



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 D5.19

Scientific paper on the impact of the underestimates of Northern Hemisphere blocking in seasonal and climate change predictions of European winter precipitation

Due date of deliverable: month 36

Actual submission date: July 2007

Start date of project: 1 September 2004

Duration: 60 Months

Organisation name of lead contractor for this deliverable: ECMWF

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	PU
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)	

ENSEMBLES

Deliverable D5.19

Authors: Antje Weisheimer, Francisco Doblas-Reyes and Tim Palmer
(ECMWF)

Due: month 36

Scientific paper on the impact of the underestimates of Northern Hemisphere blocking in seasonal and climate change predictions of European winter precipitation

The following manuscript has been submitted to Bull. Amer. Meteorol. Soc.:

Towards Seamless Prediction: Calibration of Climate- Change Projections Using Seasonal Forecasts

T.N. Palmer, F.J. Doblas-Reyes, A. Weisheimer and M. Rodwell
European Centre for Medium-Range Weather Forecasts (ECMWF),
Reading, UK

Submitted to BAMS

Corresponding author: Dr T.N. Palmer

European Centre for Medium-Range Weather Forecasts (ECMWF)

Shinfield Park, RG2 9AX, Reading, UK

Tel: +44 (0) 118 9499600

Fax: +44 (0) 118 9869 450

Email: Tim.Palmer@ecmwf.int

Abstract

Trustworthy probabilistic projections of regional climate are essential for society to plan for future climate change, and yet, by the nonlinear nature of climate, finite computational models of climate are inherently deficient in their ability to simulate regional climatic variability with complete accuracy. How can we determine whether specific regional climate projections may be untrustworthy in the light of such generic deficiencies? A calibration method is proposed whose basis lies in the emerging notion of seamless prediction. Specifically, calibrations of ensemble-based climate-change probabilities are derived from analyses of the statistical reliability of ensemble-based forecast probabilities on seasonal timescales. The method is demonstrated by calibrating probabilistic projections from the multi-model ensembles used in the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC), based on reliability analyses from the seasonal-forecast DEMETER dataset. The focus in this paper is on climate-change projections of regional precipitation, though the method is more general.

Capsule

In a seamless prediction system, the reliability of coupled climate-model forecasts made on seasonal timescales can provide useful quantitative constraints for improving the trustworthiness of regional climate-change projections.

In weather and climate prediction circles, the phrase “Seamless Prediction” is very much in vogue - indeed the notion of seamless prediction lies at the heart of the World Climate Research Programme’s recent strategic framework (http://wcrp.wmo.int/pdf/WCRP_strategImpl_LowRes.pdf). However, what does this phrase mean, and what is its relevance for the practice of weather and climate forecasting?

As an illustration of the potential value of the concept of seamless prediction, we focus in this paper on projections of regional anthropogenic climate change (ACC) from the recent Fourth Assessment Report (AR4: Solomon et al. 2007, see also Appendix), focussing primarily on precipitation change. In AR4, integrations from the ensemble of the world’s climate models are illustrated in the form of probabilistic projections of climate change. This multi-model probabilistic representation allows some quantification of uncertainty in these climate projections. For example, one clearly cannot be confident about ACC (eg towards a drier or wetter climate) in regions where the models disagree. Conversely, one should, in principle, be more confident in regions where there is a reasonable consensus about a particular type of climate change (*Giorgi and Mearns, 2002*). If trustworthy, such probabilistic analyses clearly have great value for users of climate projections, eg in providing the basis for decisions on infrastructure investment to adapt to ACC.

Climate is a profoundly nonlinear system in which variability on different time and spatial scales interact (*Palmer, 1999*). Because of this, the global models used to make climate projections incorporate processes on as many different space and time scales as possible. Fig. 1 illustrates schematically some of the implications of such nonlinearity - in our view it is key to explaining the relevance of seamless prediction for studies of climate change. The figure shows a chain. One end of this chain represents humanity’s forcing of climate through emissions of greenhouse gases into the atmosphere. The other end of the chain represents the impact of this forcing in terms of regional climate change (temperature, precipitation, wind and so on). Each link of the chain represents a class of physical processes; the links are distinguished by the primary timescale on which the underlying processes act.

The first link on the left represents the class of fast diabatic processes acting on timescales of one day - for example, direct radiative forcings and the perturbations to the cloud systems arising directly from such radiative forcings.

As is well known, systematic changes in diabatic heating fields will perturb the planetary wave structure of the atmosphere, in both the tropics and the extratropics (*Hoskins and Karoly, 1981*). These changes in planetary wave structure are determined by the teleconnectivity of the atmosphere, and are described by patterns such as the Pacific/North-American and North Atlantic Oscillation patterns in the extratropics, and the Southern Oscillation pattern in the tropics. The timescales on which these perturbations in planetary-wave structure are established, given a persistent anomalous diabatic heating source, is typically of the order of ten days. These teleconnection patterns are represented by the second link of the chain.

On timescales of the order of hundred days, the oceans and land surface (the latter generally faster than the former) react to changes in atmospheric planetary wave structure. In some cases, these surface forcings, represented by the third link of the chain, will simply reinforce the overlying atmospheric circulation anomalies, causing such anomalies to become persistent on the seasonal timescale. In other cases, the induced changes (eg in ocean dynamics) will lead to modification of the circulation anomalies.

On timescales of a thousand days and longer, persistent circulation anomalies can lead to modifications in the cryosphere (fourth link), and in the biosphere (fifth link). Again, feedbacks from the cryosphere and biosphere can reinforce the underlying circulation anomalies, or modify them.

We claim that the all-pervasive nonlinearity of climate implies that the strength of this chain, that is to say, the accuracy with which the climate impact can be determined from the underlying climate forcing, is determined by the chain's weakest link. Hence, multi-model probabilistic projections of climate change will be untrustworthy if the models' representations of any of the processes in Fig 1 are systematically deficient. For example, consider the potential biospheric amplification of ACC on the century timescale that would arise if the tropical rainforests slowly became atmospheric carbon sources. Such feedback might occur if the frequency of occurrence of circulation patterns that divert or otherwise suppress precipitation-bearing systems over the rainforests were to increase significantly. If models have systematic deficiencies in simulating such circulation patterns, these models would not provide trustworthy quantitative estimates of the risk of such long-term biospheric feedbacks. As yet, we do not have a methodology for discounting probabilistic projections of ACC, if models show systematic weaknesses in variability associated with one or more links of the chain. (However, the notion that multi-model consensus is not a sufficient criterion for assessing forecast trustworthiness is itself not new; *Giorgi and Mearns, 2002; Räisänen, 2007*).

This is where the notion of seamless prediction can play a key role. It will be decades before climate-change projections can be verified; indeed, in probabilistic form, one could argue that they can never be verified. However, our basic premise, illustrated by the schematic in Fig. 1, is that there are fundamental physical processes in common to both seasonal forecast and climate change timescales. If essentially the same ensemble forecasting system can be validated probabilistically on timescales where validation data exist, ie on daily, seasonal and (to some extent) decadal timescales, then we can modify, or calibrate, the climate-change probabilities objectively using probabilistic forecast scores on these shorter timescales. The magnitude of this calibration reflects the weakness in those links of the chain common to both seasonal forecasting and climate change timescales. *Rodwell and Palmer (2007)* have presented an application of this idea, addressing the first link of the chain. It was shown that a class of model perturbations that led to very high climate sensitivity could be discounted due to their apparently poor representation of fast processes. If such results were replicated in a fully seamless system, the high end of the climate sensitivity probability distribution function could be discounted substantially. Even though the recent IPCC AR4 report (*Meehl et al., 2007*) acknowledges the possibility of constraining climate projections using shorter time-scale ensemble predictions for which verification data exist, a methodological illustration has not yet been put forward.

In this paper we discuss the potential value of the seamless prediction philosophy using reliability-diagram analyses of multi-model ensembles of seasonal climate forecasts of regional precipitation, thus addressing primarily the first three links of the chain. Here the word "reliability" needs some elaboration. Suppose a multi-model ensemble forecast predicts with 100% probability that it will be anomalously wet over some region in the coming winter. If it is not wet, then the forecast system is manifestly not reliable. However, more generally, if, on the occasions where an ensemble forecast predicts some climatic event with probability p , the event occurs in reality on a fraction q of times, then if p is sufficiently different from q in some statistically-meaningful sense, the ensemble forecast probabilities are not reliable (*Hsu and Murphy, 1986; Wilks, 1995; Mason, 2004*).

We claim here that the analysis of multi-model seasonal-forecast Reliability Diagrams (or Attributes Diagrams as they are sometimes known) provides a means to quantitatively discount the climate change probabilities in the light of diagnosed unreliability. We do this using calibration analysis, a well-known technique in weather forecasting and seasonal prediction (*Atger, 2003; Toth et al 2006*).

In practice the models used in our seasonal forecast study and those used in AR4, are broadly from the same class of coupled atmosphere/ocean global circulation models, but are not identical. Hence, the results discussed here are illustrative of what could be achieved in a truly seamless prediction system, ie in a system where the same models are used for seasonal prediction and climate projection. It is hoped that this paper will motivate the further development of seamless prediction systems worldwide.

Multi-model projections of future precipitation changes and systematic model deficiencies

The most complete and up-to-date source of information to quantify ACC is the public multi-model database of climate projections made for the AR4 of the IPCC (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php). Each model in the AR4 database has simulated the climate of the 20th and 21st Centuries forced by standard scenarios for projected concentrations of greenhouse gases. Throughout this paper, we refer to seasons in which precipitation falls above (below) the upper (lower) tercile of the corresponding 20th Century reference data distribution as wet (dry). Here we analyse multi-model multi-scenario ensemble ACC projections of the changing frequency of wet and dry seasons; see Appendix for details. The tercile thresholds are estimated from each individual model separately for the 20th Century reference period, and frequencies of exceedence in the 21st Century for each model are subsequently calculated based on these individual thresholds. This approach implies that the frequency of exceeding the threshold during the reference period is by definition 1/3, regardless of a) the differences of the nominal threshold values amongst the models, and b) their relation to the real world, that is observations. The alternative of basing projection probabilities on thresholds from observations is discussed below.

The ratio of frequencies of dry and wet events for two seasons (December-February, DJF, and June-August, JJA, respectively) at the end of the 21st Century relative to their reference frequencies in the 20th Century are illustrated in Fig. 2 for the entire globe.

For example, according to the AR4 ensemble, the frequency of wet winters will almost double at the end of the 21st Century compared to the 20th Century in many high-latitude northern hemisphere regions, consistent with earlier studies (<http://www.ukcip.org.uk>; *Palmer and Räisänen, 2002*). However, AR4 models are known to underestimate systematically the frequency of one of the key modes (or regimes) of variability of mid and high latitude flow: the persistent blocking anticyclone (*van Ulden and van Oldenborg, 2006*). This behaviour can also be seen in seasonal forecasts from the DEMETER (*Palmer et al, 2004*; see also Appendix) database, as illustrated in Fig. 3. The general underestimation of the frequency of wintertime blocking in 2-4 month seasonal ensemble integrations as estimated with the *Tibaldi and Molteni (1990)* index is evident.

This example of a systematic error common to all models was the motivation for this paper, and brings into focus the relevance of the “nonlinear chain” illustrated in Fig.1. Regional mid and high latitude precipitation is strongly affected by the presence of blocking activity, and models undersimulate such activity. Hence, if increased emissions of greenhouse gases force changes in tropical diabatic heating such as to increase the probability of occurrence of the blocking anticyclone, the high probabilities for increased mid and high-latitude precipitation in Fig. 2 cannot be considered trustworthy. But how should we discount quantitatively the AR4 probabilities in the light of these and other systematic model deficiencies?

Reliability analysis of seasonal forecasts

The first stage in our objective method for discounting climate-change probabilities in light of systematic model deficiencies is to analyse the statistical reliability of probabilistic seasonal climate forecasts for which validation data exist - clearly no such validation data exist on the climate-change timescale. The second stage is to use these reliability estimates to calibrate the “raw” ACC probabilities such as shown in Fig 2.

The strategy proposed here cannot be used to assess how accurately models simulate processes which occur on timescales much longer than the seasonal timescale (eg the last two links of the chain in Fig. 1). It is therefore important to stress that the strategy proposed here is viewed as necessary but not sufficient for the goal of obtaining trustworthy climate-forecast probabilities.

Here we analyse seasonal-mean forecasts based on the IPCC-class multi-model ensemble project DEMETER (*Palmer et al, 2004*: see also Appendix). Consider the binary events: precipitation exceeds the upper tercile ($E_P^+(x)$), or does not exceed the lower tercile ($E_P^-(x)$), at a particular grid point x . The corresponding events for temperature, $E_T^\pm(x)$, will also be considered. As for the multi-model ACC projections in Fig. 2, tercile thresholds based on each individual model are used; for all such events, a trivial climatological forecast has, by definition, a probability of 1/3. We illustrate DEMETER forecast reliability using reliability diagrams. These diagrams illustrate the conditional relative frequency of occurrence of events such as $E_P^\pm(x)$ as a function of their forecast probability, based on a discrete binning of many forecast probabilities taken over a specified geographic region $\langle x \rangle$. Fig. 4a-f illustrates such reliability diagrams for $E_P^\pm(x)$ for six selected standard (*Giorgi and Francisco, 2000*) land regions: Eastern North America, the Amazon Basin and Northern Europe in DJF, and South Asia, Central North America and Southeast Asia in JJA.

Ideally, the reliability curves in Fig. 4 should lie on the diagonals. In the idealised case of infinite sample and ensemble sizes, the diagonal line represents perfect probabilistic reliability where, from a sub-sample of cases where an event is forecast with probability p , the event occurs on a fraction p of occasions. If the reliability curve is mainly shallower than the diagonal, then the forecast system is overconfident. If a reliability curve were to be horizontal, then the frequency of occurrence of $E_P^\pm(x)$ would not depend on the forecast probabilities, ie would be the same for all forecast probabilities.

Of relevance for this study are those bins associated with non-climatological forecast probabilities. The population of such bins by the ensemble is an indication of potential predictability on the seasonal timescale, arising, for example, from low-frequency coupled ocean-atmosphere dynamics. As argued above, variability associated with these dynamical processes form part of the “chain” that links

greenhouse gas forcing to climatic impact. The reliability of these non-climatological forecast bins depends on how close they are to the diagonal. It can be seen from Fig. 4 that reliability varies with region and time of year.

Best-estimate linear regression lines have been fitted to the data in the reliability diagrams in Fig 4, taking into account the relative weight of the individual points according to their bin population. These are shown as red lines. In addition, a 10,000 bootstrap re-sampling procedure has been applied to estimate the effect on this regression line of sampling uncertainty due to finite sample size. The red shaded areas in the diagrams indicate the range of the linear regressions using \pm one standard deviation of the bivariate probability density function of the linear regression coefficients obtained from the bootstrap procedure. The larger this area, the more the regression line will depend on the volume of available seasonal forecast data, and hence the more this line may change as new seasonal forecast data becomes available. Based on the best-estimate linear regression, it can be seen from Fig 4, that, overall, reliability is very high for the Eastern North American region (a) and Amazon region (b) in DJF and for Southeast Asia (f) in JJA, but poor for the South Asian monsoon region (d) and the Central North American region (e) in JJA. For each of these five regions, the uncertainty in the linear regression line is relatively small. For the Northern European region in winter (c), the reliability is also relatively poor, but here there is much more uncertainty in the position of the linear regression line.

Estimation of the overall reliability of the non-climatological forecast bins can be made using a quadratic measure of probabilistic forecast error, the Brier skill score (BSS) (*Wilks, 1995; Mason, 2004*). Table 1 collates the BSS for both temperature and precipitation events for 21 standard global land regions (*Giorgi and Francisco, 2000*), including those shown in Fig 4. For almost all of the regions, BSS is not significantly negative for $E_T^\pm(x)$. On the other hand, for many of the regions and events, $BSS < 0$ for $E_P^\pm(x)$ at greater than 90% confidence. Note also the larger proportion of negative (in red) versus positive scores (in green) for precipitation when compared to temperature. For the chosen Eastern North American and Amazon regions in DJF and for Southeast Asia in JJA the BSS is positive at the 90% significance level. For the South Asian monsoon region and the Central North American region in JJA the BSS is negative at the 90% level. For the Northern European region the BSS is negative, but only at the 78% significance level. For this region, the biases associated with blocking undersimulation do not generate forecast unreliability at a highly significant level because this is also a region of relative unpredictability on the seasonal timescale. We discuss further these Brier scores in the next section.

Calibration of probabilistic climate-change forecasts

As mentioned above, reliability diagrams can be used to calibrate forecast probabilities in ensemble weather prediction. The method is illustrated schematically in Fig. 5. First, find the best-estimate regression line from the seasonal forecasts. Then consider the point on the regression line whose abscissa value corresponds to some specific raw forecast probability p_{raw} from the climate change ensemble. This point is then translated horizontally to the diagonal of the reliability diagram. The calibrated forecast probability $p_{calibrated}$ is given by the abscissa value of this translated point. For example in Fig 5, a raw probability of 80% probability becomes a calibrated probability of 50%. Both probabilities are higher than the climatological probability of 33.3%, but the calibrated probability is discounted to take into account the imperfect reliability of the seasonal forecasts.

In this paper we propose that if the same multi-model ensemble is used for seasonal prediction as for climate-change prediction, then the validation of probabilistic forecasts on the shorter timescale can be used to improve the trustworthiness of probabilistic predictions on the longer timescale. This improvement would come from assessing processes in common to both the seasonal forecast and climate projection timescales, such as the atmospheric response to sea surface temperatures. To reiterate, our basic premise, illustrated in Fig. 1, is that processes, such as air-sea coupling, that are relevant for the seasonal forecast problem, also play a role in determining the impact of some given climate forcing, on the climate system itself. The calibration technique provides a way of quantifying the weakness in those links of the chain common to both seasonal forecasting and climate change timescales. We illustrate the technique by calibrating the raw AR4 probabilities as shown in Fig. 2, using the regression lines from the DEMETER reliability diagrams, shown in Fig. 4. Based on Fig. 4 and Table 1, it can be deduced that, for regions where the BSS is significantly negative, the calibration procedure will modify the raw AR4 probabilities greatly. Conversely, where the BSS is significantly positive, the calibration procedure will not modify the AR4 probabilities. Hence, the AR4 raw temperature probabilities are not modified substantially by the calibration procedure. It is important to note that if BSS=0 because the DEMETER ensemble only has a single probability bin at climatological probability (implying lack of seasonal predictability), the corresponding AR4 probabilities will remain unchanged.

Fig 6 shows raw and calibrated AR4 changes in probabilities at the end of the 21st Century for the same six selected regions as in Fig. 4. For Eastern North America (a) and the Amazon (b) in DJF and Southeast Asia (f) in JJA, where the DEMETER forecasts are relatively reliable, the raw AR4 probabilities are not modified dramatically. On the other hand, for South Asia (d) and Central North America (e) in JJA, where the DEMETER forecasts were very unreliable, the calibration has a major impact on the probabilistic ACC projections: in the monsoon region there is a substantial discounting of the strong increase in upper tercile rainfall over Bangladesh, back towards a climatological probability of 1/3. For Central North America, the decrease in the probability of wet summers is strongly discounted back towards climatology. Over Northern Europe (c) in DJF the raw AR4 data indicate a decrease in the probability of dry winters; the calibration technique weakens this signal. However, as mentioned, of the six chosen regions, the calibration technique has weakest statistical significance in this region because the estimated level of seasonal predictability is relatively small. Nevertheless, because of the models' undersimulation of blocking anticyclones and the effect blocking anticyclones have on seasonal precipitation, we would expect the calibration technique to shift the probability of dry Northern European winters under climate change toward climatology, that is towards an increase in the frequency of dry winters with respect to the uncalibrated projection.

One caveat should be mentioned with regard to Fig. 6. The DEMETER models are not identical to the AR4 models, hence the DEMETER-based calibration cannot be assumed a priori to apply to the AR4 multi-model ensemble. For this reason, the results in Fig. 6 should be considered illustrative of what would be possible from a forecast system which was truly "seamless" across the range of weather and climate timescales. Here we note that multi-model seasonal forecast ensembles are beginning to be produced operationally with unified climate prediction models (eg EUROpean multi-model Seasonal to Inter-annual Prediction (EUROSIP) system, http://www.ecmwf.int/products/forecasts/seasonal/forecast/forecast_charts/eurosip_doc.htm). For calibration purposes, substantial back-integration datasets are being generated to allow "out-of-sample" testing of calibrations. In this sense, the ACC

calibration method we propose could be envisaged as being effectively cost free in future years.

Reliability and Calibration Based on Observed Thresholds

In the discussion above, we have based the estimation of model probabilities on individual model tercile thresholds, rather than on observed tercile thresholds. In doing this, results are empirically corrected for model bias in simulating the observed climatological mean state. For some users whose decision criteria depend on probability of exceeding absolute thresholds (eg freezing temperatures), such implicit correction might seem unacceptable. However, in this section we show that, using the calibration method proposed in this paper, such uncorrected climate change probabilities are utterly unreliable, even for regions which appeared relatively trustworthy using model thresholds.

In Fig. 7 a) is shown the frequency of dry boreal winters in the reference period of the AR4 integrations, where “wet” is now defined with respect to terciles from ERA40 (*Uppala et al, 2005*) reanalyses. The extent of the regions with frequencies different from the a priori 1/3 climatological probability is striking and illustrates the existence of significant biases in these models. In addition, the plot Fig. 7b) depicts a DEMETER seasonal forecast reliability diagram where no corrections are made to the direct model output and for which the probabilities are computed using the observed terciles from GPCP (*Adler et al, 2003*) as threshold. It can be seen that the seasonal-forecast reliability regression line is extremely flat and there is little uncertainty in this flatness. Using these lines for calibration, the AR4 probabilities would be discounted back to the 20th Century values, ie the calibrations would effectively imply that the AR4 integrations provide no useful information at all. For this reason, trustworthy probabilistic projections cannot be obtained in general, from model output whose percentiles have not first been corrected.

Conclusions

Climate change is widely recognised as one of the most serious problems facing humanity in the coming decades. In addition to reducing greenhouse gas emissions, society needs also to start planning for inevitable climate change. The issue of adaptation may require investment in new infrastructure, with potential costs of trillions of dollars worldwide. Regional probabilistic climate change projections enable informed decisions to be made regarding such investments. These projections are of value, provided the associated probabilities are trustworthy. However, as finite-dimensional estimates, models are inherently deficient in being able to simulate climate variability on all space and time scales. In this paper we gave one example of this, systematic undersimulation of the persistent blocking anticyclone. Such deficiencies can be mitigated by increased resolution but this may require substantial increases in available computing power not available for several years (*Palmer, 2005*). How can we discount the probabilities of climate change in the light of such deficiencies?

Here we have proposed a quantitative solution to this question using multi-model seasonal forecasts. It will be decades before climate-change projections can be verified; indeed, in probabilistic form, one could argue that they can never be verified. However, our basic premise, illustrated by the schematic chain in Fig. 1, is that there are fundamental physical/dynamical processes in common to both weather and seasonal forecasts on the one hand, and climate-change timescales on the other. If essentially the same ensemble forecasting system can be validated probabilistically on timescales where validation data exist, ie on daily, seasonal and (to some extent)

decadal timescales, then we can modify, or calibrate, the climate-change probabilities objectively using probabilistic forecast scores on these shorter timescales. The magnitude of this calibration reflects the weakness in those links of the chain common to both seasonal forecasting and climate change timescales.

We believe that the proposed technique will be valuable for quantitative decision making eg on infrastructure investment decisions for society to adapt to climate change. For example, in parts of the UK, new water reservoirs may be required to cope with the demands of modern society. However, any assessment for investing in new capacity will also take account of projections of wetter winters under ACC. Decisions on whether and how much extra reservoir water capacity is actually needed, depend critically on balancing the cost of building new capacity against the projected future risk of a run of dry winters. Such risk estimates multiply the societal hardships and economic damage of insufficient reservoir capacity by the probability of occurrence of periods of drought. The difference between risk estimates made using raw and calibrated probabilities could influence significantly such infrastructure investment decisions.

Our proposed method provides some justification for the development of seamless prediction systems across weather and climate timescales.

Appendices:

ACC analysis: The 14 AR4 models used here are the same as listed in Table 1 of *Weisheimer and Palmer (2005)*, except for the CSIRO-Mk3.0 model. The IPCC Special Report on Emission Scenarios (SRES) (*Nakicenovic and Swart, 2000*) suggests a wide range of different future emission scenarios, but offers no judgement as to the preference for any of the scenarios. We combine three standard emission scenarios (SRES A2, SRES A1B, and SRES B1) run by all 14 models into a multi-model multi-scenario ensemble. They correspond to relatively high (A2), moderate (A1B) and low (B1) greenhouse-gas concentrations at the end of the 21st century.

The analysis is performed globally on a grid-point basis for seasonal-mean precipitation data over a recent reference period (1971-1990) and a future climate-change (2081-2100) period applying equal weights to all models and scenarios. Firstly, all data have been interpolated on a common spectral T42 grid. Secondly, terciles for each model separately based on the control period distribution have been calculated and are subsequently used as thresholds for estimating the frequencies of wet and dry seasons in the multi-scenario ensemble over the climate change period. That is, the model frequencies of wet (dry) seasons in the control period are forced by construction to be 1/3. Similarly, the frequencies of wet (dry) seasons under climate change are calculated, again for each model separately, from the 20-year 3-member multi-scenario ensemble by computing the fraction of ensemble members and years for which the relevant model tercile is exceeded (not exceeded). The so-obtained frequencies from the 14 models are then combined into a multi-model multi-scenario frequency.

DEMETER data: The seasonal predictions used in this paper are based retrospective forecasts from the DEMETER (Development of a European Multi-model Ensemble Forecast System for Seasonal to Interannual Climate Prediction) project (*Palmer et al., 2004; Thomson et al., 2006*). The DEMETER system comprises seven 9-member IPCC-class coupled ocean-atmosphere global climate models with forecasts produced over the period 1980-2001 started four times each year. The seven models are from the following institutions: ECMWF, Centre National de Recherches Météorologiques (Météo-France, France), Met Office (The Met Office, UK), Laboratoire d'Océanographie et du Climat (LODYC), Centre Europeen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS), Max-Planck-Institut für Meteorologie (MPI), Instituto Nazionale Geofisica e Vulcanologia (INGV). More detailed information on the DEMETER project can be found on <http://www.ecmwf.int/research/demeter/>.

Data from all models have been interpolated on a common $2.5^\circ \times 2.5^\circ$ horizontal grid. Thresholds of the events $E_{P,T}^\pm(x)$ are taken from the sample terciles, separately for the verification and the multi-model data. Forecast probabilities are obtained as the relative number of ensemble members satisfying the event definition (anomalies above/below the upper/lower tercile), a simple method that assigns the same weight to every single model. Verification is taken from ERA-40 reanalyses (*Uppala et al., 2005*) for temperature and from GPCP (*Adler et al., 2003*) for precipitation (data available from <http://cics.umd.edu/~yin/GPCP/>).

Reliability diagrams and calibration method: Each forecast probability bin in the reliability diagrams in Fig. 4 and 7 is represented by a solid circle whose area is proportional to the bin sample size. Reliability "curves" are made by weighted linear regression and extrapolation beyond the extreme probability bins. The diagrams are generated using nine probability bins of width 1/9.

It can be asked whether such probabilistic calibration could be achieved from other strategies, not associated with seasonal forecasts. For example, each of the AR4 models has a 20th Century simulation, with observed 20th Century greenhouse-gas forcings. However, by design, such a multi-model ensemble would be incapable of reproducing observed interannual fluctuations in 20th Century climate and the corresponding reliability diagrams for seasonal-mean climate would collapse to a single value at the climatological frequency of occurrence of the chosen event. A reliability curve cannot be constructed from a single point on the reliability diagram. A second potential strategy would be to base a calibration on multi-model atmospheric simulations of the 20th Century with observed SSTs. However, it is now well known (*Wu and Kirtman, 2005; Wu et al., 2006*) that such 1-way coupling can produce erroneous results, particularly in the tropics.

Acknowledgements:

The work reported here is part of the EU-funded ENSEMBLES project (contract number GOCE-CT-2003-505539). The authors acknowledge helpful discussions with Dr T. Jung and the seasonal forecast section at ECMWF and considerable technical support from ECMWF staff and consultants.

References:

- Adler, R.F. et al., 2003: The Version 2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979 - present). *J. Hydrometeor.* **4**, 1147-1167.
- Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Weather Rev.*, **131**, 1509-1523.
- Giorgi, F. and R. Francisco, 2000: Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM, *Clim. Dyn.*, **16**, 169-182.
- Giorgi, F. and L.O. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "Reliability Ensemble Averaging" (REA) method, *J. Clim* **15**, 1141-1158.
- Hoskins, B.J. and D.J. Karoly, 1981: The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.* **38**, 1179-1196.
- Hsu, W.-R. and A.H. Murphy, 1986: The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting* **2**, 285-293.
- Mason, S. J., 2004: On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.* **132**, 1891-1895.
- Meehl, G.A., T.F. Stocker, W.D. Collins, P. Friedlingstein, A.T. Gaye, J.M. Gregory, A. Kitoh, R. Knutti, J.M. Murphy, A. Noda, S.C.B. Raper, I.G. Watterson, A.J. Weaver and Z.-C. Zhao, 2007: Global Climate Projections. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- Nakicenovic, N. and R. Swart, 2000: *Special Report on Emissions Scenarios*. Cambridge University Press, 612 pp.
- Palmer, T.N., 1999: Climate change from a nonlinear dynamical perspective. *J. Clim.* **12**, 575-591.
- Palmer, T.N. and J. Räisänen, 2002: Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature* **415**, 512.
- Palmer, T.N. et al., 2004: Development of a European multi-model ensemble system for seasonal to inter-annual prediction. *Bull. Amer. Meteor. Soc.* **85**, 853-872.
- Palmer, T N, 2005: Global Warming in a Nonlinear Climate: Can We Be Sure.? *Europhysics News*, 36, 42-46.
- Räisänen, J., 2007: How reliable are climate models? *Tellus A*, **59**, 2-29.
- Rodwell, M.J. and T.N. Palmer, 2007: Using numerical weather prediction to assess climate models. *Q.J.R.Meteorol.Soc.*, 133, 129-146.
- Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.). 2007. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- Thomson, M.C. et al., 2006: Malaria early warnings based on seasonal climate forecasts from multi-model ensembles, *Nature* **439**, 576 - 579.

- Tibaldi, S. and F. Molteni, 1990: On the operational predictability of blocking. *Tellus*, **42A**, 343-365
- Toth, Z., O. Talagrand and Y. Zhu, 2006: The attributes of forecast systems: a general framework for the evaluation and calibration of weather forecasts. *Predictability of Weather and Climate*, T.N. Palmer and Hagedorn, Eds., Cambridge University Press, 584-595.
- Uppala, S.M. et al., 2005: The ERA-40 re-analysis. *Quart. J. R. Meteorol. Soc.* **131**, 2961-3012. doi:10.1256/qj.04.176.
- van Ulden, A.P. and G.J. van Oldenburgh, 2006: Large-scale atmospheric circulation biases and changes in global climate model simulations and their importance for climate change in Central Europe. *Atmos. Chem. Phys.* **6**, 863-881.
- Weisheimer, A. and T.N. Palmer, 2005: Changing frequency of occurrence of extreme seasonal temperatures under global warming. *Geophys. Res. Lett.* **32**, L20721, doi:10.1029/2005GL023365.
- Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*, Academic Press, 464 p.
- Wu, R., and B.P. Kirtman, 2005: Role of Indian and Pacific Ocean air-sea coupling in tropical atmospheric variability. *Clim. Dyn.* **25**, 155-179.
- Wu, R., Kirtman, B.P. and K. Pegion, 2006: Local air-sea relationship in observations and model simulations. *J. Climate*, **19**, 4914-4932.

Table Caption:

Table 1:

Forecast quality of the DEMETER multi-model seasonal forecasts in terms of Brier Skill Scores (BSS) for near-surface temperature and precipitation in JJA and DJF for 21 standard land regions (multiplied by 100). The scores for $E_{T,P^\pm}(x)$ have been computed over the forecast period 1980-2001 using seasonal means from 1-month lead ensembles started on the 1st of May/November. Bold underlined numbers indicate scores with a probability $p \geq 0.9$ that a random sample based on a 10,000 bootstrap re-sampling procedure would yield $BSS < 0$ (significantly negative) or $BSS > 0$ (significantly positive).

Figure Captions:

Fig. 1:

A schematic figure illustrating that the link between climate forcing and climate impact involves processes acting on different timescales. The whole chain is as strong as its weakest link. The use of a seamless prediction system allows probabilistic projections of climate change to be constrained by validations on weather, or seasonal forecast timescales.

Fig. 2:

The changing frequency of a) dry DJF, and b) wet JJA under ACC. Ratio of the frequencies of the AR4 multi-model multi-scenario ensemble near the end of the 21st Century relative to their 20th Century reference frequencies.

Fig. 3:

Mean blocking frequency in the ERA-40 analyses (black) and the seven DEMETER models as a function of longitude. Northern Hemisphere DJF mean blocking frequency estimated using the *Tibaldi and Molteni (1990)* index from 1-month lead time 9-member ensemble forecasts started on the 1st of November over the period 1980-2001. The models used are ECMWF (red), Météo-France (blue), Met Office (grey), LODYC (green), CERFACS (pink), INGV (cyan) and MPI (orange). The dots in the upper part of the plot indicate the longitudes where the model climatology is not significantly different from the verification data, using a two-sample test based on 500 bootstrap estimates.

Fig. 4:

Reliability diagrams for DEMETER multi-model seasonal forecasts for selected standard land regions. The data have been calculated over the forecast period 1980-2001 for $E_P^\pm(x)$ in DJF/JJA as indicated in the sub-panel titles using 1-month lead ensembles started on the 1st of November/May for DJF/JJA. The area of the red solid circles is proportional to the bin population. The blue horizontal and vertical lines indicate the climatological frequency of the event in the observations and forecasts, respectively. The black dashed line separates skilful from unskilful regions in the diagram: points with forecast probabilities smaller (larger) than the climatological frequency which fall below (above) this line, contribute to positive BSS; otherwise they contribute negatively to the BSS. Red shaded areas indicate the uncertainty of the regression line estimation based on a 10,000 bootstrap re-sampling procedure, see text for details.

Fig. 5:

Schematic sketch of the calibration method. The uncalibrated raw forecast probabilities are corrected by precisely the amount that would be needed to bring the points in the reliability diagram to the diagonal, see text for details.

Fig. 6:

Impact of the calibration on regional AR4 projections of changing frequency of wet and dry seasons under ACC. The increase in frequency of wet/dry DJF/JJA, as in Fig. 2, before (left) and after (right) calibration is shown for the same land regions and events used in Fig. 4.

Fig. 7:

Multi-model AR4 IPCC frequencies of exceeding the tercile thresholds from observations in the reference period 1971-1990: a) dry DJF b) DEMETER Reliability diagram for dry DJF over Eastern North America based on observed thresholds.

Region	2m Temperature				Precipitation			
	JJA		DJF		JJA		DJF	
	$E_T^-(x)$	$E_T^+(x)$	$E_T^-(x)$	$E_T^+(x)$	$E_p^-(x)$	$E_p^+(x)$	$E_p^-(x)$	$E_p^+(x)$
Australia	<u>10.7</u>	<u>10.1</u>	1.3	-0.4	-1.3	-2.5	-3.1	-3.6
Amazon Basin	<u>14.4</u>	9.1	<u>23.4</u>	<u>25.7</u>	2.2	2.1	<u>9.5</u>	<u>8.9</u>
Southern South America	<u>8.5</u>	<u>8.2</u>	-1.2	1.8	<u>7.8</u>	5.0	-0.7	-2.8
Central America	<u>12.1</u>	<u>9.9</u>	<u>14.8</u>	6.3	2.6	-0.7	8.7	8.5
Western North America	<u>6.5</u>	<u>7.7</u>	3.9	2.3	3.2	<u>5.5</u>	-0.6	0.0
Central North America	-4.1	-3.6	<u>-7.5</u>	0.3	-1.8	<u>-7.0</u>	3.7	5.3
Eastern North America	0.6	5.7	4.1	9.5	<u>-4.5</u>	<u>-8.3</u>	<u>9.2</u>	6.0
Alaska	3.0	2.1	0.0	-0.7	-0.1	0.3	2.4	4.9
Greenland	3.6	4.2	<u>8.0</u>	5.8	<u>-1.4</u>	-0.5	-2.1	-2.0
Mediterranean Basin	<u>7.6</u>	<u>10.7</u>	3.2	3.2	-0.5	0.1	1.6	-0.9
Northern Europe	-4.4	-4.2	4.8	2.9	-1.0	1.9	-1.1	-0.9
Western Africa	<u>10.4</u>	<u>11.8</u>	<u>18.1</u>	<u>17.2</u>	-1.6	-2.0	<u>-4.9</u>	<u>-3.5</u>
Eastern Africa	<u>12.6</u>	5.8	<u>13.3</u>	<u>10.3</u>	0.1	-0.3	1.2	0.6
Southern Africa	5.6	-1.1	<u>15.9</u>	<u>15.7</u>	0.7	-1.2	5.4	3.6
Sahara	<u>7.6</u>	<u>7.4</u>	6.9	3.9	<u>-2.6</u>	<u>-4.8</u>	<u>-2.7</u>	<u>-2.7</u>
Southeast Asia	10.7	5.9	8.7	<u>18.1</u>	<u>14.7</u>	<u>10.3</u>	3.4	2.5
East Asia	<u>4.7</u>	<u>7.9</u>	<u>10.8</u>	<u>10.0</u>	0.6	-1.0	-1.6	-0.9
South Asia	4.9	<u>13.1</u>	<u>7.6</u>	<u>8.6</u>	-1.6	<u>-3.0</u>	2.0	0.5
Central Asia	0.8	3.8	1.3	-0.4	0.5	0.1	-3.1	-3.6
Tibet	<u>10.7</u>	<u>10.1</u>	<u>23.4</u>	<u>25.7</u>	-1.1	0.0	<u>9.5</u>	<u>8.9</u>
North Asia	<u>14.4</u>	9.1	-1.2	1.8	-1.3	-2.5	-0.7	-2.8

Table 1

Figures

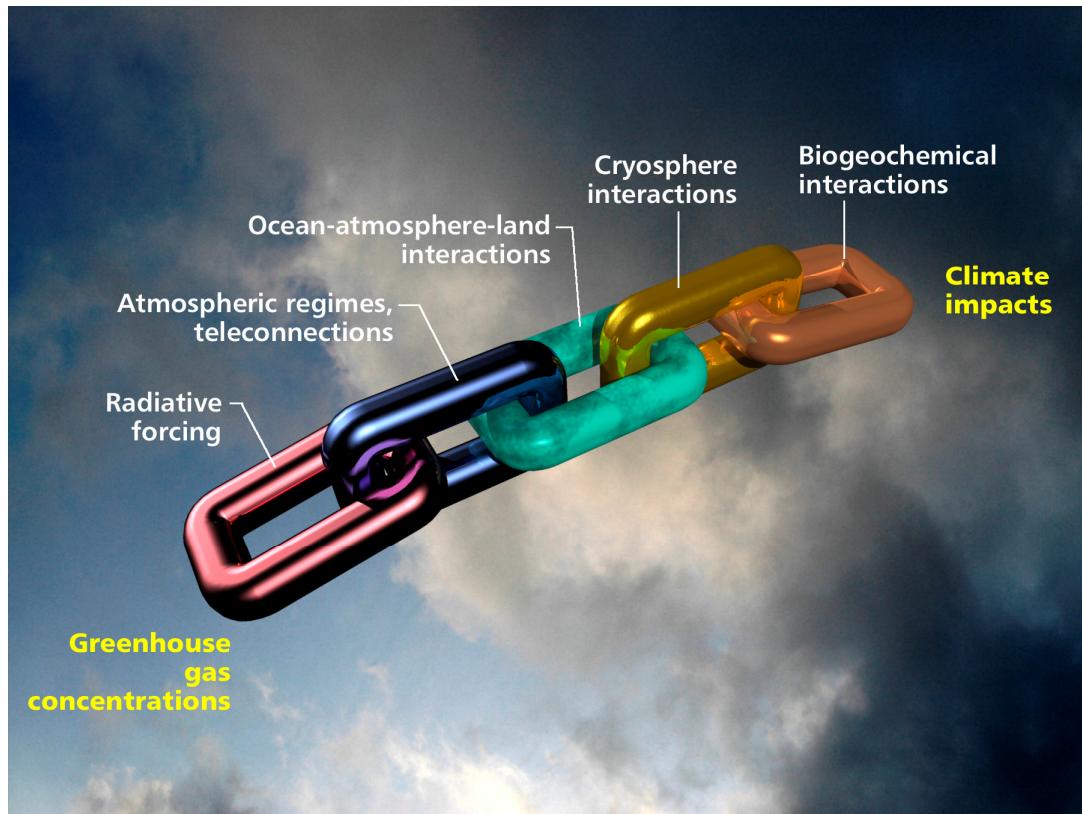
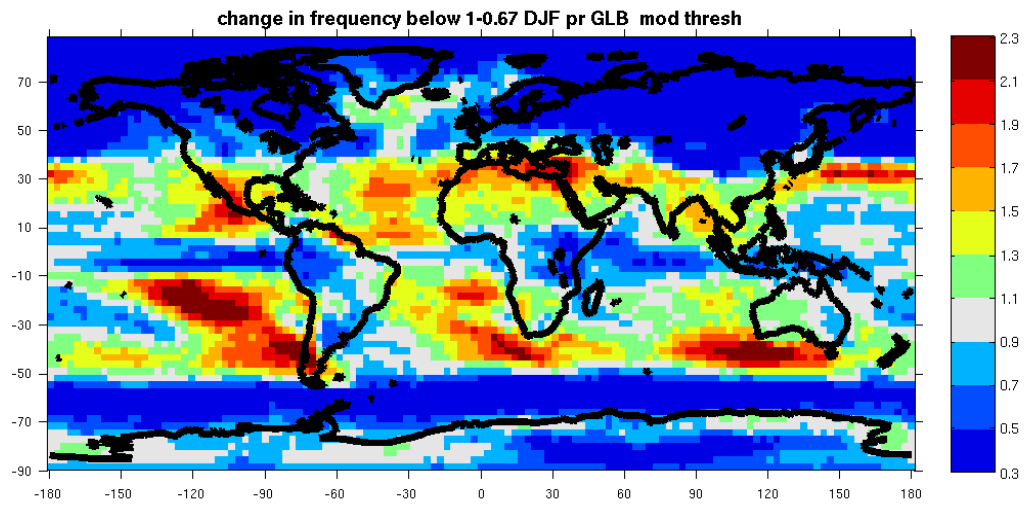


Fig. 1

a)



b)

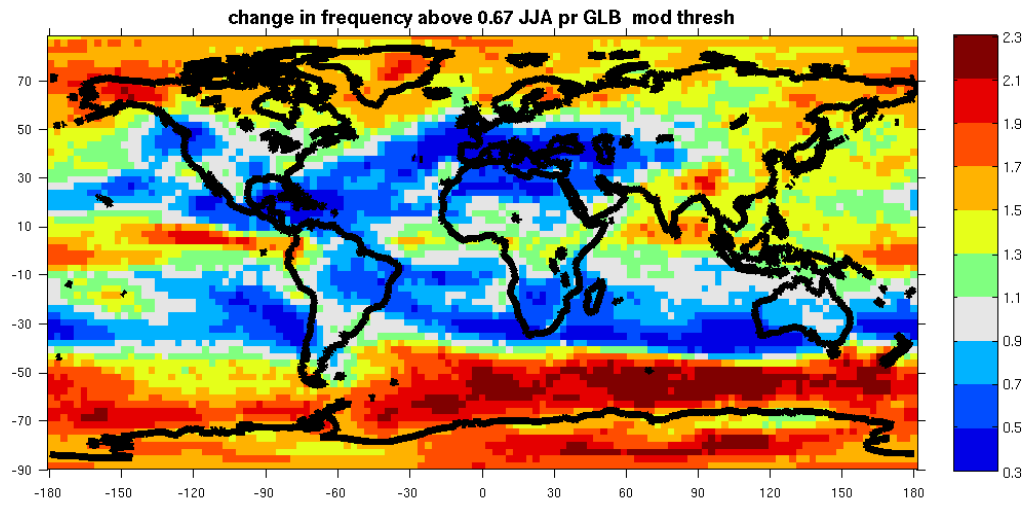


Fig. 2

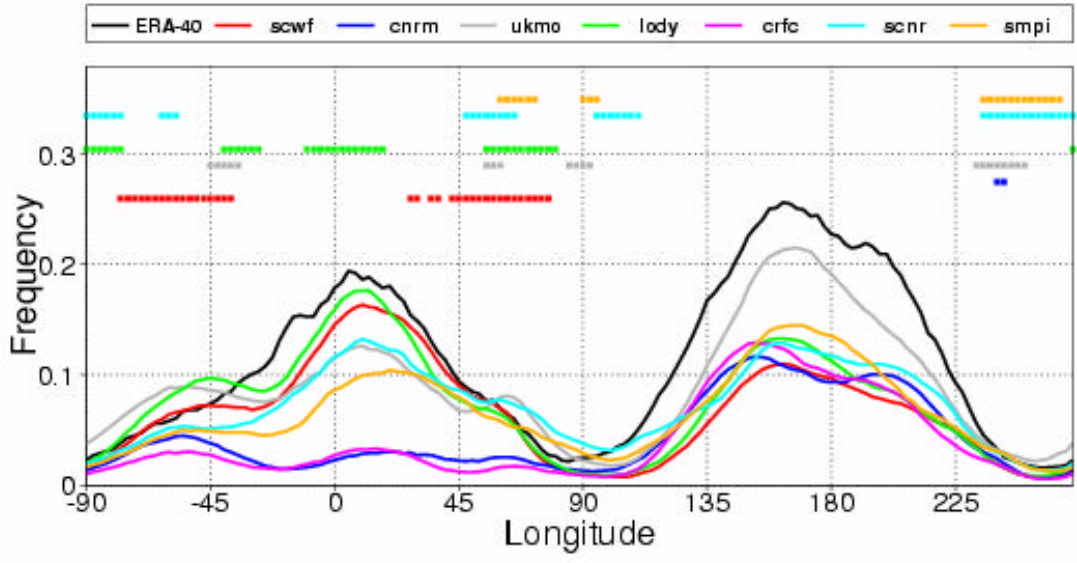
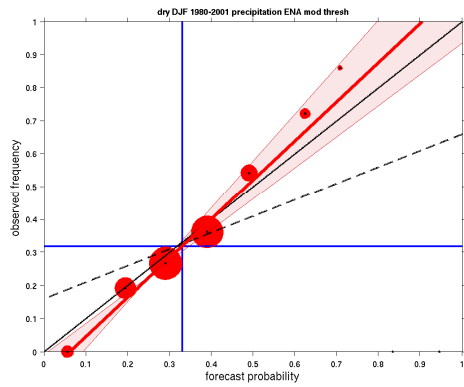
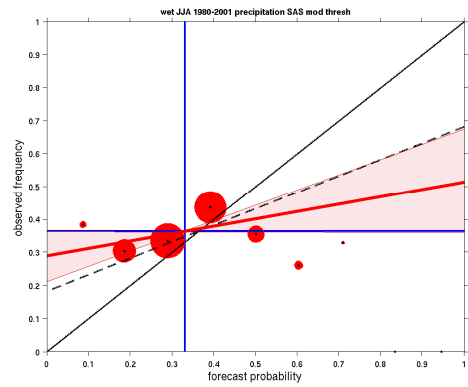


Fig. 3

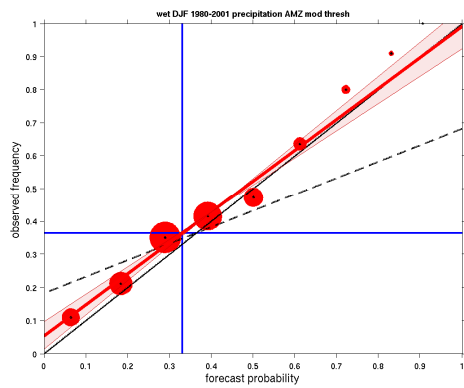
a) Eastern North America dry DJF



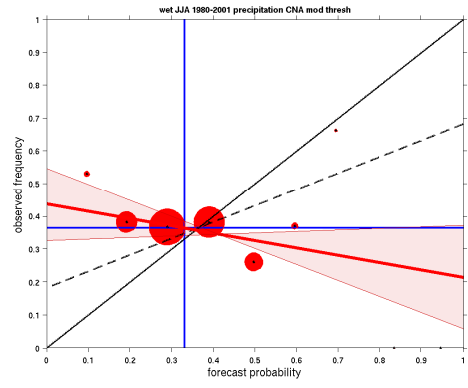
d) South Asia wet JJA



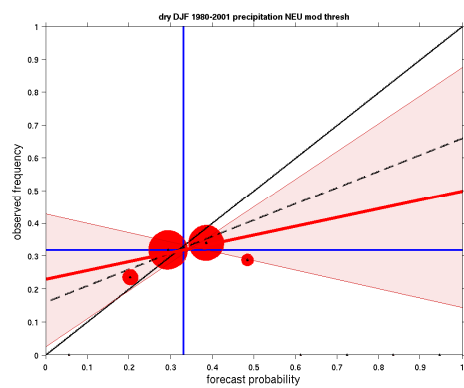
b) Amazon Basin wet DJF



e) Central North America wet JJA



c) Northern Europe dry DJF



f) Southeast Asia dry JJA

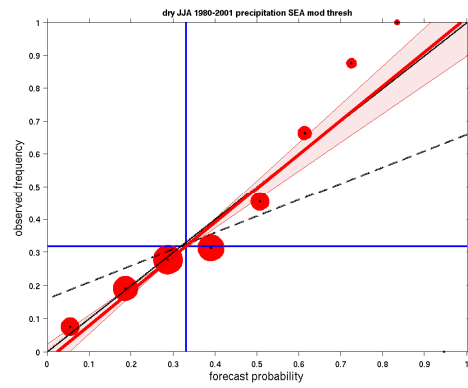


Fig. 4

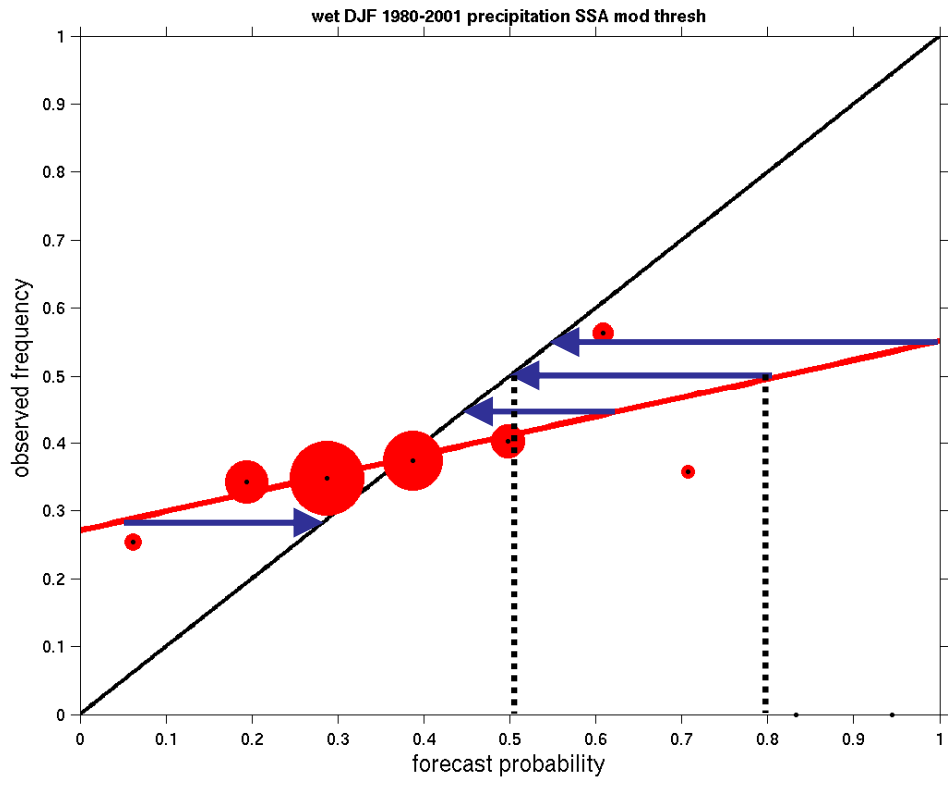
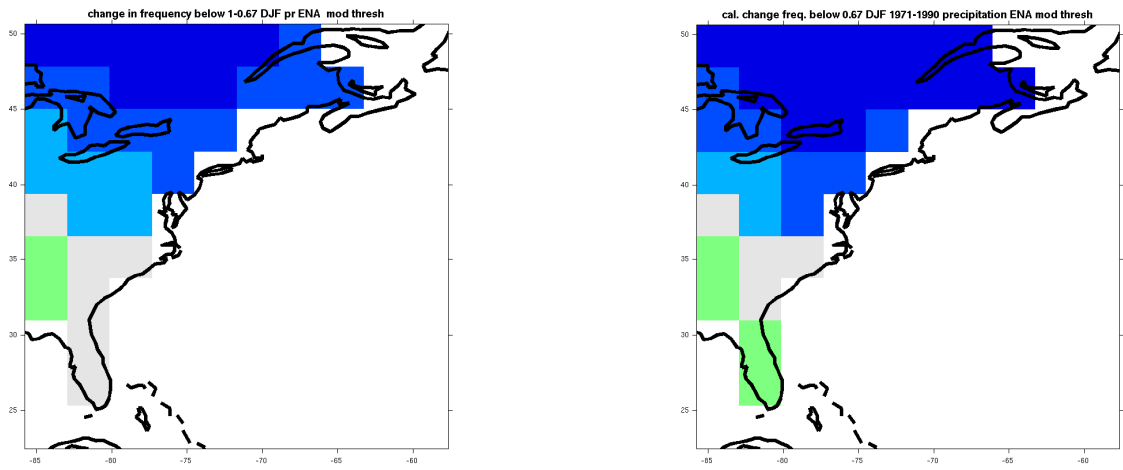
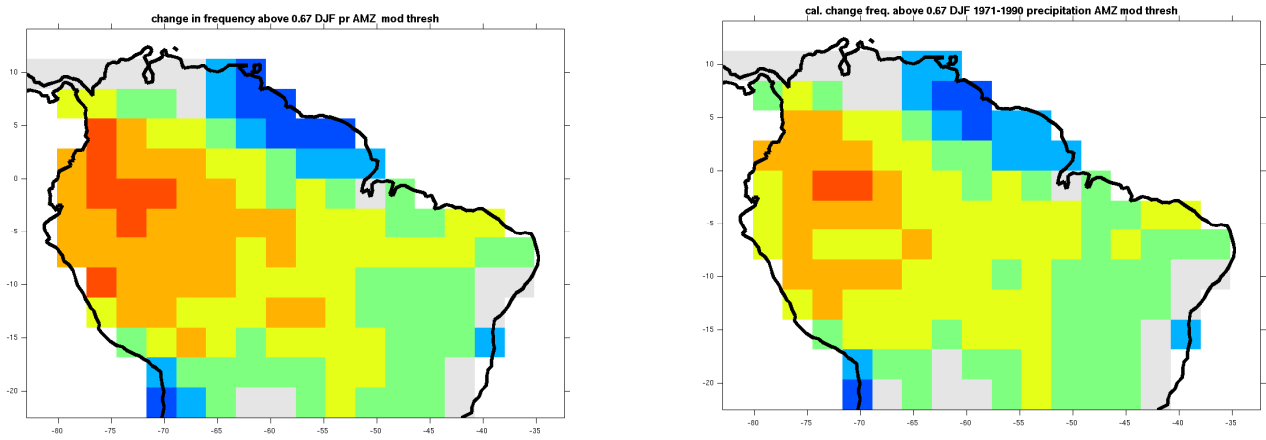


Fig. 5

a)



b)



c)

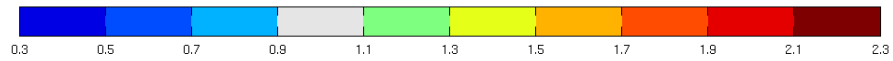
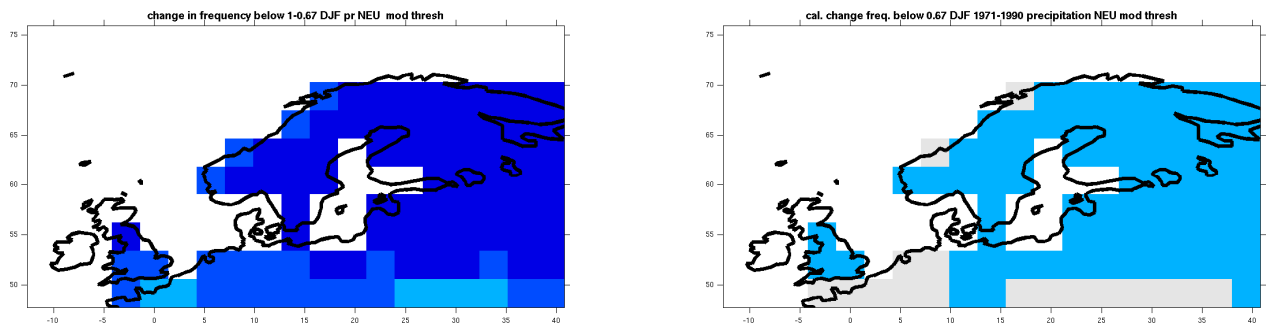
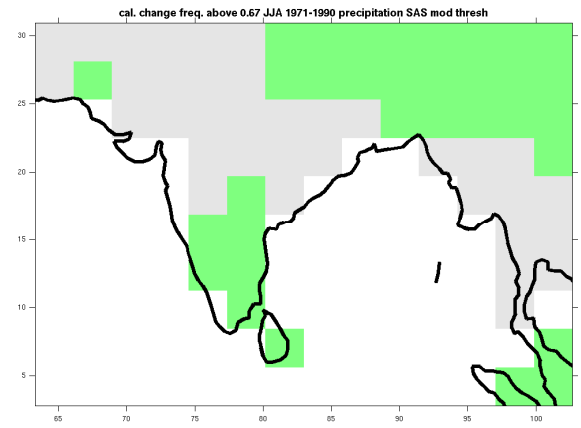
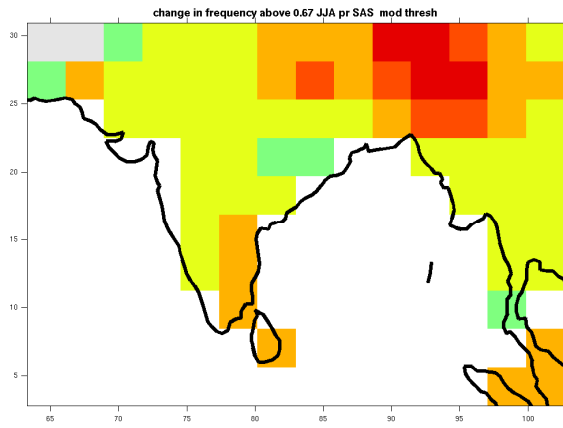


Fig. 6

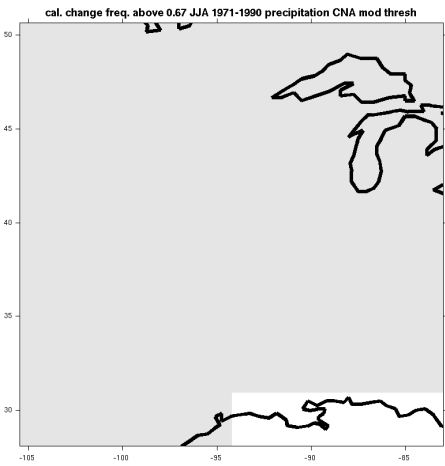
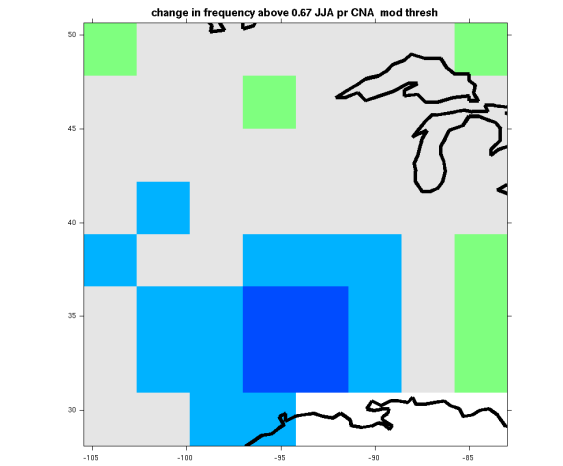
d)

d)



e)

e)



f)

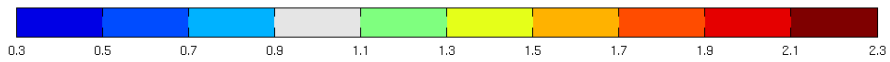
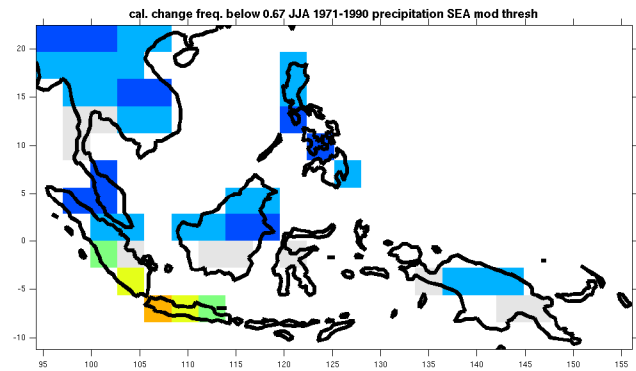
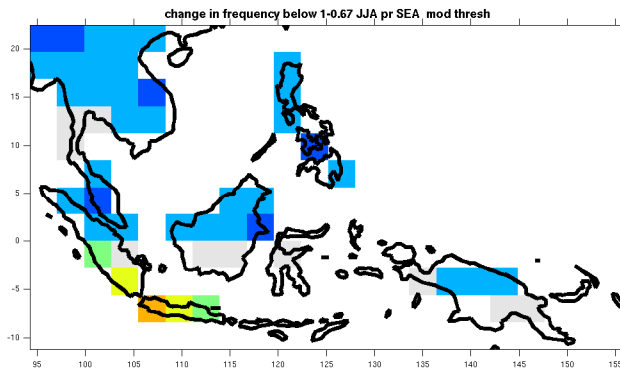
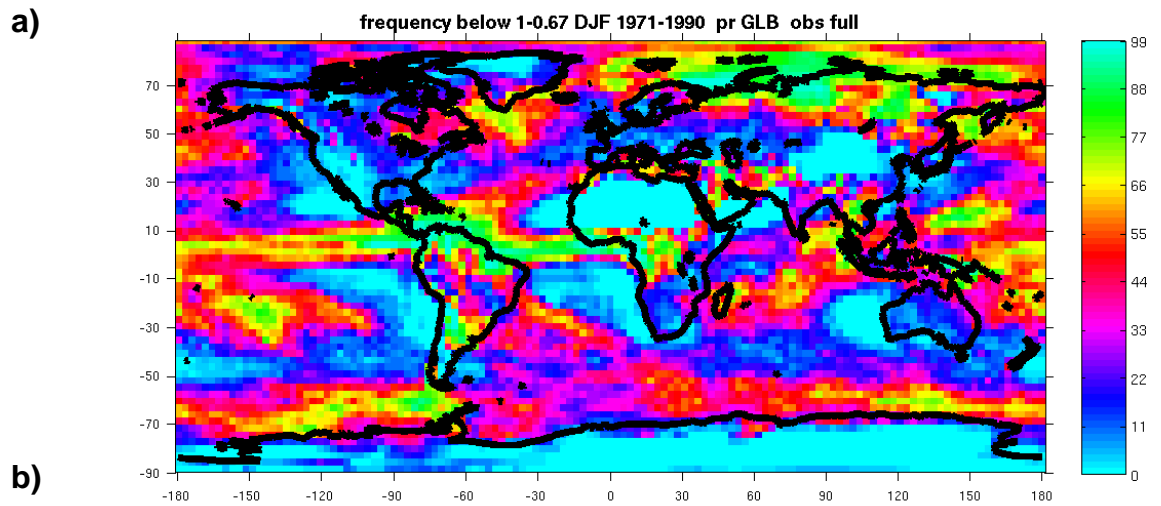


Fig. 6 (cont.)



b)

Eastern North America dry DJF

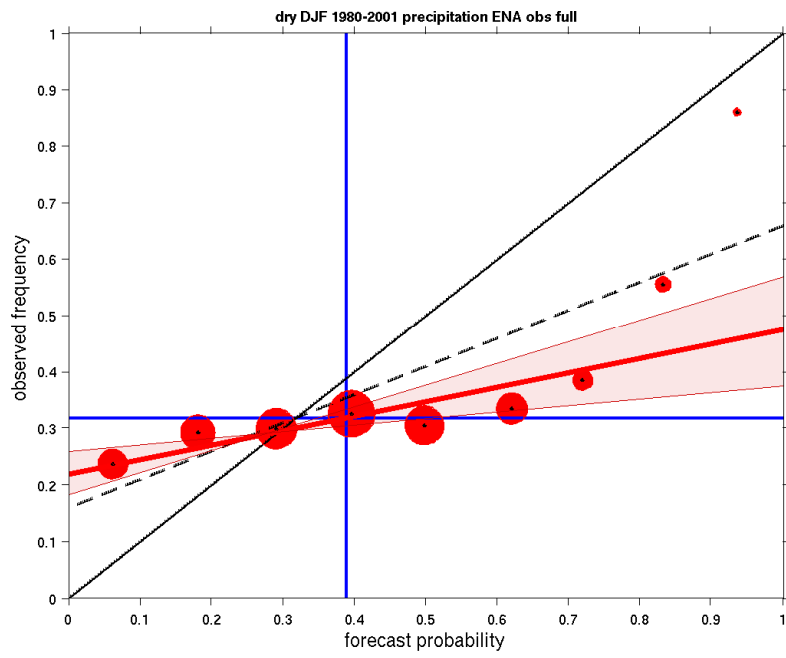


Fig. 7