Project no. GOCE-CT-2003-505539

Project acronym: ENSEMBLES

Project title: ENSEMBLE-based Predictions of Climate Changes and their
Impacts

Instrument: Integrated Project

Thematic Priority: Global Change and Ecosystems

Deliverable Reference Number and Title

**D5.7: Assessment of the skill of seasonal NAO and PNA using multi-
model seasonal integrations from DEMETER**

Due date of deliverable: 28 February 2006

Actual submission date: 28 February 2006

Start date of project: 1 September 2004   Duration: 60 Months

Organisation name of lead contractor for this deliverable: ECMWF

Revision [draft, 1, 2, ..]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the Consortium (including the Commission Services)	

1

ENSEMBLES D5.7: "Assessment of the skill of seasonal NAO and PNA using multi-model seasonal integrations from DEMETER"

European Centre for Medium-Range Weather Forecasts, Reading, UK

1. Introduction

Atmospheric extra-tropical low-frequency variability has long been studied in terms of standing, time fluctuating, large-scale structures known as teleconnection patterns (Wallace and Guztler, 1981; Barnston and Livezey, 1987). These patterns have had a large impact on the understanding of the variability on interannual and longer time scales because they play an important role in the global climate system (Thompson and Wallace, 2001; Franzke et al., 2004). The fundamental mechanisms determining the evolution of many teleconnections have been widely analysed (e.g., Panagiotopoulos et al., 2002; Quadrelli and Wallace, 2004). However, their predictability has been the subject of a significantly lower number of studies (Pavan and Doblas-Reyes, 2000; Doblas-Reyes et al., 2003; Müller et al., 2005a).

The North Atlantic Oscillation (NAO) and the Pacific North American (PNA) teleconnection patterns are the two most important modes of variability in the Northern Hemisphere extratropical atmosphere. The two patterns account for increasingly more of the variability as the averaging period is increased (Feldstein, 2000), and are thus more important at the monthly and seasonal time scales than for daily analyses. Their importance is furthermore most pronounced during winter, when the overall variability is highest. Therefore, the focus of this note will be for this season.

The NAO and PNA are commonly considered as atmospheric phenomena at the seasonal time scale, although the interaction with the ocean is supposed to play a role too (Bretherton and Battisti, 2000). In the extra-tropics, on time scales shorter than a decade, the atmosphere tends to force the ocean and the oceanic influence back to the atmosphere is rather weak (Kushnir et al., 2002). On longer time scales (not considered here) there are indications of coupled atmosphere-ocean variability, where the time scale is set by the longer time scales of ocean dynamics (Czaja et al., 2003). There are also remote fast influences from the tropical oceans that are communicated via the atmosphere. This is especially true for the PNA which, due to the proximity to the tropical Pacific, is being influenced by the El Niño-Southern Oscillation (ENSO) (Robertson et al., 2000).

This note deals with the representation of the patterns in a set of coupled models used for seasonal forecasting, so that all sorts of interactions are allowed. The forecast quality of these seasonal time-scale simulations is also assessed.

2. Seasonal climate prediction

Seasonal time scale dynamical 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 (e.g. Stockdale et al., 1998; Mason et al., 1999; Kanamitsu et al., 2002). Seasonal forecasts are clearly of value to a wide cross-section of society, for personal, commercial and humanitarian reasons (e.g. Thomson et al., 2006). Preliminary assessments also indicate that there are signs of ensemble-mean skill also in multi-annual time scales (Smith et al., 2006), which is partly due to the impact of the increase of greenhouse gases in the atmosphere. Recent results point out that the effects of anthropogenic climate forcing needs to be considered in both seasonal and multi-annual forecast systems (Doblas-Reyes et al., 2006).

In spite of the fact that predictable signals can arise from atmosphere-land-ocean interaction, the overlying atmosphere is intrinsically chaotic. This implies that predicted day-to-day evolution of weather is necessarily sensitive to initial conditions (Palmer, 1993). In practice, the impact of such sensitivity can be determined by integrating forward in time ensembles of forecasts of a model, the individual members of the ensemble differing by small perturbations to the starting conditions. However, if uncertainties in initial conditions are the only perturbations represented in a climate forecast ensemble, the resulting measures of predictability will not be reliable because 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 inevitably introduces inaccuracy. Inaccuracies at the smallest resolved scales can in principle propagate upscale and infect the entire spectrum of scales being predicted by the model.

At present, there is no underlying theoretical formalism from which a probability distribution of model uncertainty can be estimated (Palmer et al., 2005) and more pragmatic approaches must be sought. One such approach relies on the fact that global climate models have been developed somewhat independently at different climate institutes, using different numerical approaches to represent the climate dynamics and applying different parameterizations of physical processes. An ensemble comprising such quasi-independent models is referred to as a multi-model

ensemble. Other ways to represent model uncertainty are based on the stochastic physics (Palmer, 2001) or the perturbed-parameter approaches (Murphy et al., 2004). All these methods are being investigated in ENSEMBLES by performing co-ordinated experiments that mimic a real-time forecasting setting.

The advantages of the multi-model approach have been thoroughly investigated in the DEMETER project (Development of a European Multi-model Ensemble System for Seasonal to Interannual Prediction; Palmer et al., 2004; http://www.ecmwf.int/research/demeter/index.html). The principal aim of DEMETER was to advance the concept of multi-model ensemble seasonal prediction by installing a number of state-of-the-art global coupled ocean-atmosphere models on a single supercomputer, and to produce a series of six-month multi-model ensemble hindcasts with common archiving and common diagnostic software.

3. The DEMETER project

The DEMETER system comprises 7 global coupled ocean-atmosphere models. A brief summary of the different coupled models used in DEMETER is given in Table 1 and more details are available in Palmer et al. (2004). 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 (Figure 1). 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 (Uppala et al., 2005; 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 SST analyses. Atmospheric and land-surface initial conditions are taken directly from ERA-40. A separate ensemble initialization procedure is used for the MPI model. Ocean data assimilation has been used in the MetOffice experiment after 1987.

The performance of the DEMETER system has been evaluated from a comprehensive set of hindcasts over a substantial part of the ERA-40 period. Most of the discussion in this document will be for the period 1980 to 2001. This is the period

which all seven coupled models participating in the project have generated hindcasts for. Longer time series (up to 43 years) are available for a smaller number of models.

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. The method of unweighted multi-model ensembles has been preferred in this work because the sample length available (22 years) does not allow for the calculation of robust coefficients in a multi-model combination method (Doblas-Reyes et al., 2005).

Hindcast anomalies are computed by removing the model climatology in cross-validation 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. The main verification data set used in this system is ERA-40.

4. Forecast Quality Assessment

Time series of sea level pressure anomaly correlation coefficient (ACC) for all single-models and the multi-model ensemble, for winter (December to February, November start date) over Europe and North America are shown in Figure 2. There is a large variability in prediction skill in both regions, both from year to year and between different single models. In general, the identity of the most skilful single model varies with both the region and the year, although on average the multi-model ensemble performs better than most single models. Overall, average skill over North America is superior to the skill over Europe. Note the higher correlation for this region during some ENSO events, especially 1997.

The previous result indicates that, in addition to other sources of predictability for surface variables over these continental areas such as persistence of soil anomalies, large-scale circulation might have some skill at the seasonal time scale. To illustrate this point, Figure 3 shows indices of the winter (December to February, November start date) PNA and NAO patterns for the multi-model ensemble. The indices are computed following the method described in Doblas-Reyes et al. (2003). Values are obtained by projecting every ensemble member anomaly onto the leading empirical orthogonal function (EOF) of the 500-hPa geopotential height (computed

over the winter monthly mean anomalies using NCEP re-analyses for the period 1949-2000). The EOF analysis was carried out using data over the regions 20º-87.5ºN and 110ºE-90ºW for the PNA and 20º-87.5ºN and 90ºW-60ºE for the NAO. The spatial covariance between the monthly anomaly patterns and the reference pattern was computed for every single member of the hindcast ensemble. Monthly covariances were then averaged to produce seasonal means.

Other methods to estimate the teleconnection indices were also tested, such as the computation of EOFs separately for each single model using either monthly or seasonal mean data. All ensemble members (and not the ensemble mean) were used and data were weighted with the cosine of the latitude. The results and conclusions were very similar, in good agreement with Hurrell et al. (2003). The single-model EOF analysis allows the estimation of the teleconnection patterns for each model, which are shown in Figures 4 and 5. Although there are noticeable differences with regard to the reference pattern, all the models capture the main features of both the NAO and PNA, the differences being within the range exhibited in McHugh and Rogers (2005).

Figure 3 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 appears in red. The multi-model ensemble is constructed by pooling together the ensembles from every single model after their interannual variability has been corrected to have similar statistical properties as the ERA-40 time series. The verification lies within the multi-model ensemble range in all but two cases for both indices, which is a very simple measure of the reliability of the multi-model predictions. Skill measures indicate a higher reliability for the multi-model (not shown). Comparison of the interannual variations of ERA-40 and ensemble-mean values gives a visual impression of the possible ensemble-mean hindcast skill. In a more quantitative sense, Table 2 shows the correlation between the two time series for the multi-model and the seven single-model ensembles. Although the correlations are not high, they are similar to the skill obtained in previous independent experiments (e.g., Doblas-Reyes et al., 2003). The multi-model ensemble shows one of the highest correlations among all the models for both indices. The correlation can be considered non-zero and positive with a 95% confidence level using a two-sided t-test only in four cases (three of them for the

NAO). This is consistent with very large confidence intervals at the 95% level (Table 2).

Note that, while the PNA index hindcast skill tends to be relatively high, the NAO skill is lower but always positive. Figure 3b suggests that the multi-model ensemble simulates a skilful signal in years when the observed NAO index is large in magnitude, such as 1985, 1988 and 1997. These years may already account for the relatively high correlation coefficient in Table 2, as already suggested in Doblas-Reyes et al. (2003) and Müller et al. (2005a). Müller et al. (2005a) found a tendency to higher skill of the seasonal NAO probabilistic predictions when the model signal was larger than average. Nevertheless, the model signal in some years is very weak. For instance, notice the small shift of the predicted index away from zero in 1992 and 1995, when the observed index was large in magnitude.

The formulation and verification of probability forecasts offers a better illustration of the benefits of the multi-model. The dashed blue and red lines in Figure 3 correspond to the simulation and ERA-40 climatological terciles. Probability forecasts for the three categories defined by the two terciles have been formulated using a simple estimate of the ensemble relative frequency (Harter, 1984). The probabilistic skill measure used to assess the forecast quality is the ranked probabilistic skill score (RPSS, Epstein 1969). Hindcast performance for winter is summarized in Table 2. RPSS is defined so that positive values imply higher skill than climatology forecasts (the reference score used in this case) and perfect forecasts have a skill score of 1. The skill of the multi-model ensemble for the PNA index is close to the skill of the best models, in good agreement with the ensemble-mean correlation results, and is statistically significant at the 95% confidence level. The RPSS statistical confidence has been assessed by computing the distribution of the skill score from a random set of hindcasts obtained by scrambling the available hindcasts and verifications. However, it has to be taken into account that RPSS confidence intervals tend to be negatively skewed (see Müller et al. (2005b) for examples), which makes very low values to be statistically significant.

A single skill measure such as the RPSS aggregates on a single score many contributions to forecast quality. Forecast quality in separate categories gives a more user-oriented perspective of the forecast system quality. For this purpose, skill measures other than RPSS are needed, a good example of which is the ROC score or area under the ROC curve. This is a probabilistic skill measure for dichotomous

forecasts ranging from 0 to 1 (Swets 1988). The ROC skill score is constructed as twice the area minus one, with possible values ranging between -1 and +1. Values below 0 imply lower skill than climatology, whilst a perfect forecast has a ROC skill score of 1. Figure 6 displays the ROC skill score of the winter NAO (November start date) for three events (anomalies above the upper tercile, above the median and below the lower tercile). Although the performance of the different systems changes with the event defined, the multi-model offers a good compromise across the different categories. Similar conclusions apply to the PNA.

The level of skill described above diminishes slightly for longer lead times of the hindcasts started in November. However, forecast quality for other seasons is very low and non significant. Besides, when longer series of hindcasts are considered (up to 43 years) the skill scores take almost negligible values, which is mainly due to a lack of representation of the very low frequency variability of the teleconnection patterns (Pavan et al., 2005). More experimentation is required to determine the reasons for the lack of very low frequency extra-tropical variability in seasonal forecast experiments.

In spite of the relative improvement shown by the multi-model ensemble performance an important question arises. The improvement could be due either to the benefits of the multi-model approach itself or to the increased ensemble size resulting from collecting all members of the single-model ensembles to construct the multi-model ensemble, or to both (Hagedorn et al., 2005). In order to separate the benefit of the multi-model approach that derives from combining models of different formulation from the benefit due to the increase in ensemble size, a 54-member ensemble hindcast has been generated with the ECMWF model alone for the period 1987-1999. Not surprisingly for such a small sample of a very low signal phenomenon, the skill of both the single-model and multi-model ensembles is very similar, so that no definite conclusions can be drawn.

5. Summary and Conclusions

The ability to simulate and predict the interannual variations of the NAO and PNA with a multi-model ensemble system based on seven European global coupled ocean-atmosphere models has been described. The individual models simulate realistic NAO and PNA patterns. Both deterministic and probabilistic forecast quality measures point out that seasonal predictions of the winter indices for the last 20 years

of the last century have some positive skill, with the multi-model approach showing some hints of improvement. However, the skill scores obtained are low and it remains to be proved whether they have socio-economic value. Besides, no skill is found for a longer sample starting in 1960, which seems to be mainly a consequence of the lack of very low frequency variability.

Whilst the results shown in this paper clearly indicate the need to represent model uncertainty when forecasting climate, the improvements shown by the multi-model approach suggest that it cannot be considered the final solution. The DEMETER results have motivated more theoretical approaches to the representation of model uncertainty based on cellular automaton stochastic subgrid models (Palmer et al., 2005). The benefits of this method will be assessed against the multi-model approach within ENSEMBLES.

References

Barnston, A. G. and R. E. Livezey, 1987. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.*, **115**, 1083-1126.

Bretherton, C. S. and D. S. Battisti, 2000. An interpretation of the results from atmospheric general circulation models forced by the time history of the observed sea surface temperature distribution. *Geophys. Res. Lett.*, **27**, 767-770.

Czaja, A., A. W. Robertson and T. Huck, 2003. The role of Atlantic Ocean-Atmosphere coupling in affecting North Atlantic variability. *The North Atlantic Oscillation: Climate Significance and Environmental Impact*. J. W. Hurrell et al., Eds., Amer. Geophys. Union. 147-172.

Doblas-Reyes, F. J., V. Pavan and D. B. Stephenson, 2003. The skill of multi-model seasonal forecasts of the North Atlantic Oscillation. *Climate Dyn.*, **21**, 501-514.

Doblas-Reyes, F.J., R. Hagedorn and T.N. Palmer, 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part II: Calibration and combination. *Tellus A*, **57**, 234-252.

Doblas-Reyes, F. J., R. Hagedorn, T. N. Palmer and J.-J. Morcrette, 2006. Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. *Geophys. Res. Let.*, in press.

Epstein, E. S., 1969. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985-987.

Feldstein, S. B., 2000. The timescale, power spectra and climate noise properties of teleconnection patterns. *J. Climate*, **13**, 4430.4440.

Franzke C., S. Lee and S. B. Feldstein, 2004. Is the North Atlantic Oscillation a breaking wave? *J. Atmos. Sci.*, **61**, 145-160.

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

Harter, H. L., 1984. Another look at plotting positions. *Comm. Stat.-Theory and Methods*, **13**, 1613-1633.

Hurrell, J. W., Y. Kushnir, G. Ottersen and M. Visbeck, 2003. An overview of the North Atlantic Oscillation. *The North Atlantic Oscillation: Climate Significance and Environmental Impact*. J. W. Hurrell et al., Eds., Amer. Geophys. Union. 1-35.

Kanamitsu, M., A. Kumar, H.-M Juang, J.-K. Schemm, W. Wang, F. Yang, S.-Y. Hong, P. Peng, W. Chen, S. Moorthi and M. Ji, 2002. NCEP dynamical seasonal forecast system 2000. *Bull. Amer. Meteor. Soc.*, **83**, 1019-1037.

Kushnir, Y., W. A. Robinson, I. Bladé, N. M. J. Hall, S. Peng and R. T. Sutton, 2002. Atmospheric GCM response to extratropical SST anomalies: Synthesis and evaluation. *J. Climate*, **15**, 2233-2256.

Mason, S. J., L. Goddard, N. E. Graham, E. Yulaeva, L. Sun and P. A. Arkin, 1999. The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bull. Amer. Meteor. Soc.*, **80**, 1853-1873.

McHugh, M. A. and J. C. Rrogers, 2005. Multi-model representation of the North Atlantic Oscillation in the 20th and 21st centuries. *Geophys. Res. Lett.*, **32**, doi:10.1029/2005GL023679.

Müller, W. A., C. Appenzeller and C. Schär, 2005a. Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature. *Climate Dyn.*, **24**, 213-226.

Müller, W. A., C. Appenzeller, F. J. Doblas-Reyes and M. Liniger, 2005b. 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, D. A. Stainforth, 2004. Quantifying uncertainties in climate change from a large ensemble of general circulation model predictions. *Nature*, **430**, 768-772.

Pavan, V., S. Marchesi, A. Morgillo, C. Cacciamani and F. J. Doblas-Reyes, 2005. Downscaling of DEMETER winter seasonal hindcasts over Northern Italy. *Tellus A*, **57**, 424-434.

Quadrelli, R. and J. M. Wallace, 2004. A simplified linear framework for interpreting patterns of Northern Hemisphere wintertime climate variability. *J. Climate*, **17**, 3728-3744.

Palmer, T. N., 1993. Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.*, **74**, 49-65.

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

Palmer, T. N., G. J. Shutts, R. Hagedorn, F. J. Doblas-Reyes, T. Jung and M. Leutbecher, 2005. Representing model uncertainty in weather and climate prediction. *Annual Review of Earth and Planetary Sciences*, **33**, 163-193.

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.

Panagiotopoulos F., M. Shahgedanova and D. B. Stephenson, 2002. A review of Northern Hemisphere winter-time teleconnection patterns. *J. Phys. IV France*, **12**, 27-47, doi:10.1051/jp4:20020450.

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

Robertson, A. W., C. R. Mechoso and Y. J. Kim, 2000. The influence of Atlantic sea surface temperature anomalies on the North Atlantic Oscillation. *J. Climate*, **13**, 122-138.

Smith, D. M., A. W. Colman, S. Cusack, C. K. Folland, S. Ineson and J. M. Murphy, 2006. Predicting surface temperature for the coming decade using a global climate model. *Nature*, submitted.

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

Swets, J. A., 1988. Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-1293.

Thomson, M. C., F. J. Doblas-Reyes, S. J. Mason, R. Hagedorn, S. J. Connor, T. Phindela, A. P. Morse and T. N. Palmer, 2006. Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature*, **439**, 576-579.

Thompson D. W. J. and J. M. Wallace, 2001. Regional climate impacts of the Northern Hemisphere annular mode. *Science*, **293**, 85-89.

Uppala, S. M, P. W. Kållberg, A. J. Simmons, U. Andrae, V. Da Costa Bechtold, M. Fiorino, J. K. Gibson, J. Haseler, A. Hernandez, G. A. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R. P. Allan, E. Andersson, K. Arpe, M. A. Balmaseda, A. C. M. Beljaars, L. Van De Berg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Hólm, B. J. Hoskins, L. Isaksen, P. A. E. M. Janssen, R. Jenne, A. P. McNally, J.-F. Mahfouf, J.-J. Morcrette, N. A. Rayner, R. W. Saunders, P. Simon, A. Sterl, K. E. Trenberth, A. Untch, D. Vasiljevic, P. Viterbo and J. Woollen, 2005. The ERA-40 re-analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961-3012.

Wallace, J. M. and D. S. Gutzler, 1981. Teleconnections in the geopotential height field

during the Northern Hemisphere winter. *Mon. Wea. Rev.*, **109**, 784-812.
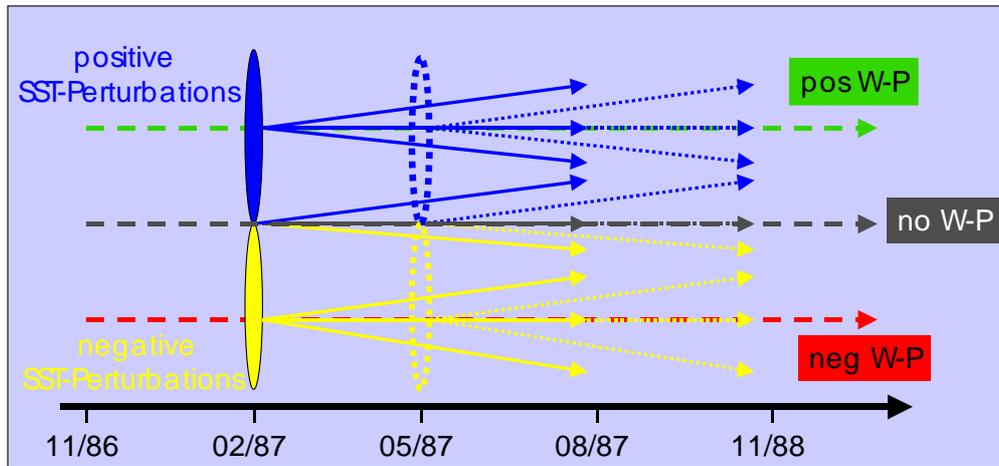
Figures

Figure 1: Schematic representation of the ensemble generation and hindcast production strategy. Dashed lines represent the three continuous runs of ocean analyses forced by ERA-40 data, the control analysis without any windstress perturbations (grey) and two additional analyses with positive/negative (green/red) daily windstress perturbations applied. In order to generate 9 different initial conditions for the coupled hindcasts, four SST-perturbations (represented by the ellipses) are added (blue ellipse) and subtracted (yellow ellipse) to the ocean analyses. Thus, there is one member with no windstress or SST-perturbations applied on and 8 perturbed ensemble members. This procedure is performed every three months at every start date of the hindcasts.
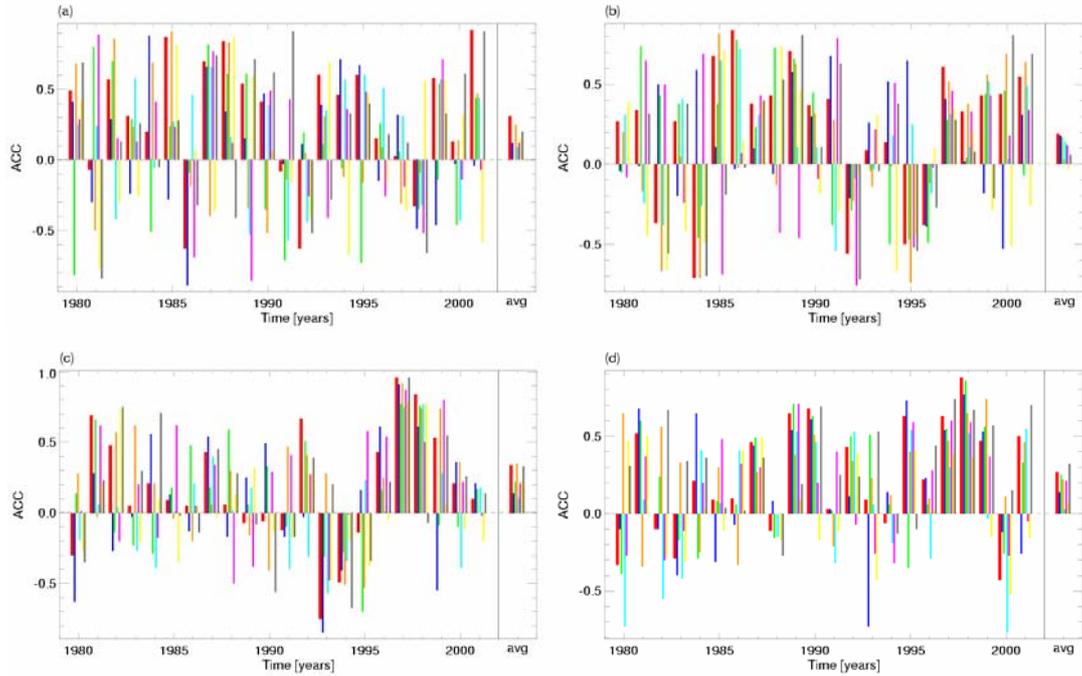
Figure 2: Time series of the 1-month lead winter (DJF, November start date) ensemble-mean sea level pressure and 2-metre temperature anomaly correlation coefficients for the multi model (thick red bars) and all individual models (thin bars; ECMWF: blue, Met Office: green, Météo-France: orange, MPI: cyan, LODYC: pink, INGV: yellow, CERFACS: grey). (a) European (35ºN-75ºN, 12.5ºW-42.5ºE) sea level pressure and (b) 2-metre temperature, (c) North American (30ºN-70ºN, 130ºW-60ºW) sea level pressure and (d) 2-metre temperature. Additionally, the average over the whole period 1980-2001 is shown at the end of each plot.
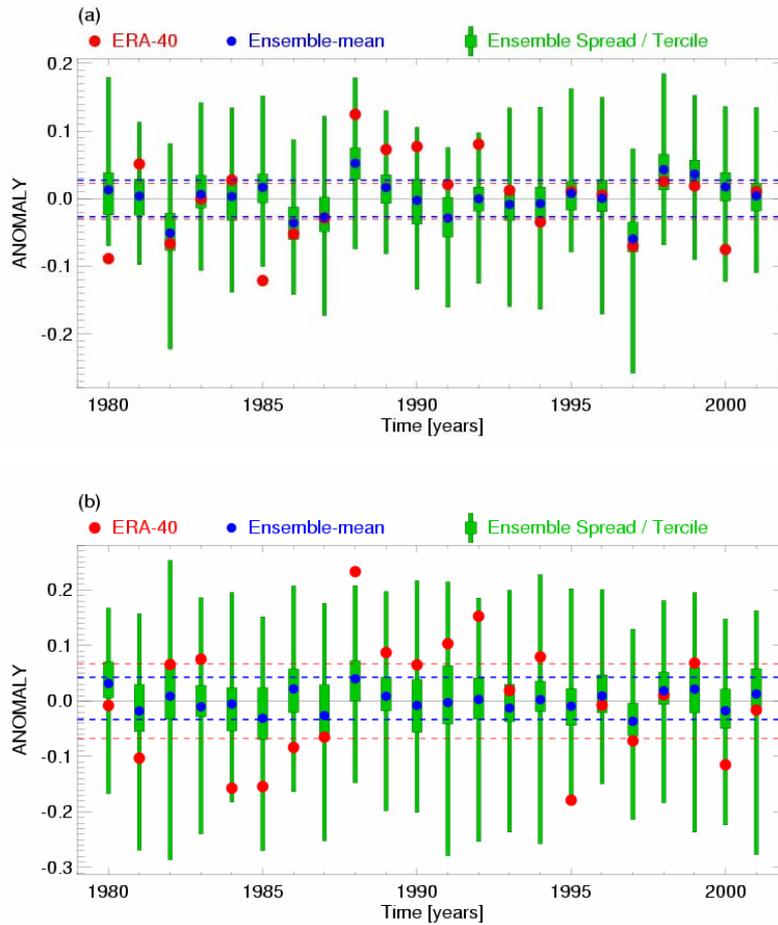
Figure 3: Time series of the 1-month lead winter (DJF, November start date) PNA (a) and NAO (b) index for the period 1980-2001. The multi-model ensemble spread is depicted by the box-and-whisker representation with the whiskers containing the lower and upper tercile of the ensemble. The blue dots represent the ensemble mean, the ERA-40 anomalies being displayed by slightly bigger red bullets. The horizontal lines around the solid zero line mark the terciles of the ERA-40 (red) and hindcast data (blue).
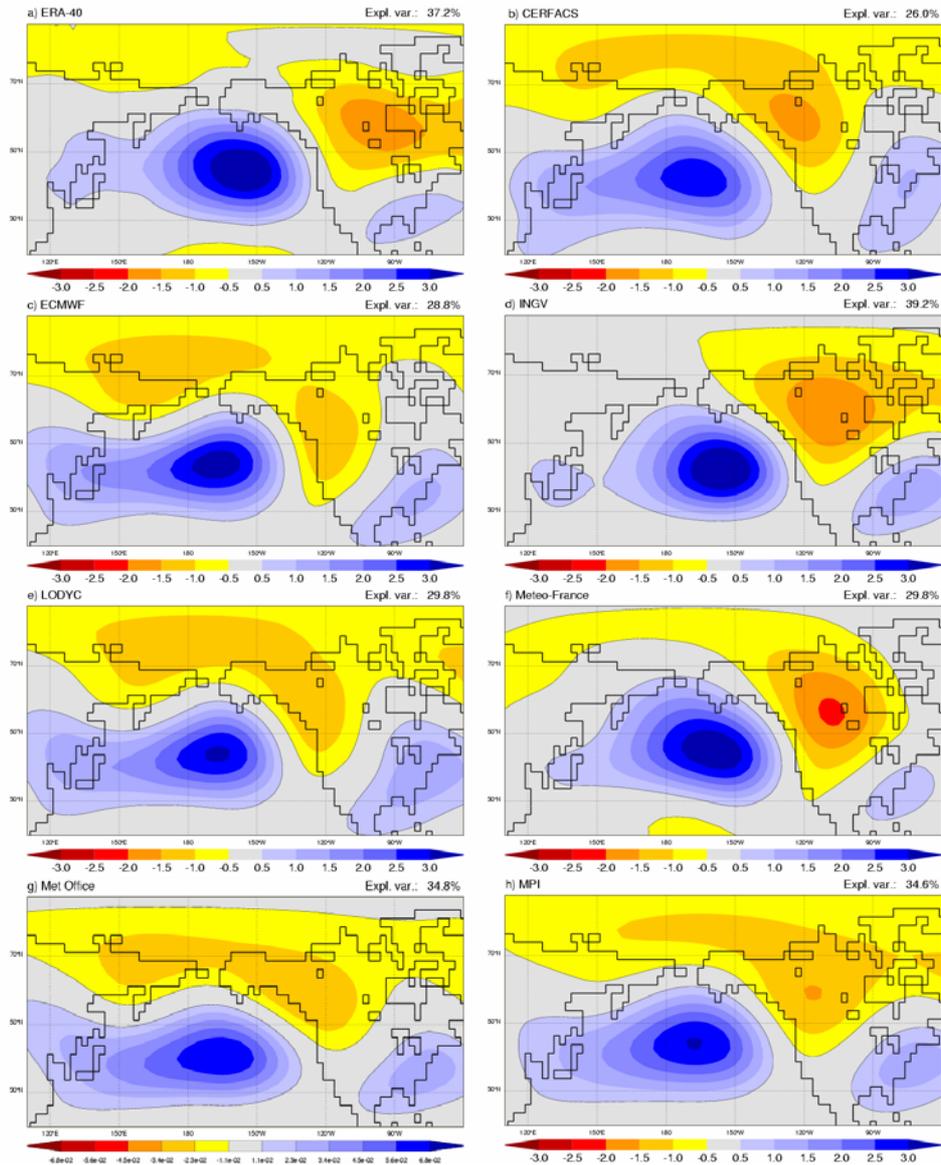
Figure 4: Leading empirical orthogonal function of the winter 500 hPa geopotential height over the Pacific North America region for a) ERA-40, b) CERFACS, c) ECMWF, d) INGV, e) LODYC, f) Météo-France, g) Met Office, and h) MPI. Monthly data for December, January and February (November start date for the hindcasts) were used. The percentage of explained variance is displayed in the top right corner of every panel.
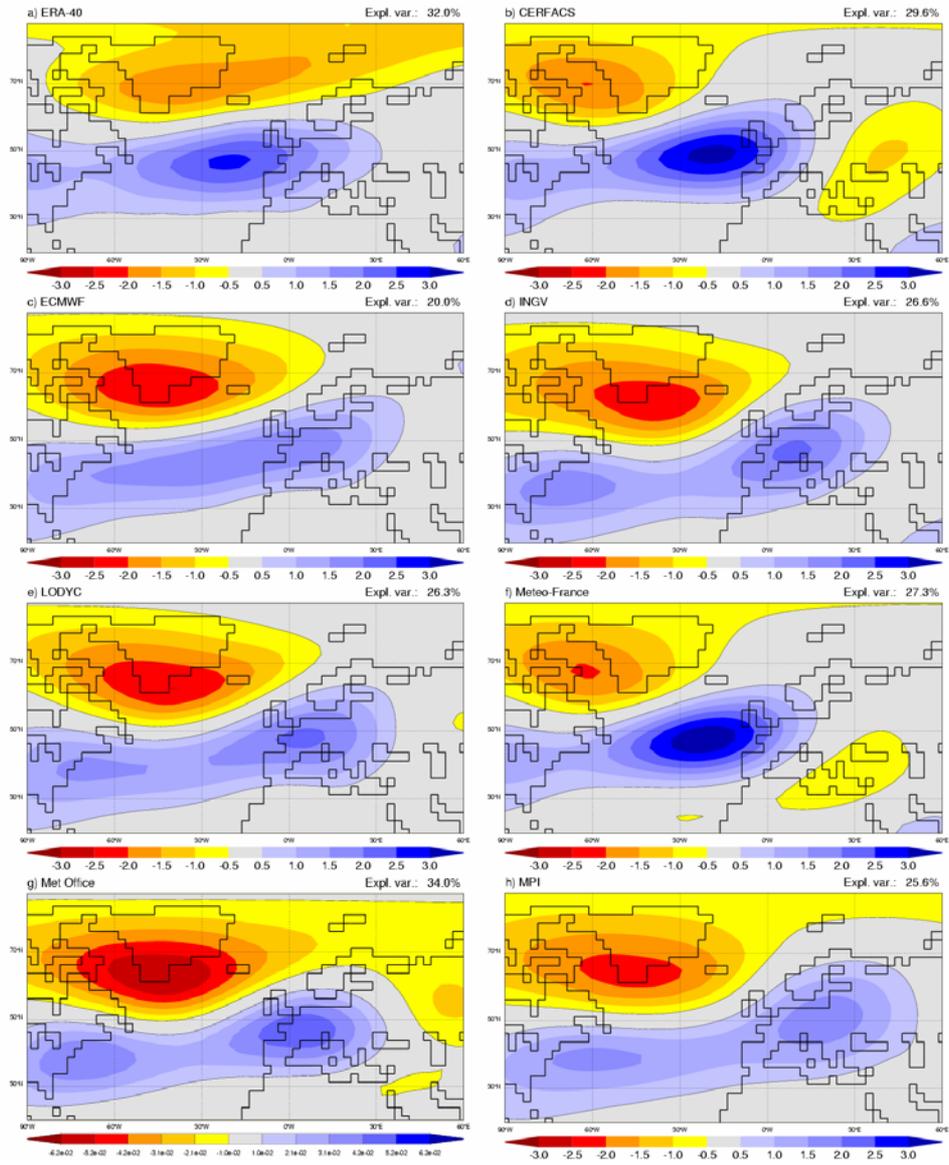
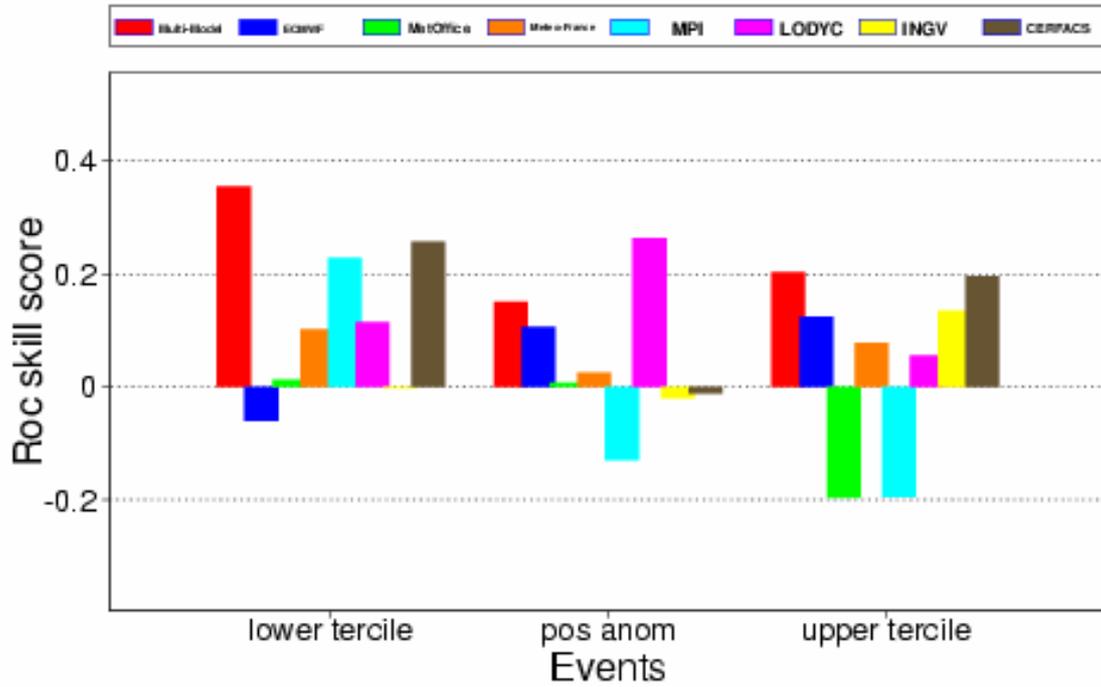**Figure 5**: As Figure 4, but for the North Atlantic Europe region.

Figure 6: ROC skill score of the 1-month lead winter (DJF, November start date) NAO index for the period 1980-2001 for the multi model (thick red bars) and all individual models (thin bars; ECMWF: blue, Met Office: green, Météo-France: orange, MPI: cyan, LODYC: pink, INGV: yellow, CERFACS: grey).

Tables

	CERFACS	ECMWF	INGV	LODYC	Météo-France	Met Office	MPI
atmosphere component	ARPEGE	IFS	ECHAM-4	IFS	ARPEGE	HadAM3	ECHAM-5
resolution	T63 31 Levels	T95 40 Levels	T42 19 Levels	T95 40 Levels	T63 31 Levels	2.5º x 3.75º 19 Levels	T42 19 Levels
atmosphere initial conditions	ERA-40	ERA-40	coupled AMIP-type experiment	ERA-40	ERA-40	ERA-40	coupled run relaxed to observed SSTs
ocean component	OPA 8.2	HOPE-E	OPA 8.1	OPA 8.2	OPA 8.0	GloSea OGCM, based on HadCM3	MPI-OM1
resolution	2.0º x 2.0º 31 Levels	1.4º x 0.3º-1.4º 29 Levels	2.0º x 0.5º-1.5º 31 Levels	2.0º x 2.0º 31 Levels	182 GP x 152 GP 31 Levels	1.25º x 0.3º-1.25º 40 Levels	2.5º x 0.5º-2.5º 23 Levels
ocean initial conditions	ocean analyses forced by ERA-40	ocean analyses forced by ERA-40	ocean analyses forced by ERA-40	ocean analyses forced by ERA-40	ocean analyses forced by ERA-40	ocean analyses forced by ERA-40	coupled run relaxed to observed SSTs
ensemble generation	windstress and SST perturbations	windstress and SST perturbations	windstress and SST perturbations	windstress and SST perturbations	windstress and SST perturbations	windstress and SST perturbations	9 different atmospheric conditions from the coupled initialization run (lagged method)

Table 1: Combinations of atmosphere and ocean models used by the seven partners contributing with coupled models to DEMETER. The resolution of the models and the initialization strategy is outlined as well. The modeling partners are: CERFACS (European Centre for Research and Advanced Training in Scientific Computation, France), ECMWF (European Centre for Medium-Range Weather Forecasts, International Organization), INGV (Istituto Nazionale de Geofisica e Vulcanologia, Italy), LODYC (Laboratoire d'Océanographie Dynamique et de Climatologie, France), Météo-France (Centre National de Recherches Météorologiques, Météo-France, France), Met Office (The Met Office, UK) and MPI (Max-Planck Institut für Meteorologie, Germany).

	Multi-model	CERFACS	ECMWF	INGV	LODYC	Météo-France	Met Office	MPI
Correlation (PNA)	0.41 (0.94)	0.16 (0.52)	0.39 (0.93)	0.23 (0.70)	0.46 (0.97)	0.23 (0.70)	0.31 (0.84)	0.32 (0.85)
	(-0.01,0.71)	-0.28,0.54	(-0.04,0.70)	(-0.21,0.59)	(0.05,0.74)	(-0.21,0.59)	(-0.13,0.65)	(-0.12,0.65)
Correlation (NAO)	0.54 (0.99)	0.30 (0.83)	0.10 (0.34)	0.20 (0.63)	0.43 (0.95)	0.55 (0.99)	0.18 (0.58)	0.14 (0.47)
	(0.15,0.78)	(-0.14,0.64)	(-0.34,0.50)	(-0.24,0.57)	(0.01,0.72)	(0.17,0.79)	(-0.26,0.56)	(-0.30,0.53)
RPSS (PNA)	0.18 (1.00)	-0.19 (0.00)	0.24 (1.00)	-0.02 (0.00)	0.10 (0.99)	-0.11 (0.00)	0.22 (1.00)	0.10 (1.00)
RPSS (NAO)	0.10 (1.00)	0.16 (0.99)	-0.12 (0.00)	0.04 (0.90)	0.00 (0.00)	0.01 (0.89)	-0.12 (0.00)	0.02 (0.91)

Table 2: Ensemble-mean correlation and ranked probability skill score for tercile categories of the Pacific North American (PNA) and North Atlantic Oscillation (NAO) indices calculated from the 1-month lead hindcasts started in November (DJF seasonal average) over the period 1980-2001. Confidence levels for rejecting the hypothesis of the skill score being zero are indicated in brackets. The second row in the correlation cells displays the confidence intervals for a 0.95 probability computed from the z-transform of the correlation, with the intervals given by the expression tanh(z±1.96/sqrt(n-3)), where n is the sample size. The model acronyms are defined in Table 1