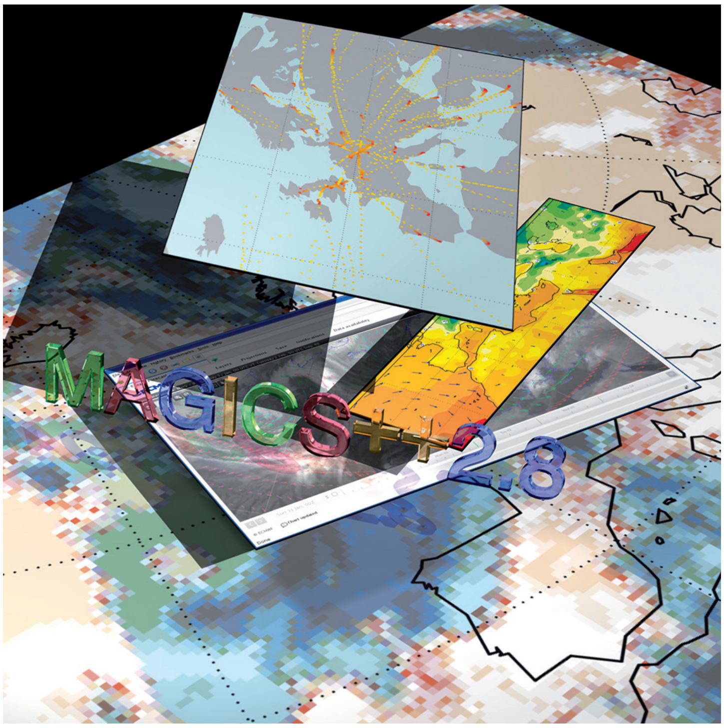
ECMWF Feature article

from Newsletter Number 122 – Winter 2009/10

METEOROLOGY

Model uncertainty in seasonal to decadal forecasting – insight from the ENSEMBLES project



www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/5hayrg0p

This article appeared in the Meteorology section of ECMWF Newsletter No. 122 – Winter 2009/10, pp. 21–26.

Model uncertainty in seasonal to decadal forecasting – insight from the ENSEMBLES project

Antje Weisheimer, Francisco Doblas-Reyes, Tim Palmer

Ensembles is a five-year EU FP6 project concerned with ensemble-based predictions of climate changes and their impacts. The project has more than 60 European partners. It came to an end in December 2009. One of its main objectives was to develop an ensemble prediction system based on global models developed in Europe to produce probabilistic estimates of uncertainty in future climate at the seasonal to decadal and longer timescales. This article describes some of the results obtained from the ensemble prediction system on seasonal to decadal time scales where ECMWF contributed with a set of coupled hindcast experiments and diagnostics of these simulations.

The key question that we were trying to address is how to best account for model uncertainty in dynamical forecasts of the climate from a few seasons to a decade ahead. Can new approaches, like perturbing physical model parameters or stochastically modifying physical parametrizations, be considered powerful alternatives to the well-established but somewhat pragmatic and *ad hoc* multi-model ensemble? To this end, a coordinated set of experiments exploring three methodologies was run and the relative merits of these approaches were assessed (the methodologies are described later). We find that, overall the multi-model ensemble gives the best forecast scores on seasonal to annual time scales, in agreement with preliminary findings (*Doblas-Reyes et al.*, 2009). The perturbed parameter and stochastic parametrization techniques are competitive new physical approaches to the traditional but intrinsically *ad hoc* assembling of single-model ensembles. These two new techniques provide promising indications that a similar level of performance to the multi-model ensemble can potentially be achieved through the application of systematic techniques for the sampling of uncertainties in a single-model system. The optimal strategy for decadal forecast production in the presence of model biases remains an open question for future work.

Ensemble techniques and methodology

The non-linear chaotic nature of the climate system makes dynamical climate model forecasts sensitive to small perturbations introduced to both the initial state of forecasts and parts of the model (e.g. changes to the structure or parameter values in a parametrization scheme). Individual forecasts with one fixed model are thus of limited value and ensembles of forecasts are used to assess the range of possible evolutions of future climate for different timescales. In the ENSEMBLES project, three different techniques to represent model uncertainties and generate ensembles for seasonal-to-decadal forecasting have been explored.

- ***Multi-model ensemble (MME).*** This combines five single-model ensembles from quasi-independent forecasting models to sample uncertainty due to differences in model formulations and in errors between the individual models (see Box A). An overview of the models contributing to the MME is given in Table 1.
- ***Perturbed physical parameter ensemble (PPE).*** This uses perturbations to numerical parameter values in physical parametrization schemes and accounts for some aspects of physical model uncertainty (see Box B). The forecasts were generated using the UK Met Office Decadal Prediction System.
- ***Stochastic physical parametrization ensemble (SPE).*** This 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 (see Box C). ECMWF's Integrated Forecast System (IFS) coupled to the HOPE ocean model was used.

Each of these three approaches is combined with a single-model ensemble of perturbed initial conditions to address uncertainty in the initial state of the system.

 doi:10.21957/5hayrg0p

Retrospective forecasts (re-forecasts or hindcasts) that emulate real-time seasonal to decadal forecast situations for the past were performed in a coordinated experiment using the MME, PPE and SPE. The full hindcast period (called Stream 2) covers the 46 years 1960–2005. For each year, 7-month-long seasonal forecasts starting on 1st of February, May, August and November have been issued. Additionally, the November forecasts from all single-model ensembles (except for those from INGV) were extended to provide a 14-month-long annual forecast. Decadal 10-year long hindcasts were initalised every five years on 1 November, that is they were started in November 1960, November 1965, November 1970 etc. The last decadal forecast was started for November 2005 and will partly be a 'real' forecast (as is the November 2000 start date). By the time of writing, the SPE has completed a subset of the Stream 2 simulations consisting of seasonal hindcasts over the 15-year hindcast period 1991–2005 with start dates in May and November.

Each of the individual model ensembles contributing to the MME was run with 9 initial condition ensemble members. Thus, the MME uses 45 members for the seasonal hindcasts and 36 for the annual-range hindcasts. The PPE and SPE were run with 9 ensemble members each.

Partner	Atmospheric model and resolution	Ocean model and resolution	Initialization		Additional components and comments
			Atmosphere and land	Ocean	
ECMWF	IFS Cy31r1 T159/L62	HOPE 0.3°–1.4°/L29	ERA-40/ operational analysis, atmospheric singular vectors	Wind stress perturbations to generate ensemble of ocean reanalyses; SST perturbations at initial time	Indentical to the operational seasonal forecasting system S3
UK Met Office	HadGEM2-A N96/L38	HadGEM2-O 0.33°–1°/L20	ERA-40/ operational analysis, anomaly assimilation for soil moisture	Wind stress perturbations to generate ensemble of ocean reanalyses; SST perturbations at initial time	Fully interactive sea ice module
Météo-France	ARPEGE4.6 T63	OPA8.2 2°/L31	ERA-40/ operational analysis	Wind stress, SST and water flux perturbations to generate ensemble of ocean reanalyses	GELATO sea ice model
Leibniz Institute of Marine Sciences at Kiel University (IFM)	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		
Euro-Mediterranean Centre for Climate Change (INGV) in Bologna	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	Dynamical snow-sea ice model and land-surface model

Table 1 Overview of models contributing to the new ENSEMBLES multi-model ensemble.

doi:10.21957/5hayrg0p

The ENSEMBLES multi-model ensemble (MME) A

The MME for seasonal forecasts comprises global coupled atmosphere-ocean climate models from the UK Met Office, Météo-France, European Centre for Medium-Range Weather Forecasts (ECMWF), Leibniz Institute of Marine Sciences at Kiel University (IFM) and the Euro-Mediterranean Centre for Climate Change (INGV) in Bologna. All models include major radiative forcings. None of the coupled models has flux adjustments. The atmosphere and ocean were initialized using realistic estimates of their observed states.

Table 1 summarises the main model components and their initialization strategies. Additional details on the initial condition perturbations can be found in *Weisheimer et al.* (2009).

We have applied the simplest approach to constructing an MME by combining individual models using equal weights to all contributing models and ensemble members. On the annual-range, forecasts from UK Met Office, Météo-France, ECMWF and IFM contributed to the MME.

The perturbed parameter ensemble (PPE) B

The PPE samples model uncertainty in poorly constrained cloud physics and surface parameters. It was generated with the UK Met Office Decadal Prediction System (DePreSys) which is based on the HadCM3 climate model. The model uses flux adjustments to restrict the development of regional biases in SST and salinity. Eight versions of the model with simultaneous perturbations to 29 parameters were used in addition to the unperturbed version so that each member of the PPE samples a different set of parameter values (*Doblas-Reyes et al.*, 2009).

In order to generate initial conditions for the hindcasts, each model version was run in assimilation mode with atmospheric and oceanic anomalies assimilated. The assimilation integration was itself started with an initial state taken from a simulation of the 20th century climate.

The stochastic physical parametrization ensemble (SPE) C

Conventional physical parametrization schemes describe the effects of subgrid-scale processes in models of weather and climate by deterministic bulk formulae which depend on 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 contributes to model errors on different spatial and temporal scales. Stochastic physical parametrization ensembles provide a methodology for representing model uncertainty due to variability of the unresolved scales.

ECMWF has recently revised its stochastically perturbed parametrization tendency (SPPT) scheme and developed the stochastic backscatter scheme (SPBS) (see *Palmer et al.*, 2009). For the SPE, both these schemes have been included in the preliminary set of seasonal hindcasts based on the IFS Cy35r2 coupled to the HOPE ocean model.

- The SPPT scheme applies univariate Gaussian perturbations to the total parameterised tendency 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 total dissipation rate and uses vertical phase correlations.

doi:10.21957/5hayrg0p

Hindcast skill on seasonal time scales

The scientific basis for seasonal predictability lies in the slowly evolving components of the climate system, like the ocean or land surface, 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 (El Niño/ Southern Oscillation) event in the tropical Pacific, which is the dominant mode of seasonal and interannual climate variability. Because ENSO has, via its well-known teleconnection patterns, remote effects on the weather and climate, assessing the skill of forecasting the sea surface temperatures (SSTs) in the tropical Pacific is essential also for the predictability on seasonal time scale in other parts of the world.

Model drift

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 1a shows the mean model drift for SST, estimated from all ensemble members and hindcasts, in the Niño3 region (5°S–5°N, 150°W–90°W) for the individual models contributing to the MME for each of the four start months. For comparison, Figure 1b shows the SST drift for a set of previous-generation models from the DEMETER project (an EU-funded project for the development of a European multi-model ensemble system for seasonal to interannual prediction).

It is clear that 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 varied between +2° C and –7° C for a lead time of up to 6 months, the ENSEMBLES models have a much reduced drift with overall value of less than ±1.5° C, see also *Weisheimer et al.* (2009).

Results not shown here indicate that the individual model versions/ensemble members of PPE have only a small drift, which is not surprising as its initialization uses observed anomalies rather than full fields. SPE develops a slightly warm drift during the first couple of months and a weak cold drift thereafter. The drift does not exceed ±0.5° C for the 7-month forecast range.

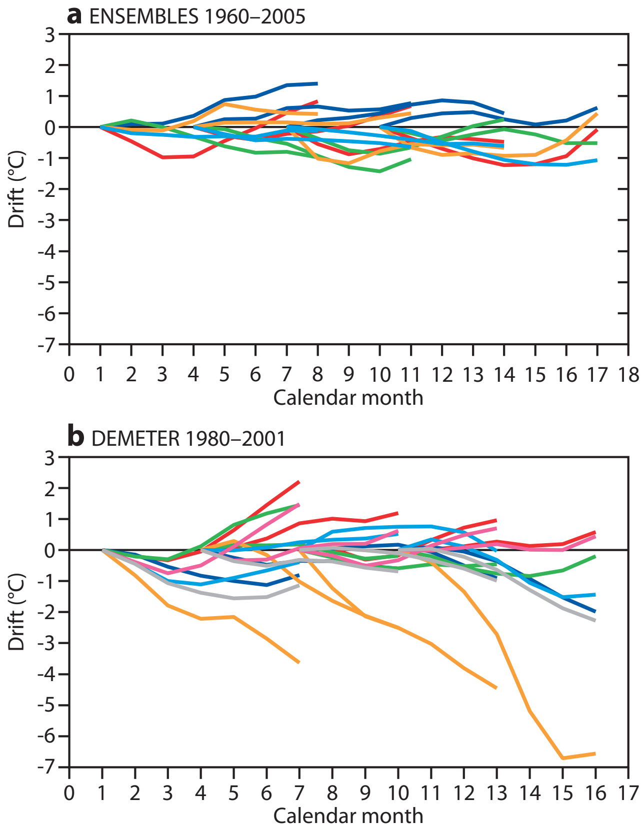


Figure 1 Systematic model errors in Niño3 SST indicated by the drift from the verification over the 7-month forecast time for the individual seasonal forecast models contributing to (a) the ENSEMBLES MME and (b) the DEMETER MME. Results for all four start dates are shown. The drift has been estimated from all available ensemble members for each start date separately over the hindcast periods 1960–2005 (ENSEMBLES) and 1980–2001 (DEMETER). The colour codes are red – Météo-France, dark blue – ECMWF, green – UK Met Office, orange – IFM, light blue – INGV, grey – LODYC and pink – CERFACS (the latter two only used for DEMETER). The abbreviations for the forecasting centres are defined in Table 1. Figure from *Weisheimer et al.* (2009).

Forecast skill – tropical Pacific perspective

The systematic errors have been corrected for computing forecast anomalies 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 2 shows the temporal evolution of ensemble-mean root-mean square error (RMSE) and ensemble spread for the SST hindcast anomalies over the tropics and over the Niño3 region for MME, PPE and SPE. For a well calibrated (or reliable) system there should be a close match between forecast error and ensemble spread.

As can be seen in Figure 2, the MME has the smallest forecast errors over all lead times. It is also the best calibrated ensemble in terms of the match between forecast error and ensemble spread. While PPE is systematically under-dispersive for all forecast times (i.e. there is not enough spread in the ensemble), SPE has a good match between the errors and ensemble spread throughout the forecast range. The main improvement of SPE over the corresponding control version of the IFS/HOPE system without stochastic physics consists of a significant increase in spread to an otherwise considerably under-dispersive forecasting system. This increase in spread leads to more reliable forecasts and thus better probabilistic skill scores.

The results from the ENSEMBLES MME confirm earlier findings from the DEMETER project (*Hagedorn et al.*, 2005) that the MME, compared to the single-model ensembles, effectively reduces the RMSE while the ensemble spread is increased leading to overall improved forecast skill.

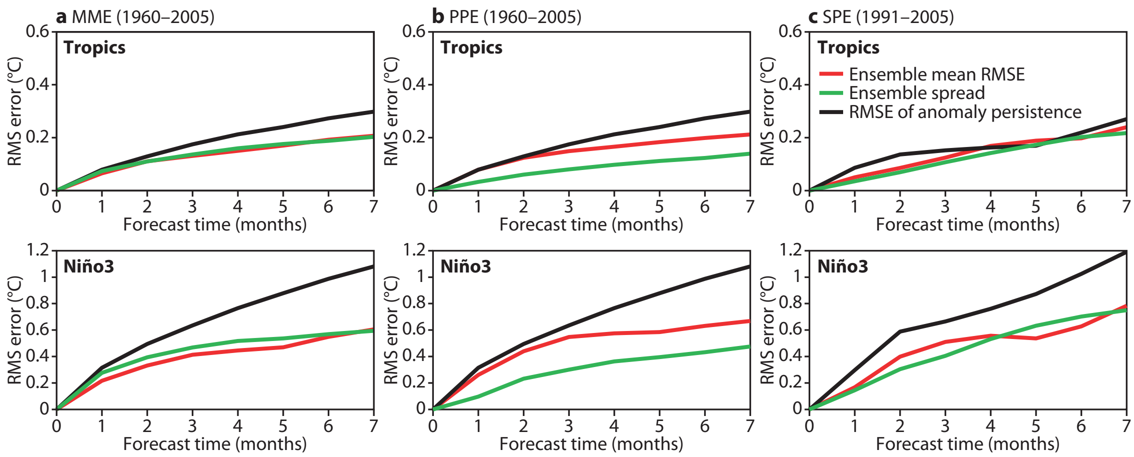


Figure 2 Seasonal-range RMS forecast error and ensemble spread as a function of lead time for (a) MME, (b) PPE and (c) SPE based on all available start dates. Scores are shown for SSTs averaged over the whole tropics (top row) and over the Niño3 region in the tropical Pacific (bottom row). Results are shown for the ensemble-mean RMSE (red), ensemble spread (green) and RMS error based on a simple statistical model of anomaly persistence (black). Scores for MME and PPE have been estimated from the Stream 2 hindcast period 1960–2005 whereas for SPE the reduced period 1991–2005 was used. The bottom panel of (a) is from *Weisheimer et al.* (2009).

doi:10.21957/5hayrg0p

Forecast skill – global perspective

As a measure of probabilistic forecast skill Figure 3 shows global maps of the Brier Skill Score (BSS) for near-surface temperature and two lead times. Because the BSS estimate is affected by the different ensemble sizes of the three forecasting systems (it has a negative bias for small ensemble sizes), an analytical expression that extrapolates the score to a hypothetical infinite ensemble size, $BSS(\infty)$, will be used in the following (*Ferro et al.*, 2008). As with the standard Brier Skill Scores, $BSS(\infty)=1$ indicates perfect forecasts, $BSS(\infty)=0$ for forecasts that have as much skill as the reference, and $BSS(\infty)<0$ indicates forecasts that are less skilful than the reference. We use the climatological forecast from a reanalysis as the reference.

It can be seen that the MME has, on average, the highest skill, in particular in the tropics at shorter lead times (Figure 3a). The tropical Pacific is an area of very high skill for all three systems during the December, January and February season. While the pattern of positive skill in PPE for lead times of 2–4 months (Figure 3b) has a large-scale structure, the hindcasts based on SPE (Figure 3c) show regions of higher skill than PPE but over somewhat smaller areas. However, the skill estimates for MME, PPE and SPE are based on different hindcast periods (see the figure caption), which implies larger sampling uncertainty for SPE. The general drop of skill in the tropical Pacific at lead time 5–7 months during the boreal spring season of March, April and May (Figures 3d to 3f) cannot only be attributed to the generic loss of prediction skill at longer forecasting ranges, but is also related to the spring barrier, a seasonal dependence of ENSO forecast skill with substantially lower skill during and after the spring months.

A more detailed analysis of the skill in PPE versus SPE over their common hindcast period revealed that, while in general the performance of the two systems is comparable in absolute terms, forecasts based on SPE tend to be more reliable and have slightly better resolution than forecasts issued by PPE.

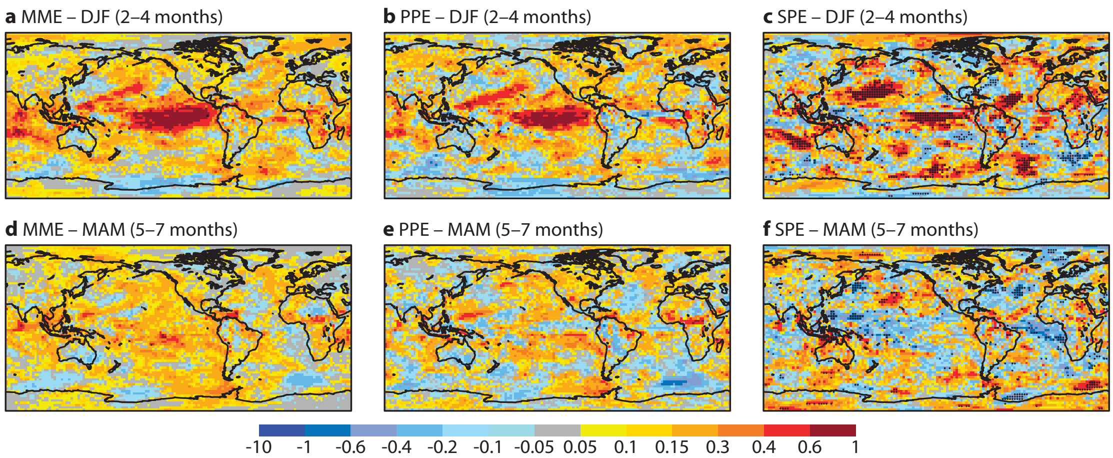


Figure 3 Brier Skill Score (BSS) for an infinite-sized ensemble of near-surface temperature anomalies falling in the upper tercile for (a) MME, (b) PPE and (c) SPE based on hindcasts initialised on 1 November for the December, January and February (DJF) season (lead time 2–4 months). (d), (e), (f). As (a), (b), (c) but for the March, April and May (MAM) season (lead time 5–7 months). As a reference forecast to compute the skill score the climatological forecast from a reanalysis (ERA-40) was used. Scores for MME and PPE have been estimated from the Stream 2 hindcast period 1960–2005 whereas for SPE the reduced period 1991–2005 was used.

Forecast skill – European perspective

As Figure 3 gives a global perspective of the level of skill in the three systems, one might also be interested in the actual seasonal forecast skill for the region we live in, that is Europe. As an example of two standard European regions, Figure 4 has a comparison of the Brier Skill Scores for an infinite-sized ensemble for near-surface temperature and precipitation over Southern Europe. Here, only land points over the region 30° N to 48° N and 10° W to 40° E have been used.

Figure 4a show *BSS*(∞) for the lower tercile (left panel) and upper tercile (right panel) temperature events, while Figure 4b shows the corresponding scores for precipitation. As can be seen, temperature is, on the whole, more predictable than precipitation with more skill than a simple climatological forecast in summer. The significant skill in forecasting summer temperature in all three systems can be partially explained by the long-term warming trend in the observations that is well captured in the seasonal hindcasts (not shown). For the winter temperature forecasts, PPE has a lower skill relative to MME and SPE. The level of skill for predicting precipitation over Europe is similar to a climatological forecast for all three systems.

Northern Europe is a less predictable region than Southern Europe with skill scores that are often not much better than a climatological forecast (not shown). The lower skill in predicting temperature compared to Southern Europe is partly due to the fact that temperature in Northern Europe, in contrast to Southern Europe, does not show any obvious long-term trend over the hindcast period.

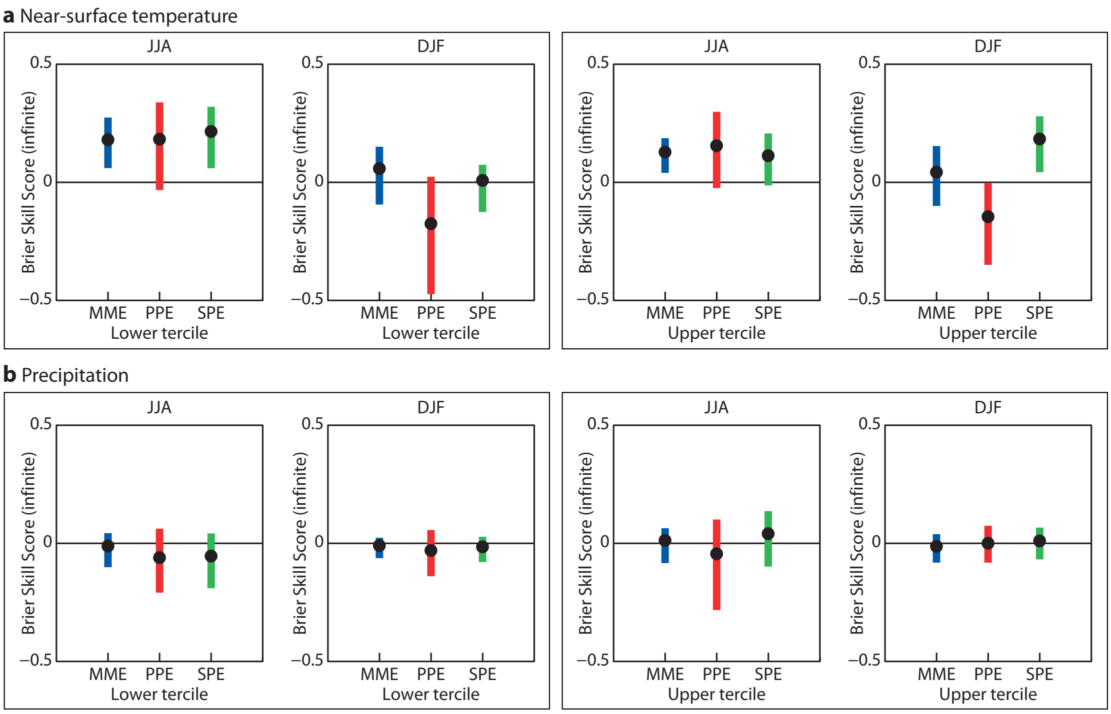


Figure 4 Brier Skill Score for an infinite-sized ensemble of (a) temperature and (b) precipitation over Southern Europe land points for MME (blue), PPE (red) and SPE (green). Scores for the event ‘anomalies in the lower tercile’ are in the left boxes and scores for ‘anomalies in the upper tercile’ are in the right boxes. Two start dates (May and November) and lead times of 2–4 months have been used over the common hindcast period 1991–2005. The error bars (95%) have been computed using a bootstrapping method with replacement. ERA-40/operational analysis has been used for the verification of temperature and GPCP for the verification of precipitation.

doi:10.21957/5hayrg0p

Annual-range forecasts

Hindcasts of the MME and PPE starting in November have been extended to 14 months to explore predictability on annual time scales. Corresponding runs with the SPE are not available as of now. Some positive skill has been found on these long lead-times for Niño3 SSTs. The anomaly correlation drops to 0.5 and 0.4 at month 9 for MME and PPE, respectively, and remains nearly constant thereafter. Remarkably, the above-mentioned good match between the RMSE and spread of the ensemble in the MME is further sustained over the extended forecast lead-time with an approximately linear error and spread growth. The PPE becomes under-dispersive after about month 6.

Prospects for decadal predictions

As part of the ENSEMBLES activities to explore the potential of decadal predictions using coupled atmosphere-ocean model initialised from observed states, we have been, for the first time at ECMWF, testing the IFS/HOPE coupled model in ten-year long integrations. Our model does not use any techniques (e.g. anomaly initialisation, nudging or flux corrections) to avoid the coupled system drifting away from the observed state. It is based on the atmospheric IFS cycle 33r1 and also includes new monthly evolving two-dimensional climatologies for green house gases like carbone dioxide, ozone, methane, and for sulphate aerosols.

Figure 5a shows the global mean 2-metre temperature for the ten start dates of the Stream 2 hindcast period. During the first years of the simulations, the model develops a global mean cold bias of approximately 1° C. The spatial structure of the bias averaged over the forecast years 2–5 and estimated from all available start dates is displayed in Figure 5b. It can be seen that while the tropical and subtropical oceans undergo a strong cooling, the system builds up a substantial warm bias over the northern hemisphere extra-tropical continents.

In decadal forecasting, the forecast signals are often much smaller than the biases we currently have in our system. At this stage it is not clear how these relatively large biases can be accounted for. The approach used in seasonal forecasting, where *a posteriori* corrections to remove the bias are applied to the raw model output, relies on the assumption of a quasi-linear behaviour of the atmosphere and ocean anomalies. This is clearly not the case for our decadal forecasts. Reducing the model biases by testing the system in coupled long-term mode and continuing to improve the physics of the coupled model will have to be the ways forward in the future.

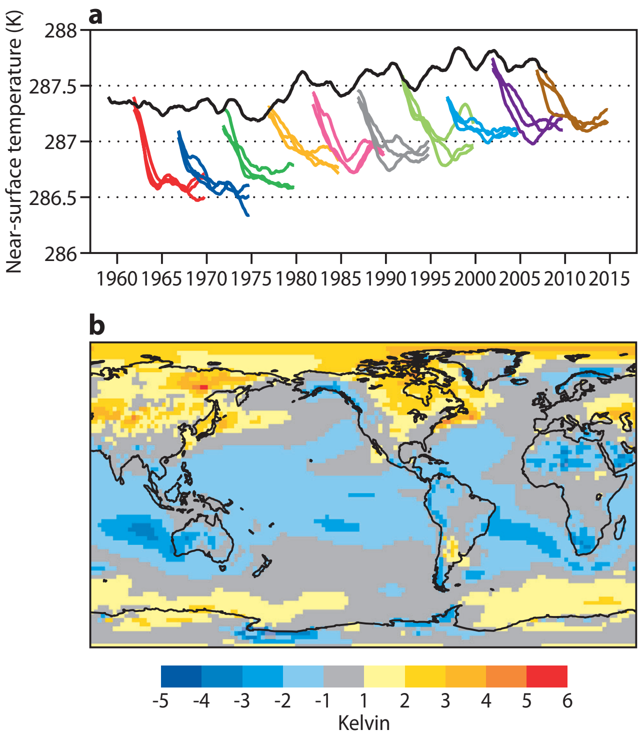


Figure 5 Decadal hindcasts in the coupled IFS/HOPE system. (a) The global mean of 2-metre temperature (filtered with a two-year running mean) in each of the ten-year long hindcast simulations with three ensemble members. The continuous black line shows the corresponding values from reanalyses (ERA-40/ERA-Interim). (b) 2-metre temperature bias (K) with respect to the reanalyses for the forecast range 2–5 years estimated from the 1960–2000 hindcasts.

Public data dissemination

A common set of hindcast data provided by all ENSEMBLES partners has been archived and is available for public download without charge for use in research, education and commercial work. Both daily and monthly data are available for the atmospheric variables. The ocean output includes monthly means of ocean analyses and forecasts. Further details can be found in *Weisheimer et al.* (2009).

The ECMWF Meteorological Archival and Retrieval System (MARS) and a system based on the Open-source Project for a Network Data Access Protocol (OPeNDAP) provide users with access to the ENSEMBLES data in the most efficient way for their specific requirements, see http://www.ecmwf.int/research/EU_projects/ENSEMBLES/data/data_dissemination.html.

The ENSEMBLES data is also available through the KNMI Climate Explorer, an interactive tool to analyze climate data.

We hope that making the data publically available will enable the international community to explore the full scientific potential of the ENSEMBLES data.

Further Reading

Doblas-Reyes, F.J., A. Weisheimer, M. Déqué, N. Keenlyside,M. McVean, J.M. Murphy, P. Rogel, D. Smith & **N. Palmer** (2009): Addressing model uncertainty in seasonal and annual dynamical seasonal forecasts. *Q. J. R. Meteorol. Soc.*, **135**, 1538–1559.

Ferro, C.A.T., D.S. Richardson & **P. Weigel** (2008): On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorol. Appl.*, **15**, 19–24.

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

Palmer, T.N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G.J. Shutts, M. Steinheimer & **A. Weisheimer** (2009): Stochastic parametrization and model uncertainty. *ECMWF Tech. Memo. No. 598*

Weisheimer, A., F.J. Doblas-Reyes, T.N. Palmer, A. Alessandri, A. Arribas, M. Deque, N. Keenlyside, M. MacVean, A. Navarra & **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.

doi:10.21957/5hayrg0p