DEVELOPMENT OF A EUROPEAN MULTIMODEL ENSEMBLE SYSTEM FOR SEASONAL-TO-INTERANNUAL PREDICTION (DEMETER)

BY T. N. PALMER, 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, M. C. THOMSON

Ensembles of leading European global coupled climate models show impressive reliability for seasonal climate prediction—including useful output for probabilistic prediction of malaria incidence and crop yield.

Seasonal time-scale climate predictions are now made routinely at a number of operational meteorological centers 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; Alves et al. 2002). This development can be traced back to a revolution in our understanding of the coupled ocean–atmosphere system in the second half of the twentieth century (Neelin et al. 1998), to the development and deployment of specialized buoys to observe and measure the evolution of near-surface waters in the

AFFILIATIONS: PALMER, DOBLAS-REYES, AND HAGEDORN—European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom; ALESSANDRI AND GUALDI—Istituto Nazionale de Geofisica e Vulcanologia, Bologna, Italy; ANDERSEN AND FEDDERSEN—Danmarks Meteorologiske Institut, Copenhagen, Denmark; CANTELAUBE AND TERRES—Land Management Unit, Institute for Environment and Sustainability, Joint Research Centre, Ispra, Italy; DAVEY AND GRAHAM—Met Office, Bracknell, United Kingdom; DÉLÉCLUSE AND LAZAR—Laboratoire d'Océanographie Dynamique et de Climatologie, Paris, France; DÉQUÉ AND GUÉRÉMY—Centre National de Recherches Météorologiques, Météo-France, Toulouse, France; DÍEZ AND ORFILA—Instituto Nacionale de Meteorología, Madrid, Spain; HOSHEN AND MORSE—Department of Geography, University of Liverpool, Liverpool, United Kingdom; KEENLYSIDE AND LATIF—Max-Planck Institut für Meteorologie, Hamburg, Germany; MAISONNAVE AND ROGEL—European Centre for Research and Advanced Training in Scientific Computation, Toulouse, France; MARLETTO—Agenzia Regionale Prevenzione Ambiente dell'Emilia Romagna, Bologna, Italy; THOMSON—International Research Institute for Climate Prediction (IRI), The Earth Institute at Columbia University, Columbia University, New York, New York
CORRESPONDING AUTHOR: Dr. T. N. Palmer, ECMWF, Shinfield Park, RG2 9AX, Reading, United Kingdom
E-mail: t.palmer@ecmwf.int
DOI: 10.1175/BAMS-85-6-853

tropical Pacific (McPhaden et al. 1998), and, most importantly, to successful predictions of El Niño by prototype coupled ocean–atmosphere models (Zebiak and Cane 1987). In turn, these developments were fostered by the outstandingly successful international Tropical Ocean Global Atmosphere (TOGA) program (World Climate Research Programme 1985).

Seasonal forecasts are clearly of value to a wide cross section of society, for personal, commercial, and humanitarian reasons (e.g., Stern and Easterling 1999; Thomson et al. 2000; Pielke and Carbone 2002; Hartmann et al. 2002a; Murnane et al. 2002). However, 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 sensitivity can be determined by integrating forward-in-time ensembles of forecasts of a coupled ocean–atmosphere model, the individual members of the ensemble 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. Such inaccuracies 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 (see Palmer 2001); as such, a more pragmatic approach must be sought. One such 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 multimodel ensemble. The ability of multimodel ensembles to produce more reliable probability forecasts of seasonal climate risk over single-model ensembles has been addressed by the Prediction of Climate Variations on Seasonal to Interannual Timescales (PROVOST) project funded by the European Union IVth Framework Environment Programme; a similar "sister" project, Dynamical Seasonal Prediction (DSP), was undertaken in the United States (Palmer and Shukla 2000, and references therein).

As part of the PROVOST project, several atmospheric general circulation models were integrated over 4-month time scales with prescribed observed sea surface temperatures (SSTs). Each model was itself run in ensemble mode, based on nine different initial conditions from each start date; results were stored in a common archive in identical format. One of the key results from PROVOST and DSP is that, despite identical SSTs, ensembles showed considerable model-to-model variability in the estimates both of the SST-forced seasonal-mean signal, and the seasonal-mean "noise" generated by internal dynamics (Straus and Shukla 2000; Pavan and Doblas-Reyes 2000). As a result, single-model ensemble estimates of the response to observed SSTs were generally not reliable. Within the PROVOST project, both single-model and multimodel ensembles were treated as potential forecasts (assuming, in some sense, oracular knowledge of the ocean), and scored using probability forecast skill scores (Doblas-Reyes et al. 2000; Graham et al. 2000; Palmer et al. 2000). A key result was that probability scores based on the full multimodel ensemble were generally higher than those from any of the single-model ensembles.

Based on such results, the Development of a European Multimodel Ensemble System for Seasonal to Interannual Prediction project (DEMETER) was conceived, and funded under the European Union Vth Framework Environment Programme. The principal aim of DEMETER was to advance the concept of multimodel ensemble 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 6-month multimodel ensemble hindcasts with common archiving and common diagnostic software. Such a strategy posed substantial technical problems, as well as more mundane but nevertheless important issues (e.g., agreeing on units in which model variables were archived).

A description of the DEMETER coupled models, the DEMETER hindcast integrations, the archival structure, and the common diagnostics package used to evaluate the hindcasts, is given in section 2. Some meteorological and oceanographic results, comparing these single and multimodel ensemble hindcasts, are described in section 3.

The DEMETER project has applications partners in agronomy and in tropical disease prediction. These

applications are described in section 4. A general methodology for assessing the value of ensemble forecasts for such users was discussed in Palmer (2002). In particular, if users have quantitative application models requiring forecast weather information as input (Hartmann et al. 2002b), 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 or climate, but of a variable directly relevant to the user, for example, in the case of the agronomist, a forecast probability distribution of crop yield. The potential usefulness of the DEMETER system can then be judged by asking whether the forecast probability distributions of crop yield are sufficiently different from climatological probability distributions, and sufficiently reliable for the agronomist to be able to make decisions or recommendations, for example, on the types of crop to plant. As such, the design of DEMETER was based on the concept of an "end to end" system (Buizer et al. 2000; Pielke and Carbone 2002), in which users feed information back to the forecast producers. (To obtain more information on seasonal-to-interannual prediction applications, visit the WMO–CLIPS program Web site at www.wmo.ch/web/wcp/clips2001/html/index.html.)

Quantitative application models of the sort used in DEMETER have been derived using data from specific meteorological stations. By contrast, the output from global climate models represents averages over a relatively coarse grid. As such, the statistics of model variables, especially precipitation in regions of steep orography, can differ substantially from the statistics of station data. It is therefore necessary to downscale the climate model output (Goddard et al. 2001), either by some statistical/empirical scheme, or by embedding a high-resolution limited-area model into the climate model. Both approaches have been followed in DEMETER, and are described in section 4 before discussing the application results.

Section 5 describes some experiments intending to analyze the sensitivity of the results to different factors involved in ensemble seasonal prediction with coupled models.

As a result of DEMETER, real-time multimodel 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 section 6.

THE DEMETER SYSTEM. *Coupled models and initialization procedures.* The DEMETER system com-

prises seven global coupled ocean–atmosphere models. A brief summary (detailed information on the models and the initialization procedures can be found on the DEMETER Web site www.ecmwf.int/research/demeter/general/docmodel/index.html) of the different coupled models used in DEMETER is given in Table 1. From 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 (Fig. 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-yr Reanalysis (ERA-40 henceforth; 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 online at www.ecmwf.int/research/era), 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 sepa-
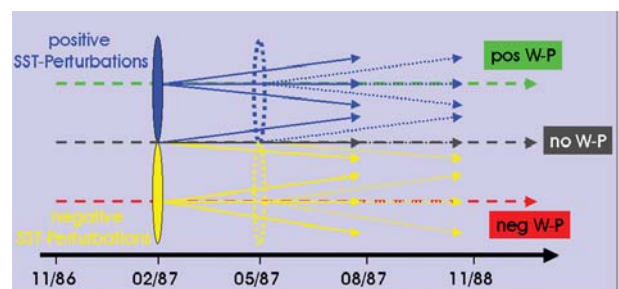

FIG. 1. Schematic representation of the ensemble generation and hindcast production strategy. Dashed lines represent the three continuous runs of ocean analyses forced by ERA-40 data, the control analysis without any wind stress perturbations (gray) and two additional analyses with positive/negative (green/red) daily wind stress perturbations applied. In order to generate nine 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 wind stress or SST perturbations applied and eight perturbed ensemble members. This procedure is performed every 3 months at every start date of the hindcasts.

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 are 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 (United Kingdom), and MPI (Max-Planck Institut für Meteorologie, Germany).**

	CERFACS	ECMWF	INVG	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	2.5 × 3.75 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
Reference	Déqué (2001)	Gregory et al. (2000)	Roeckner (1996)	Gregory et al. (2000)	Déqué (2001)	Pope et al. (2000)	Roeckner (1996)
Ocean component	OPA 8.2	HOPE-E	OPA 8.1	OPA 8.2	OPA 8.0	GloSea OGCM, based on HadCM3	MPI-OMI
Resolution	2.0° × 2.0° 31 levels	1.4° × 0.3°–1.4° 29 levels	2.0° × 0.5°–1.5° 31 levels	2.0° × 2.0° 31 levels	182 GP × 152 GP 31 levels	1.25° × 0.3°–1.25° 40 levels	2.5° × 0.5°–2.5° 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
Reference	Delecluse and Madec (1999)	Wolff et al. (1997)	Madec et al. (1998)	Delecluse and Madec (1999)	Madec et al. (1997)	Gordon et al. (2000)	Marsland et al. (2003)
Ensemble generation	Wind stress and SST perturbations	Wind stress and SST perturbations	Wind stress and SST perturbations	Wind stress and SST perturbations	Wind stress and SST perturbations	Wind stress and SST perturbations	Nine different atmospheric conditions from the coupled initialization run (lagged method)

rate ensemble initialization procedure is used for the MPI model. Ocean data assimilation has been used in the Met Office experiment after 1987 as well as in a sensitivity experiment carried out with the ECMWF model.

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 1980–2001 will be discussed in this paper. This is the period for which all seven coupled models participating in the project have generated hindcasts. Longer time series (up to 43 yr) are available for a smaller number of models.

In order to assess seasonal dependence on skill, the DEMETER hindcasts have been started from 1 February, 1 May, 1 August, and 1 November initial conditions. Each hindcast has been integrated for 6 months and comprises an ensemble of nine members. In its simplest form, the multimodel ensemble is formed by merging the ensemble hindcasts of the seven models, thus comprising 7×9 ensemble members. To enable a fast and efficient postprocessing and analysis of this complex dataset, 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 large subset of atmosphere and ocean variables (a complete description of the atmosphere and ocean archived variables can be found on the DEMETER Web site www.ecmwf.int/research/demeter/news/var_list.html), with both daily data and monthly means, has been stored in 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 dataset (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. (Model hindcasts can be retrieved in GRIB and NetCDF formats from the Web site www.ecmwf.int/research/demeter/data. A tool to display the fields is also available.)

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. Based upon this standard, a comprehensive verification system to evaluate all DEMETER single models as well as the multimodel DEMETER ensemble system has been set up at ECMWF. It is run periodically to monitor hindcast production, to quality control the data (and correct archival) and to calculate a common set of diagnostics.

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. 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, for example, 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, such as the 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 multimodel ensembles using scatter diagrams of area-averaged skill measures and probability density functions (PDFs) of gridpoint skill scores.

Both anomalies and scores have been computed using a cross-validation "leave one out" method (Wilks 1995; Livezey 1999). 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 dataset 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 to validate the model data with more than one verification dataset. In fact, precipitation has been verified against the

	Multimodel	CERFACS	ECMWF	INGV	LODYC	Météo-France	Met Office	MPI	Persistence
1-month lead									
ACC	0.95	0.94	0.93	0.92	0.95	0.93	0.92	0.86	0.80
Bias	—	−0.34	−0.87	−0.60	−0.96	−0.03	−0.53	−2.07	—
3-month lead									
ACC	0.89	0.86	0.86	0.82	0.89	0.83	0.81	0.66	0.62
Bias	—	0.07	−1.50	−0.76	−1.52	0.43	1.45	−3.42	—

TABLE 2. Ensemble-mean ACC and bias 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 1980–2001. The persistence hindcast is formulated as the anomaly in the month previous to the starting date. Hindcasts verify on the seasons MAM, JJA, SON, and DJF for the lead time of 1 month, and MJJ, ASO, NDJ, and FMA for the lead time of 3 months. Note that the bias for the multimodel ensemble and the persistence hindcast are not defined because the multimodel ensemble is based on single-model anomalies, which are constructed with regard to the single-model bias, while persistence uses observed anomalies.

Global Precipitation Climatology Project (GPCP) dataset. (The GPCP dataset can be found online at http://cics.umd.edu/~yin/GPCP/.)

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 Web site. (A selection of the verification results is displayed and automatically updated at www.ecmwf.int/research/demeter/verification/index.html.)

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 (e.g., Palmer and Anderson 1994). Thus, a prerequisite for successful seasonal forecasts is an 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 time scales, 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 multimodel ensemble for the SSTs averaged over the Niño-3.4 area. The correlation has been computed for the 1-month and 3-month lead seasonal hindcasts starting in February, May, August, and November. Therefore, the values verify during the seasons MAM, JJA, SON, and DJF for the 1-month lead hindcasts, and MJJ, ASO, NDJ, and FMA for the 3-month lead hindcasts. Results suggest that the single-model ensembles generally perform well as El Niño–Southern Oscillation (ENSO) prediction systems. For the sake of comparison, the ACC for a persisted-SST hindcast has been included. The SST anomaly in the month previous to the starting date has been taken as a persistence hindcast. For instance, the 6-month-long persistence hindcasts starting on the first of February are obtained from the anomaly on the previous January. Both the multimodel ensemble and the single models perform at levels comparable to dedicated ENSO prediction models and much better than persistence, especially in the 3-month lead-time range. In addition, note the high correlation of the multimodel ensemble for both lead times, proving it to be the most skillful system in an overall assessment. The bias

of the single models is generally in the range of ±1 K (Table 2). These are typical figures for current state-of-the-art coupled models. As is the case for most variables and areas, there appears to be no clear relationship between bias and skill, though this is a topic that needs further investigation.

Figure 2 shows 1980–2001 time series of precipitation ACC for all single models and the multimodel ensemble, for summer (June to August, May start date) over the Tropics (Fig. 2a) and winter (December to February, November start date) over the northern extratropics (Fig. 2b). The skill in the northern extratropics is considerably less than in the Tropics. In both regions the variability in prediction skill, both from year to year and between different single models, is clearly evident. The occurrence of higher skill during ENSO events is consistent with relatively large ACC for 1982/83, 1987/88, and 1997/98 (Fig. 2). This, in turn, is consistent with the link between ENSO activity and seasonal predictability found in many studies (for PROVOST see Brankovi and Palmer 2000). In general, the identity of the most skillful single model varies with region and year. Finally, this figure illustrates the relatively skillful performance of the multimodel ensemble. In spite of the multimodel ensemble not being distinctly superior to the best single-model ensemble for individual events or small regions, in agreement with Peng et al. (2002), it is systematically the most skillful when the scores are averaged over large regions or long time series.

To further summarize atmospheric hindcast skill, Fig. 3 shows indices of the winter (December to February, November start date) Pacific–North American (PNA) and North Atlantic Oscillation (NAO) patterns for the multimodel ensemble. The indices are computed following the method described in the Doblas-Reyes et al. (2003). Values are obtained by projecting every ensemble member anomaly onto the leading empirical orthogonal function (EOF) of the
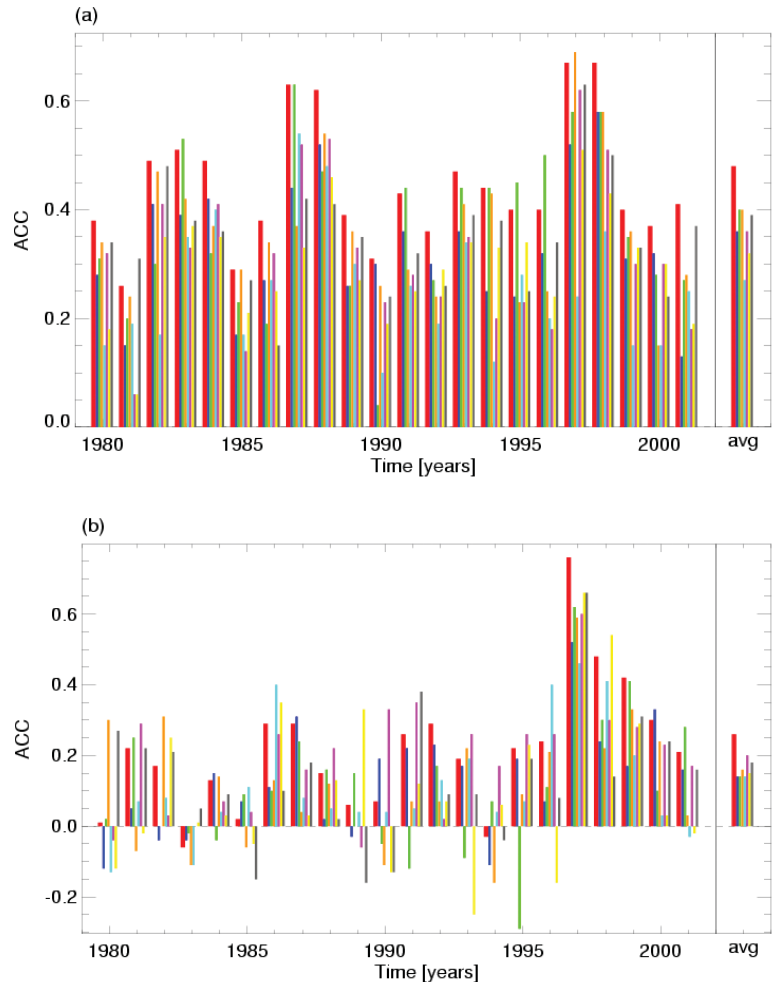

FIG. 2. Time series of the ensemble-mean precipitation anomaly correlation coefficients for the multimodel (thick red bars) and all individual models (thin bars; ECMWF: blue, Met Office: green, Météo-France: orange, MPI: cyan, LODYC: pink, INGV: yellow, CERFACS: gray). (a) One-month lead summer (JJA) precipitation in the Tropics (latitudinal band of 30°S–30°N); (b) 1-month lead winter (DJF) precipitation in the northern extratropics (latitudinal band of 30°–87.5°N). Additionally, the average over the whole period 1980–2001 is shown at the end of each plot.

500-hPa geopotential height [computed over the winter monthly mean anomalies using National Centers for Environmental Prediction (NCEP) reanalyses for the period 1949–2000]. The EOF analysis was carried out using data over the regions 20°–87.5°N and 110°–90°W for the PNA and 20°–87.5°N and 90°W–60°E for the NAO. The spatial covariance between the monthly anomaly patterns and the reference pattern was then computed for every single member of the hindcast ensemble. Monthly covariances were averaged to produce seasonal means. Figure 3 displays the index against time using a box-and-whisker representation in which the central box and each whisker con-

TABLE 3. Ensemble-mean correlation and RPSS for tercile categories for the PNA and NAO indices calculated from the 1-month lead hindcasts started in November (DJF seasonal average) for the years 1980–2001. Confidence levels are indicated in parentheses. The model acronyms are defined in Table 1.

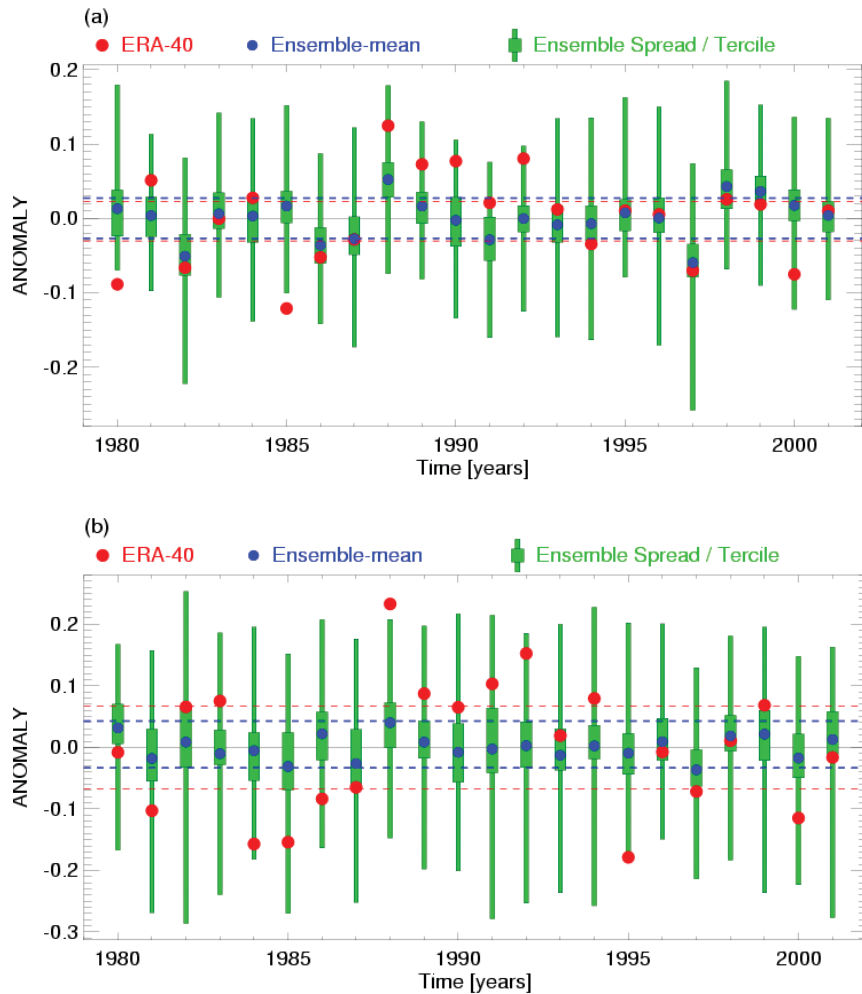|	Multimodel	CERFACS	ECMWF	INGV	LODYC	Météo-France	Met Office	MPI
Correlation (PNA)	0.41 (0.98)	0.16 (0.00)	0.39 (0.78)	0.23 (0.01)	0.46 (1.00)	0.23 (0.01)	0.31 (0.09)	0.32 (0.10)
Correlation (NAO)	0.54 (1.00)	0.30 (0.05)	0.10 (0.00)	0.20 (0.00)	0.43 (1.00)	0.55 (1.00)	0.18 (0.00)	0.14 (0.00)
RPSS (PNA)	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)

FIG. 3. Time series of the 1-month lead winter (DJF) (a) PNA and (b) NAO index for the period 1980–2001. The multimodel 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).

tain 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 multimodel ensemble range in all but two cases for both indices. In addition, skill measures indicate a higher reliability for the multimodel. Table 3 shows the correlation between the two time series for the multimodel and the single-model ensembles. The multimodel ensemble shows one of the highest correlations among all the models for both indices. In addition, the multimodel ensemble correlation can be considered nonzero with a 95% confidence level using a two-sided t test, which is not always the case for the single-model ensembles. However, it should be noted that scores based on indices are less robust than scores based on large-area correlations, when calculated with relatively short time series. For example, the high PNA correlation for some single models may be explained by good predictions in 1982 and 1997.

Note that, while PNA index hindcast skill tends to be quite satisfactory (Fig. 3a), NAO index skill is lower but always positive. Figure 3b indicates that the multimodel 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 very weak (little shift of the predicted index away from zero) as in 1992 and 1995, when the observed index was large in magnitude.

Considerable effort has been devoted to the validation of the ensembles as probability forecasts. The dashed blue and red lines in Fig. 3 correspond to the ensemble and ERA-40 terciles. The probabilistic skill measure used is the RPSS (Epstein 1969) based on these terciles. 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 1. The skill of the multimodel ensemble for the PNA index is close to the skill of the best models and is statistically significant at the 95% confidence level, in good agreement with the correlation results. Also, for the NAO index, RPSS values are generally 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 by scrambling the available hindcasts and verifications).

In order to get a more comprehensive assessment of the single-model versus multimodel ensemble skill, a number of ROC skill scores are collected together (Fig. 4). The value of the area under the ROC curve is a probabilistic skill measure ranging from 0 to 1 (Swets 1988). The ROC skill score is constructed as twice the area minus one, ranging between –1 and +1. Values below 0 imply lower skill than climatology, while a perfect forecast has a ROC skill score of 1. The comparison of all ROC skill scores for 2-m temperature, calculated over different regions, start dates, lead times, and events, shows that, although in some cases single-model ensembles have a higher ROC score than the multimodel ensemble, in the vast majority of cases (90%) the ROC skill score of the multimodel ensemble exceeds the score of the single models. Furthermore, the number of cases with less skill than climatology is greatly reduced for the multimodel ensemble; for the latter there are no cases with negative ROC skill score compared to 25 cases for the single models.

The greater probabilistic skill of the multimodel ensemble compared to the single-model skill leads to
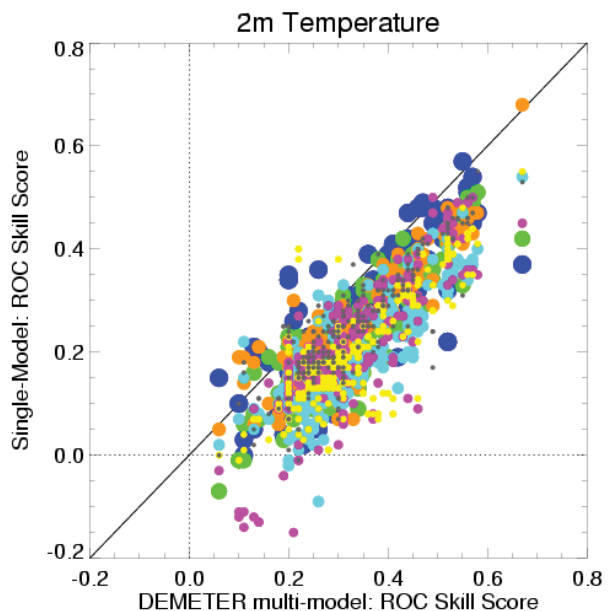

FIG. 4. Scatterplot of single-model (ECMWF: blue, Met Office: green, Météo-France: orange, MPI: cyan, LODYC, pink, INGV: yellow, CERFACS: gray; bullets of different size have been used for better visibility of all models) vs multimodel ROC skill scores of the 2-m temperature hindcasts from 1980–2001. The plot comprises results from seasonal hindcast scores for eight different areas (northern extratropics, Tropics, southern extratropics, North America, Europe, West Africa, East Africa, southern Africa), four start dates (Feb, May, Aug, Nov), two lead times (1 month, 3 month), and four events (anomaly above/below 0.43 standard deviation, anomaly above/below 0).

an increased potential economic value (Richardson 2000; Palmer et al. 2000; Atger 2001). For instance, it has been found that, for predictions of positive tropical winter (December to February, November start date) precipitation anomalies, the multimodel 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 multimodel ensemble performance an important question arises. This improvement could be due either to the multimodel approach itself or to the increased ensemble size resulting from collecting all members of the single-model ensembles to construct the multimodel ensemble, or to both. In order to separate the benefit of the multimodel approach that is derived 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 1987–99 using only a single start date (1 May). The

ensemble was generated using additional wind and SST perturbations in order to have a better sampling of the initial condition uncertainty. As a skill measure, the Brier skill score for tropical summer (June to August) precipitation positive anomalies for the multimodel ensemble (red bars) and the ECMWF model (blue bars) is shown in Fig. 5a. For each ensemble size, the corresponding ensemble was constructed by randomly selecting the members from the 63 available in the multimodel ensemble and the 54 in the single-model ensemble. Results show that the skill score grows faster with ensemble size for ensembles with less than about 30 members, although this threshold changes with the region, variable, and event considered (not shown). The skill saturates for large ensembles, as expected from the conceptual model described in Kumar and
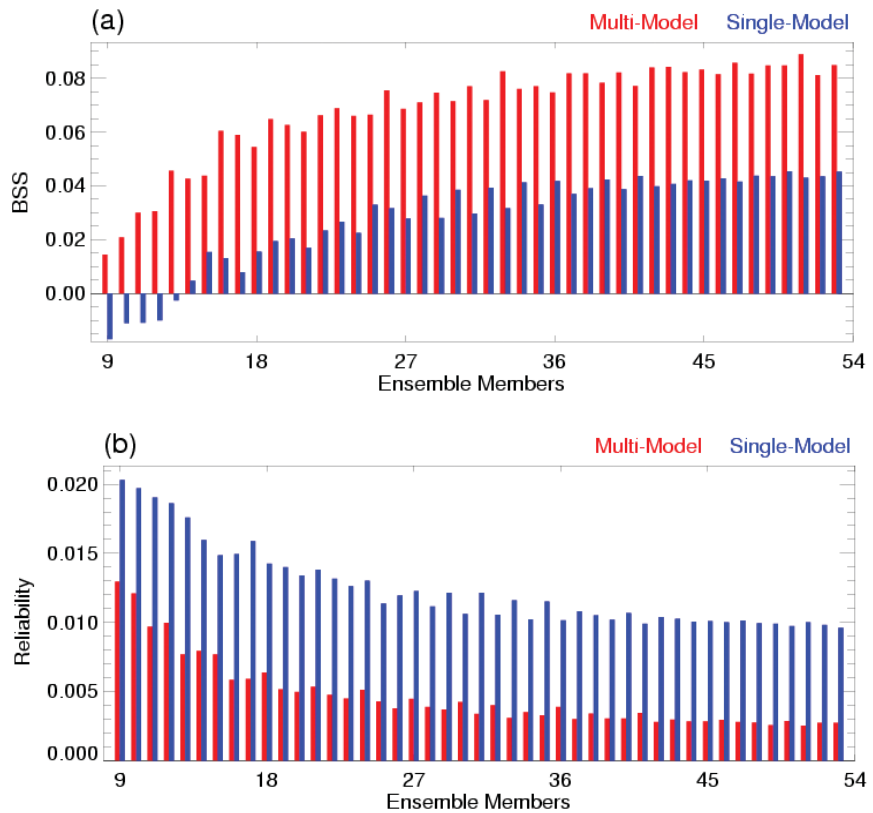

FIG. 5. (a) Brier skill score and (b) reliability component of Brier score for the 1-month lead tropical summer (JJA) precipitation 1987–99 for the single ECMWF control model (blue) and the DEMETER multimodel (red). The event is "precipitation anomalies above zero." Results are shown for different ensemble sizes from 9 to 54 members. Note that lower values of the reliability term mean better reliability.

Hoerling (2000), though a slight increase with size is still found. Based on a decomposition of the Brier score (Murphy 1973), results show that the largest contribution to the multimodel 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 Fig. 5b. The multimodel ensemble performs better than the single-model ensemble for every ensemble size, despite the ECMWF model having particularly good behavior over the Tropics. The increase in Brier skill score and reliability is similar for both the single-model and the multimodel ensembles, so that their difference remains approximately constant as the ensemble size increases. This implies that the multimodel ensemble advantage over a given single-model ensemble, as shown in Figs. 2–4, is not an artifact of the usually large ensemble size of the multimodel, but rather is due to the multimodel approach itself. Similar results are found for other variables and regions. This suggests that the multimodel

ensemble provides a better sampling of forecast uncertainty in the sense that it contains the verification more often than a given single-model ensemble.

SEASONAL FORECAST 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. However, existing application models typically require weather input (precipitation, temperature, wind, radiation, etc.) at a substantially higher spatial resolution than is available from the global models. We therefore begin this section with a brief discussion of the downscaling techniques used in DEMETER.

Downscaling. In DEMETER, both statistical/empirical methods and dynamical regional climate models have been tested and applied for downscaling purposes. The methods used and some illustrative results are described in the following.

a: Direct multi-model output b: Observations c: Synoptic network

d: Analog-based downscaling e: SVD-based downscaling f: Dynamical downscaling

mm/day
0 - 1
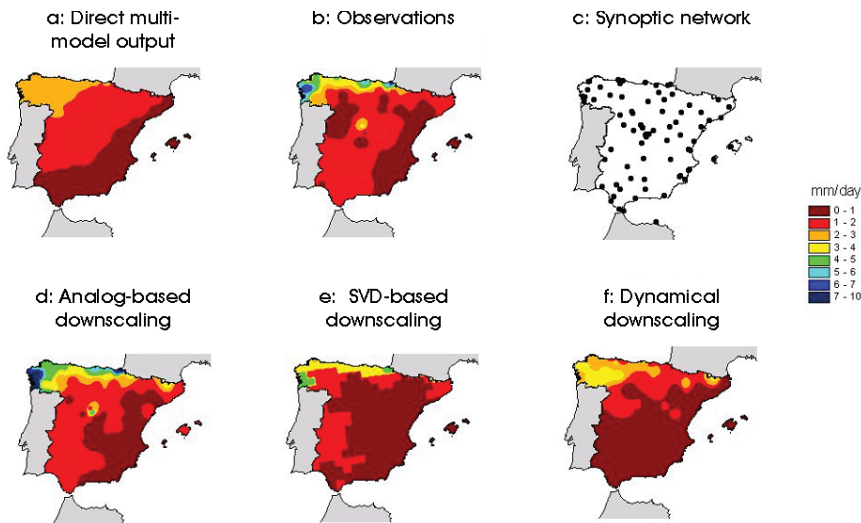1 - 2
2 - 3
3 - 4
4 - 5
5 - 6
6 - 7
7 - 10

Fig. 6. (a) Three-month (Feb–Apr) average precipitation over Spain for the ensemble mean of the multimodel hindcasts (ECMWF and Met Office models) initialized in Nov 1986. Panel (b) shows the corresponding observed precipitation, which has been computed using the stations of the Spanish synoptic network shown in (c). The downscaled precipitation with the (d) analog-based, (e) MCA-based, and (f) dynamical methods is shown in the lower row.

piled by the European Commission Joint Research Centre (EC-JRC) (van der Voet et al. 1994; Terres 1999). Both downscaling methods emphasize the anomalously wet area over northwestern Spain (of which there is only a slight indication in the raw model output).

As a second type of downscaling method, dynamical models have also been used in DEMETER. Dynamical downscaling has the potential to outperform statistical/empirical methods, particularly in the prediction of extreme events or in areas where observed data needed to train the statistical/empirical models, are not available. However, there are outstanding problems, including propagation of systematic biases from the global to the regional model (Giorgi and Mearns 1999). In addition, the computational expense of running a high-resolution regional climate model can be comparable to that of running a global seasonal prediction model.

Regional climate models have been applied to study local effects of global climate change resulting from increasing concentration of greenhouse gases in the atmosphere (e.g., Christensen et al. 2001). By comparison, little has been done for the seasonal time scale (Misra et al. 2003).

The dynamical downscaling method tested in DEMETER is based on the Rossby Centre Atmosphere (RCA) model, which is a climate version of the High-Resolution Limited-Area Model (HIRLAM; Rummukainen et al. 2001) regional weather prediction model. The RCA model has been nested to the ECMWF model output and run in climate mode for 6 months. To demonstrate the feasibility of the approach, a set of runs was started the first of November for 1986, 1987, and 1988 and the first of May for 1987, 1988, and 1989. The area covered by the RCA integration was 15.5°–65.0°N and 67.5°W–31.0°E (Atlantic–European region). The horizontal resolution was 0.5° with 31 levels in the vertical.

Figure 6f shows the hindcast for February to April 1987 average precipitation (November 1986 start date, three-member ensemble) over Spain. The precipita-

In the statistical/empirical methods, a mapping (e.g., based on regression methods, analogue techniques, or neural networks; Wilby and Wigley 1997, 2000) is derived from one or more large-scale fields to the fine scale required by the application models. The statistical methods are relatively straightforward to apply and are computationally cheap.

As an illustrative example, Fig. 6 displays total precipitation over Spain for the February to April seasonal average. The raw multimodel ensemble output for the hindcast started in November 1986 (Fig. 6a, showing a multimodel ensemble mean field from the ECMWF and Met Office coupled models), and the downscaled precipitation based on three different downscaling methods are shown in the figure. The observed precipitation is displayed in Fig. 6b, while Fig. 6c shows the network used to create this database. The first downscaling method (Fig. 6d) is based on the search for analogues of 1000- and 500-hPa geopotential and low-level relative humidity fields of the ECMWF and Met Office models. This method is currently used operationally for short-range precipitation forecasts (Fernández et al. 2001) at the Spanish Instituto Nacional de Meteorologia. The second method (Fig. 6e) uses a stochastic weather generator (Richardson 1981) and is based on a regression of the leading modes of a maximum covariance analysis (MCA; Feddersen et al. 1999) between model predictions of seasonal precipitation and gridded observations of precipitation on a 50 km × 50 km grid com-

tion pattern is more realistic than that of the global model ensemble, especially over northwestern Spain, although for this case the statistical/empirical downscaling methods appear to perform as well or better. The RCA model has also been run for a region covering Africa to test the ability to provide downscaled climate forecast information for the malaria prediction application described below.

Crop yield prediction. The total value of European Union (EU) agriculture is around 200 billion euro, while EU countries are collectively the largest importers of agricultural products in the world, and the second largest exporters (European Commission 2000). Moreover, the EU production of wheat is around 100 million tons, setting the EU as the second largest producer in the world after Asia (FAOSTAT 2001). 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.

Crop simulation models that estimate crop growth and crop yield, as a function of environmental conditions and management practices, are important tools for decision makers (Boote et al. 1996; Hoogenboom 1997). These models can be used for simulating and forecasting crop yields, and in addition for computing the effects of management practices (such as irrigation, sowing time, or nitrogen fertilization; Hoogenboom 2000), and assessing the pressures of these practices on the environment (e.g., through nitrate leaching or crop water requirement; Singh and Thornton 1992). They can also be applied for the assessment of climatic change risk in agriculture (Harrison et al. 1995).

The EC-JRC Crop Growth Monitoring System uses a crop model called the World Food Studies Model (WOFOST), and performs crop yield forecasting through a regression analysis comparing simulated crop indicators and historical yield series for the main crops at the 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 (a) the technological time trend and (b) a number of simulated crop 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 predictions brings additional information for the remaining crop season. 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 transport simulation system called CRITERIA (Marletto et al. 2001).

The potential importance of seasonal predictions for crop yield estimation was demonstrated by forcing the crop model with ERA-15 reanalysis used as a "perfect forecast" and comparing with the current operational system in crop yield forecasting at EC-JRC (Terres and Cantelaube 2003). Results showed that simulated crop indicators had a higher contribution in the estimated yield, as one of them was selected as the best predictor (compared with the technological time trend factor) in a greater number of cases than the current operational system, and moreover yield estimates were closer to the reference crop yield based on the Eurostat value (Eurostat is the Statistical Office of the European Union with the Eurostat value being the official yield used as a reference for yield comparison) in nearly twice as many cases.

As stated above, given the low spatial resolution of the coupled models, a need to downscale the global model output is evident. 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 with the result obtained using raw model output. This resulted in a 10% reduction in root-mean-square error and an increase of regression coefficient from 0.63 to 0.71.

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 PDF of the crop yield. Based on this PDF, the end user can directly quantify the benefits and risks of specific weather-sensitive decisions.

Wheat yield hindcasts for European countries were carried out over a 4-yr period (1995–98) using DEMETER multimodel ensemble downscaled data. Figure 7 shows the predictions for Germany and France, the largest wheat producers in Europe, along with those for Denmark and Greece. For the sake of comparison, a set of runs were also carried out with the crop model using downscaled ERA-40 data, these results being depicted with red dots. This reference differs in some cases from the official Eurostat figures (depicted with horizontal black lines) for several reasons not only related to the quality of ERA-40 data, but also related to the uncertainty associated with the crop yield recorded by the member states and the lack of impact from pests and conditions at
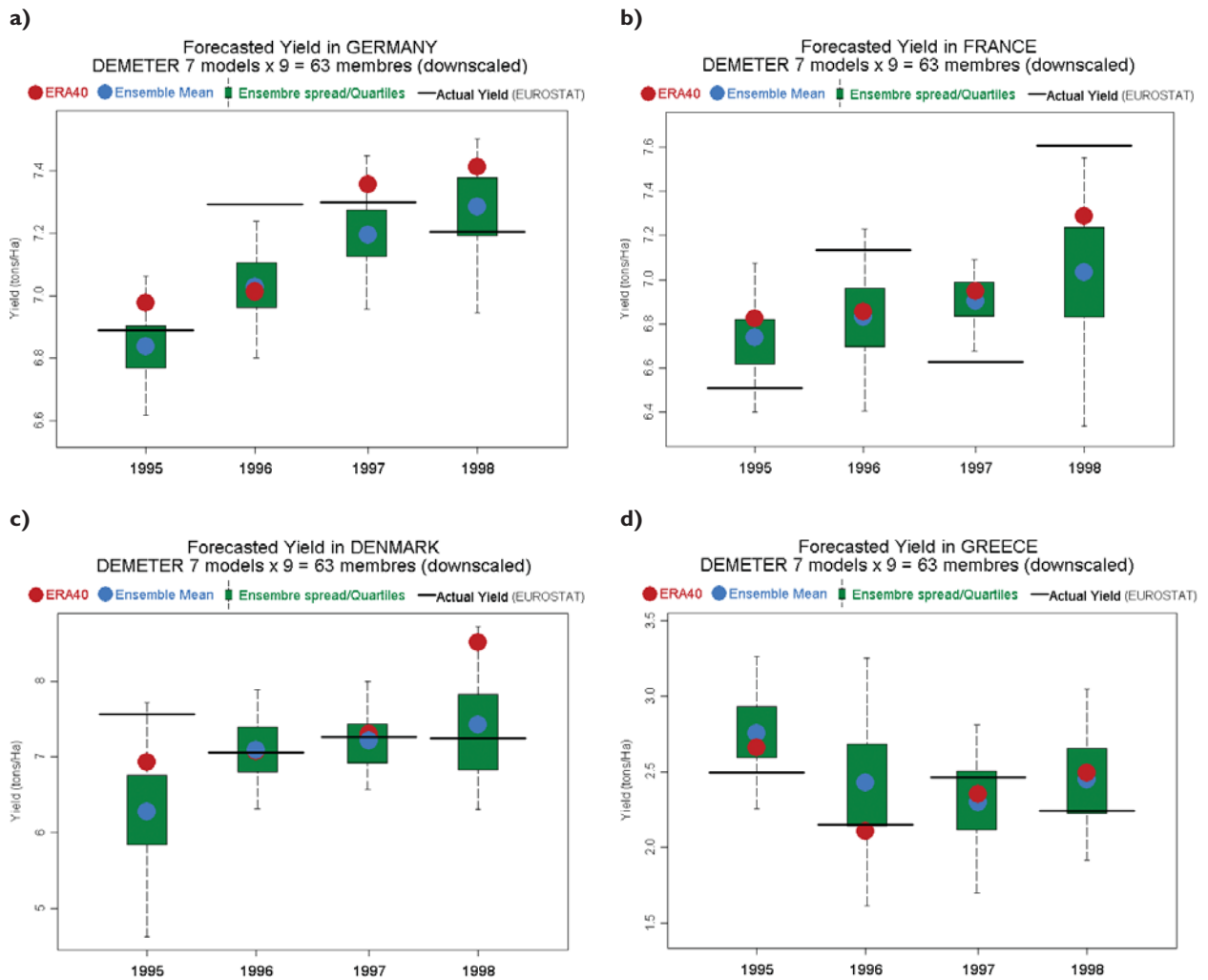
a)

Forecasted Yield in GERMANY
DEMETER 7 models x 9 = 63 membres (downscaled)

● ERA40 ● Ensemble Mean ■ Ensembre spread/Quartiles — Actual Yield (EUROSTAT)

b)

Forecasted Yield in FRANCE
DEMETER 7 models x 9 = 63 membres (downscaled)

● ERA40 ● Ensemble Mean ■ Ensembre spread/Quartiles — Actual Yield (EUROSTAT)

c)

Forecasted Yield in DENMARK
DEMETER 7 models x 9 = 63 membres (downscaled)

● ERA40 ● Ensemble Mean ■ Ensembre spread/Quartiles — Actual Yield (EUROSTAT)

d)

Forecasted Yield in GREECE
DEMETER 7 models x 9 = 63 membres (downscaled)

● ERA40 ● Ensemble Mean ■ Ensembre spread/Quartiles — Actual Yield (EUROSTAT)

FIG. 7. Time series of the wheat yield predictions from downscaled data for (a) Germany, (b) France, (c) Denmark, and (d) Greece over the period 1995–1998. The multimodel ensemble spread is depicted by the box-and-whisker representation, with the whiskers containing the lower and upper quartile of the ensemble. The blue dots represent the ensemble mean, the yield obtained by forcing the crop model with ERA-40 data being displayed by slightly bigger red bullets. The black horizontal line corresponds to the reference value (Eurostat).

harvest. Substantial interannual variations are found for the ERA-40 estimates in every country. They are satisfactorily matched by the multimodel ensemble prediction. In fact, the ERA-40 yield estimate, in every single case, lies within the multimodel ensemble range. Quantitative forecast quality estimates are not available given the short length of the time series, although these preliminary results are promising and have led the European Commission to consider the possibility of including multimodel seasonal forecasts as part of its routine monitoring and forecast system.

As an illustration of the potential for probability forecasts of the yield anomaly, Fig. 8 shows the cumulative probability function of wheat yield obtained with the downscaled multimodel ensemble hindcasts. For example, the official yield (Eurostat) was 0.5 tons ha^{-1}

above the trend (positive anomaly) in 1988, while the system using the DEMETER hindcasts predicted the yield to be greater than 0.4 tons ha^{-1} above the trend with a probability of 60%.

Malaria prediction. Malaria kills 1 million to 2.7 million people per year, the vast majority of these deaths occurring in children in sub-Saharan Africa (Breman et al. 2001). Malaria endemicity in Africa varies along a continuum. In areas of intense transmission, adults are generally immune (unless in their first pregnancy) and children are vulnerable to severe disease and death. In areas of moderate endemicity, transmission is often highly seasonal with the number of cases increasing each year following the seasonal rains. A typical lag in the peak case numbers compared with the

Cumulative Distribution Function
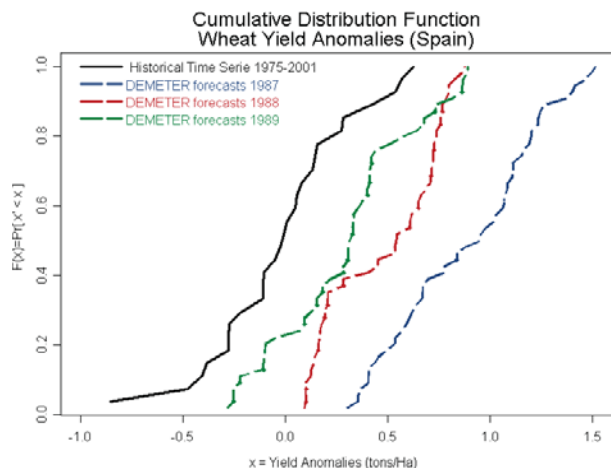**Wheat Yield Anomalies (Spain)**

FIG. 8. Cumulative distribution function of the wheat yield anomalies for Spain predicted for 1987 (blue), 1988 (red), and 1989 (green) with the multimodel ensemble downscaled data. The black solid line corresponds to the historical cumulative distribution function.

peak in the rainfall of between 2 and 4 months is observed. In semiarid or high-altitude areas, transmission is highly unstable with the disease normally absent, although epidemics with high death rates in all age groups may occur in climatically anomalous years, controlled by wetter and/or warmer-than-average conditions. In such areas, malaria epidemics result from the interplay of the vector (Anopheline mosquitoes), parasite (*Plasmodium spp.* with *P. falciparum* being the most dangerous and the cause of African epidemics), and human host. The ambient temperature drives the nonhuman stage of the disease, in both the vector and parasite, provided sufficient water or rainfall is present to provide pools for mosquito breeding sites and high humidity levels that are essential for the survivorship of adult mosquitoes.

It has been proposed that malaria early warning systems, which incorporate vulnerability assessment, seasonal climate forecasts, weather monitoring, and case surveillance, be developed for unstable areas in Africa (WHO 2001). However, since seasonal climate forecasts are inherently probabilistic, it is important to develop models that represent the associated PDF of epidemic risk in order to give decision makers confidence in the use of such forecasts (Thomson et al. 2000).

Models for use in epidemic malaria prediction may be either statistical (Hay et al. 1998) or mathematical–biological. The biological processes may be included in the model as a series of interlinked submodels, each represented as coupled differential equations with a probabilistic transition between

model stages. The equations' parameters are deduced from empirical data (when available) or by fitting the model to data. The biological model permits the impact of malaria control interventions (such as insecticide spraying of houses) to be analyzed directly in the model. It can also be used to assess the potential impact of future climate change on malaria transmission (McCarthy et al. 2001). Furthermore, the timing of interventions in relation to their impact on the seasonal transmission of infection can be assessed, and thereby the cost effectiveness of early intervention is established.

In this context, a numerical dynamic biological model that predicts the onset of potential epidemics and is driven by temperature and precipitation daily time series has been developed (Hoshen et al. 2004). The meteorological data used in the model may come from station observations, ERA-40, or seasonal hindcasts. The model simulates the population dynamics of cohorts of mosquitoes, and thus predicts the behavior of the total mosquito population. The lag that is seen between extreme climatic conditions and the actual peak in any malaria epidemic is a result of three processes: the time taken for mosquitoes to breed and develop, the time required for the parasite to develop in the mosquito, and the time it takes for a human to become infectious.

The model was first tested using either station or ERA-40 data. For example, the model has been run on continental ERA-40 data to assess its value in terms of depicting the normal seasonality of the disease across Africa, and showed the known patterns of the epidemiology of malaria. It has also been compared with a short time series of hospital malaria case data using local meteorological station data. However, bias correction and downscaling techniques need to be applied to the ensemble outputs to maximize their potential in a malaria forecast model. Unfortunately, statistical/empirical downscaling will be harder to perform in Africa than in other regions due to the paucity of long-term local observations. Therefore, alternative modeling-based downscaling methods are required. As an initial step, a simple bias correction scheme has been used. An estimate of the seasonal cycle using daily data has been constructed for both ERA-40 and the different models. A seasonal cycle estimate has been obtained separately for each hindcast initial date and for each model. The difference between these seasonal cycles has been added to each member of an ensemble hindcast. Figure 9 shows an example for the grid point 20.0°S–27.5°E (the closest to Hwange, an area with high malaria incidence in Botswana) using data corrected for the daily bias

of precipitation and temperature as described above. The 3-month average of monthly incidence, as simulated by the biological model using ERA-40 data for the period 1987–2001, is shown in red. Incidence is defined as the percentage of people in a population who are infected over a given period (a month in this particular case) by the malaria parasite. Values obtained by forcing the model with ERA-40 data have been considered as "truth" given the scarcity of clinical case reports. The hypothesis that an ensemble of multimodel seasonal hindcasts could predict probabilistically the temporal structure of malaria prevalence has been assessed. For this purpose, the biological model was run using temperature and precipitation from each of the 63 bias-corrected members of the multimodel ensemble. The resulting monthly incidence was then averaged over months 2–4 and displayed in Fig. 9 using a box-and-whisker representation, with the whiskers containing the lower and upper terciles of the ensemble and the ensemble mean as a blue dot. The multimodel predictions represent the seasonal cycle with the maximum of incidence in boreal spring (March to May). In addition, the multimodel ensemble hindcasts match the interannual evolution of the ERA-40 estimates. As a measure of skill, the RPSS of the malaria prevalence time series is 0.24 for the DJF major epidemic season, which has a confidence level of 0.95. Hence, the interannual variability of the reference malaria incidence is predicted with some success, the reference value being always within the multimodel ensemble range for the MAM period. These promising results have led to further research in which the malaria model will be forced with downscaled seasonal prediction data.

SENSITIVITY EXPERIMENTS. Various sensitivity studies have been carried out, falling into four categories. In the first, methods for generating ensemble spread (i.e., use of perturbed SST and wind stress patterns and stochastic physics) are assessed. In the second category, the impact of ocean initial data and model resolution on

the hindcasts is investigated. In the third, experiments to determine the nature of equatorial SST biases are performed. Finally, the benefits of the coupled system over a prediction system using an atmosphere-only model forced by persisted SST anomalies are examined. In the interests of brevity, only this last sensitivity study is discussed in more detail. Further details on the other experiments will be published elsewhere.

Given the biases that are present in coupled models, it is important to assess the benefits of using a fully coupled system. The Met Office has performed a parallel set of hindcasts using the HadAM3 atmospheric model (the atmospheric component of the coupled GloSea model, see Table 1) forced by persisted SST anomalies. The initial conditions for individual ensemble members are taken from ERA-40 at 6-hourly intervals starting 2 days before the hindcast date. Full results can be viewed on the DEMETER Web site (www.ecmwf.int/research/demeter/verification/index.html).

Although the advantages of a coupled model over an atmosphere-only integration with persisted SST anomalies are likely to vary between El Niño and non–El Niño conditions (Graham et al. 2000), the sample of years available is not sufficiently large to make such a distinction. Here instead, an overall comparison based on 1987–99 hindcast is given, based on ROC scores obtained with the GloSea and HadAM3 models for probabilistic predictions of
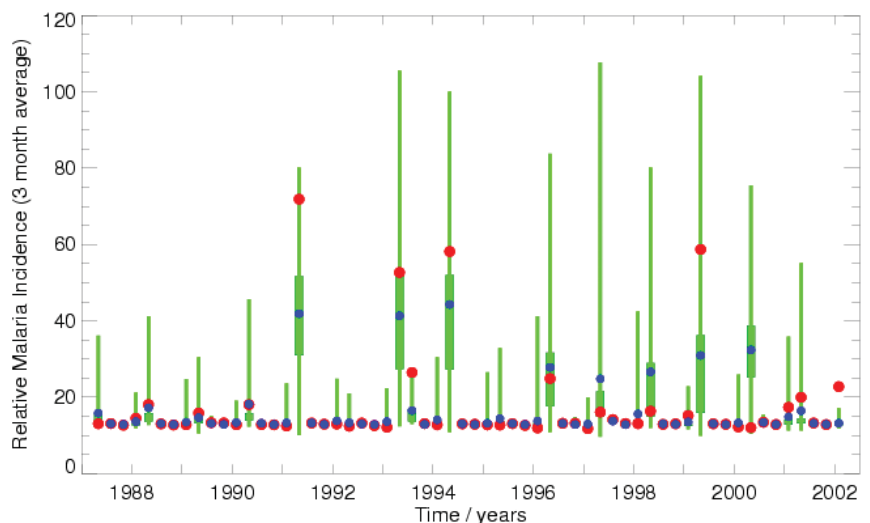

FIG. 9. Three-month average of relative monthly malaria incidence for the 1-month lead hindcasts at the grid point 20.0°S–27.5°E (Botswana). The red bullets correspond to the incidence obtained forcing the malaria model with bias-corrected ERA-40 data. The multimodel ensemble results calculated using all start dates for the period 1987–2001 are depicted using a box-and-whisker representation, with the whiskers containing the lower and upper terciles of the ensemble, and the ensemble mean as a blue dot.

above-normal seasonal mean 2-m temperature (Fig. 10). Scores are calculated for 3-month mean values corresponding to early and conventional seasonal periods (i.e., for boreal spring the early season is defined as February to April and the conventional season as March to May). The prediction for the early season is at 3-month lead (i.e., a long-lead outlook), while the prediction for the conventional season is at 1-month lead (i.e., corresponding to an "update" on the long-lead prediction). Scores for six geographical regions are shown. The number of times each model is more skillful than the other is indicated next to the appropriate axis. In counting the "wins" for each model, only cases where the difference in scores exceeds 2%, and the winning model has a score exceeding 0.5 are considered.

Comparisons for 2-m temperature show a degree of scatter about the diagonal in most seasons. Nevertheless (as indicated by the number of wins) the overall performance of GloSea and HadAM3 is similar in spring, summer, and autumn, while for the winter season GloSea appears notably more successful, with seven wins versus one. Scores are highest for the tropical region in all seasons with GloSea achieving the best scores in all cases and, as expected, the update scoring better than the long-lead hindcast. Notable successes for GloSea at 3-month lead include forecasts of February-to-April temperature over North America (Fig. 10a), and November-to-January temperature in the Tropics, West Africa, and northern extratropics (Fig. 10d). In contrast, HadAM3 currently appears to perform better at both lead times over Europe in summer and autumn. In winter, scores for Europe are similar with both models, while in spring GloSea has better 1-month lead forecasts, while HadAM3 is notably better at 3-month lead (Fig. 10a).

CONCLUSIONS. As part of a European Union–funded DEMETER project, a multimodel ensemble system based on seven European global coupled ocean–atmosphere models has been described and validated in hindcast mode using ERA-40 data and GPCP for precipitation. Output from the DEMETER multimodel system, suitably downscaled, has been applied to crop yield and malaria prediction models. Results indicate that the multimodel 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. An additional outcome of the project is the production of
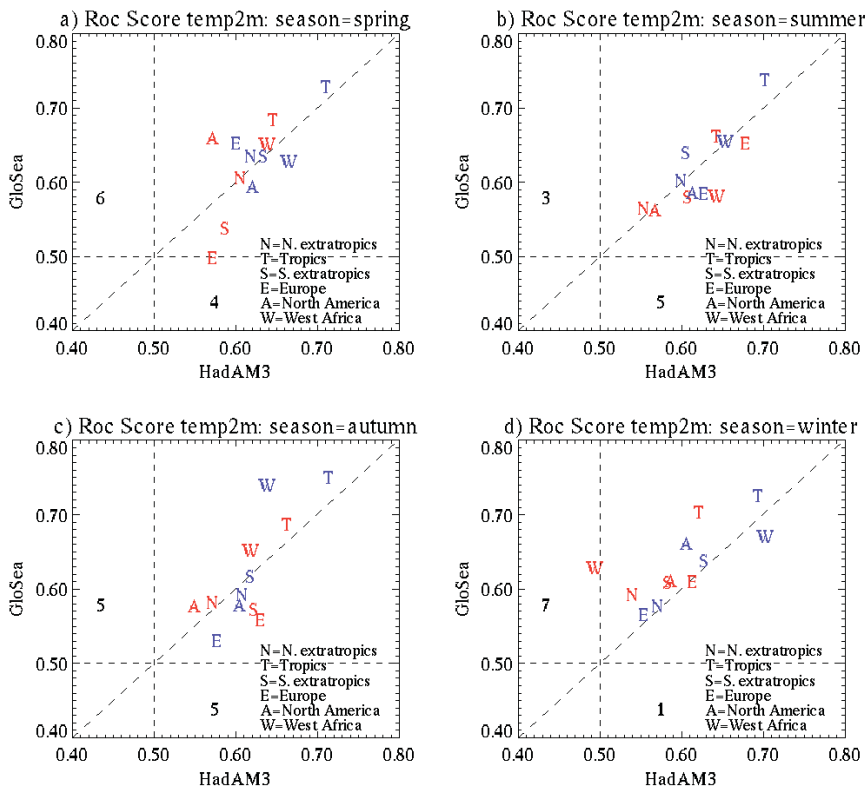

FIG. 10. Comparison of ROC scores for probabilistic predictions of the event "2-m temperature above the climate normal" obtained with the GloSea and HadAM3 models for six geographical regions [northern extratropics (N); Tropics (T); southern extratropics (S); Europe (E); North America (A); West Africa (W)]. Scores for two lead times are shown, a 3-month lead (long lead) forecast for the "early" season (early spring/summer/autumn/winter = FMA/MJJ/ASO/NDJ) in red, and a 1-month lead "update" forecast for the conventional season (MAM/JJA/SON/DJF) in blue. The number of times the score obtained with each model exceeds that of the other model is provided next to the appropriate axis. Only cases when the winning model achieves a ROC score that exceeds 0.5 (the "no skill" threshold), and exceeds that of the other model by 2% or more are counted. Scores are calculated over the period 1987–99. Forecasts are verified against anomalies from ERA-40.

a unique seasonal hindcast dataset. This dataset offers the potential to improve our understanding of the mechanisms of intraseasonal and interannual climate variations.

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 (www.ecmwf.int/research/demeter/verification), 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 dataset 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 dataset for education and training purposes, both in the developed and developing worlds.

As a result of the success of DEMETER, real-time multimodel ensemble 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 Météo-France coupled systems to be included in this multimodel mix. It is possible that other DEMETER 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 multimodel ensembles not only for seasonal-to-interannual time scales, but also for decadal time scales for which scientific evidence of predictability has emerged in recent years (Grötzner et al. 1999). For this purpose, 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 time scale anthropogenic climate change. In this way, the reliability of century time-scale climate change projections can be assessed by running essentially the same ensemble systems on time scales for which verification data exists. A unification and rationalization of research and development across these time scales will enhance enormously the credibility of climate science.

Finally, while the results shown in this paper clearly indicate the need to represent model uncertainty when forecasting climate, the multimodel approach cannot be considered the final solution. For example, there is manifestly no representation of common model deficiencies in a multimodel ensemble. Rather, the DEMETER results have motivated a more theoretical approach to the representation of model uncertainty using stochastic-dynamic subgrid models (Palmer 2001; Khonider et al. 2003). Practical schemes based on cellular automaton stochastic dynamics subgrid models are currently in development, and it is hoped to assess these against the multimodel approach in coming years.

REFERENCES

Alves, O., G. Wang, A. Zhong, N. Smith, G. Warren, A. Marshall, F. Tzeitkin, and A. Schiller, 2002: POAMA: Bureau of Meteorology operational coupled model seasonal forecast system. *Proc. ECMWF Workshop on the Role of the Upper Ocean in Medium and Extended Range Forecasting,* Reading, United Kingdom, ECMWF, 22–32.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes Geophys.,* **8,** 401–417.

Boote, K. J., J. W. Jones, and N. B. Pickering, 1996: Potential uses and limitations of crop models. *Agron. J.,* **88,** 704–716.

Branković, C., and T. N. Palmer, 2000: Seasonal skill and predictability of ECMWF PROVOST ensembles. *Quart. J. Roy. Meteor. Soc.,* **126,** 2035–2068.

Breman, J. G., A. Egan, and G. T. Keusch, 2001: The intolerable burden of malaria: A new look at the numbers. *Amer. J. Trop. Med. Hyg.,* **64** (Suppl.), 4–7.

Buizer, J. L., J. Foster, and D. Lund, 2000: Global impacts and regional actions: Preparing for the 1997–98 El Niño. *Bull. Amer. Meteor. Soc.,* **81,** 2121–2139.

Christensen, J. H., J. Raisanen, T. Iversen, D. Bjorge, O. B. Christensen, and M. Rummukainen, 2001: A

synthesis of regional climate change simulations: A Scandinavian perspective. *Geophys. Res. Lett.,* **28,** 1003–1006.

Délécluse, P., and G. Madec, 1999: Ocean modeling and the role of the ocean in the climate system. *Modeling the Earth's Climate and Its Variability,* W. R. Holland, S. Joussaume, and F. David, Eds., Elsevier Science, 237–313.

Déqué, M., 2001: Seasonal predictability of tropical rainfall: Probabilistic formulation and validation. *Tellus,* **53A,** 500–512.

Doblas-Reyes, F. J., M. Déqué, and J.-Ph. Piedelièvre, 2000: Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Quart. J. Roy. Meteor. Soc.,* **126,** 2069–2088.

——, V. Pavan, and D. B. Stephenson, 2003: The skill of multi-model seasonal forecasts of the North Atlantic Oscillation. *Climate Dyn.,* **21,** 501–514.

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

European Commission (DG Agriculture), 2000: *La Politique Agricole Commune: Synthèse 1999.* Office des Publications Officielles des Communautés Européennes, Luxembourg, KF-29-00-262-FR-C, 30 pp. [Available from Office des Publications Officielles des Communautés Européennes, 2 rue Mercier, L - 2985 Luxembourg.]

FAOSTAT, cited 2001: FAO statistical databases. United Nations Food and Agriculture Organization, Rome, Italy. [Available online at http://apps.fao.org/page/collections?subset=agriculture.]

Feddersen, H., A. Navarra, and M. N. Ward, 1999: Reduction of model systematic error by statistical correction for dynamical seasonal predictions. *J. Climate,* **12,** 1974–1989.

Fernández, J., J. del Hoyo, A. Mestre, and C. Peral, 2001: Local probabilistic forecast of precipitation by the use of an analogical approach. *Proc. Fifth European Conf. on Applications of Meteorology (ECAM 2001),* Budapest, Hungary, European Meteorological Society, 28–29.

Giorgi, F., and L. O. Mearns, 1999: Introduction to special section: Regional climate modeling revisited. *J. Geophys. Res.,* **104,** 6335–6352.

Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001: Current approaches to seasonal-to-interannual climate predictions. *Int. J. Climatol.,* **21,** 1111–1152.

Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood, 2000: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Cen-

tre coupled model without flux adjustments. *Climate Dyn.,* **16,** 147–168.

Graham, R. J., A. D. L. Evans, K. R. Mylne, M. S. J. Harrison, and K. B. Robertson, 2000: An assessment of seasonal predictability using atmospheric general circulation models. *Quart. J. Roy. Meteor. Soc.,* **126,** 2211–2240.

Gregory, D., J. J. Morcrette, C. Jakob, A. C. M. Beljaars, and T. Stockdale, 2000: Revision of convection, radiation and cloud schemes in the ECMWF Integrated Forecasting System. *Quart. J. Roy. Meteor. Soc.,* **126,** 1685–1710.

Grötzner, A., M. Latif, A. Timmermann, and R. Voss, 1999: Interannual to decadal predictability in a coupled ocean–atmosphere general circulation model. *J. Climate,* **12,** 2607–2624.

Harrison, P. A., R. E. Butterfield, and T. E. Downing, 1995: Climate change and agriculture in Europe: Assessment of impacts and adaptation. Environmental Change Unit, University of Oxford, Research Rep. 9, 411 pp.

Hartmann, H. C., T. C. Pagano, S. Sorooshian, and R. Bales, 2002a: Confidence builders: Evaluating seasonal climate forecasts for user perspectives. *Bull. Amer. Meteor. Soc.,* **83,** 683–698.

——, R. Bales, and S. Sorooshian, 2002b: Weather, climate, and hydrologic forecasting for the US Southwest: A survey. *Climate Res.,* **21,** 239–258.

Hay, S. I., R. W. Snow, and D. J. Rogers, 1998: Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data. *Trans. Roy. Soc. Trop. Med. Hyg.,* **92,** 12–20.

Hoogenboom, G., 1997: AEGIS/WIN: A computer program for the application of crop simulation models across geographic areas. *Agron. J.,* **89,** 919–928.

——, 2000: Contribution of agrometeorology to the simulation of crop production and its applications. *Agric. For. Meteor.,* **103,** 137–157.

Hoshen, M. B., E. Worrall, S. J. Connor, A. P. Morse, and M. C. Thomson, 2004: Weather based modelling of the epidemiology of malaria. *GEOMED,* in press.

Kanamitsu, M., and Coauthors, 2002: NCEP dynamical seasonal forecast system 2000. *Bull. Amer. Meteor. Soc.,* **83,** 1019–1037.

Khonider, B., A. J. Majda, and M. A. Katsoulakis, 2003: Coarse-grained stochastic models for tropical convection and climate. *Proc. Natl. Acad. Sci. USA,* **100,** 11 941–11 946.

Kumar, A., and M. P. Hoerling, 2000: Analysis of a conceptual model of seasonal climate variability and implications for seasonal prediction. *Bull. Amer. Meteor. Soc.,* **81,** 255–264.

Livezey, R. E., 1999: Field intercomparison. *Analysis of Climate Variability: Applications of Statistical Tech-*

niques, H. von Storch and A. Navarra, Eds., Springer, 176–177.

Madec, G., P. Delecluse, M. Imbard, and C. Levy, 1997: OPA release 8, ocean general circulation model reference manual. LODYC Internal Rep., Paris, France, 200 pp. [Available from LODYC/IPSL, 4 Place Jussieu 75252 Paris Cedex 05, France.]

——, ——, ——, and ——, 1998: OPA version 8.1 ocean general circulation model reference manual. LODYC Tech. Rep. 11, Paris, France, 91 pp. [Available from LODYC/IPSL, 4 Place Jussieu, 75252 Paris Cedex 05, France.]

Marletto, V., L. Criscuolo, and M. R. M. Van Soetendael, 2001: Implementation of WOFOST in the framework of the CRITERIA geographical tool. *Proc. Second Int. Symp. on Modeling Cropping Systems,* Florence, Italy, European Society for Agronomy, 219–220.

Marsland, S. J., H. Haak, J. H. Jungclaus, M. Latif, and F. Röske, 2003: The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. *Ocean Modell.,* **5,** 91–127.

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.

McCarthy, J. J., O. F. Canziani, N. A. Leary, D. J. Dokken, and K. S. White, Eds., 2001: *Climate Change 2001: Impacts, Adaptation & Vulnerability, Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC).* Cambridge University Press, 1000 pp. [Available online at http://www.grida.no/climate/ipcc_tar/wg2/359.htm.]

McPhaden, M. J., and Coauthors, 1998: The tropical ocean–global atmosphere observing sytem: A decade of progress. *J. Geophys. Res.,* **103,** 14 169–14 240.

Misra, V., P. A. Dirmeyer, and B. P. Kirtman, 2003: Dynamic downscaling of seasonal simulations over South America. *J. Climate,* **16,** 103–117.

Murnane, R. J., M. Crowe, A. Eustis, S. Howard, J. Koepsell, R. Leffler, and R. Livezey, 2002: The weather risk management industry's climate forecast and data needs. *Bull. Amer. Meteor. Soc.,* **83,** 1193–1198.

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

Neelin, J. D., D. S. Battisti, A. C. Hirst, F.-F. Jin, Y. Wakata, T. Yamagata, and S. E. Zebiak, 1998: ENSO theory. *J. Geophys. Res.,* **103,** 14 261–14 290.

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

——, 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.

——, 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.,* **128,** 747–774.

——, and D. L. T. Anderson, 1994: The prospects for seasonal forecasting. *Quart. J. Roy. Meteor. Soc.,* **120,** 755–793.

——, and J. Shukla, 2000: Editorial to DSP/PROVOST special issue. *Quart. J. Roy. Meteor. Soc.,* **126,** 1989–1990.

——, Č. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quart. J. Roy. Meteor. Soc.,* **126,** 2013–2034.

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.

Peng, P., A. Kumar, and H. van den Dool, 2002: An analysis of multimodel ensemble prediction for seasonal climate anomalies. *J. Geophys. Res.,* **107,** 4710, doi:10.10129/2002JD002712.

Pielke, R., Jr., and R. E. Carbone, 2002: Weather impacts, forecasts, and policy. *Bull. Amer. Meteor. Soc.,* **83,** 393–403.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton, 2000: The impact of new physical parameterizations in the Hadley Centre climate model: HadAM3. *Climate Dyn.,* **16,** 123–146.

Richardson, C. W., 1981: Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resour. Res.,* **17,** 182–190.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–668.

Roeckner, E., 1996: The atmospheric general circulation model ECHAM-4: Model description and simulation of present-day climate. Max-Planck-Institut für Meteorologie Tech. Rep. 218, Hamburg, Germany, 90 pp. [Available from Max-Planck Institut für Meteorologie, Bundesstr. 55, D-20146 Hamburg, Germany.]

Rummukainen, M., J. Räissänen, B. Bringfelt, A. Ullerstig, A. Omstedt, U. Willen, U. Hansson, and C. Jones, 2001: A regional climate model for northern Europe: Model description and results from the downscaling of two GCM control simulations. *Climate Dyn.,* **17,** 339–359.

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

Singh, U., and P. K. Thornton, 1992: Using crop models for sustainability and environmental quality assessment. *Outlook Agric.,* **21,** 209–218.

Stern, P. C., and W. E. Easterling, Eds., 1999: *Making Climate Forecasts Matter.* National Academy Press, 175 pp.

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.

Straus, D. M., and J. Shukla, 2000: Distinguishing between the SST-forced variability and internal variability in mid-latitudes: Analysis of observations and GCM simulations. *Quart. J. Roy. Meteor. Soc.,* **126,** 2323–2350.

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

Terres, J.-M., 1999: The Crop Growth Monitoring System implemented by JRC/ARIS unit for the information needs of the EC DG VI—Agriculture. *Proc. Int. Symp. on Modeling Cropping Systems,* Cataluña, Spain, European Society for Agronomy, Division Agroclimatology and Agronomic Modeling, EX 18746/C ORA/PRO 60194R, 261–264.

——, and P. Cantelaube, 2003: DEMETER project, second annual report. Contribution from JRC, application to crop yield forecasts. Institute for Environmental and Sustainability Tech. Rep., Joint Research Centre of the European Commission, S.P.I.03.20, 58 pp.

Thomson, M. C., T. N. Palmer, A. P. Morse, M. Cresswell, and S. J. Connor, 2000: Forecasting disease risk with seasonal climate predictions. *Lancet,* **355,** 1559–1560.

van Diepen, C. A., and T. 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, Rep. EUR 16008 EN, 143–157.

von der Voet, P., C. A. van Diepen, and J. O. Voshaar, 1994: Spatial interpolation of daily meteorological data. Winand Staring Centre for Integrated Land, Soil and Water Research Rep. 53.3, 105 pp.

Vossen, P., 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, Rep. EUR 16008 EN, 21–51.

Wilby, R. L., and T. M. L. Wigley, 1997: Downscaling general circulation model output: A review of methods and limitations. *Prog. Phys. Geogr.,* **21,** 530–548.

——, and ——, 2000: Downscaling general circulation model output: A reappraisal of methods and limitations. Climate prediction and agriculture. *Proc. START/WMO Int. Workshop,* Geneva, Switzerland, WMO, 39–68.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences,* 1st ed. Academic Press, 467 pp.

Wolff, J. E., E. Maier-Reimer, and S. Legutke, 1997: The Hamburg Ocean primitive equation model. Deutsches Klimarechenzentrum Tech. Rep. 13, Hamburg, Germany, 13 pp. [Available from Model and Data Group c/o Max-Planck Institut für Meteorologie, Bundesstr. 55, D-20146 Hamburg, Germany.]

WHO, 2001: Malaria early warning systems: Concepts, indicators and partners. Roll Back Malaria Cabinet Project, World Health Organization, Geneva, Switzerland, 80 pp.

World Climate Research Programme, 1985: Scientific plan for the tropical ocean and global atmosphere program. World Meteorological Organisation Tech. Doc. WMO/TD-64, 146 pp.

Zebiak, S. E., and M. A. Cane, 1987: A model of El Niño–Southern Oscillation. *Mon. Wea. Rev.,* **115,** 2262–2278.