Assessment of Representations of Model Uncertainty in Monthly and Seasonal Forecast Ensembles

Antje Weisheimer

ECMWF, Reading, UK
Antje.Weisheimer@ecmwf.int
and
National Centre for Atmospheric Sciences (NCAS),
Department of Physics, Atmospheric, Oceanic and Planetary Physics, Oxford University, UK
Weisheimer@atm.ox.ac.uk

Abstract

The probabilistic skill of ensemble forecasts for the first month and the first season of the forecasts is assessed, where model uncertainty is represented by the a) multi-model, b) perturbed parameters, and c) stochastic parameterisation ensembles. The main foci of the assessment are the Brier Skill Score for near-surface temperature and precipitation over land areas and the spread-skill relationship of sea surface temperature in the tropical equatorial Pacific. On the monthly timescale, the ensemble forecast system with stochastic parameterisation provides overall the most skilful probabilistic forecasts. On the seasonal timescale the results depend on the variable under study: for near surface temperature the multi-model ensemble is most skilful for most land regions and for global land areas. For precipitation, the ensemble with stochastic parameterisation most often produces the highest scores on global and regional scales. Our results indicate that stochastic parameterisations should now be developed for multi-decadal climate predictions using earth-system models.

1. Introduction

The inevitable approximations needed to solve the equations of weather and climate are a major source of forecast error and uncertainty on all timescales, from hours to centuries and longer. Over recent years, the multi-model ensemble has emerged in weather (TIGGE, *Bougeault et al.*, 2010), seasonal (DEMETER, *Palmer et al.*, 2004) and climate (CMIP, *Solomon et al.*, 2007) prediction, as a pragmatic tool for representing the effects of model uncertainty. However, such ensembles are limited by the number of models available and their assumed independence, and moreover there is no prior guarantee that the available models faithfully represent true model uncertainty. The first of these problems, that of limited ensemble size, is mostly solved in the alternative approach of perturbing free sub-grid scale parameters within a single model framework (*Murphy et al.*, 2004, *Stainforth et al.*, 2005). However, the second problem, that of ensuring true model uncertainty is properly represented, is even more acute in the perturbed parameter framework, since uncertainty in the structural form of the parameterisations is not addressed. Some of these issues are illustrated in recent analyses of multi-model and perturbed parameter ensembles (*Masson and Knutti*, 2011).

A third approach to representing model uncertainty has emerged in recent years and relies on the idea of stochastic parameterisation (*Palmer*, 2001, *Palmer and Williams*, 2009, Berner at al., 2011). In this approach, the underlying deterministic sub-grid bulk-formulae are replaced by an inherently stochastic formulation, recognising that the problem of representing sub-grid tendencies as a function

of the resolved variables may not be consistent with underlying scaling symmetries of the dynamical equations or with observations of power law structure in the real atmosphere.

In this paper we assess the skill of the different representations of model uncertainty (multi-model ensemble MME, perturbed parameter ensemble PPE, stochastic physics ensemble SPE) as outlined above, using IPCC-class global coupled ocean atmosphere models, in monthly and seasonal forecast mode. These ensembles were made as part of the EU-FP7 ENSEMBLES project, which promoted the concept of probabilistic and seamless prediction across a range of timescales from seasons to a century. We focus on the analysis of probabilistic predictions of precipitation and near-surface temperature over land areas on time scales of one month to one season, and on seasonal forecasts of sea surface temperature (SST) in the tropical equatorial Pacific.

2. Description of models and experiments

A co-ordinated set of retrospective seasonal forecasts (re-forecasts) using the MME, the PPE and the SPE was carried out to study the impact on forecast skill of these three approaches to represent model uncertainty. The ENSEMBLES MME comprises five coupled atmosphere-ocean general circulation models developed quasi-independently in Europe and initialised using realistic estimates of the observed states (*Weisheimer et al.,* 2009a). Each model was run from an ensemble of nine initial conditions which results in an overall MME size of 45 members.

The PPE seasonal re-forecasts were generated with the UK Met Office Decadal Prediction System (DePreSys, *Smith et al.,* 2007, *Doblas-Reyes et al.*, 2009, *Smith et al.,* 2010) by perturbing poorly constraint atmospheric and surface parameters. Eight model variants with simultaneous perturbations to 29 parameters (see www.ecmwf.int/research/EU_projects/ENSEMBLES/table_experiments/pert_param_desc.html) and one standard unperturbed version of the model were used in the nine-member PPE re-forecasts.

For the nine-member SPE, two different stochastic physical parameterisation schemes were applied to the atmospheric part of ECMWF's coupled seasonal forecast model using version IFS-CY35R2/HOPE T159L62/1°. The stochastically perturbed parameterisation tendency scheme (*Buizza et al.,* 1999; *Palmer et al.,* 2009) applies univariate Gaussian perturbations to the wind, temperature and humidity tendencies of physical processes in the form of multiplicative noise with a smoothly varying pattern in space and time. A two-scale version of the perturbations with a shorter characteristic spatio-temporal scale on the order of 6 hours and 500 km together with a longer scale of 30 days and 2500 km has been used. The stochastic backscatter scheme (*Shutts, 2005*; *Palmer et al.*, 2009) is based on the idea of backscatter of kinetic energy from unresolved scales. It is formulated in terms of a spectral stream function forcing field estimated from the numerical, convective and orographically induced dissipation rates and uses vertical phase correlation. For comparison, a set of control re-forecasts with the ECMWF model without any stochastic physics parameterisations has also been generated (CTRL).

The re-forecasts comprise the ENSEMBLES stream 2 set of experiments for boreal summer (June - August, JJA) and winter (December - February, DJF) seasons initialised on 1 May and 1 Nov, respectively, over the re-forecast period 1991-2005.

3. Temperature and precipitation forecast quality over land

Here we look at forecast skill for predicting the binary events that monthly or seasonal-mean temperature/precipitation anomalies exceed the upper tercile (i.e. warm/wet events) or does not exceed the lower tercile (i.e. cold/dry events) of the bias-corrected model climatological distribution for a set of land regions. As a probabilistic measure of skill we use the Brier skill score (BSS, *Jolliffe and Stephenson*, 2003). Positive (negative) BSS values indicate (no) forecast skill compared with trivial forecasts of climatological probabilities of 1/3. A modified version of the BSS for hypothetical infinite ensemble sizes (*Ferro*, 2007) has been used to minimise the effect on the skill measure of sampling uncertainty due to the different ensemble sizes of the three systems. The verification over the re-forecast period was carried out in cross-validation mode and confidence intervals for the BSS were estimated from a 1,000 sample bootstrapping with replacement (for further detail of the computation see *Palmer et al.* (2008) and *Doblas-Reyes et al.* (2009)). ERA-40 (*Uppala et al.*, 2005) and GPCP (*Adler et al.*, 2003) have been used as verification data for temperature and precipitation, respectively.

The mean skill over all (global) land areas in terms of BSS for temperature and precipitation forecasts on lead times of 1 month and 2-4 months is summarised in Table 1. The SPE system scores highest for almost all temperature and precipitation events during the first month of the forecasts. The results also clearly show that the MME gives the highest scores for cold and warm temperature events in JJA and DJF. For precipitation, however, it is the SPE system that scores best for dry and wet JJA and wet DJF events. In the case of dry DJF events the PPE has the highest scores. In no situation (warm/cold, wet/dry, monthly/seasonal) does the control ensemble without representation of model uncertainty perform best. A two-sampled one-sided t-test applied to the bootstrap resamples of each event in Table 1 supported the hypothesis that the highest BSSs are statistically significantly larger than the second largest BSSs with all p-values < 0.02.

A similar analysis has been performed for a set of 21 standard land regions. By way of summary, Figures 1 and 2 show which of the four forecasting systems has highest BSS scores during the first month and first season of the forecast respectively. Figure 1 indicates the forecasting system that produces the overall highest BSS during the first month of the forecasts for cold May (a), warm November (b), dry May (c) and wet November (d) events. Here, bold colors indicate BSS>0 and light colors stand for BSS<0. As can be seen, the best-performing system varies with the variable, event and season. For the events shown in Fig. 1, the SPE scores highest in the majority of regions, in agreement with the findings for global land areas in Table 1. The ECMWF control ensemble CTRL produces the highest BSS especially for large parts of Eurasia in November, while the MME is superior for cold May conditions in the tropics and during November over Africa.

Lead time: 1 month	T2m				precip			
	MAY		NOV		MAY		NOV	
	cold	warm	cold	warm	dry	wet	dry	wet
CTRL	0.147	0.148	0.126	0.148	0.044	0.061	0.058	0.075
MVE	0.178	**0.195**	0.141	0.159	0.085	0.079	0.080	0.099
PPE	0.059	0.054	−0.012	0.033	0.031	0.009	0.031	0.000
SPE	**0.194**	0.192	**0.149**	**0.172**	**0.104**	**0.118**	**0.095**	**0.114**
Lead time: 2-4 months	T2m				precip			
	JJA		DJF		JJA		DJF	
	cold	warm	cold	warm	dry	wet	dry	wet
CTRL	−0.024	−0.002	−0.011	0.063	−0.032	−0.020	0.037	0.042
MVE	**0.084**	**0.082**	**0.037**	**0.090**	0.023	0.030	0.041	0.039
PPE	0.004	0.046	−0.001	0.064	0.013	0.006	**0.046**	0.035
SPE	0.059	0.054	0.019	0.076	**0.037**	**0.037**	0.040	**0.062**

Table 1: *Brier Skill Scores for predicting eight events of global land area 2m temperature and precipitation for the four forecasting systems. Bold figures indicate the system with the highest score. The highest scores for all events are significantly different from the second highest scores with p-values < 0.02, see text for details.*

Corresponding results for seasonal-mean temperature and precipitation events on forecast lead times of 2-4 months are displayed in Fig. 2. Although the performance of the individual systems varies and no clear "winner" can be identified, there is a tendency for the MME to produce highest scores for temperature events and for the SPE to produce highest scores for precipitation events. These conclusions are supported by the findings of Table 1 where we identified MME and SPE as the forecasting systems producing the highest scores for global land areas for temperature and precipitation, respectively.

One potential reason for the different relative performance of SPE for near-surface temperature and precipitation lays in the vertical structure of the physical tendency perturbations with no perturbations applied in the lowest ~300 m of the atmosphere. The impact of this would be expected to be larger for near-surface temperature than it would for precipitation which is more linked to processes higher up in the atmosphere.

The Brier score can be decomposed into the three components reliability, resolution and uncertainty (*Murphy*, 1973) that describe certain attributes of the forecasts. In order to understand why some of the forecasting systems perform better than others, it would be desirable to be able to compare the reliability and resolution components for the different systems. However, all these statistical measures strongly depend on the size of the ensemble system (see *Doblas-Reyes et al.*, (2010)) even though some attempts were discussed to de-bias the BSS (*Müller et al.*, 2005; *Weigel et al.*, 2007). To our knowledge, the only methodology to overcome the problem is the one suggested by Ferro (*2007*) which derives an analytical expression of the BSS for a hypothetical infinite ensemble size. Unlike the full BSS, we know of no method to correct the reliability and resolution components of the skill score for ensemble size.

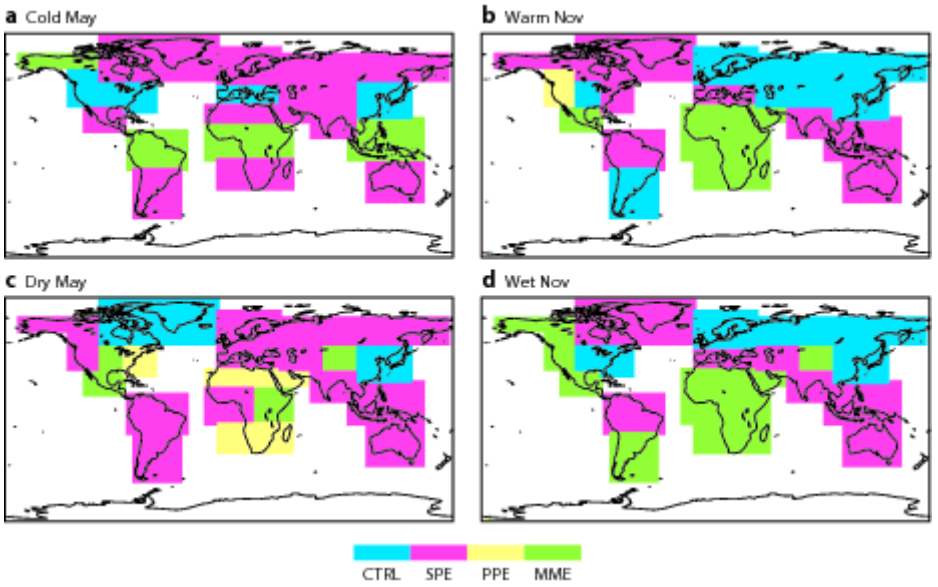

***Figure 1**: Probabilistic skill comparison for cold May (a), warm November (b), dry May (c) and wet November (d) between the ENSEMBLES multi-model ensemble (MME), perturbed physics ensemble (PPE), stochastic physics ensemble (SPE) and the control simulation for the stochastic physics ensemble (CTRL) on forecast lead time 1 month. The forecasting system with the highest Brier Skill Score is indicated with bold (light) colors for skill scores larger (smaller) than zero. Definition of the regions after Giorgi and Francisco (2000).*

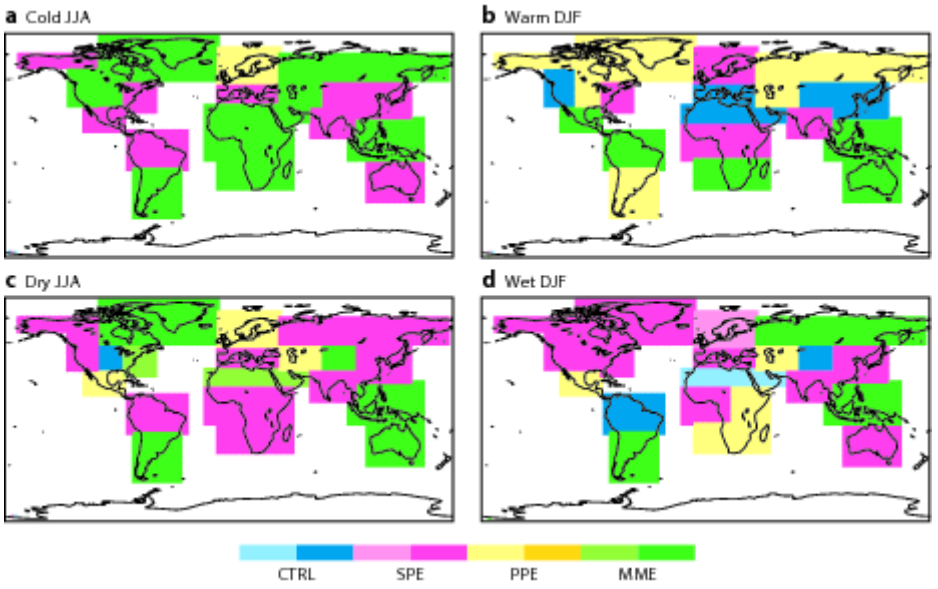

***Figure 2**: As in Fig 1 but for forecast lead times of 2-4 months and the following events: cold JJA (a), warm DJF (b), dry JJA (c) and wet DJF (d).*

4. Tropical Pacific SST re-forecasts

The El Niño/Southern Oscillation (ENSO) in the tropical Pacific is the prime example of a coupled atmospheric and oceanic phenomenon on seasonal and inter-annual timescales and a skill comparison between the MME, PPE, SPE and CTRL in terms of predicting SSTs in the Niño3 region (5°S–5°N, 150°W–90°W) has been carried out. Figure 3 shows the temporal evolution over forecast time of the ensemble-mean root-mean square error (RMSE) and ensemble spread of the bias-corrected re-forecast anomalies. For comparison, results from a simple statistical persistence forecast model are also plotted. In a perfect ensemble, over a large number of ensemble forecasts, the RMSE of the ensemble mean would equal the ensemble spread about the ensemble mean. A general feature of single-model ensembles is that their ensemble spread is substantially smaller than their RMSE, that is, each individual ensemble is underdispersive, or overconfident (see Fig 2 from *Weisheimer et al.* (2009a) for the spread-skill relationship of the five individual model ensembles contributing to the MME). Figures 3a and 3c show this deficiency for the CTRL and PPE systems (neither take account of any structural model uncertainties).

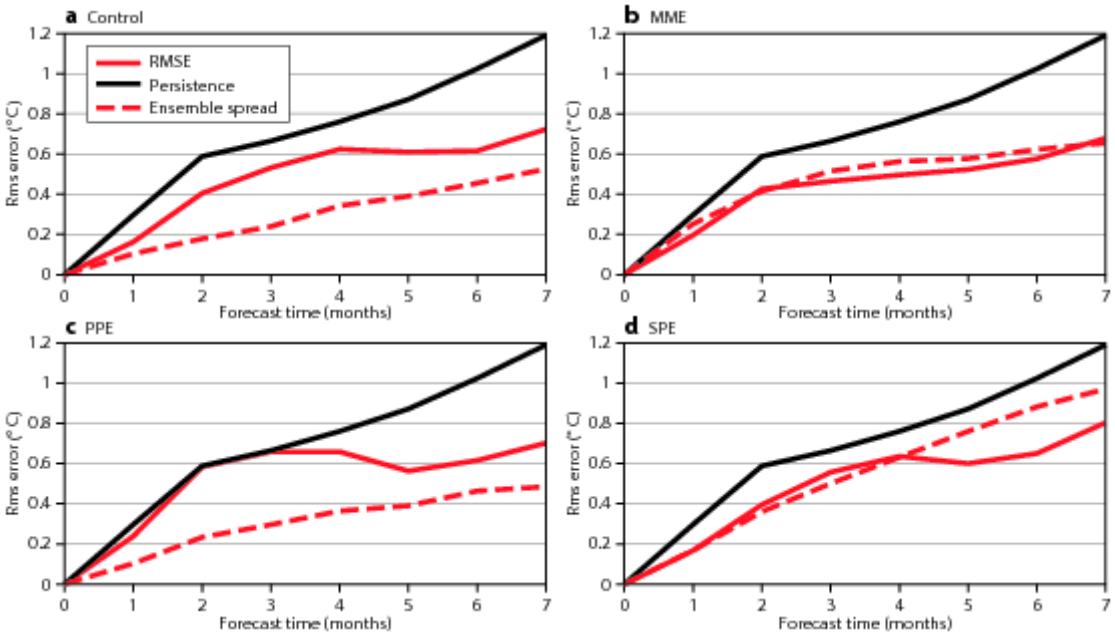

***Figure 3**: Skill comparison for predicting Niño3 SST anomalies in a) the control simulation (CTRL), b) the ENSEMBLES multi-model ensemble (MME), c) the perturbed physics ensemble (PPE) and d) the stochastic physics ensemble (SPE) showing the ensemble mean RMSE (red solid), the ensemble spread (red dashed) and the RMSE of a simple persistence reference forecast (black dash-dotted) as a function of forecast lead time.*

As has been demonstrated in numerous studies (e.g., *Palmer et al.*, 2004; *Weigel et al.*, 2008), the multi-model combination effectively reduces the RMSE while the ensemble spread is increased leading to overall improved skill. For the MME SSTs this leads to an almost perfect match between the RMSE and spread (Fig 3b). In contrast, the CTRL and PPE reveal a substantial mismatch between the RMSE and ensemble spread on all lead times. Similar results have been found for the PPE over

the tropics and for a longer re-forecast set (*Weisheimer et al.*, 2009b). SPE performs better than the CTRL system (Fig. 3d) due to a pronounced increase of the ensemble spread so that a good spread-skill relationship is obtained on forecast lead times of 1 to 4 months. The spread increase also leads to more reliable forecasts and thus better probabilistic skill scores.

5. Conclusions

A goal for the development of a forecasting system, for either weather or climate time scales, is the provision of reliable predictions. This necessarily implies that predictions must take explicit account of inherent uncertainties in the prediction process, and therefore be probabilistic in nature. A key source of uncertainty in the prediction of weather, and even more climate, arises from the computational approximations needed to solve the underlying equations of motion. Here, following an earlier analysis by *Doblas-Reyes et al.* (2009), we compare the performance in monthly and seasonal forecast mode, of three different approaches to the representation of model uncertainty: the multi-model ensemble, the perturbed parameter ensemble, and the method of stochastic parameterisation. We also compare these with a default ensemble with no representation of model uncertainty. On the monthly timescale the results are fairly clear-cut that the system with stochastic parameterisation generally provides the most skilful probabilistic forecasts and the system without representation of model uncertainty never provides the best forecasts. On the seasonal timescale the results are more mixed: the multi-model ensemble provides more often the best forecasts of near surface temperature, whilst the stochastic parameterisation ensemble provides more often the best forecasts of precipitation. This result may indicate the need to extend the notion of stochastic parameterisation into the ocean and land surface model.

In discussing the PPE system, it should be noted that no corresponding control ensemble based on the same underlying model as PPE but with fixed parameter values was available for comparison. Hence the results above should not be interpreted as implying that the PPE method is overall worst of the three techniques studied – the relative poor results from PPE (e.g. in the first month) could reflect that the underlying model itself was relatively poor. In fact, it would seem plausible to speculate that it may be possible to systematically outperform an MME using a single model system that combined aspects of PPE and SPE. Also, it might be possible that an ensemble forecast system which combines all three approaches studied here may prove optimal, but an analysis of this is beyond the scope of this paper.

The representation of model uncertainty in climate change projections is particularly problematic as there is as yet little verification data to assess potential representations. As a result, the concept of seamless prediction has arisen (*Palmer et al.*, 2008, *Hazeleger et al.*, 2009), that climate models can and should be tested in weather and seasonal prediction mode. The notion of seamless prediction suggests that the results presented here may be relevant on longer multi-decadal timescales and that stochastic parameterisations should now be developed for multi-decadal climate predictions using earth-system models.

Acknowlegements

The presentation I gave at the workshop in June 2011, and thus this contribution to the proceedings of the workshop, are based on *Weisheimer et al. (2011).* I would like to thank the following colleagues for their contributions to this study: Francisco Doblas-Reyes, Tim Stockdale, Thomas Jung, Glenn Shutts, Hannah Cloke, Florian Pappenberger and Tim Palmer.

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.

Alessandri, A. et al. (2011), Evaluation of probabilistic quality and value of the ENSEMBLES multi-model seasonal forecasts: comparison with DEMETER, *Mon. Wea. Rev* **139(2),** 581-607.

Berner, J., S.-Y. Ha, J.P. Hacker, A. Fournier and C. Snyder (2011), Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972-1995.

Bougeault, P. et al. (2010), The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072.

Buizza, R., M. Miller and T.N. Palmer (1999), Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*,**125**, 2887-2908.

Doblas-Reyes, F.J., et al. (2009), Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts, *Q.J.R.Meteorol.Soc.* **135**, 1538-1559.

Doblas-Reyes, F.J., et al. (2010), Forecast quality assessment of the ENSEMBLES seasonal-to-decadal Stream 2 hindcasts, ECMWF Tech. Memo. 621, 45pp.

Ferro, C.A.T. (2007), Comparing probabilistic forecasting systems with the Brier Score. *Wea. Forecasting*, **22**, 1076-1088.

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.

Hazeleger, W., C. Jones, R. McGrath and J. Hesselbjerg-Christensen (2009), EC-Earth: a seamless prediction approach to earth system modelling. *IOP Conf. Series: Earth and Environmental Science 6 052002*, doi:10.1088/1755-1307/6/5/052002.

Jolliffe, I.T. and D.B. Stephenson (2003), Forecast Verification: A Practitioner's Guide in Atmospheric Science. *Wiley and Sons*, London, UK.

Masson, D. and R. Knutti (2011), Climate model genealogy, *Geophys. Res. Lett.*, **38**, L08703, doi:10.1029/2011GL046864.

Müller, W.A. et al. (2005), A debiased ranked probability skill scores to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Climate*, **18**, 1513-1523.

Murphy, J.M. et al (2004), Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768-772.

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

Palmer, T.N. et al. (2004), Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), *Bull. Am. Meteorol. Soc.*, **85**, 853-872.

Palmer, T.N., F. Doblas-Reyes, A. Weisheimer, and M. Rodwell (2008), Reliability of climate change projections of precipitation: Towards "seamless" climate predictions. *Bull. Amer. Meteor. Soc.*, **89**, 459-470.

Palmer, T.N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G.J. Shutts, M. Steinheimer and A. Weisheimer (2009). Stochastic parametrization and model uncertainty. *ECMWF Tech. Memo.* **598**, 42pp.

Palmer, T.N. and P. Williams (eds.) (2009), Stochastic Physics and Climate Modelling, Cambridge University Press, 496 pages.

Shutts, G. J. (2005). A kinetic energy backscatter algorithm for use in ensemble prediction systems, *Quart. J. Roy. Meteorol. Soc.*, **131**, 3079–3102.

Smith, D., S. Cusack, A. Colman, C. Folland, G Harris and J. Murphy (2007), Improved surface temperature prediction for the coming decade from a global climate model, *Science,* **317**, 796-799.

Smith, D. R. Eade, N. Dunstone, D. Fereday, J. Murphy, H. Pohlmann, A. Scaife (2010), Skilful multi-year predictions of Atlantic hurricane frequency, *Nature Geoscience*, **3**, 846-849.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.) (2007), 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.

Stainforth, D. et al. (2005), Uncertainty in predictions of the climate response to rising levels of greenhouse gases, *Nature*, **433**, 403-406.

Uppala, S.M. et al (2005), The ERA-40 reanalysis. *Q. J. R. Meterol. Soc.*, **131**, 2961-3012.

Weigel, A.P., M.A. Liniger and C. Appenzeller (2007), The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118-124.

Weisheimer, A., et al. (2009a), ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions – Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs, *Geophys. Res. Lett.*, **36**, L21711, doi:10.1029/2009GL040896.

Weisheimer, A., F.J. Doblas-Reyes and T.N. Palmer (2009b), Model uncertainty in seasonal to decadal forecasting – insight from the ENSEMBLES project, *ECMWF Newsletter*, **122**, 21-26.

Weisheimer, A., T.N. Palmer and F. Doblas-Reyes (2011), Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles. *Geophys. Res. Lett.*, **38**, L16703, doi:10.1029/2011GL048123.