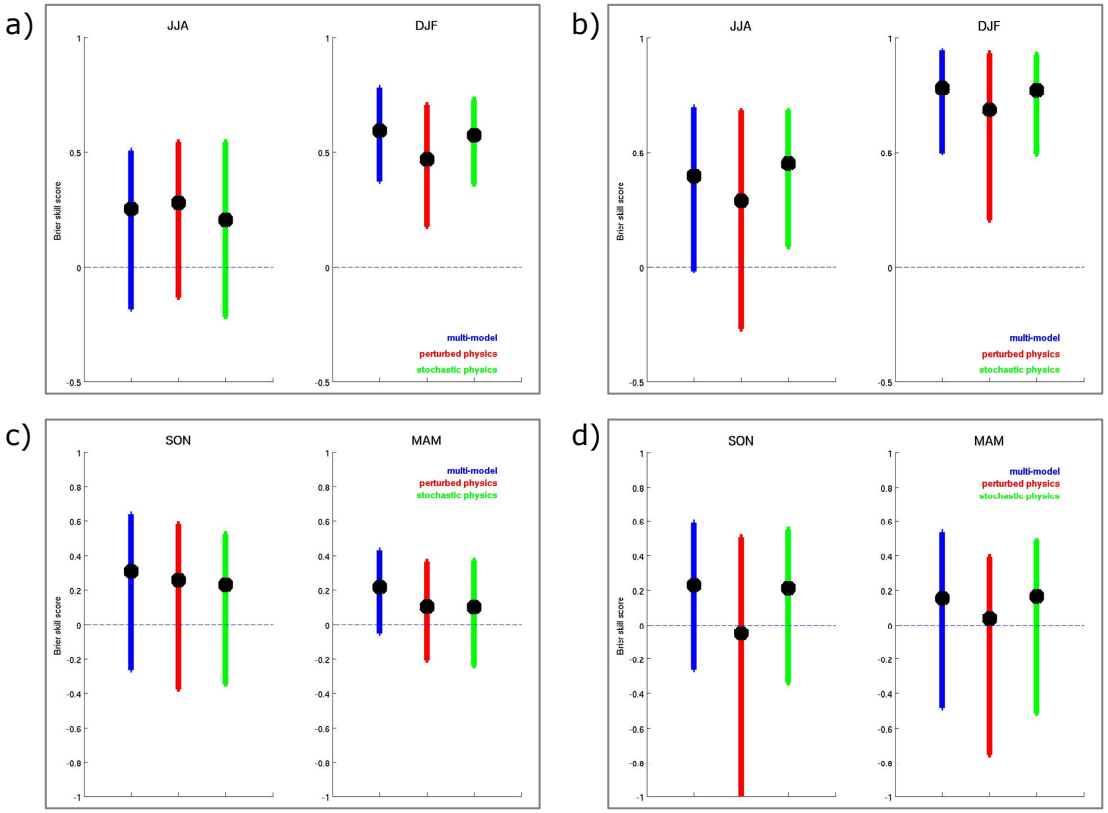

Figure 15: Brier skill score for SSTs over the Niño3 region for the multi-model ensemble (blue), the perturbed parameter ensemble (red) and the stochastic physics ensemble (green). The plots in the top row show the scores for forecast lead times of 2-4 months, initialised in May and November. The bottom row shows the scores for the 5-7 month lead time. The plots on the left are for lower tercile events while the plots on the right are for upper tercile events. The uncertainty ranges were estimated using a bootstrap method.

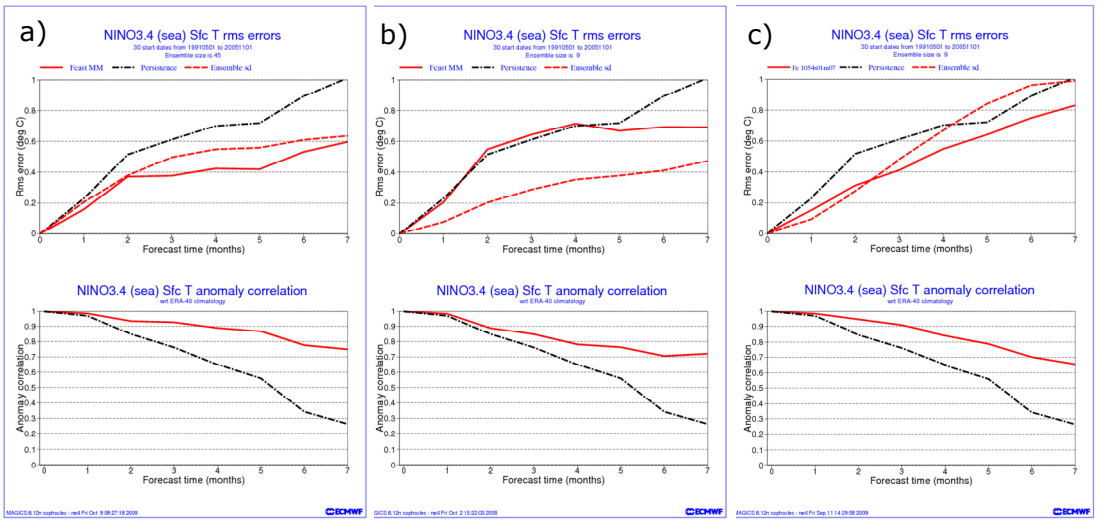

Figure 16: As in Fig. 13, but for Niño3.4 SSTs.

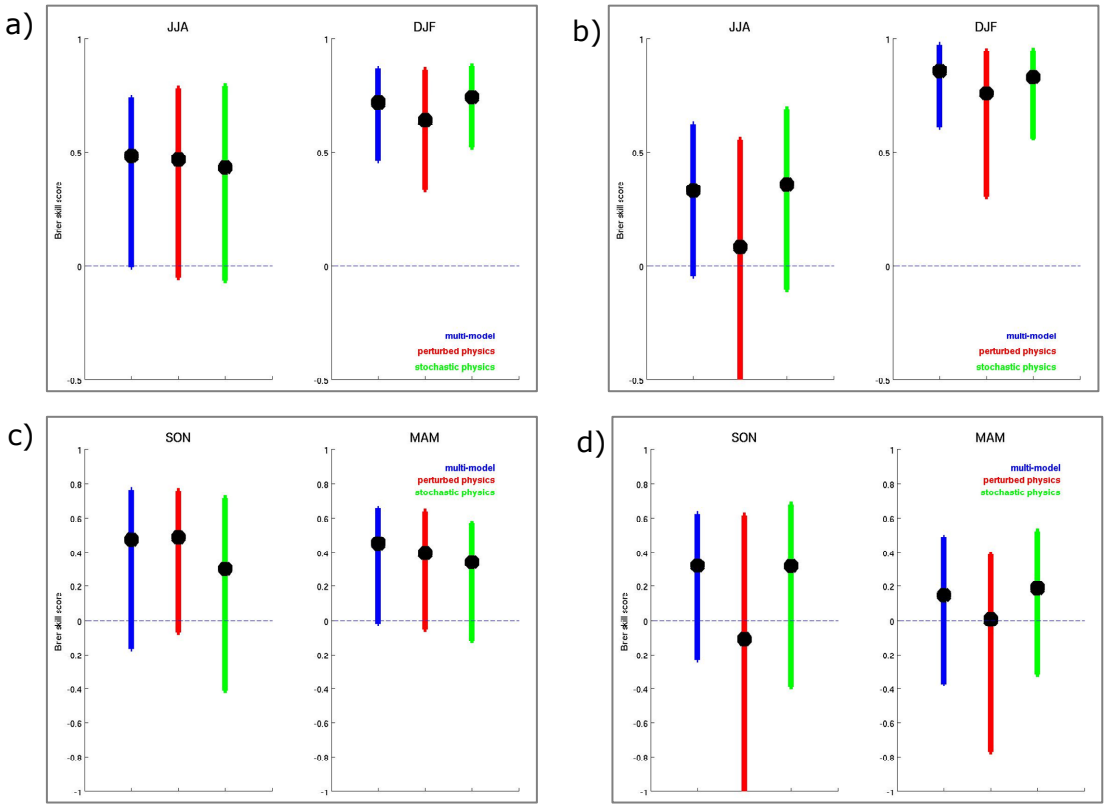

Figure 17: As in Fig. 15, but for the Niño3.4 region.

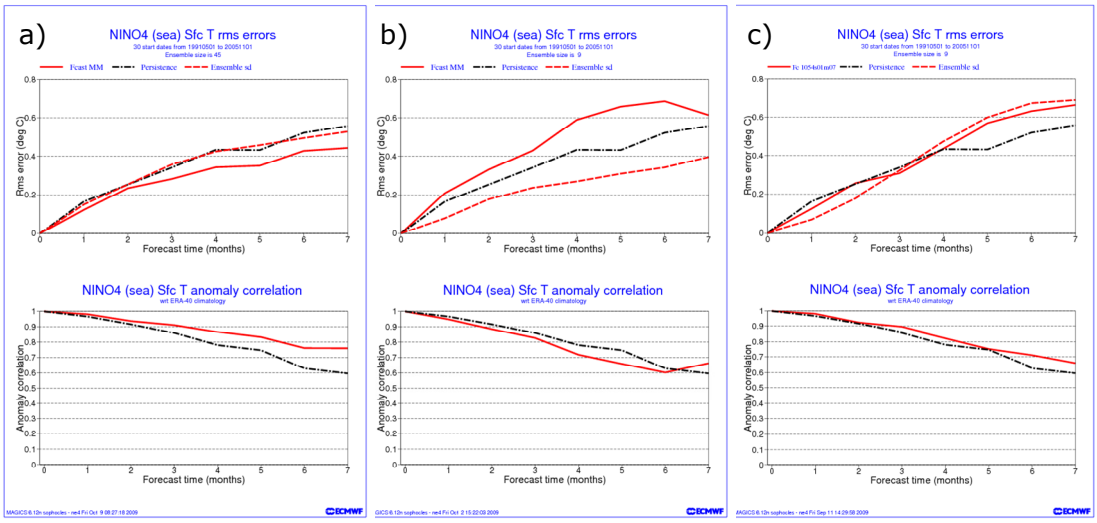

Figure 18: As in Fig. 13, but for Niño4 SSTs.

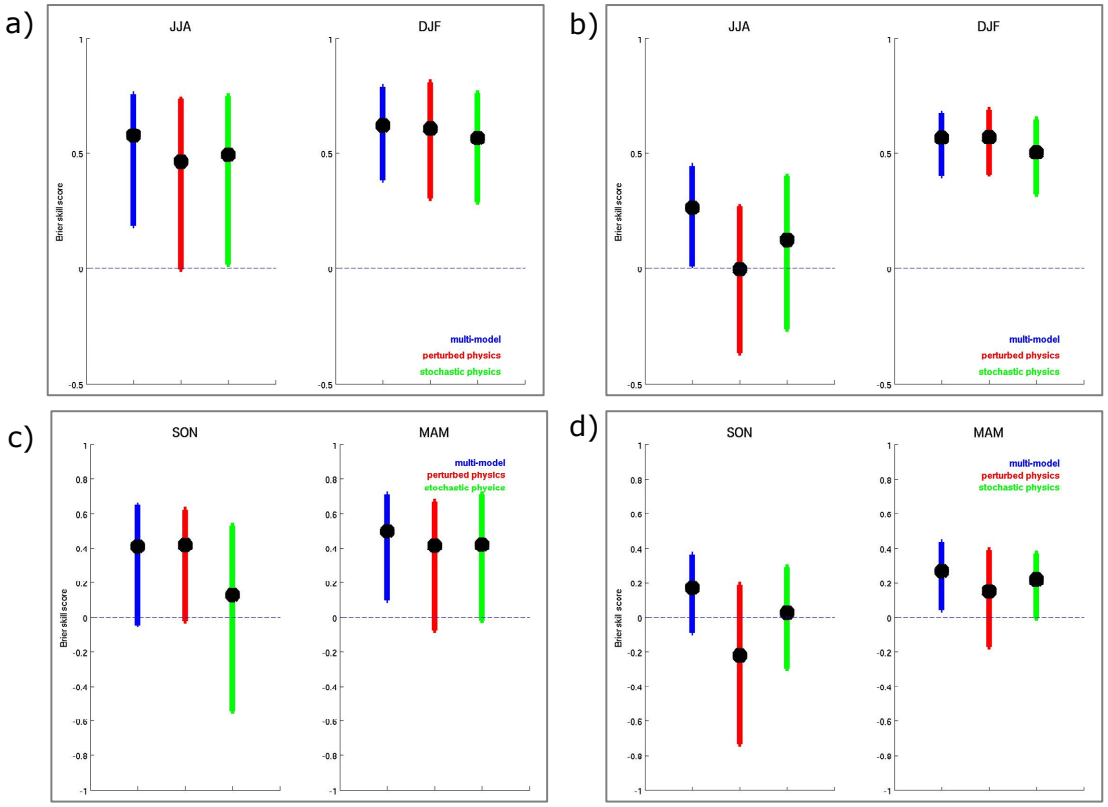

Figure 19: As in Fig. 15, but for the Niño4 region.

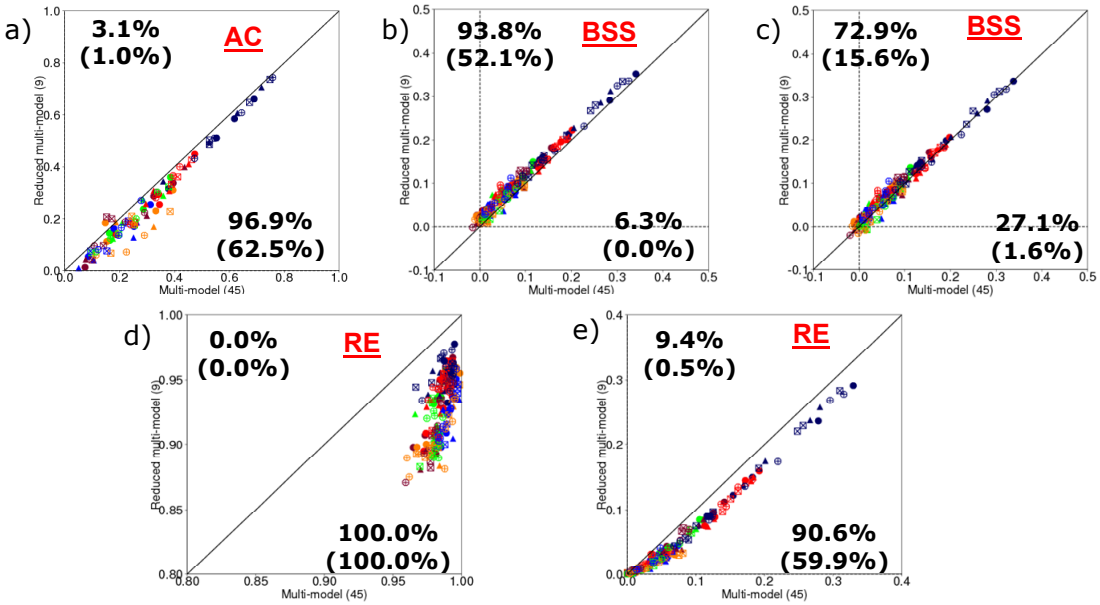

Figure 20: Comparison of seasonal hindcast skill between the full multi-model ensemble with 45 members (abscissa) and the reduced multi-model ensemble with nine members (ordinate). Each dot indicates the anomaly correlation coefficient (a), the debiased Brier skill score (b), the infinite size Brier skill score (c), the reliability (d) and the resolution skill scores (e) for one of the six large-scale regions (tropical land, tropical ocean, NH land, NH ocean, SH land and SH ocean), for the two lead times 2-4 and 5-7 months, for each of the four start dates and for near-surface temperature and precipitation. The dichotomous events used for the computation of this probabilistic skill scores are "anomaly below the lower tercile" and "anomaly above the upper tercile". The verification for temperature was done over the full hindcast period 1960-2005, whereas for precipitation it was done for the 1980-2005 period. The numbers in the corners indicate the percentage of cases where each system is superior (in brackets with 95% confidence).

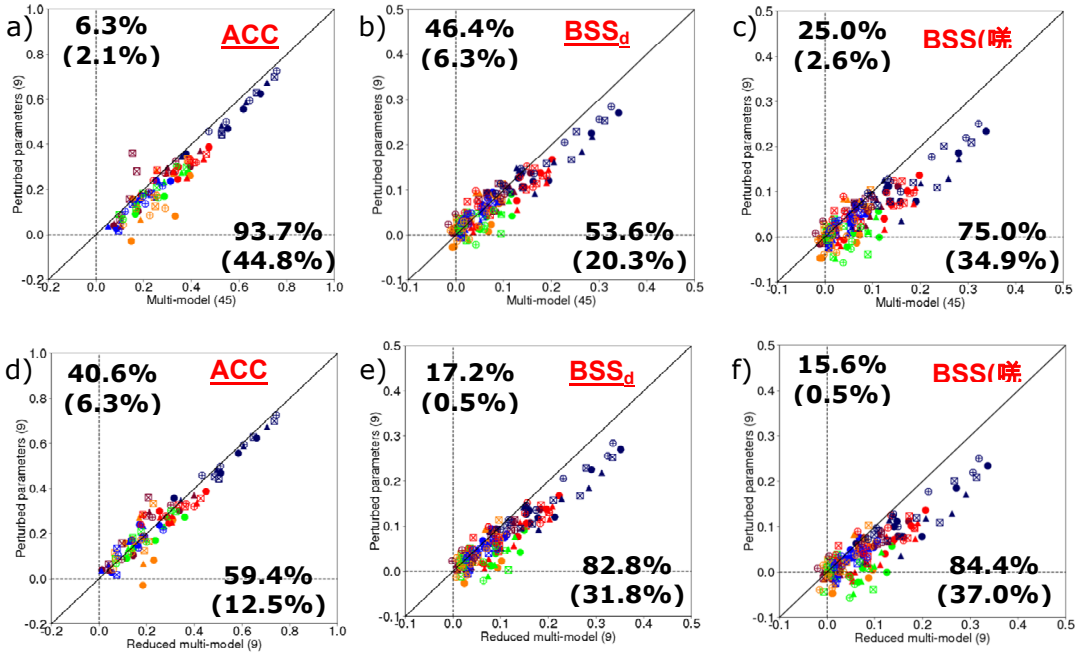

Figure 21: Comparison of seasonal hindcast skill between the multi-model ensemble (abscissa) with the perturbed parameter ensemble (ordinate). Scores shown are the anomaly correlation coefficient (a and d), the debiased Brier skill score (b and e) and the infinite size Brier skill score (c and f), see Fig. 20.The plots in the top row consider the full multi-model ensemble (45 members); the plots in the bottom row use a reduced multi-model ensemble with 9 members.

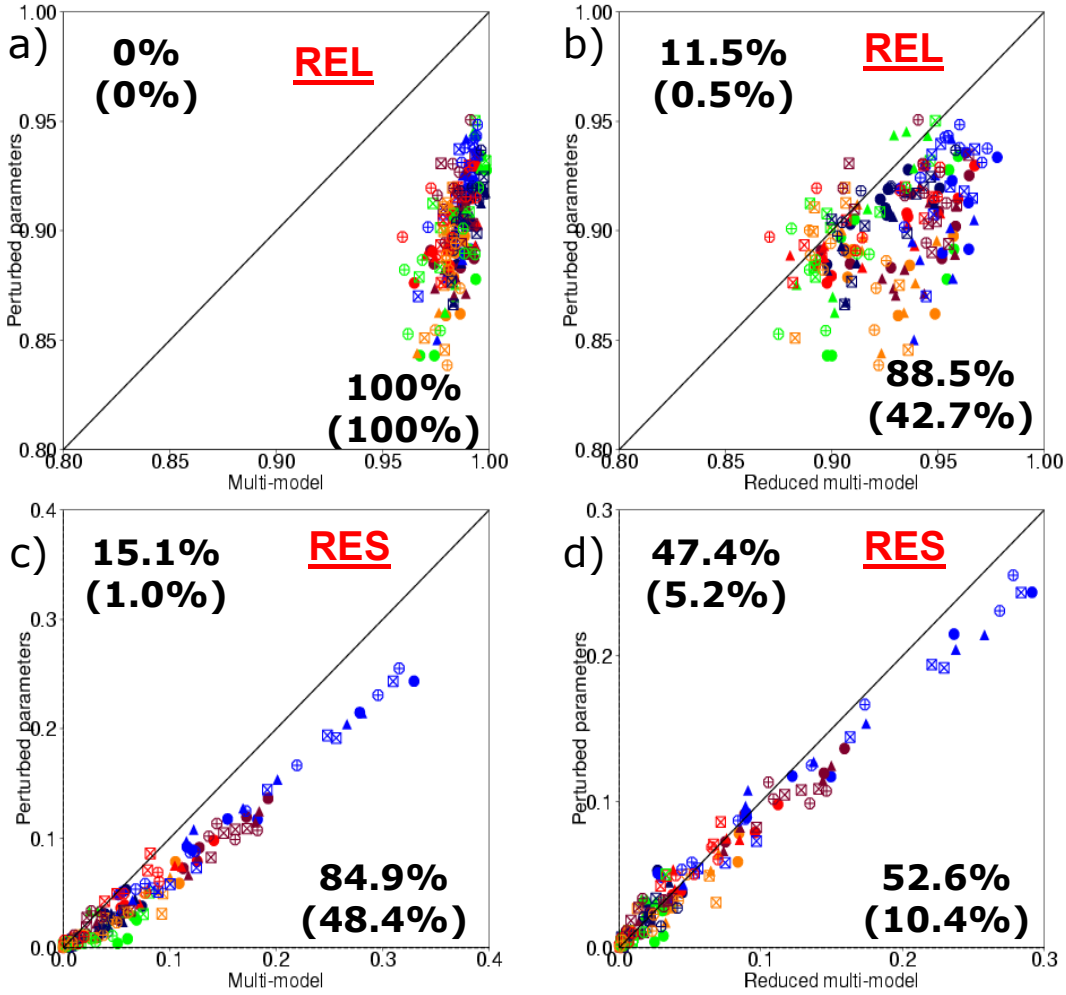

Figure 22: As Fig. 21, but for the reliability skill score (a and b) and the resolution skill score (c and d).

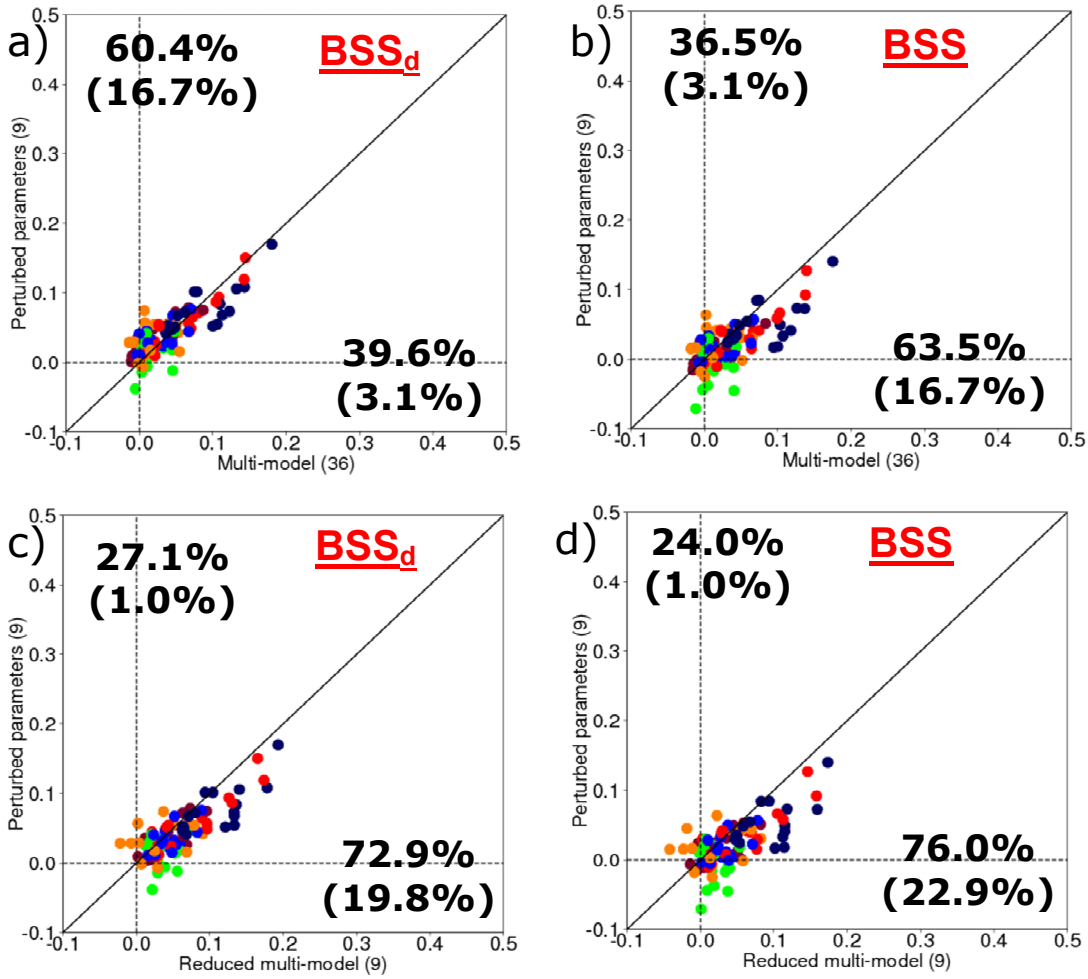

Figure 23: Comparison of seasonal to annual skill between the multi-model ensemble (abscissa) and the perturbed parameter ensemble (ordinate). The scores shown are the debiased Brier skill score and the infinite size Brier skill score, as in Fig. 20, but for lead times 6-8, 8-10, 10-12 and 12-14 months. The plots in the top row consider the full multi-model ensemble (36 members); the plots in the bottom row use a reduced multi-model ensemble for annual forecasts with nine members.

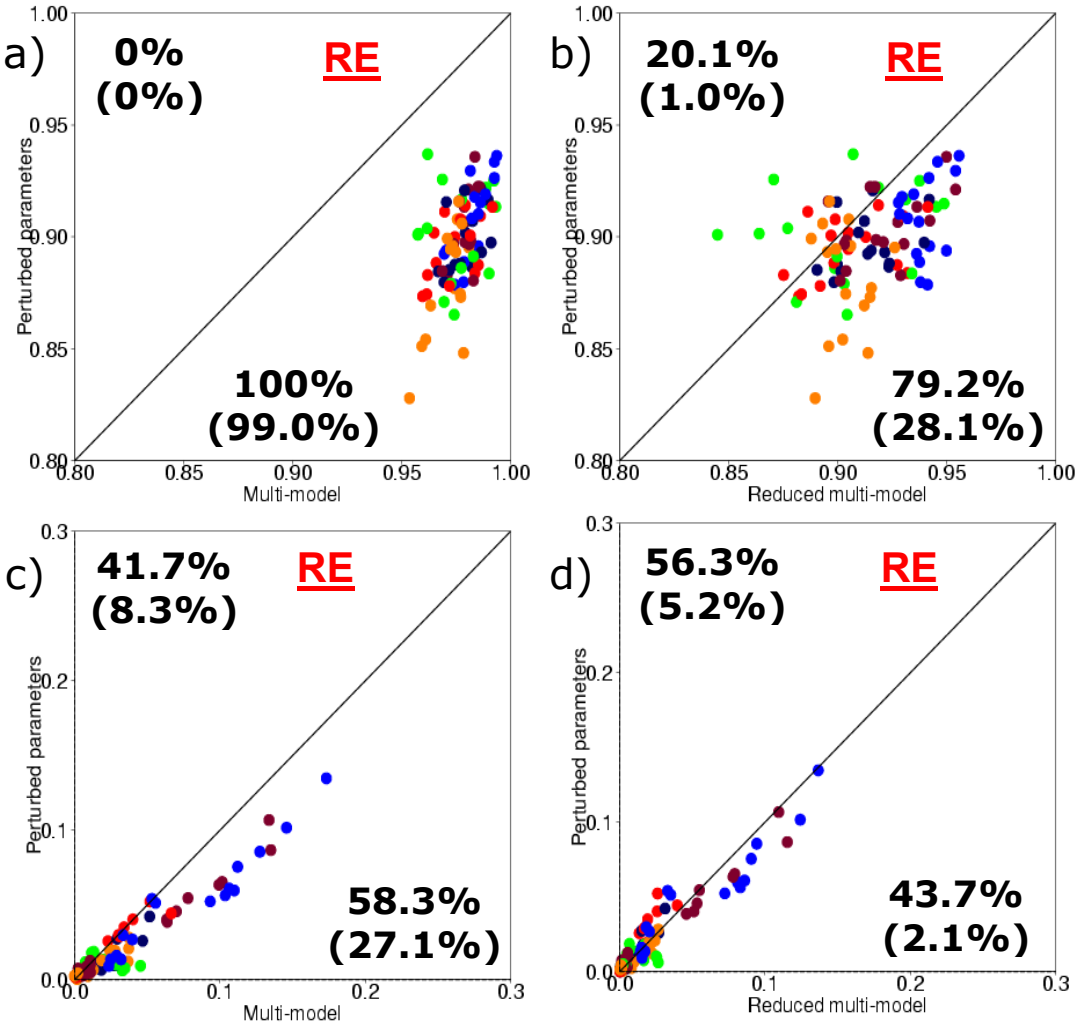

Figure 24: As Fig. 22, but for the seasonal to annual hindcasts.

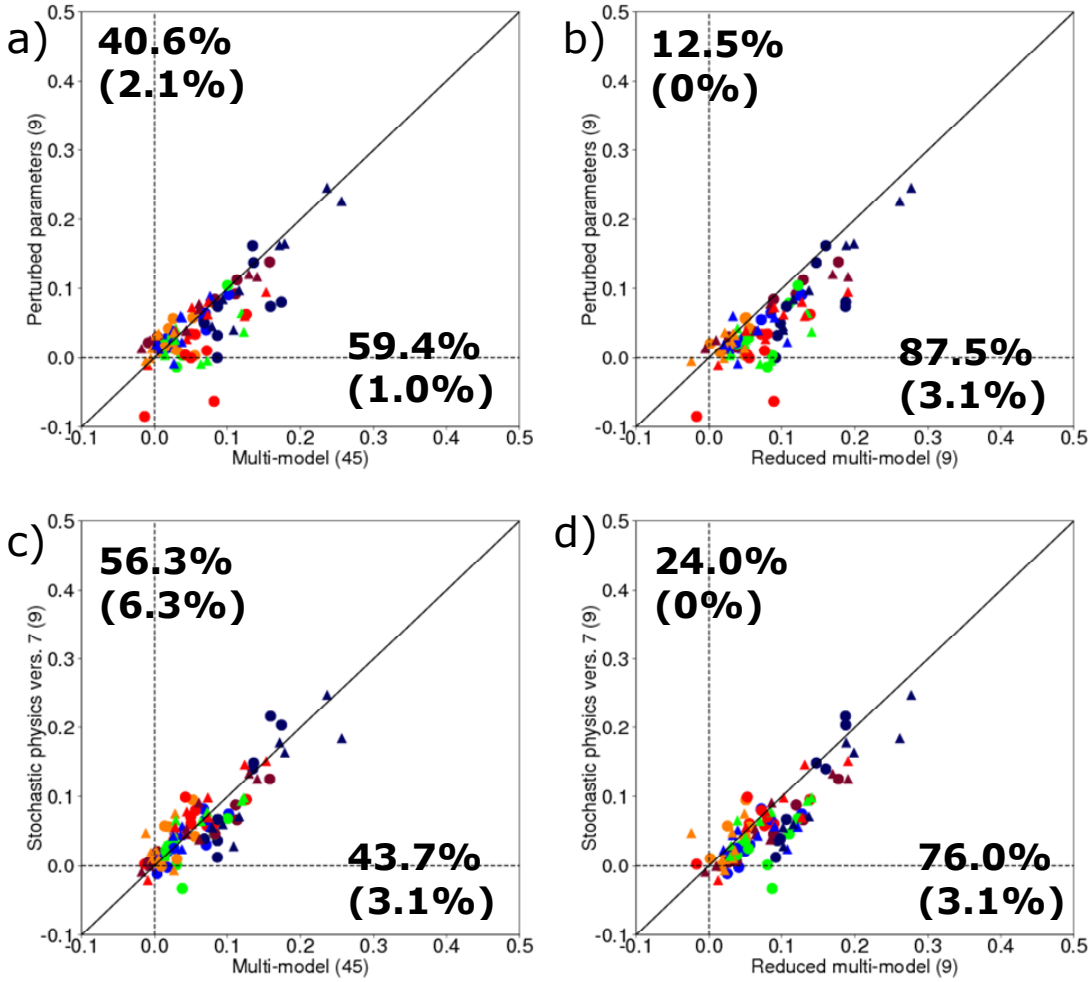

Figure 25: Comparison of seasonal hindcast skill between the multi-model ensemble (abscissa) and the perturbed parameter (a and b) or stochastic physics ensembles (c and d). The score shown is the debiased skill score, for lead times 2-4 and 5-7 months and the November and May start dates over the 1991-2005 hindcast period. The plots on the left consider the full multi-model ensemble (45 members); the plots on the right use a reduced multi-model ensemble with nine members.

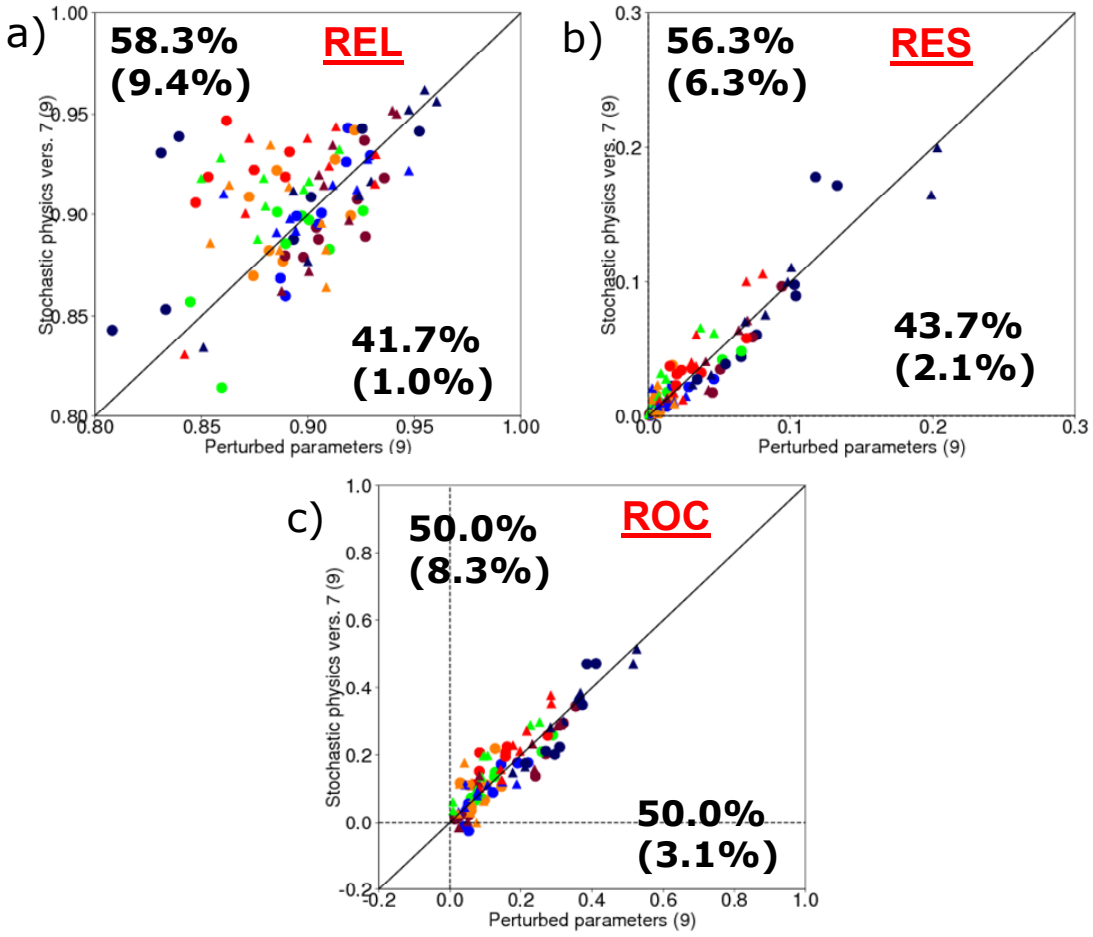

Figure 26: Comparison of seasonal hindcast skill between the perturbed parameter ensemble (abscissa) and the stochastic physics ensemble (ordinate). The scores shown are the reliability skill score (a), the resolution skill score (b) and the ROC skill score (c), for lead times 2-4 and 5-7 months and the November and May start dates over the 1991-2005 hindcast period.

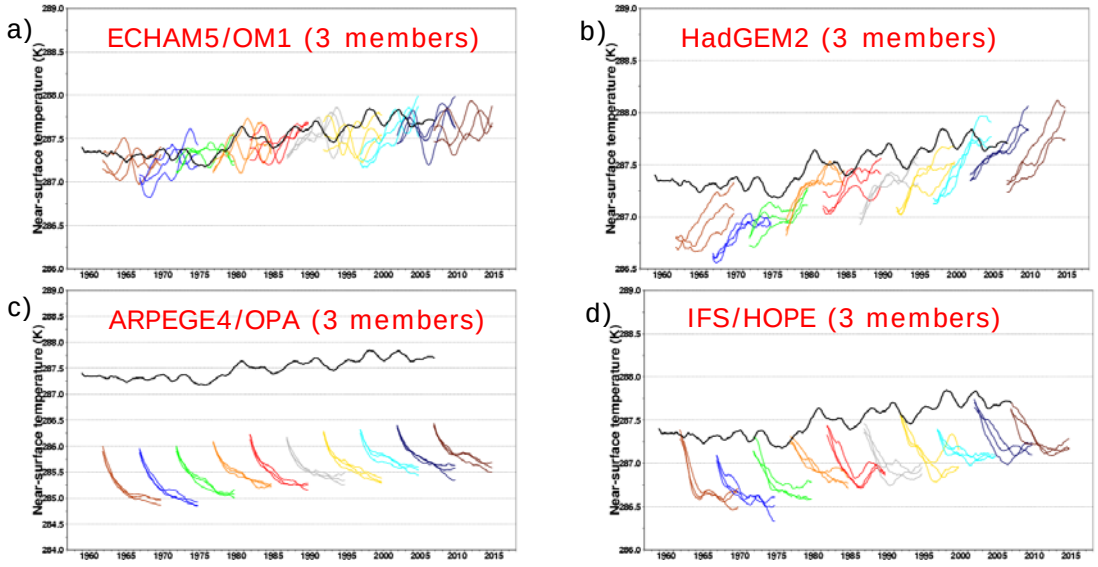

Figure 27: Global mean near-surface air temperature (2-year running mean applied) from the multi-model ENSEMBLES re-forecasts: a) IfM, b) HadGEM2, c) CERFACS and d) ECMWF. ERA40/ERAInt is used as a reference and depicted with a solid black line. Each re-forecast is plotted with lines of a different colour to illustrate the forecast setup.

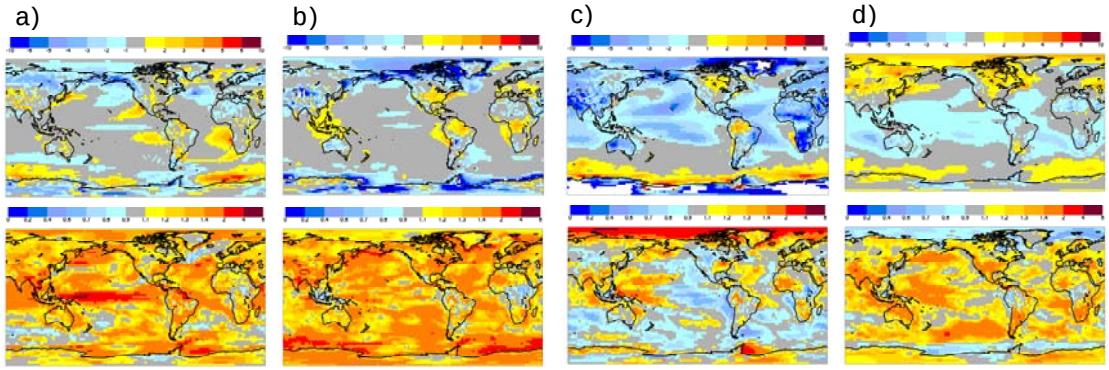

Figure 28: Near-surface air temperature biases (K, top row) and ratio between of standard deviation (bottom row) with respect to ERA40/ERAInt for a) IfM, b) HadGEM2, c) CERFACS and d) ECMWF for the forecast period 2-5 years. The climatologies are computed with the re-forecasts of the period 1960-2000.

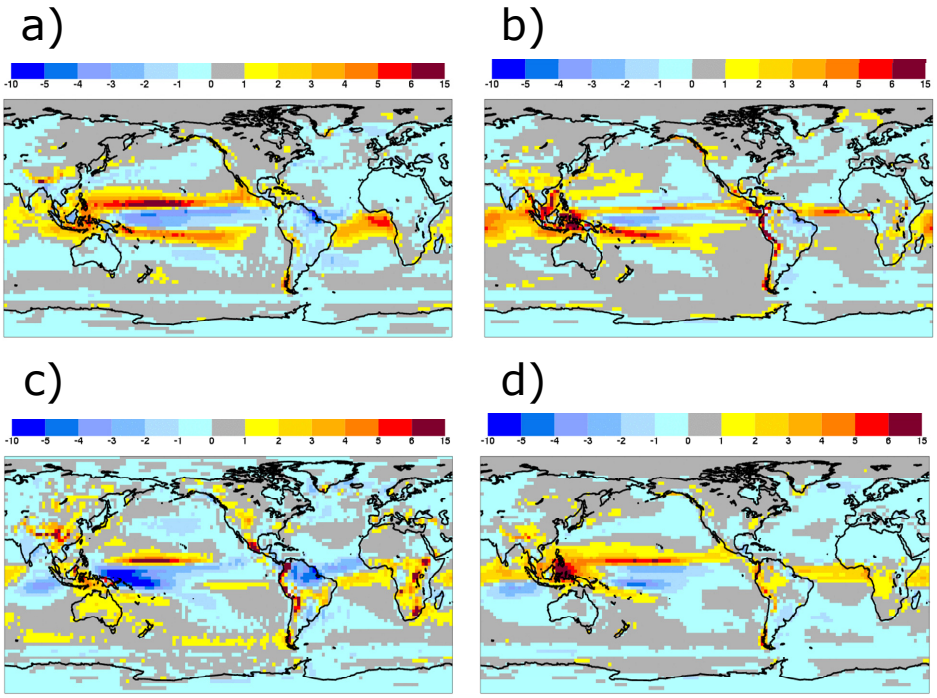

Figure 29: Precipitation (mm/day) biases with respect to GPCP for a) IfM, b) HadGEM2, c) CERFACS and d) ECMWF for the forecast period 2-5 years. The climatologies are computed with the re-forecasts of the period 1980-2000.

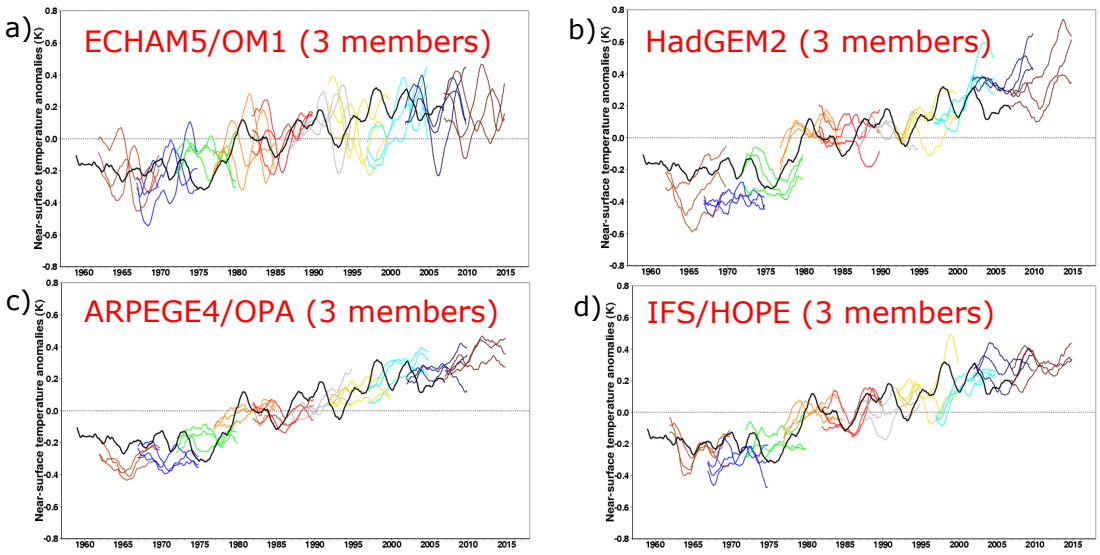

Figure 30: Global mean near-surface air temperature anomaly (2-year running mean applied) from the multi-model ENSEMBLES re-forecasts: a) IfM, b) HadGEM2, c) CERFACS and d) ECMWF. ERA40/ERAInt is used as a reference and depicted with a solid black line. The mean systematic error has been removed over the period 1960-1995. Each re-forecast is plotted with lines of a different colour to illustrate the forecast setup.

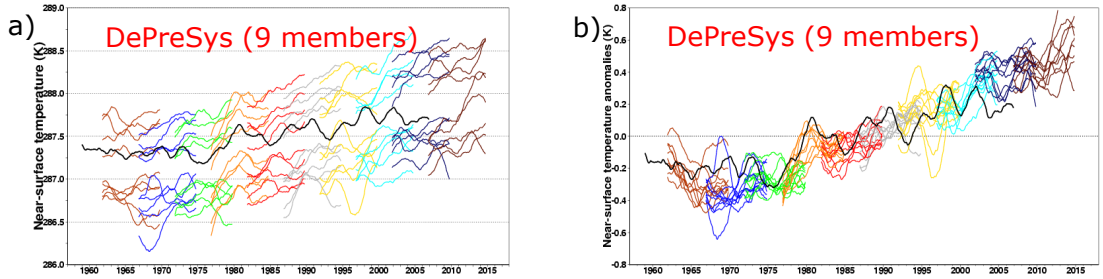

Figure 31: Global mean near-surface air temperature anomaly (2-year running mean applied) from the DePreSys_PP re-forecasts. ERA40/ERAInt is used as a reference and depicted with a solid black line. The mean systematic error has been removed over the period 1960-2005 separately for each model version. Each re-forecast is plotted with lines of a different colour to illustrate the forecast setup.

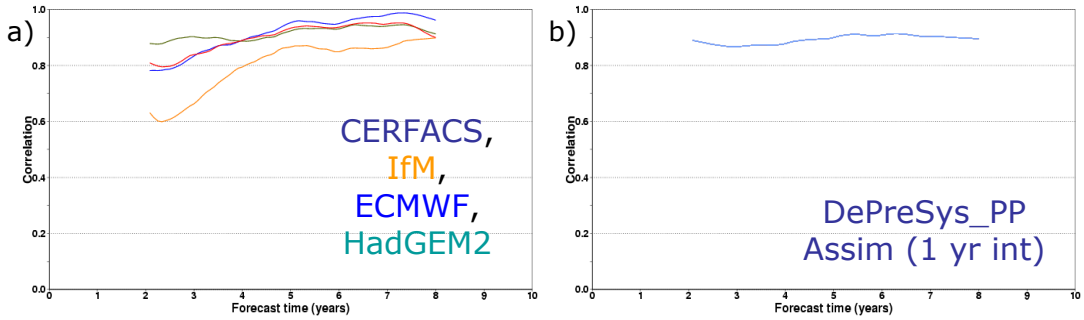

Figure 32: Ensemble-mean correlation as a function of the forecast time of the global mean near-surface air temperature anomaly (four-year running average predictions) from the a) multi-model ENSEMBLES (5-year interval between start dates) and b) DePreSys_PP (one-year interval between start dates) re-forecasts. ERA40/ERAInt is used as a reference and depicted with a solid black line. The correlation has been computed with re-forecast started over the period 1960-1995.

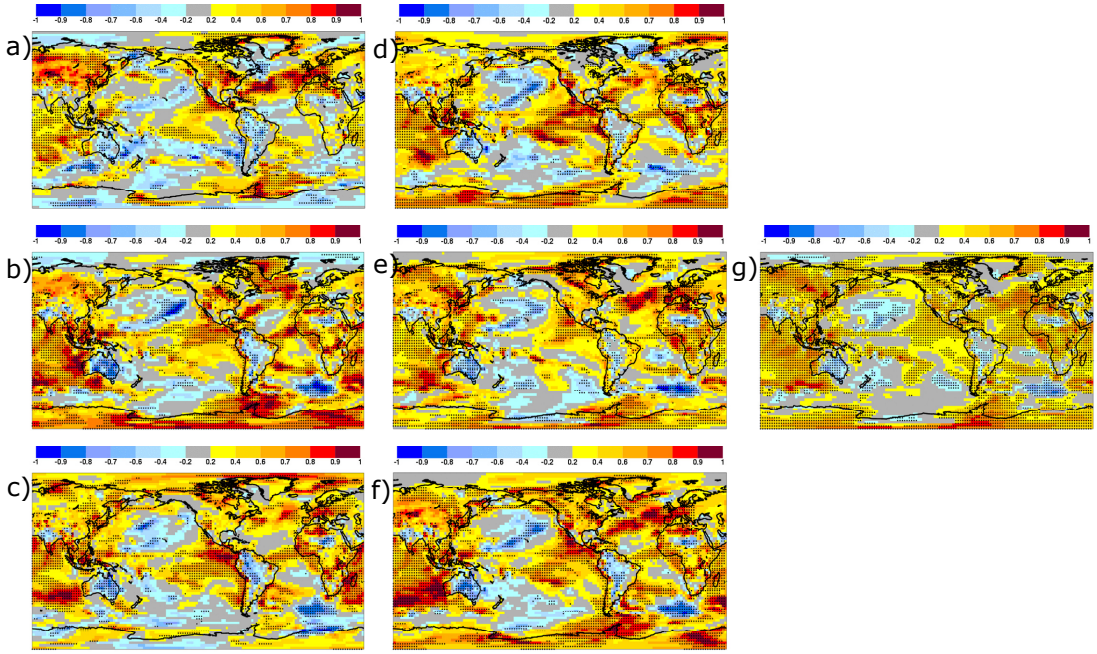

Figure 33: Ensemble-mean correlation for a) IfM, b) HadGEM2, c) CERFACS, d) ECMWF, e) DePreSys_PP (five-year interval between start dates) and f) multi-model for the forecast period 2-5 years. Panel g) shows the results for DePreSys_PP with a one-year interval between start dates. ERA40/ERAInt is used as a reference. The correlation has been computed with re-forecast started over the period 1960-1995.

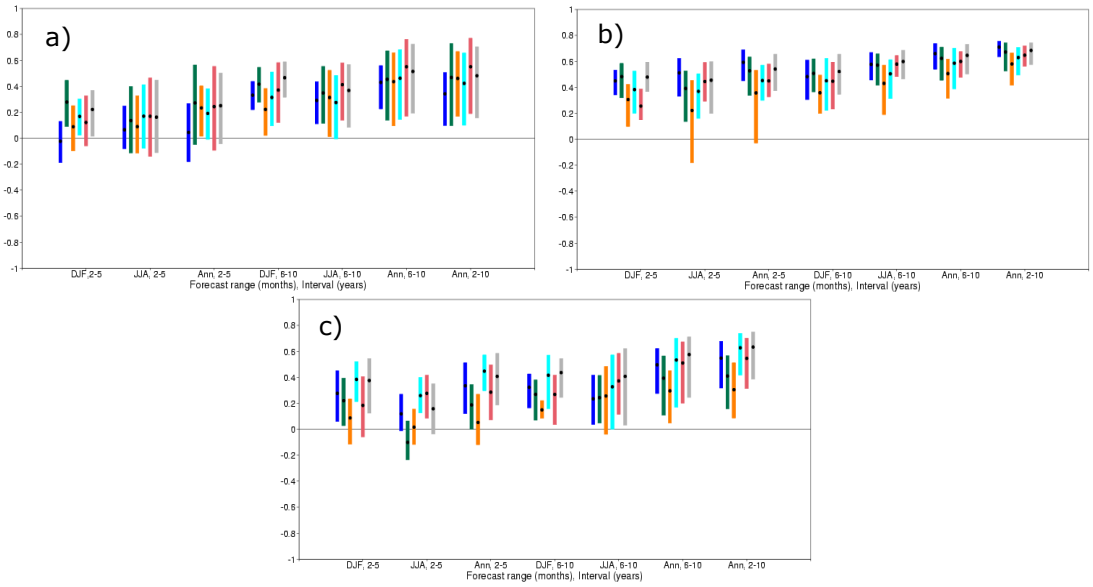

Figure 34: Ensemble-mean correlation for 850 hPa temperature over a) the Northern Hemisphere (north of 30°N), b) the tropical band (20°N-20°S) and c) the Southern Hemisphere (south of 30°S), for different forecast periods of the ECMWF (blue bars), CERFACS (green), IfM (orange), HadGEM2 (cyan), DePreSys_PP (pink) and multi-model (grey) experiments. The groups of bars correspond, from left to right, to the winter, summer and annual mean of the 2-5 year forecast period (computed with re-forecasts for the period 1960-2000), the winter, summer and annual mean of the 6-10 year forecast period and the annual mean of the 2-10 year forecast period (all computed with re-forecasts for the period 1960-1995) . The black dots depict the sample values and the bars show the 95% confidence intervals. The estimates have been computed using ERA40/ERA-Interim.

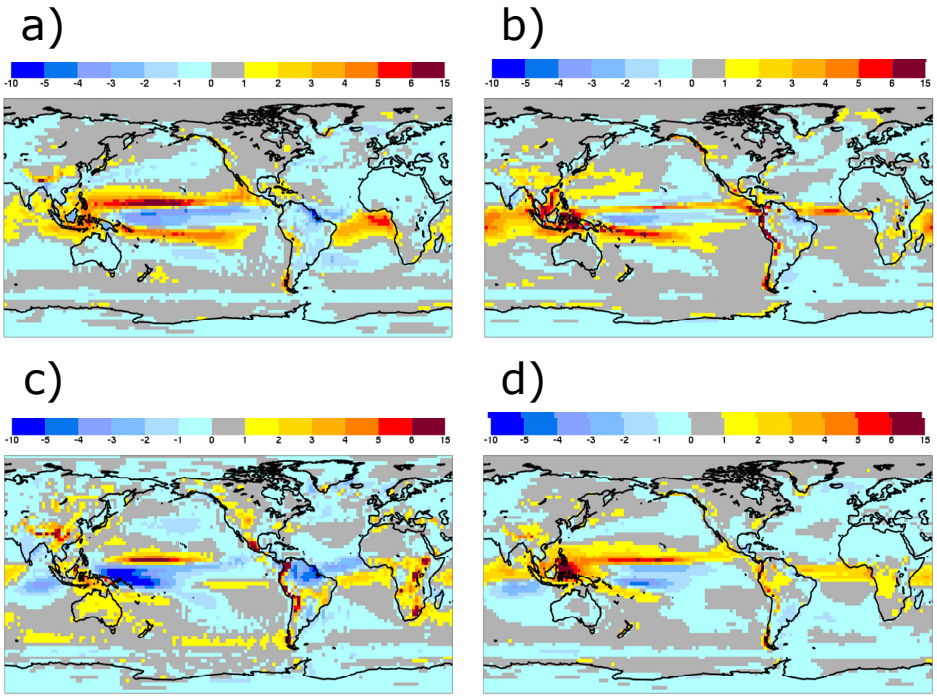

Figure 35: Upper ocean heat content (top 300 metres, K) biases for a) IfM, b) HadGEM2, c) CERFACS and d) ECMWF for the forecast period 2-5 years. The climatologies are computed with the re-forecasts of the period 1960-2000 and the biases with respect to the ocean re-analysis used to initialize the re-forecasts of each system.

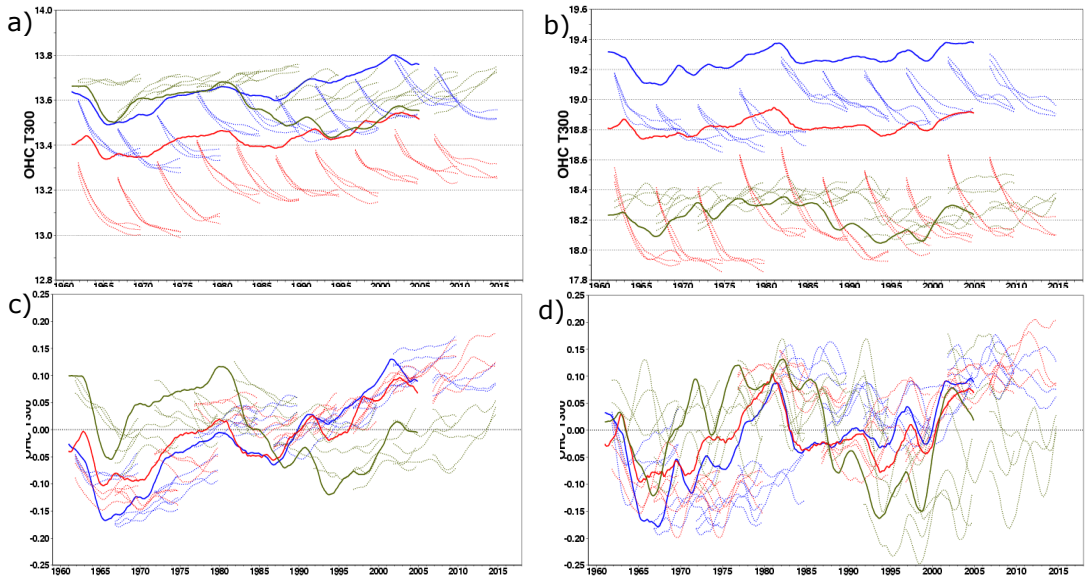

Figure 36: Upper ocean heat content (top 300 metres, K) averaged a) globally and b) over the tropics (20°N-20°S) for ECMWF (blue lines), IfM (green) and CERFACS (red) decadal re-forecasts and the corresponding ocean re-analysis (thicker lines of the corresponding colour). c) and d) Re-forecast and re-analysis upper ocean heat content anomalies. A two-year running mean has been applied to each time series.

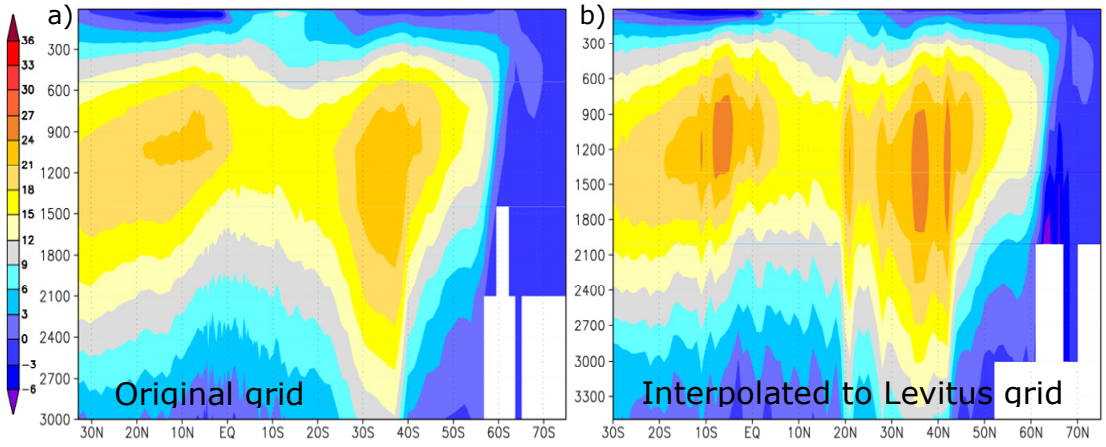

Figure 37: Mean Atlantic meridional overturning circulation intensity (Sv) as a function of depth and latitude for the ECMWF ORA-S3 ocean re-analysis computed using data a) in the original and b) interpolated to a 1 degree regular grid.

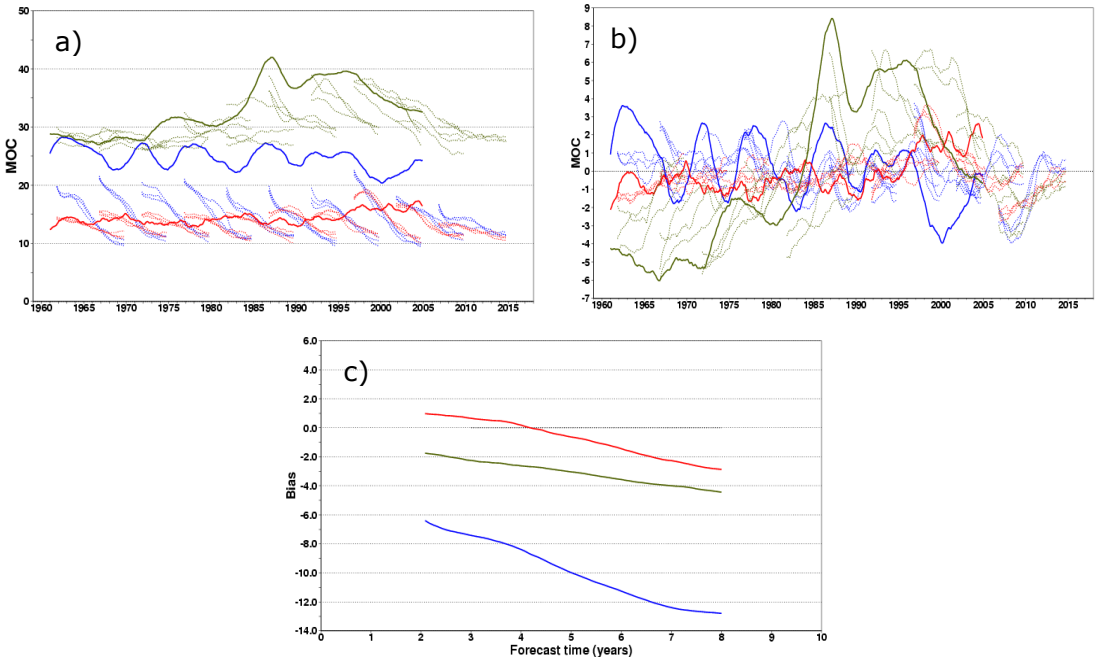

Figure 38: a) Atlantic meridional overturning circulation (AMOC) estimates (Sv) for ECMWF (blue lines, average of 10° around 37°N), IfM (green, average around 35°N) and CERFACS (red, average around 40°N) decadal re-forecasts wrt their own re-analysis (thicker lines). b) Re-forecast and re-analysis AMOC anomalies. c) Bias of the re-forecasts as a function of forecast time. A two-year running mean has been applied to each time series in panels a) and b), while panel c) shows four-year running averages.

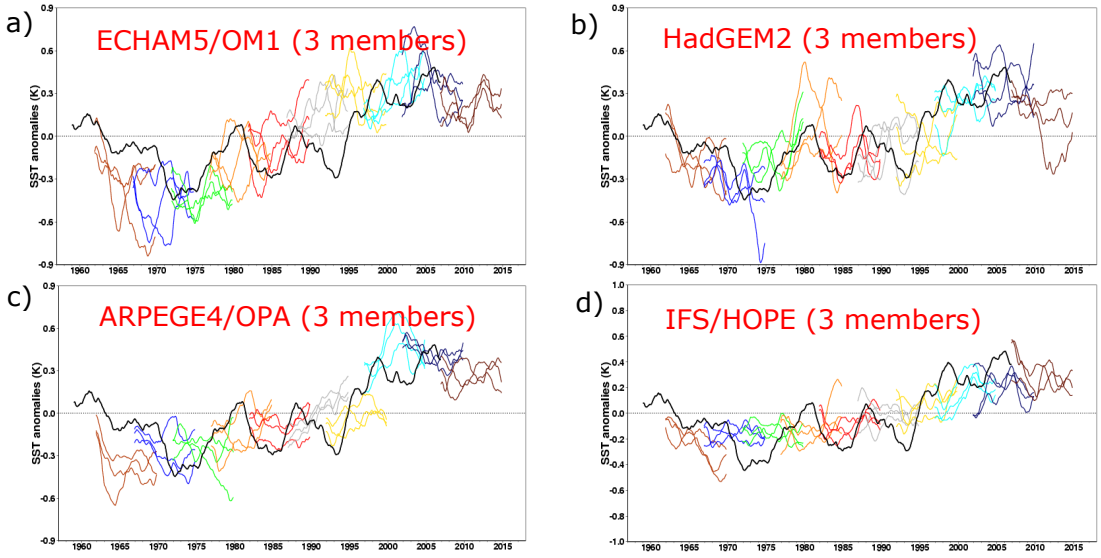

Figure 39: North Atlantic average (north of 10°N) SST anomalies for the ENSEMBLES multi-model re-forecasts: a) IfM, b) HadGEM2, c) CERFACS and d) ECMWF. Reynolds OI v2 is used as a reference (solid black line). A two-year running mean has been applied to each time series. The mean systematic error has been removed over the period 1960-1995. Each re-forecast is plotted with a lines of a different colour.

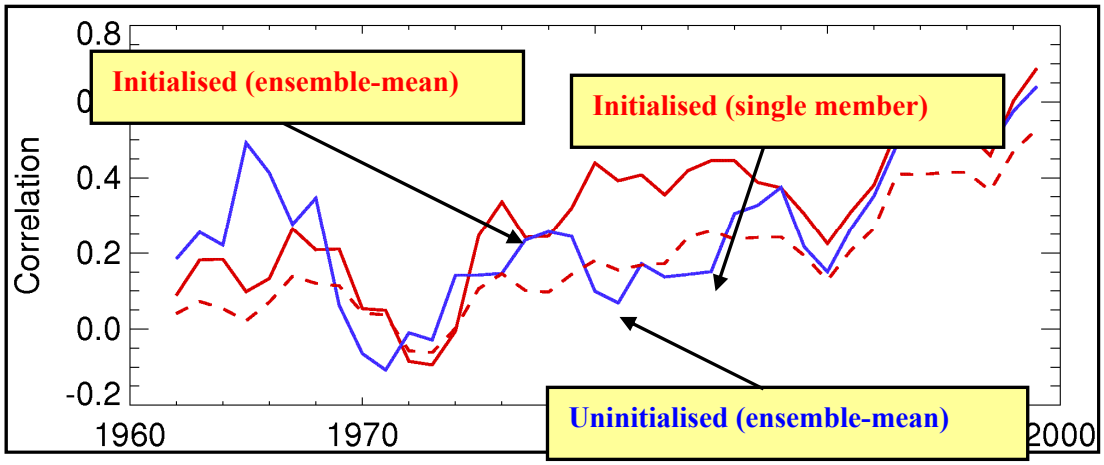

Figure 40: Time series of correlation between hindcast and observed global patterns of near-surface temperature anomalies, for hindcasts of nine-year means during the Stream 2 period. Red dashed curve shows average scores for individual variants of HadCM3 included in the DePreSys_PP ensemble, and solid red line shows scores for the ensemble mean of the nine constituent variants. Blue curve shows scores for the ensemble mean of a corresponding "No_Assim" perturbed parameter ensemble in which hindcasts are driven by the same time-dependent specification of external radiative forcing anomalies, but lacking the initialisation from analyses of atmosphere and ocean observations used in the DePreSys hindcasts.

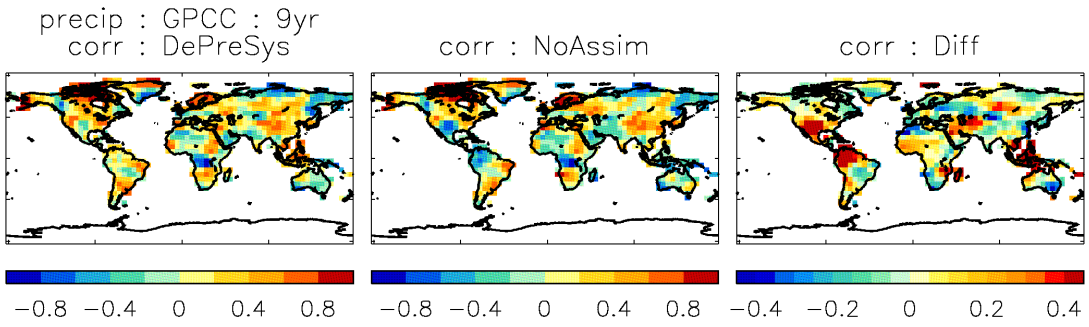

Figure 41: Correlations between hindcast and observed anomalies of nine-year mean precipitation anomalies over land, for ensemble means of the Stream 2 DePreSys_PP hindcasts initialised every year from 1960-2005 (left panel), and the corresponding uninitialised NO_Assim hindcasts (middle panel) The right panel shows the differences between the DePreSys_PP and No_Assim scores. At each location, precipitation anomalies were spatially averaged over 15x15 degree latitude-longitude boxes prior to the calculation of the correlation coefficients.