DEMETER and the application of seasonal forecasts

Renate Hagedorn, Francisco Doblas-Reyes, Tim Palmer

ECMWF, Shinfield Park, Reading
RG2 9AX, United Kingdom
hagedorn@ecmwf.int

ABSTRACT

A multi-model ensemble-based system for seasonal-to-interannual prediction has been developed in a joint European project known as DEMETER (Development of a European Multi-Model Ensemble System for Seasonal to Interannual Prediction). The DEMETER system comprises seven global atmosphere-ocean coupled models, each running from an ensemble of initial conditions. Comprehensive hindcast evaluation demonstrates the enhanced reliability and skill of the multi-model ensemble over a more conventional single- model ensemble approach. In addition, innovative examples of the application of seasonal ensemble forecasts in malaria and crop forecasting processes are discussed. The strategy followed in DEMETER deals with important problems as communication across disciplines, downscaling of climate simulations, and use of probabilistic forecast information. This illustrates the economic value of seasonal-to-interannual prediction for society as a whole.

1 Introduction

Seasonal-timescale climate predictions are now made routinely at a number of operational meteorological centres around the world, using comprehensive coupled models of the atmosphere, oceans, and land surface (Alves et al., 2002; Kanamitsu et al., 2002; Mason et al., 1999; Stockdale et al., 1998). They are clearly of value to a wide cross section of society, for personal, commercial and humanitarian reasons (Hartmann et al., 2002b; Thomson et al., 2000). However, the successful transition from research activity to full operational practice has led some potential users of seasonal forecasts to have unrealistic expectations of what is practicable (”My daughter is getting married in six months time - should I order a marquee for the wedding reception, or will it be dry that day?”). Notwithstanding predictable signals arising from atmosphere-ocean coupling, the overlying atmosphere is intrinsically chaotic, implying that predicted day-to-day evolution of weather is necessarily sensitive to initial conditions (Palmer, 1993; Shukla, 1998). In practice, the impact of such initial-condition sensitivity can be determined by integrating forward in time ensembles of forecasts of coupled ocean atmosphere models; the individual members of the forecast differing by small perturbations to the starting conditions of the atmosphere and underlying oceans. The phase-space dispersion of the ensemble gives a quantifiable flow-dependent measure of the underlying predictability of the flow.

However, if uncertainties in initial conditions are the only perturbations represented in a seasonal-forecast ensemble, then the resulting measures of predictability will not be reliable; the reason being that the model equations are also uncertain. More specifically, although the equations for the evolution of climate are well understood at the level of partial differential equations, their representation as a finite-dimensional set of ordinary differential equations, for integrating on a digital computer, inevitably introduces inaccuracy.

At present, there is no underlying theoretical formalism from which a probability distribution of model uncertainty can be estimated – as such a more pragmatic approach must be sought. One approach relies on the fact that global climate models have been developed somewhat independently at different climate institutes. An ensemble comprising such quasi-independent models is referred to as a multi-model ensemble.

In order to advance the concept of multi-model ensemble prediction and to explore the utility of such a forecast sytem for potential end-user, the DEMETER project (Development of a European Multi-Model Ensemble System for Seasonal to Interannual Prediction) was conceived, and successfully funded under the European Union Vth Framework Environment Programme. A description of the DEMETER coupled models, the DEMETER hindcast integrations, the archival structure, and the common diagnostics package used to evalulate the hindcasts, is described in section 2. Some meteorological and oceanographic results, comparing these single and multi-model ensemble hindcasts are described in section 3. As mentioned at the beginning of this paper, there is considerable interest amongst a wide cross section of society, for seasonal climate forecast information. However, as also mentioned, some users will be disappointed in what can realistically be achieved, whilst others may find great economic value in the predictions. How can one distinguish viable applications from unrealistic applications? Whilst it is easy to dismiss as unrealistic the potential customer who wants to know whether it will rain in the afternoon six months from today, is the demand of an agronomist who wants to use seasonal predictions to predict crop yield six months ahead, and whose crop models require daily weather parameters as input, also unrealistic?

A general methodology for assessing the value of ensemble forecasts for such users was discussed in Richardson (2000). In particular, if these users have quantitative application models requiring forecast weather information as input (Hartmann et al., 2002a), these models can be directly linked to the output of individual members of the forecast ensemble. The net result is a probability forecast, not of weather as such, but of a variable directly relevant to the user. Hence, in the case of the agronomist, the ensemble will produce a probability distribution of crop yield. The potential usefulness of the ensemble forecasts can then be judged by asking whether the forecast probability distributions of crop yield are sufficiently different from climatological probability distributions, for the agronomist to be able to make decisions or recommendations e.g. on the types of crop to plant. In the DEMETER project, there are applications partners both in agronomy and also in tropical disease prediction. Some of the results of these end-users in DEMETER are described in section 4. As a result of DEMETER, real-time multi-model ensemble seasonal predictions are now routinely made at the European Centre for Medium- Range Weather Forecasts (ECMWF). This development, and other plans that derive from DEMETER, are outlined in the concluding section of this paper.

2 The DEMETER system

2.1 Coupled models and initialization procedures

The DEMETER system comprises 7 global coupled ocean-atmosphere models. A brief summary[1] of the different coupled models used in DEMETER is given in Table 1.

For each model, except that of the Max Planck Institute (MPI), uncertainties in the initial state are represented through an ensemble of nine different ocean initial conditions. This is achieved by creating three different ocean analyses; a control ocean analysis is forced with momentum, heat and mass flux data from the ECMWF 40-year Re-Analysis[2] (ERA-40 henceforth), and two perturbed ocean analyses are created by adding daily wind stress perturbations to the ERA-40 momentum fluxes. The wind stress perturbations are randomly taken from a set of monthly differences between two quasi-independent analyses. In addition, in order to represent the uncertainty in SSTs, four SST perturbations are added and subtracted at the start of the hindcasts. As in the case of the wind perturbations, the SST perturbations are based on differences between two quasi-independent

[1] Detailed information on the models and the initialization procedures can be found on the DEMETER web site: http://www.ecmwf.int/research/demeter/general/docmodel/index.html.

[2] ERA-40 intends to produce a global analysis of variables for the atmosphere, land and ocean surface for the period 1958-2001. More information is available in http://www.ecmwf.int/research/era.

Partner	atmospheric component			ocean component			
	model	resolution		model	resolution		
		horizontal	vertical		longitudinal	latitudinal	vertical
CERFACS	ARPEGE	T63	19 L	OPA 8.2	2.0°	2.0°	31 L
ECMWF	IFS	T95	40 L	HOPE-E	1.4°	0.3° - 1.4°	29 L
INGV	ECHAM-4	T42	19 L	OPA 8.1	1.4°	0.3° - 1.4°	29 L
LODYC	IFS	T95	40 L	OPA 8.2	2.0°	2.0°	31 L
M*é*t*é*o-France	Arpege	T63	31 L	OPA 8.0	182 GP	152 GP	31 L
Met Office	HadCM3	2.5° x 3.75°	19 L	HadCM3	1.25°	0.3° - 1.25°	40 L
MPI	ECHAM-5	T42	19 L	MPI-OM1	2.5°	0.5° - 2.5°	23 L

Table 1: Horizontal and vertical resolution of the atmospheric and ocean components of the seven individual global coupled models forming the DEMETER multi-model system.

SST analyses. Atmospheric and land-surface initial conditions are taken directly from the ERA-40 re-analyses. A separate ensemble initialization procedure is used for the MPI model.

2.2 Definition of hindcast experiments

The performance of the DEMETER system has been evaluated from a comprehensive set of hindcasts over a substantial part of the ERA-40 period. Only hindcasts for the period 1987 to 1999 will be discussed in this paper.

In order to assess seasonal dependence on skill, the DEMETER hindcasts have been started from 1st February, 1st May, 1st August, and 1st November initial conditions. Each hindcast has been integrated for 6 months and comprises an ensemble of 9 members. In its simplest form, the multi-model ensemble is formed by merging the ensemble hindcasts of the seven models, thus comprising 7x9 ensemble members. To enable a fast and efficient post-processing and analysis of this complex data set, much attention was given to the definition of a common archiving strategy for all models; the ECMWF's Meteorological Archival and Retrieval System (MARS) was used for this purpose. A subset of atmosphere and ocean variables, both daily data and monthly means have been stored into MARS. Special attention was given to the time-consuming task of ensuring that all model output complies with agreed data formats and units.

A significant part of the DEMETER data set (monthly averages of a large subset of surface and upper-air fields) is freely available for research purposes through an online data retrieval system installed at ECMWF[3].

2.3 Diagnostics and evaluation tools

The need to provide a common verification methodology has been recognized by the World Meteorological Organization Commission for Basic Systems (WMO-CBS), and an internationally accepted standardized verification system (SVS) is being prepared. A comprehensive verification system to evaluate all DEMETER single models as well as the multi-model DEMETER ensemble system has been set up at ECMWF. It is run periodically to monitor hindcast production, to check correct archiving and to calculate a common set of diagnostics.

[3] Model hindcasts can be retrieved in GRIB and NetCDF formats from http://www.ecmwf.int/research/demeter/data. A tool to display the fields is also available.

The DEMETER verification system is designed with a modular structure so as to easily incorporate new evaluation tools provided by project partners or other sources. The basic set of diagnostics is summarized as follows:

- Global maps and zonal averages of the single-model bias are shown relative to a model climatology. Hindcast anomalies are computed by removing the model climatology for each grid point, each initial month, and each lead time from the original ensemble hindcasts. A similar process is used to produce the verification anomalies.
- Time series of specific climate indices, e.g. related to area averaged SSTs, precipitation and circulation patterns are displayed.
- Standard deterministic ensemble mean scores, such as anomaly correlation coefficient (ACC), root mean square skill score (RMSSS), and mean square skill score (MSSS) are shown.
- Probabilistic skill measures: reliability diagrams, relative operating characteristic (ROC) score, Brier score, ranked probability skill score (RPSS), and potential economic value curves are calculated and displayed. Significance tests are applied to most of the skill measures.
- The skill of single-model ensembles is compared with that of multi-model ensembles using scatter diagrams of area-averaged skill measures and probability density functions (PDFs) of grid-point skill scores.

Both anomalies and scores have been computed using a cross-validation ”leave-one-out” method. To generate the anomaly or the score for a particular time t, only data at other times different from t have been used.

The main verification data set used in this system is ERA-40. This is consistent with the general concept of producing the DEMETER hindcasts, in which ERA-40 is used as forcing for the ocean analyses and as atmospheric and land-surface initial conditions. Effectively, it is assumed that we are ”living in the ERA-40 world”. However, because of the modularity of the validation system, it is possible in principle to validate the model data with more than one verification data set. This is particularly useful in the case of precipitation.

3 Hindcast skill assessment

A sample of results from the DEMETER standard verification system is presented in this section. To view a more comprehensive set of verification diagnostics the reader is referred to the DEMETER website[4].

The scientific basis for seasonal atmospheric prediction relies on the premise that the lower boundary forcing, in particular SST, can impart significant predictability on atmospheric development (Palmer and Anderson, 1994). Thus, one of the pre-requisites for successful seasonal forecasts is the ability to represent and predict accurately the state of the ocean. A basic problem, faced when attempting to predict SST with coupled models, is the bias in the model forecasts, which may be comparable to the magnitude of the interannual anomalies to be predicted. Since SSTs in the tropical Pacific are a major source of predictability in the atmosphere on seasonal timescales, model performance in the tropical Pacific is of particular interest. To demonstrate the typical level of skill in this area, Table 2 shows the anomaly correlation coefficient (ACC) of the ensemble mean for the single-model ensembles and the multi-model ensemble for the SSTs averaged over the Niño-3.4 area. The correlation has been computed for the 1-month and 3-month lead hindcasts starting in February, May, August, and November. Results suggest that the single-model ensembles generally perform well as ”ENSO prediction”

[4] http://www.ecmwf.int/research/demeter/verification/index.html

Model	1-month lead		3-month lead	
	Bias / K	ACC	Bias / K	ACC
DEMETER multi-model	–	0.95	–	0.91
CERFACS	-0.38	0.93	0.00	0.89
ECMWF	-0.36	0.96	-0.70	0.90
INGV	-0.05	0.90	-0.03	0.84
LODYC	-1.03	0.94	-1.53	0.89
Météo-France	-0.10	0.93	0.38	0.90
Met Office	-0.48	0.91	-0.52	0.89
MPI	-1.99	0.85	-3.27	0.72
Persistence	–	0.80	–	0.56

Table 2: Ensemble-mean bias and anomaly acorrelation coefficient (ACC) for the 1-month and 3-month lead seasonal average of sea surface temperature over the Niño 3.4 area calculated using all start dates for the years 1987-1999. Note that the bias for the multi-model ensemble and the persistence hindcast are not defined since the multi-model ensemble is based on single-model anomalies, which are constructed with regard to the single-model bias, and persistence uses observed anomalies.

systems. For the sake of comparison, the ACC for a persisted-SST hindcast has been included. This hindcast is made by persisting initial SST anomaly for the six months corresponding to the coupled model integration. Both, the multi-model ensemble and the single-models perform better than persistence, especially in the 3-month lead time range. In addition, note the high correlation of the multi-model ensemble for both lead times, proving it to be the most skilful system for the 3-month lead hindcasts. The coupled model climate may differ from the observed climatology as a result of model ocean-atmosphere interactions. The bias of the single-models is generally in the range of ±1 K (Table 2). These are typical figures for present leading coupled models. As is the case for most variables and areas, there appears to be no clear relationship between bias and anomaly forecast skill, though this is a topic that needs further investigation.

Figure 1 shows 1987-1999 time series of ACC of precipitation for all single-models and the multi-model ensemble, for summer (JJA, May start date) over the tropics (Fig. 1a) and winter (DJF, November start date) over the northern extra-tropics (Fig. 1b). The skill in the northern extra-tropics is considerably less than in the tropics. In both regions the variability in prediction skill, both from year to year and between different models is clearly evident. Evidence of higher skill during ENSO events is provided by relatively large ACC for 1988 and 1997 (Fig. 1). The skill in the northern extra-tropics is considerably less than in the tropics. In both regions the variability in prediction skill, both from year to year and between different models is clearly evident. Evidence of higher skill during ENSO events is provided by relatively large ACC for 1988 and 1997 (Fig. 1). This is consistent with the link between ENSO activity and seasonal predictability found in many studies (e.g. Brankovic and Palmer (2000)). In general, the identity of the most skilful single model varies with region and year. Finally, this figure illustrates the relatively skilful performance of the multi-model ensemble. In most years the multi-model ensemble skill is close to the best single-model skill and is the most skilful when performance is averaged over all years. This highlights the greater reliability of the multi-model ensemble system.

To further summarize atmospheric hindcast skill, Figure 2 shows indices of the winter (DJF, November start date) Pacific North American (PNA) and North Atlantic Oscillation (NAO) patterns for the multi-model ensemble. The indices are computed by projecting every ensemble member anomaly onto a pre-defined pattern. To compute the reference patterns, an empirical orthogonal function (EOF) analysis of the 500-hPa geopotential height has been performed for the winter monthly mean anomalies using NCEP re-analyses for the period

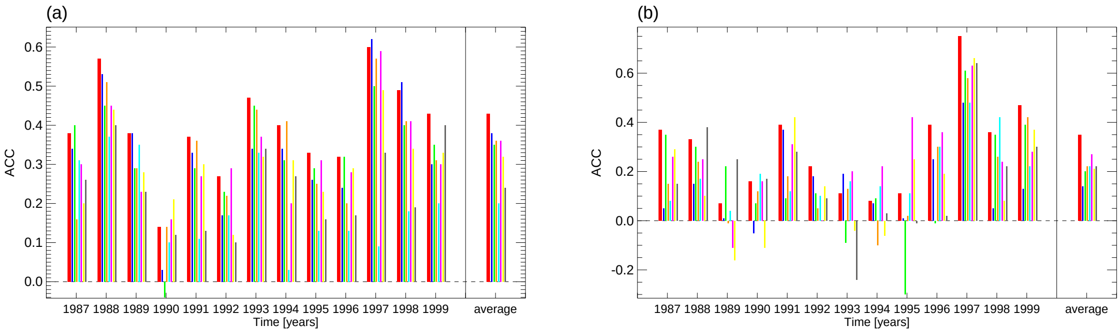

Figure 1: Time series of the ensemble mean anomaly correlation coefficients for the multi model (thick red bars) and all individual models (thin bars; ECMWF: blue, Météo-France: orange, Met Office: green, MPI: cyan, LODYC: pink, CERFACS: grey, INGV: yellow). a: 1-month lead summer (JJA) precipitation in the tropics (latitudinal band of 30°S - 30°N), b: 1-month lead winter (DJF) precipitation in the northern extra-tropics (latitudinal band of 30°N - 90°N). Additionally, the time average over the whole period 1987-1999 is shown at the end of each plot.

1949-2000. The EOF analysis was carried out using data over the regions 20°N - 87.5°N and 110°E-90°W for the PNA and 20°N - 87.5°N and 90°W -60°E for the NAO, and the leading EOF retained. The spatial covariance between the monthly anomaly patterns was then computed for every single member of the hindcast ensemble and the reference pattern was computed. The monthly covariances were averaged to produce seasonal means. Figure 2 displays the index against time using a box-and-whisker representation in which the central box and each whisker contain one third of the ensemble members. The value obtained computing the spatial covariance between the reference pattern and the ERA-40 anomalies is also displayed. Comparison of the interannual variations of ERA-40 and ensemble-mean values gives a visual impression of ensemble mean hindcast skill. The verification lies within the ensemble range in every case for the PNA index and in all but one case for the NAO index. Table 3 shows the correlation between the two time series for the multi-model and the single-model ensembles. The multi-model ensemble shows the highest correlation of all the models for the NAO index and one of the highest for the PNA index. In addition, the multi-model ensemble correlation can be considered non-zero with a 95% confidence level using a t-test, which is not always the case for the single-model ensembles. It should be taken into account that scores based on indices are less robust than scores based on large area correlations, when calculated with short time series. The high PNA correlation for some single models may be explained by the exceptionally good predictions for the 90s, in particular 1994 and 1997. In fact, scores for hindcasts carried out over periods longer than 13 years suggest that scores are lower and more in agreement with the area- averaged ACC over the Pacific-North American region (between 0.2 and 0.5).

Note that, while PNA index hindcast skill tends to be quite satisfactory (Fig. 2a), NAO index skill is lower but always positive. Figure 2b indicates that the multi-model ensemble can produce a useful signal in years when the observed NAO index is large in magnitude, such as 1987, 1988 and 1997. These years may in themselves account for the high correlation coefficient obtained in Table 3. Nevertheless, the model signal in some years is weak (little shift of the predicted index away from zero) and is effectively contrary to observations in some other years when the observed index was large in magnitude (1992 and 1995).

Considerable effort has been devoted to the validation of the ensembles as probability forecasts. The dashed blue and red lines in Figure 2 correspond to the ensemble and ERA40 tercile boundaries. The corresponding

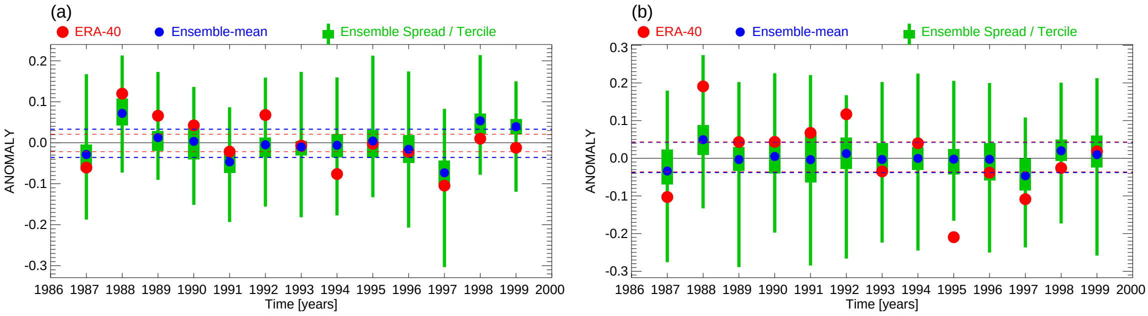

Figure 2: Time series of the 1-month lead winter (DJF) PNA (a) and NAO (b) index. The multi-model ensemble spread is depicted by the box-and-whisker representation with the whiskers containing the lower and upper tercile of the ensemble. The blue dots represent the ensemble mean, the ERA-40 anomalies being displayed by slightly bigger red bullets. The horizontal lines around the solid zero line mark the tercile boundaries of the ERA-40 (red) and hindcast data (blue). Note that in the right panel (NAO) the model and ERA-40 tercile boundaries are coincident.

probabilistic skill measure used is the ranked probabilistic skill score (RPSS) based on these tercile categories. Hindcast performance is summarized in Table 3. RPSS is defined so that positive values imply higher skill than climatology forecasts and perfect forecasts have a skill score of 100. The skill of the multi-model ensemble for the PNA index is close to the skill of the best models and statistically significant at the 95% confidence level, in good agreement with the correlation results. The situation is slightly different for the NAO index. RPSS values are high and tend to be statistically significant, which was not the case for the correlation. RPSS statistical confidence has been assessed by computing the distribution of the skill score from a random set of hindcasts obtained from scrambling the available hindcasts and verifications.

In order to get a more comprehensive assessment of single-model versus multi-model ensemble skill, a wide range of results has been collected (Fig. 3) for different cases. The value of the area under the ROC curve is a probabilistic skill measure ranging from 0 to 1. Values below 0.5 imply lower skill than climatology, whilst a perfect forecast has a ROC score of 1. The comparison of all ROC scores for 2-metre temperature, calculated over different regions, start dates, lead times and events shows that, although in some cases the single-models have a higher ROC score than the multi-model ensemble, in the vast majority of cases (90%) the ROC score of the multi-model ensemble exceeds the score of the single models. Furthermore, the number of cases with less skill than climatology is greatly reduced for the multi-model ensemble; for the latter there are no cases with ROC score smaller than 0.5 compared to 36 cases for the single models. The greater probabilistic skill of the multi-model ensemble compared to the single-model skill leads to an increased potential economic value (Richardson, 2000). For instance, it has been found that, for predictions of positive tropical winter (DJF, November start date) precipitation anomalies, the multi- model ensemble improves the potential economic value from 15% to 80%, depending on the single model taken as reference (not shown).

In spite of the clear improvement of the multi-model ensemble performance an important question arises. This improvement could be due to either to the multi-model approach itself or to the increased ensemble size resulting from collecting all members of the single-model ensembles, or both. In order to separate the multi-model approach benefits that derive from combining models of different formulation to those derived simply from the accompanying increase in ensemble size, a 54-member ensemble has been generated for a single start date (boreal summer hindcasts) using the ECMWF model alone. The ensemble was generated using additional

Model	Correlation		RPSS	
	PNA	NAO	PNA	NAO
DEMETER multi-model	**0.75**	**0.68**	**19.3**	**23.3**
CERFACS	0.19	0.48	-1.8	**15.3**
ECMWF	**0.80**	0.43	**21.8**	**21.3**
INGV	**0.57**	0.37	-4.6	**7.4**
LODYC	**0.75**	0.49	**27.3**	**18.5**
Météo-France	0.20	0.44	-26.9	**17.1**
Met Office	**0.62**	0.31	0.9	8.8
MPI	**0.68**	0.23	**37.5**	7.9

Table 3: Ensemble-mean correlation and ranked probability skill score for the Pacific North American (PNA) and North Atlantic Oscillation (NAO) indices calculated from the 1-month lead hindcasts started in November (DJF) seasonal average) for the years 1987-1999. Statistically significant values (95% confidence level) are printed in bold letters.

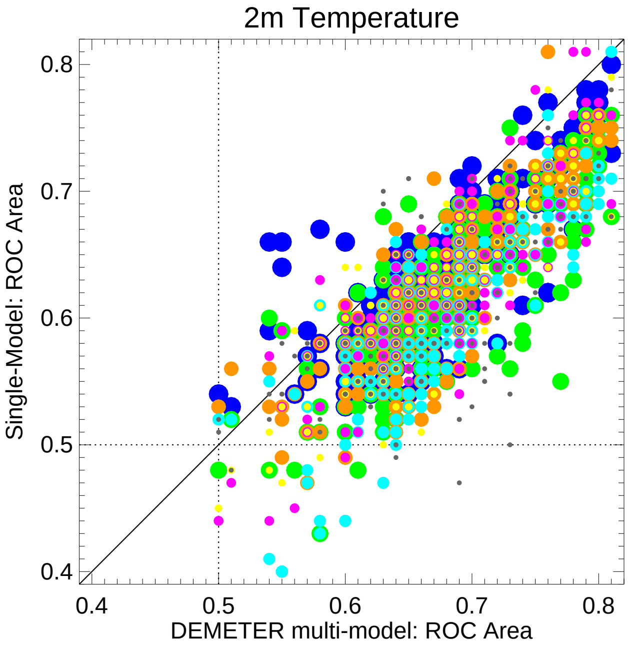

Figure 3: Scatter plot of single-model (ECMWF: blue, Met Office: green, Météo-France: orange, Max-Planck-Institute: cyan, LODYC: pink, CERFACS: yellow, INGV: grey; different size of bullets for better visibility of all models) versus multi-model ROC scores of the 2-metre temperature hindcasts from 1987 - 1999. The plot comprises results from seasonal hindcast scores for 8 different areas (Northern extra-tropics, tropics, southern extra-tropics, north America, Europe, west Africa, east Africa, south Africa), 4 start dates (Feb, May, Aug, Nov), 2 lead times (1 month, 3 month), and 4 events (anomaly above/below 0.43 standard deviation, anomaly above/below 0).

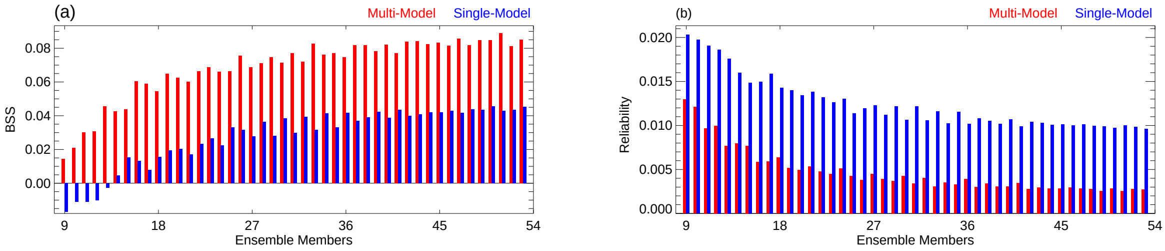

Figure 4: Brier skill score (a) and reliability component of Brier score (b) for the 1-month lead tropical summer (JJA) precipitation 1987 - 1999 for the single ECMWF-control model (blue) and the DEMETER multi-model (red). The event is ”precipitation anomalies above zero”. Results are shown for different ensemble sizes from 9 to 54 members. Note that lower values of the reliability term mean better reliability.

wind and SST perturbations in order to have a better sampling of the initial condition uncertainty. As skill measure, the Brier skill score for tropical summer (JJA) precipitation positive anomalies for the multi-model ensemble (red bars) and the ECMWF model (blue bars) is shown in Figure 4. For each ensemble size, the corresponding ensemble was constructed by randomly selecting the number of members from the 63 available in the multi-model ensemble and the 54 in the single-model ensemble. It turns out that the skill grows faster with ensemble size for ensembles with less than about 30 members, though this threshold changes with region, variable, and event considered. The skill seems to saturate for bigger ensembles, though a slight increase is still found. The figure also demonstrates that the multi-model ensemble probabilistic hindcasts are more skilful than the single-model hindcasts, regardless of the ensemble size. Similar results are found for other variables and regions. Based on a decomposition of the Brier score (Murphy, 1973), results show that the largest contribution to the multi-model ensemble skill improvement is due to increased reliability (smaller values of the reliability term in the Murphy decomposition imply greater reliability of the ensemble), as shown in Figure 4b. This indicates that the multi-model ensemble provides a better sampling of the phase space so that the multi-model ensemble contains the verification more often than the single model used in this test.

4 Applications

One of the main objectives of DEMETER is a demonstration of the utility of seasonal climate forecasts through the coupling of quantitative application models, such as crop yield models, to the global climate prediction models. Western European agriculture is highly intensive and weather is a principal source of uncertainty for crop yield assessment and for crop management (Vossen, 1995). As such, seasonal weather forecasts have high potential value for European agriculture.

The Crop Growth Monitoring System of the European Commission Joint Research Centre (EC-JRC) uses a crop model called WOFOST (WOrld FOod STudies), and performs crop yield forecasting through a regression analysis comparing simulated crop indicators and historical yield series for the main crops at national / European level (van Diepen and van der Wal, 1995). To estimate the yield at the end of the season, the regression analysis module computes the best predictor equation from 2 sets of parameters: (i) the technological time trend, (ii) the simulated crops indicators. However, in the current system, at the time when a crop yield forecast is issued, the weather conditions leading up to harvest time are unknown and are therefore a major source of uncertainty. The provision of seasonal weather forecast would in principal bring additional information for the remaining crop season. Also at the local level, the Regional Meteorological Service of the Emilia-Romagna environmental agency (ARPA-SMR, Italy) also uses WOFOST as part of a geographical soil water flow and

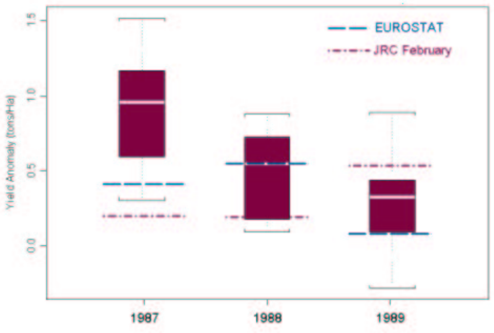

Figure 5: Distribution (25% quartile) of wheat yield anomaly predictions from the multi-model ensemble downscaled data for Spain. The blue dashed line correspond to the reference value (Eurostat), the red dash dotted line to the operational system at JRC with the prediction issued at the end of February, and the white solid line inside the central box to the ensemble median.

transport simulation system called CRITERIA (Marletto et al., 2001).

Based on the system described above, an innovative method to supply seasonal forecast information to crop simulation models has been developed in DEMETER. It consists of running the crop model on each individual member of the ensemble to derive a probability distribution of crop yield. Based on the spread of the probability distribution, the end-user can directly quantify the benefits and risks of specific weather-sensitive decisions.

The potential importance of seasonal predictions for crop yield estimation was demonstrated by forcing the crop model with ERA15 re-analysis used as a ”perfect forecast” and comparing with the current operational practice in crop yield forecasting at EC-JRC. Results showed (i) that simulated crop indicators had a higher contribution in the yield estimation, as they were selected as the best predictor (compared to the technological time trend factor) in a greater number of cases than the current operational system, and (ii) that yield estimates were closer to the reference crop yield (based on the Eurostat[5] value) in nearly twice as many cases.

The multi-model ensemble seasonal hindcasts are performed at global scale but at low spatial resolution for the crop model requirements. Therefore, even correctly predicted, large-scale weather systems are not yielding a reliable representation of local weather conditions. This leads to the necessity of downscaling the output from the prediction system to represent more accurately the weather at the local scale. For example, in terms of the ”perfect forecast” experiment described above, the downscaled precipitation over the Iberian Peninsula improved substantially the simulated plant biomass compared to the result obtained using the raw model output. For example, a 10% reduction in root mean square error and an increase of regression coefficient from 0.63 to 0.71 was observed.

Wheat yield hindcasts were carried out over three years (1987-1989) using DEMETER multi-model ensemble downscaled data. For 1988 Figure 5 shows that the median of the ensemble (white solid line) precisely coincides with the official Eurostat wheat yield (dashed blue line) and the ensemble dispersion is relatively low, implying high confidence in the prediction. For 1989, the median of the ensemble is closer to Eurostat official figure than the hindcast made using the JRC operational method (red dash-dotted line). Only for 1987, the median of the ensemble was farther than the traditional forecast, although the official Eurostat figure was within the range of the ensemble forecast. In addition, the positive sign of the yield anomaly was well predicted in all three cases.

[5] Eurostat: Statistical Office of the European Union. Eurostat value being the reference for yield comparison (official yield value).

5 Summary

As part of a European-Union funded DEMETER project, a multi-model ensemble system based on seven European global coupled ocean-atmosphere models has been described and validated in hindcast mode using ECMWF ERA-40 reanalysis data. Output from the DEMETER system, suitably downscaled, has been applied to crop yield and malaria prediction models. Results indicate that the multi-model ensemble is a viable pragmatic approach to the problem of representing model uncertainty in seasonal-to-interannual prediction, and will lead to a more reliable forecasting system than that based on any one single model.

In the limited space available in this paper, a few illustrative examples of results from the DEMETER project have been given. However, we invite readers to visit the DEMETER web site[6] where an extensive range of diagnostics and skill scores used to evaluate the DEMETER system are presented.

In addition to these specific diagnostics and skill scores, visitors to the DEMETER web site can download (in GRIB or NetCDF format) gridded data from a large data set comprising monthly mean fields for a large number of variables from the DEMETER hindcasts, including ERA-40 verification. We thus encourage scientists and potential users of seasonal forecasts to perform their own analysis of the DEMETER data (perhaps to assess skill for specific regions and variables of interest not covered in our standard analysis). More generally, we offer this DEMETER data set for education training purposes, both in the developed and developing world.

As a result of the success of DEMETER, real-time multi-model forecasting is now being established as part of the operational seasonal forecast suite at ECMWF. At the time of writing, plans are well established for the ECMWF, Met Office and M*été*o-France coupled systems to be included in this multi-model mix. It is possible that other models may be included at a later stage.

In future research it is hoped to use a successor system to DEMETER to explore the use of multi-model ensembles not only for seasonal-to-interannual timescales, but also for decadal timescales for which scientific evidence of predictability has emerged in recent years. For this puropose it is planned to ensure that the model components used for seasonal-to-decadal ensemble prediction, are, as far as practicable, identical to those used for century-timescale anthropogenic climate change. In this way, the reliability of century-timescale climate change projections can be assessed by running essentially the same ensemble systems on timescales for which verification data exists. We believe that a unification and rationalization of research and development across these timescales will enhance enormously the credibility of our science.

Acknowledgments

The DEMETER project has been funded by the European Union under the contract EVK2-CT-1999-00024. The authors would like to thank M. Alonso-Balmaseda, D. Anderson, L. Ferranti, M. Fuentes, C. Gibert, D. Lucas, T. Stockdale, J. Vialard, F. Vitart, for their invaluable help and support. The project would not have been possible without the technical support of the ECMWF staff and consultants.

References

Alves, Oscar; Guomin Wang; Aihong Zhong; Neville Smith; Graham Warren; Andrew Marshall; F. Tzeitkin and Andreas Schiller; 2002: POAMA: Bureau of Meteorology operational coupled model seasonal forecast system. In *Seminar Proceedings: The Role of the Upper Ocean in Medium and Extended Range Forecasting*, ECMWF.

[6] http://www.ecmwf.int/research/demeter/verification

Brankovic, Cedo and Timothy N. Palmer; 2000: Seasonal skill and predictability of ECMWF PROVOST ensembles. *Q. J. R. Meteorol. Soc.*, **126**, 2035–2068.

Hartmann, Holly C.; Roger Bales and Soroosh Sorooshian; 2002a: Weather, climate, and hydrologic forecasting for the US Southwest: a survey. *Climate Research*, **21**, 239–258.

Hartmann, Holly C.; Thomas C. Pagano; Soroosh Sorooshian and Roger Bales; 2002b: Confidence builders: Evaluating seasonal climate forecasts for user perspectives. *Bulletin of the American Meteorological Society*, **83**, 683–698.

Kanamitsu, Masao; Arun Kumar; Hann-Ming Henry Juang; Jae-Kyung Schemm; Wanqui Wang; Fanglin Yang; Song-You Hong; Peitao Peng; Wilber Chen; Shrinivas Moorthi and Ming Ji; 2002: NCEP dynamical seasonal forecast system 2000. *Bulletin of the American Meteorological Society*, **83**, 1019–1037.

Marletto, Vittorio; Luca Criscuolo and Margot Van Soetendael; 2001: Implementation of WOFOST in the framework of the CRITERIA geographical tool. In *Proceedings of the Second International Symposium on Modeling Cropping Systems*, pp. 219–220, European Society for Agronomy, Florence, Italy.

Mason, Simon J.; Lisa Goddard; Nicholas E. Graham; Elena Yulaeva; Liqiang Sun and Phillip A. Arkin; 1999: The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bulletin of the American Meteorological Society*, **80**, 1853–1873.

Murphy, Allan H.; 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

Palmer, Timothy N.; 1993: Extended-range atmospheric prediction and the Lorenz model. *Bulletin of the American Meteorological Society*, **74**, 49–65.

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

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

Shukla, Jagadish; 1998: Predictability in the midst of chaos: A scientific basis for climate forecasting. *Science*, **282**, 728–731.

Stockdale, Tim N.; David L. T. Anderson; J. Oscar S. Alves and Magdalena A. Balmaseda; 1998: Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. *Nature*, **392**, 370–373.

Thomson, Madeleine C.; Timothy N. Palmer; Andy P. Morse; Mark Cresswell and Steve J. Connor; 2000: Forecasting disease risk with seasonal climate predictions. *Lancet*, **355**, 1559–1560.

van Diepen, Cees A. and Tamme van der Wal; 1995: Crop growth monitoring and yield forecasting at regional and national scale. Agrometeorological models: Theory and applications in the MARS Project. Office for Official Publications of the European Communities, Luxembourg. Report No. EUR 16008 EN, 143-157.

Vossen, Paul; 1995: Early crop production assessment of the European Union, the system implemented by the MARS-STAT Project. Agrometeorological models: Theory and applications in the MARS Project. Office for Official Publications of the European Communities, Luxembourg. Report No. EUR 16008 EN, 21-51.