Centre for
Climate Change
Economics and Policy

The Munich Re Programme: *Evaluating the Economics of Climate Risks and Opportunities in the Insurance Sector*

Grantham Research Institute on
Climate Change and
the Environment

Probabilistic skill in ensemble seasonal forecasts

Leonard A. Smith, Hailiang Du, Emma B. Suckling and Falk Niehörster

27th February 2014

Centre for Climate Change Economics and Policy
Working Paper No. 168
Munich Re Programme Technical Paper No. 20
Grantham Research Institute on Climate Change and the Environment
Working Paper No. 151

The Grantham Foundation for the Protection of the Environment
Global Green Growth Institute
E·S·R·C Economic & Social Research Council
Munich RE
University of Leeds
LSE The London School of Economics and Political Science

The Centre for Climate Change Economics and Policy (CCCEP) was established by the University of Leeds and the London School of Economics and Political Science in 2008 to advance public and private action on climate change through innovative, rigorous research. The Centre is funded by the UK Economic and Social Research Council and has five inter-linked research programmes:

1. Developing climate science and economics
2. Climate change governance for a new global deal
3. Adaptation to climate change and human development
4. Governments, markets and climate change mitigation
5. The Munich Re Programme - Evaluating the economics of climate risks and opportunities in the insurance sector (funded by Munich Re)

More information about the Centre for Climate Change Economics and Policy can be found at: http://www.cccep.ac.uk.

The Munich Re Programme is evaluating the economics of climate risks and opportunities in the insurance sector. It is a comprehensive research programme that focuses on the assessment of the risks from climate change and on the appropriate responses, to inform decision-making in the private and public sectors. The programme is exploring, from a risk management perspective, the implications of climate change across the world, in terms of both physical impacts and regulatory responses. The programme draws on both science and economics, particularly in interpreting and applying climate and impact information in decision-making for both the short and long term. The programme is also identifying and developing approaches that enable the financial services industries to support effectively climate change adaptation and mitigation, through for example, providing catastrophe insurance against extreme weather events and innovative financial products for carbon markets. This programme is funded by Munich Re and benefits from research collaborations across the industry and public sectors.

The Grantham Research Institute on Climate Change and the Environment was established by the London School of Economics and Political Science in 2008 to bring together international expertise on economics, finance, geography, the environment, international development and political economy to create a world-leading centre for policy-relevant research and training in climate change and the environment. The Institute is funded by the Grantham Foundation for the Protection of the Environment and the Global Green Growth Institute, and has five research programmes:

1. Global response strategies
2. Green growth
3. Practical aspects of climate policy
4. Adaptation and development
5. Resource security

More information about the Grantham Research Institute on Climate Change and the Environment can be found at: http://www.lse.ac.uk/grantham.

This working paper is intended to stimulate discussion within the research community and among users of research, and its content may have been submitted for publication in academic journals. It has been reviewed by at least one internal referee before publication. The views expressed in this paper represent those of the author(s) and do not necessarily represent those of the host institutions or funders.

This paper is currently under review for the Quarterly Journal of the Royal Meteorological Society.

Probabilistic skill in ensemble seasonal forecasts*

Leonard A. Smith, Hailiang Du, Emma B. Suckling and Falk Niehörster

Centre for the Analysis of Time Series,
London School of Economics, London WC2A 2AE, UK
E-mail: L.Smith@lse.ac.uk

February 26, 2014

Abstract

Operational seasonal forecasting centres employ simulation models to make probability forecasts of future conditions on seasonal to annual lead times. Skill in such forecasts is reflected in the information they add to purely empirical statistical models, or to earlier versions of simulation models. An evaluation of seasonal probability forecasts from the DEMETER and the ENSEMBLES multi-model ensemble experiments is presented. Two particular regions are considered (Nino3.4 in the Pacific and Main Development Region in the Atlantic); these regions were chosen before any spatial distribution of skill were examined. The ENSEMBLES models are found to have skill against the climatological distribution on seasonal time scales; for models in ENSEMBLES which have a clearly defined predecessor model in DEMETER the improvement from DEMETER to ENSEMBLES is discussed. Due to the long lead times of the forecasts and the evolution of observation technology, the forecast-outcome archive for seasonal forecast evaluation is small; arguably evaluation data for

*Under review for Quarterly Journal of the Royal Meteorological Society

seasonal forecasting will always be precious. Issues of information contamination from in-sample evaluation are discussed, impacts (both positive and negative) of variations in cross-validation protocol are demonstrated. Other difficulties due to the small forecast-outcome archive are identified. The claim that the multi-model ensemble provides a "better" probability forecast than the best single model is examined and challenged. Significant forecast information beyond the climatological distribution is also found in a probability forecast based on persistence. On seasonal time scales, the ENSEMBLES simulation-based probability forecasts add significantly more information to empirical probability forecasts than on decadal scales. It is suggested most skillful operational seasonal forecasts available would meld information both from simulation models and empirical models.

1 Introduction

Skillful probabilistic forecasting of seasonal weather and climate statistics would be of value in many fields including agriculture, health and insurance. Since the late nineties seasonal forecasting using dynamical models of the coupled atmosphere, ocean and land surface system has become common in operational weather forecasting centres around the world. In recent years, multi-model ensembles have become popular tools to investigate and account for shortcomings due to structural model error in dynamical model-based predictions on time scales from days to seasons and centuries ([21, 34, 36]). The resources allocated to operational seasonal dynamical models, and the potential use of multi-model ensembles rather than a single model, depend critically on the forecast information simulation models add beyond statistical approaches.

The need for a consistent experimental design for the assessment of skill in multi-model seasonal forecasting was embraced by two large European projects in the last decade. These projects provided the basis for subsequent multi-model designs for operational seasonal-to-decadal forecasting ([33, 17]). The earlier European project, initiated in 2000, was DEMETER ([21, 8, 11]), in which a consistent framework was developed to conduct multi-model seasonal forecasting with a set of general circulation models (GCMs). A similar framework was adopted in ENSEMBLES ([13, 36, 9]), which produced the next generation of seasonal hindcast (or retrospective forecast) simulations, using updated model versions. Further details of the ENSEMBLES &

DEMETER experiments can be found in Table 1 & 2 in the Supplementary Material.

The multi-model ensemble simulations from these projects provide a basis for the quantification of skill in GCM forecasts and an opportunity to assess the benefit of using multi-model ensembles ([36, 2]) over other approaches, such as forecasts based on statistical models ([7, 20, 27, 30, 32]). Furthermore, the consistency between the experimental design of the DEMETER and ENSEMBLES seasonal forecasts makes it possible to quantify the improvement of skill, or in other words, the additional information gained from the forecasts due to model development in the intervening period between the two projects. While evaluations of skill between individual model versions may exist in-house at forecast centres, the authors are unaware of any systematic comparison across centres and model versions. The analysis presented below allows direct comparisons between both the relative performance of and the improvement in different models.

Two particular regions are considered. As a coupled atmospheric and oceanic phenomenon, the El Niño/Southern Oscillation (ENSO) in the tropical Pacific is the dominant mode of seasonal and interannual climate variability. Sea surface temperatures (SSTs) in the Nino3.4 region at seasonal timescales provides an indicator for the ENSO phenomenon. SSTs in the Main Development Region (MDR), over the North Atlantic, provide an indicator for hurricane activity over the coming season. This paper focuses on probability forecast skill in these two regions.[1] Probabilistic skill of seasonal forecasts from both DEMETER and ENSEMBLES are evaluated and contrasted. In each case, ensembles of GCM simulations are transformed into probabilistic distributions via kernel dressing (see [6]) and blended with the climatological distribution to provide calibrated seasonal forecasts; an approach which is becoming common in operational forecasting ([31]). Evaluating probability forecasts as probability forecasts, rather than computing summary statistics of the ensemble mean, allows clearer consideration of the uncertainties sampled by the multi-model ensemble. It is also more easily interpreted in terms of the value, or information content, of the forecast from a decision-makers perspective.

An overview of the DEMETER and ENSEMBLES multi-model exper-

[1] Attention was restricted to these two regions prior to examination of any other regions. This approach eases interpretation of the statistical significance of the results obtained over studies that examine the entire globe and then focus analysis on areas with "significant" skill.

iments used to evaluate seasonal forecast skill over the Nino3.4 and MDR regions are given in section 2 and the approach to generating probabilistic forecasts and evaluating them is described in Section 3. In Section 4, probabilistic skill above that of the climatological distribution is demonstrated up to a lead time of seven months for SSTs over the Nino3.4 region and up to a lead time of two months for SSTs over the MDR. In Section 5 forecasts from the ENSEMBLES models show improvements in skill compared to those from DEMETER for each of the models that are common to both projects. Broadly speaking these results are consistent with previous evaluations of skill from the DEMETER and ENSEMBLES projects ([36, 2]), in which improvements in the anomaly correlation, RMS and Brier scores from DEMETER to ENSEMBLES were reported for SSTs over the tropical Pacific and some other regions up to six months ahead. Section 6 shows that somewhat surprisingly competitive results can be formed from purely empirical probability forecasts based on persistence. The illustrations presented in Section 7 suggest that increasing the ensemble size of future multi-model experiments could provide an efficient way of improving forecast skill, while Sections 8 and 9 highlight the motivation for using proper scoring rules and the challenges involved in model combination to produce multi-model ensemble forecasts, respectively. Section 10 discusses the issues of information contamination when data are precious. The key results and conclusions are summarized in section 11.

2 The seasonal multi-model ENSEMBLES forecasts

The ENSEMBLES multi-model ensemble experiment for seasonal-to-annual forecasting comprises global coupled atmosphere-ocean climate models from the UK Met Office (UKMO), Météo France (MF), the European Centre for Medium-Range Weather Forecasts (ECMWF), the Leibniz Institute of Marine Sciences at Kiel University (IFM-GEOMAR) and the Euro-Mediterranean Centre for Climate Change (CMCC-INGV) in Bologna ([9]). In each case the ensemble simulations include all the major radiative forcings; none of the coupled models has flux adjustments ([13, 36, 9]). A set of seasonal hindcast simulations cover the 46 year period from 1960 to 2005. For each launch date the atmosphere and ocean for each model were initialized using realistic estimates of their observed states, providing an ensemble consisting of nine initial condition ensemble members for each model. Hindcast simulations were launched on the first day of February, of May, of August and of November each year over the hindcast period and run for seven months. This set of 46 seasonal forecasts for each launch date is analysed below. Additionally each model, with the exception of CMCC-INGV, was run for an extended period up to a lead time of 14 months from the November launch.

Improvements made in the ENSEMBLES multi-model forecasting system include a better representation of sub-gridscale physical processes in the simulation models, the inclusion of interannual variability in the greenhouse gas forcing and the use of improved ocean data assimilation, based on quality-controlled *in situ* ocean temperature and salinity profiles for the construction of the initial conditions ([14, 36]). Given two simulation models from the same modelling centres, the experimental designs are sufficiently consistent to allow a direct comparison between the skill of seasonal forecasts from each version of the system. Further details of the models used for the DEMETER and ENSEMBLES projects are provided in Tables 1 and 2 of the Supplement Material.

3 Defining probabilistic forecast skill

Simulations from dynamical models are often used to make probabilistic predictions with the aim of providing useful information for decision support. Evaluating the performance of these predictions, as well as understanding the sources of skill, is crucial for guiding decision-makers in which regions and on what timescales of interest the models are likely to be informative. And perhaps more importantly clarifying when they are likely to be misinformative. Only proper scoring rules offer appropriate, clear measures of probabilistic forecast skill ([5, 37]).

I. J. Good's logarithmic score (Ignorance) (see [10, 25, 5]), is unique among several scoring rules ([37]) designed for evaluating the skill of probabilistic forecasts. It is the only proper and local score[2] for continuous vari-

[2]Proper meaning that it cannot be optimized by hedging the probabilistic forecasts toward other values against the forecasters true belief ([5, 35]). Local meaning that the score depends solely on the probability assigned to the outcome, rather than being rewarded for other features of the forecast distribution, such as its shape.

ables (see [3, 23, 5]). The Ignorance Score is defined by:

$$S(p(y), Y) = -\log_2(p(Y)), \tag{1}$$

where Y is the observed outcome and $p(y)$ is the density function of the forecast distribution. Ignorance has a clear interpretation in terms of gambling returns (see [10, 16, 25]): Under a certain betting scenario, "Kelly Betting" ([16]), the Ignorance describes the rate at which the forecaster's wealth changes with time. Through its close relation to Shannon's information entropy, Ignorance can also be related to the amount of information expected from a forecast (see [25]). It is easily communicated as an effective interest rate (see [12]).

In practice, given K forecast-outcome pairs, $(p_t, Y_t, t = 1, ..., K)$, the empirical Ignorance score is:

$$S_E(p(y), Y) = \frac{1}{K}\sum_{i=1}^{K} -\log_2(p_i(Y_i)). \tag{2}$$

Relative Ignorance reflects the performance of (a set of) forecasts p from one model relative to those of a reference forecast p_{ref}:

$$S_{rel}(p(y), Y) = \frac{1}{K}\sum_{i=1}^{K} -\log_2[(p_i(Y_i))/p_{ref}(Y_i)]. \tag{3}$$

The relative Ignorance of two forecast systems quantifies the information gain (in terms of bits) the model forecast system provides over the reference system. In other words, Ignorance reflects the (average) increase in probability density that the model forecast placed on the outcome relative to that of the reference forecast. By convention, Ignorance is a negatively oriented score, which means the smaller the score more skillful the forecasts. An Ignorance score of $S_{rel} = -1$ means that, on average, forecasts from the model assign twice the probability density to the outcome compared to the reference forecast. Suitable references could include the climatological distribution, a probability forecast from a statistical model, or forecasts from another GCM. The climatological distribution provides the primary benchmark for seasonal forecast skill in this paper, see however Section 6.

Probability forecasts are generated from the DEMETER and the ENSEMBLES simulations via kernel dressing and are blended with climatology to produce seasonal probability forecasts (for a full description see [6], and Appendix A). The climatological distribution is estimated by kernel dressing all available historical observations under cross-validation (see Appendix B). Figure 1 shows an example of the kernel dressed and blended probabilistic forecast distributions for a subset (over the period 1995-2000) of the IFS(ECMWF) hindcast simulations from ENSEMBLES for the Nino3.4 index, launched in November. The blue shaded regions indicate the forecast percentiles between 1-99% and the red line shows the observed outcome (from the ERA40 reanalysis) for comparison. The grey shaded bands show the percentiles between 1-99% for the climatological distribution.

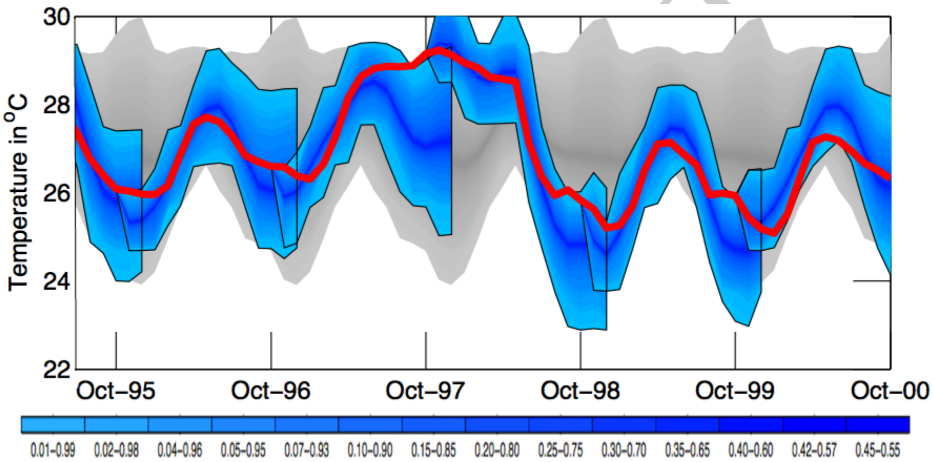

Figure 1: Probabilistic forecast distributions for the IFS(ECMWF) hindcast simulations from ENSEMBLES for the Nino3.4 index, launched in November over the period 1995-2000. The blue shaded regions indicate the forecast percentiles between 1-99% and the red line shows the observed outcome from the ERA40 reanalysis. The grey shaded intervals show the percentiles for the climatological distribution.

The empirical Ignorance score of the dressed and blended GCM forecasts is then computed as a function of lead time (in months) for SSTs over the MDR and Nino3.4 regions relative to the climatology in Section 4. Forecasts from each of the ENSEMBLES models are contrasted with those of DEMETER in Section 5.

4 ENSEMBLES seasonal forecast skill

Figures 2 and 3 show the skill of probability forecasts from each of the models and launch dates available in the ENSEMBLES seasonal forecast project. Figure 2 shows empirical Ignorance scores for forecasts of the Nino3.4 index as a function of lead time, in months, relative to climatology. Each of the four panels corresponds a different forecast launch month (as indicated). In general at short lead times all the models are substantially more skillful than climatology (that is a negative relative Ignorance) for all four initialization dates. This result is generally consistent with [36], who reported anomaly correlation skill for the multi-model ensemble mean was found to decay with lead time over the Nino3 region, to $\sim$0.5 up to fourteen months ahead. At longer lead times ENSEMBLES models show systematically less skill than at early lead times, as expected. In each case, however, the simulation models demonstrate skill above the climatology up to a lead time of seven months. For the hindcasts launched in November some skill appears up to a lead time of fourteen months (although alternative cross-validation protocol casts some doubt on this result - see Section 10). At the longer lead times relative Ignorance scores of approximately -0.25 are found for most models, which translates into the simulation models placing, on average, $\sim$ 19% more probability density on the outcome compared to the climatological distribution. The IFS(ECMWF) and HadGem2(UKMO) models often score slightly lower (are more skillful) than the other three models. The sampling uncertainty across forecast launches is represented by a bootstrap resampling procedure, which resamples the set of forecast Ignorance scores for each model, with replacement. The bootstrap resampling intervals are shown as vertical bars in each of the figures as a 5-95% interval.

Figure 3 shows the Ignorance score as a function of lead time for SSTs over the MDR relative to climatology. Compared to the Nino3.4 index, hindcasts of SSTs in the MDR are less informative at all lead times, particularly for the forecasts launched in November, whose performance decreases significantly within the first two months. Despite the higher Ignorance scores (lower skill), the GCM hindcasts for the MDR demonstrate significant skill relative to climatology up to seven months ahead for most models and launch dates, with the exception of the November launch. Comparison with alternative benchmarks, like the persistence forecast show much larger variation than altering the cross-validation scheme.

In Figures 2 and 3, two models with similar bootstrap resampling inter-

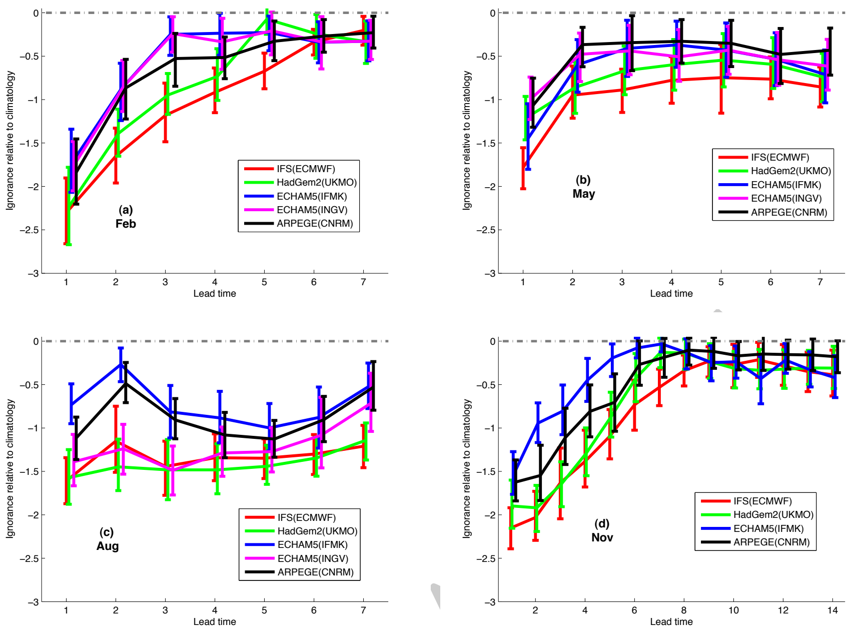

Figure 2: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to climatology as a function of lead time in months. The four different panels show the hindcasts initialized in (a) February, (b) May, (c) August and (d) November. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. All models show significantly more skill than climatology up to a lead time of five months, regardless of when the forecasts are launched. For the November launch (d) the bootstrap resampling intervals often cross the zero skill line beyond a lead time of six months.

vals might be misinterpreted to suggest that neither model is significantly better than the other. Bootstrap resampling skill against climatology is misleading if interpreted incorrectly. One model can systematically outperform a second model on every forecast yet the resample ranges in the skill relative to climatology may overlap. The relative Ignorance between two models on the other hand, provides a clear result reflected in bootstrap resampling from the model-model relative scores.

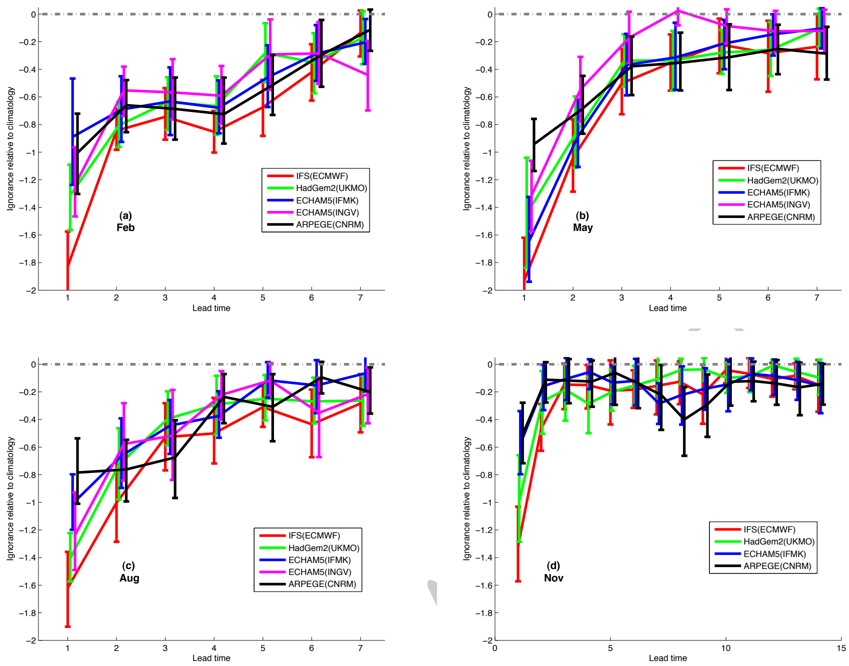

Figure 3: Ignorance score of each model from ENSEMBLES for the MDR index relative to climatology as a function of lead time in months. The four different panels show the hindcasts initialized in (a) February, (b) May, (c) August and (d) November. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. Significant skill above climatology is demonstrated for most models and launch dates at early lead times (up to six months for the February launches, for example), with the exception of the November forecast launches, where the bootstrap intervals overlap the zero-skill climatology beyond a lead time of two months.

Figure 4 shows the Ignorance of each of the ENSEMBLES models for the Nino3.4 index relative to the IFS(ECWMF) model. There are indeed some cases where the IFS(ECMWF) model outperforms all other models despite the overlapping bootstrap resampling intervals in Figure 2. For example, the IFS(ECMWF) model systematically outperforms the ARPEGE(CNRM), ECHAM5(INGV) and ECHAM5(IFMK) models particularly at early lead times for most launch dates. In the case analysed above, there is substantial information in the forecasts from the ENSEMBLES models for the Nino3.4 index even at longer lead times; the IFS(ECMWF) model shows higher skill (often exceeding 0.5 bits in the first 6 months) relative to the other seasonal forecast models used in ENSEMBLES.

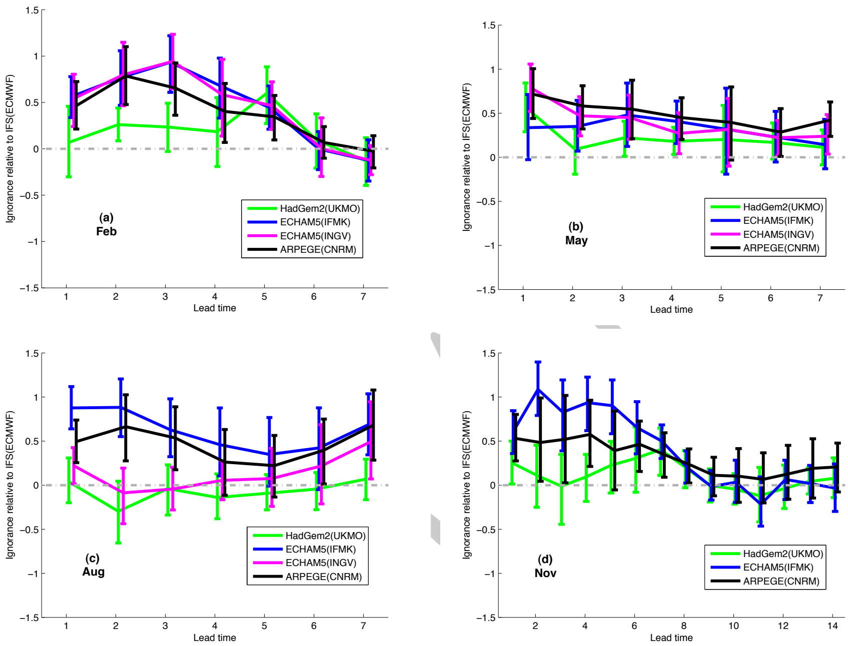

Figure 4: Ignorance score of the ENSEMBLES model forecasts for the Nino3.4 index relative to the IFS(ECWMF) model as a function of lead time in months. Zero Ignorance indicates a model has no skill relative to the IFS(ECMWF) model and negative relative Ignorance scores suggest a model is more skillful than the IFS(ECMWF) model. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. All models shown are typically less skillful than IFS(ECMWF) at all lead times and for most forecast launch dates. For launch dates in August, however, the IFS(ECMWF) model is shown neither to perform significantly better nor significantly worse than HadGem2(UKMO) and ECHAM5(INGV).

5 Contrasting skill of ENSEMBLES & DEMETER

The methods and models used for the seasonal hindcast experiments in the ENSEMBLES project were developed in light of the experience gained and models available from the DEMETER project. The DEMETER seasonal hindcasts and ENSEMBLES hindcasts for the same verification period provide an opportunity to measure the improvement of forecast skill after four years of model development. Such an evaluation is aided by the similarities in the experimental design between the two projects.

Figure 5 shows the Ignorance score of each of the DEMETER model forecasts for the Nino3.4 index relative to climatology. With the exception of ECHAM5(MPI), each model appears substantially more skillful than climatology at all lead times and for all four initialization dates. The lack of skill demonstrated by the ECHAM5(MPI) model reflects the fact that when its ensemble members are dressed and blended with climatology (see Appendix A), they are assigned relatively little weight (that is the forecast is virtually the climatological distribution). There is little or no contribution from the ECHAM5(MPI) model ensemble to the calibrated forecast) beyond a lead time of three months. This is particularly true for the November launch, in which the forecast blending parameter as a function of lead time, α, takes values $[\alpha = 0.90, 0.81, 0.02, 0.00, 0.00, 0.00]$, respectively.

In order to measure the improvement of forecast performance due to model development from the DEMETER to the ENSEMBLES project, the Ignorance of the forecast distributions derived from pairs of model simulations from each project is compared. Although seven European simulation models were used in the DEMETER project, only those models that correspond to earlier "versions" of those used in ENSEMBLES are considered.

Figure 6 shows the Ignorance for seasonal forecasts of the Nino3.4 index forecasts from the ENSEMBLES models relative to those of the corresponding DEMETER models. In general, the relative Ignorance scores in Figure 6 demonstrate improvements for ENSEMBLES (negative relative Ignorance scores) for most lead times and for most models. The ECHAM5(INGV) model is an exception to this finding; the reduction in skill for this model is consistent with [1], which it was shown that subsurface data assimilation for ocean initialization degraded prediction skill over the tropical Atlantic. The ECHAM5(IFMK) model shows substantial improvements, up to one bit, at

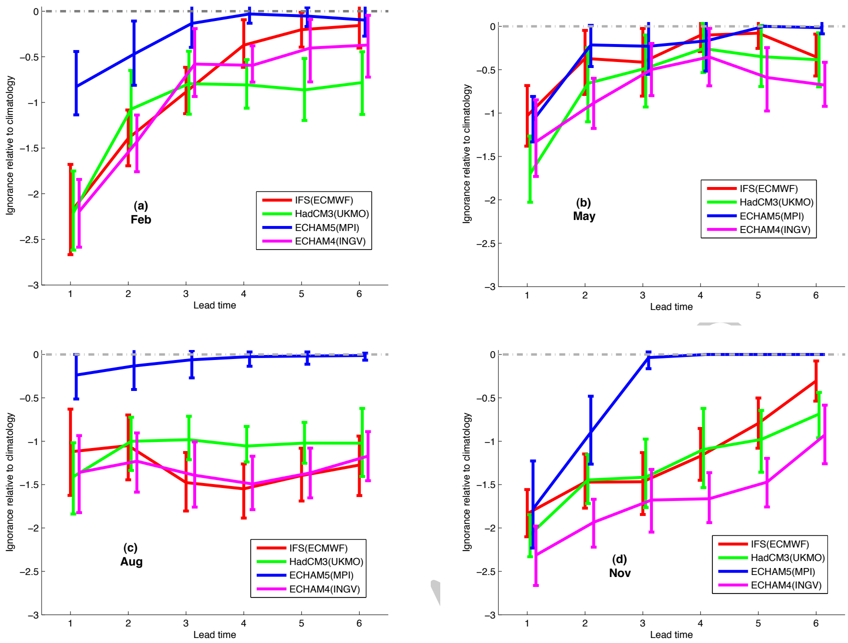

Figure 5: Ignorance score of each model from DEMETER for the Nino3.4 index relative to climatology as a function of lead time in months. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. All models, with the exception of ECHAM5(MPI) are significantly more skillful than climatology at most lead times, particularly for forecasts launched in August and November. At lead times beyond four months, for forecasts launched in November, the ECHAM5(MPI) model is given zero weight when blended with the climatological distribution.

early lead times, particularly for forecast launches in February and May (the ENSEMBLES model placing twice the probability density on the outcome compared to the DEMETER model). Improvements are also demonstrated at lead times beyond three months for forecasts launched in August, particularly for the ECHAM5(IFMK) and HadGem2(UKMO) models.

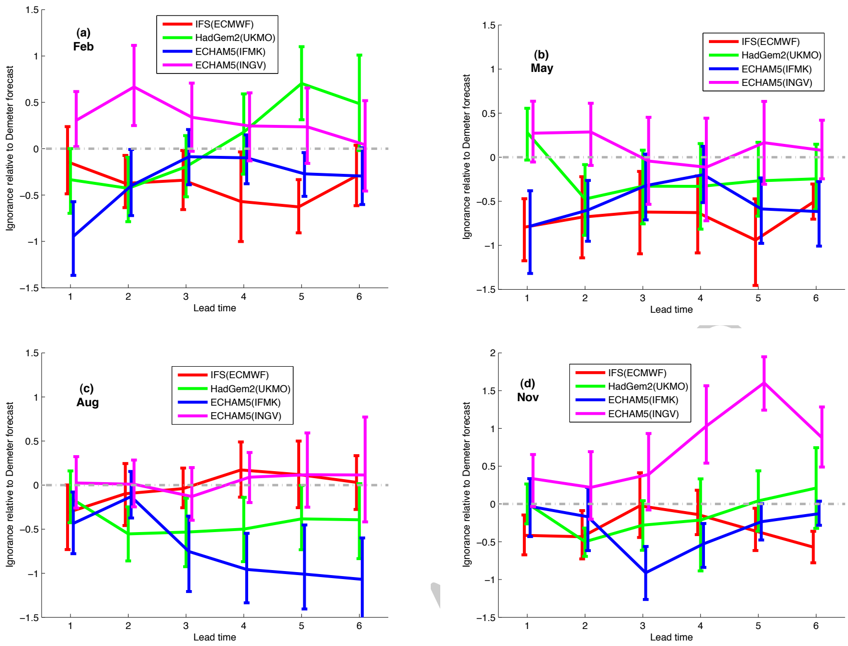

Figure 6: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to the corresponding DEMETER forecasts as a function of lead time in months. Zero Ignorance indicates an ENSEMBLES model has no added skill relative to the corresponding DEMETER model and negative relative Ignorance scores suggest the ENSEMBLES model is more skillful than that of the corresponding DEMETER model. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. The ENSEMBLES models typically demonstrate improvements, of up to one bit in some cases, over their corresponding DEMETER models. ECHAM5(INGV) is an exception to this improvement and is shown to perform worse in ENSEMBLES than its DEMETER model version.

6 Contrasting ENSEMBLES seasonal skill with persistence forecasts

In the previous sections the climatological distribution was used as a benchmark against the performance of the ENSEMBLES and the DEMETER seasonal hindcasts. Whilst comparing skill between dynamical models and climatology provides insight into the information gained from forecasting with dynamical models, there may also be other simple empirical models that can serve as appropriate benchmarks to model performance [27, 30]. A probabilistic persistence forecast provides an interesting benchmark accounting for the effects both of physical persistence and of any long term drift in the temperature of the target region. Whether the additional skill in the ENSEMBLES models over the Nino3.4 region compared to the MDR is related to the strong persistence of ENSO can be investigated by looking at the performance of forecasts over these two regions relative to a persistence model[3]. The persistence forecasts generated here use the observed SST value over the chosen region in the month prior to the forecast launch, persisted forward in time, and transformed into a probabilistic distribution using kernel dressing parameters that vary with lead time (as described in [30]).

Figure 7 shows the Ignorance score of each of the ENSEMBLES models for the Nino3.4 index relative to persistence. For forecasts launched in February most of the ENSEMBLES models are significantly more skillful than persistence at all lead times. For launch dates in August and November little if any information is added compared to the persistence forecasts for most models at any lead time. In fact at early lead times (up to three months ahead) persistence outperforms the ECHAM5(IFMK) and ARPEGR(CNRM) models. At moderate lead times for the August launch and most lead times in the May launch, on the other hand, the IFS(ECMWF) and HadGEM2(UKMO) models outperform persistence.

Figure 8 shows the corresponding results for the MDR index relative to a probabilistic persistence forecast. In this case the ENSEMBLES models and persistence have similar skill, with no one model emerging as significantly better than another. These comparable levels of skill suggest that blending statistical model output with simulation model output would add value to seasonal forecasts.

7 More models or more members?

Knowledge of the relationship between ensemble size and forecast quality aids forecast system design. The cost of increasing the number of ensemble members is typically small relative to the cost of model development. The cost of increasing the ensemble size increases only (nearly) linearly. It is often

[3] We are very grateful to an anonymous reviewer for suggesting this comparison.

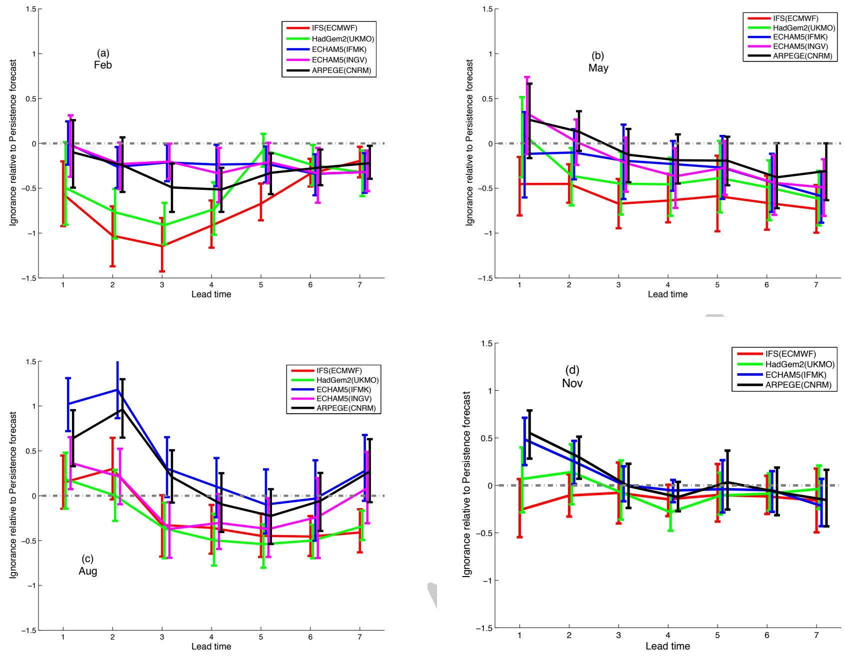

Figure 7: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to persistence forecasts as a function of lead time in months. The four different panels show the hindcasts initialized in (a) February, (b) May, (c) August and (d) November. Scores below zero indicate that an ENSEMBLES model is more skillful than the persistence forecasts. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. ENSEMBLES model forecasts launched in February are shown to be more skillful than persistence at all lead times, whereas for forecasts launched in August the models are significantly worse than persistence at early lead times.

true that the quality of the forecast increases with the number of ensemble members as well, however this improvement in forecast skill depends on both the current ensemble size and the quality of that model's ultimate distribution. The seasonal forecasts from the ENSEMBLES project provide an opportunity to investigate the relationship between ensemble size and forecast quality. This analysis would be eased, for example, had one launch date included an increased number of members so that the value of additional

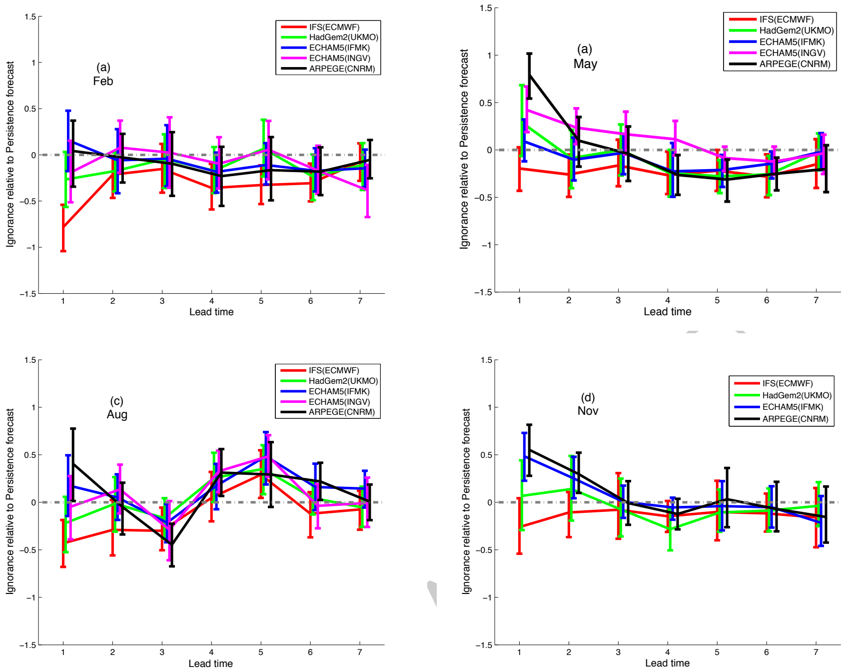

Figure 8: Ignorance score of each model from ENSEMBLES for the MDR index relative to persistence forecasts as a function of lead time in months. The four different panels show the hindcasts initialized in (a) February, (b) May, (c) August and (d) November. Scores below zero indicate that an ENSEMBLES model is more skillful than the persistence forecasts. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. While there is a tendency for Ignorance score remain negative for several months in a row, suggesting skill, the upper (95%) resampling bound is almost always greater than zero.

members could be tested more directly.

Figure 9 shows the effect of decreasing the number of ensemble members on the forecast skill for the Nino3.4 index from the IFS(ECMWF) model launched in November. The skill of two-member ensembles (red) and four-member ensembles (green) are shown relative to the full nine-member ensemble (the zero line) both as a set of random draws from the nine original members without replacement (Figure 9a) and as the average Ignorance of all two- or four-ensemble member combinations (Figure 9b). In Figure 9a most two- and four-member combinations show less skill than the full nine-member ensemble, with only a few ensemble member combinations scoring better than the original ensemble now and then. Figure 9b shows that decreasing the number of ensemble members systematically decreases the average skill (that is, increases the Ignorance score) across all lead times. This result holds both when decreasing from nine members to four members and when decreasing from four to two ensemble members. At a lead time of six months, where the IFS(ECMWF) model still has non-trivial skill relative to climatology (Figure 2), for example, the two-member forecast places $\sim 7\%$ and the four-member ensemble places $\sim 3\%$ less probability density on average on the outcome[4] relative to the nine-member ensemble (Figure 9b). This result suggests that increasing the current ensemble size of nine would further improve the forecast performance[5].

A larger ensemble could be obtained either by increasing the number of ensemble members from one particular model, or, alternatively, by combining simulations from different models to form a multi-model ensemble (see [21, 35]). Of course developing a new, ideally independent model is more costly than increasing the number of ensemble members from an existing model. Combining the output of different (independent) models might, however, have the added advantage of reducing the systematic bias of any single model[6]. One may therefore expect to obtain significantly more information by using multi-model outputs than by increasing the number of ensemble members from a single model.

Figure 10 shows the Ignorance score for a set of multi-model forecasts, in which ensemble members from each of the different ENSEMBLES models are treated equally (that is each ensemble member is assigned equal weight). Here the nine-member IFS(ECMWF) forecasts define the zero line. Figure 10a shows the Ignorance score for forecasts built from multi-model

[4]Under true cross-validation (see Section 10) the effect increases: a two-member forecast places $\sim 15\%$ less probability on the observed outcome.

[5]Operational systems typically consist of 40 to 50 ensemble members. Without hindcast sets, representative of operational systems, however, it is impossible to fully test this hypothesis.

[6]In practice, numerical models developed for weather and climate simulations are far from independent because they share common parametrizations and numerical schemes, and are typically tuned towards the same training dataset. And they face the same technological (computation) limitation. This leads to structural similarities the models and, consequently, to common shortcomings, (e.g. in "blocking").

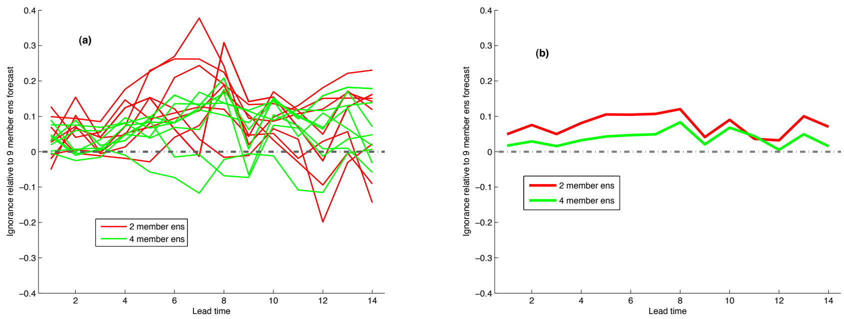

Figure 9: (a) Ignorance of the IFS(ECMWF) model as a function of lead time in months for the Nino3.4 index. The green (red) lines represent the skill of a subset of four-member (two-member) ensemble forecasts relative to the full nine-member ensemble forecast. Each four-member and two-member ensemble consist of random draws from the original nine-member ensemble; (b) Average Ignorance of all possible combinations of two-member (red) and four-member (green) ensembles. On average the four-member ensembles are more skillful than the two-member ensemble, while both ensemble sizes are shown to perform worse on average than the full nine-member ensemble (that is Ignorance score are all above zero).

ensembles containing four members randomly drawn from the 36 available ensemble members (nine members from each of four models) without replacement. Similarly, Figure 10b shows the skill of multi-model ensembles containing nine randomly drawn members. The blue line in each case shows the skill of the full multi-model ensemble, containing 36 members from simulations of the IFS(ECMWF), HadGem2(UKMO), ECHAM5(IFMK) and ARPEGE(CNRM) models. The four-member multi-model forecasts are shown to perform substantially worse than the nine-member IFS(ECMWF) ensemble (indicated by positive Ignorance scores), particularly over short lead times (up to eight months). The skill of the nine-member multi-model forecasts are generally increased compared to the four-member forecasts, however, the single-model, IFS(ECMWF), forecast is still shown to be more skillful[7]

[7]As noted by a referee, in this study the "best" model has been identified in-sample. In this particular study, the ECMWF model is by far the highest scoring model across forecasts (see Supplement Material), and is typically ranked first or second in over half of all skillful forecasts. Rather than resample to show ECMWF is the best, the fraction

than the multi-model forecast at short lead times. This is also true for the full 36-member multi-model forecast, although at longer lead times (beyond eight months) the full multi-model ensemble is shown to outperform the IFS(ECMWF) ensemble. This result in this case suggests that increasing the ensemble size of the "best" model is most likely to improve forecast skill in these regions.

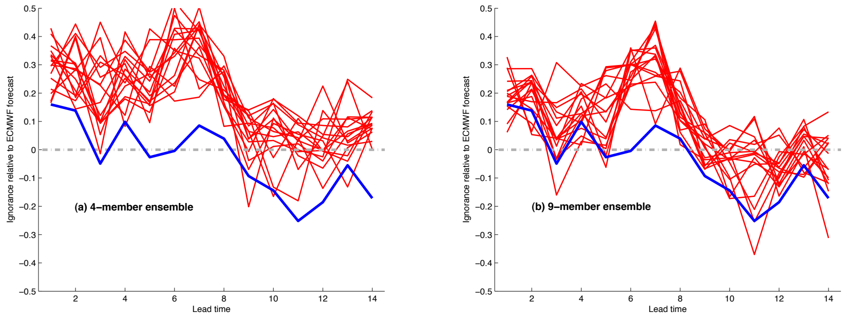

Figure 10: Ignorance of multi-model forecasts as a function of lead time in months for the Nino3.4 index, launched in November, relative to the nine-member IFS(ECMWF) forecast. The blue line represents the multi-model forecast using all 36 ensemble members from the four ENSEMBLES models, equally weighted. The red lines are multi-model forecasts using randomly drawn combinations of four-members (a) and nine-members (b) from the full ensemble. The four-member multi-model forecasts are shown to perform substantially worse than the nine-member IFS(ECMWF) ensemble (that is Ignorance scores are often above zero) and worse than the full 36-member multi-model ensemble. The nine-member multi-model forecasts perform better in general than the four-member forecasts, and to a similar level of skill as the nine-member IFS(ECMWF) ensemble at lead times beyond eight months.

---

of times it is best or second is shown in supplement material. Note also Table 1 and Table 2 in this context. In practice, determining the best model a priori, either for a given purpose, or in a multidimensional sense, is not straightforward (if possible at all). In-sample evaluations of past model performance over relatively short hindcast periods further hinder this task.

8 The importance of being proper

It is sometimes said that a multi-model ensemble forecast is more skillful than any of its constituent single-model ensemble forecasts. This may be the case in terms of reducing root-mean-square (RMS) like scores (see [21, 11, 4, 35, 36, 2]). For probability forecasts, the definition of skill should reflect the characteristics of the forecast problem. While RMS scores are effectively optimal in linear stochastic systems, they are misleading in evaluating non-linear forecast systems, even when the data is not precious. Indeed RMS scores can be misleading even in the limit of an infinite forecast-verification archive (see [19]). Improvements in RMS skill when using multi-model ensembles may be due to error cancellation from independent model contributions (see [11, 15, 4]). For example, if some of the single-model ensembles lie below the observations and some lie above then the ensemble mean could lie closer to the observed outcome than any single ensemble member. While such an error cancellation would reduce the RMS score, rewarding the multi-model forecast more than any single model contribution, a proper skill score ([5]) would not credit this "false" skill. Similarly, combining ensemble members from different models may serve to reduce the variance of ensemble mean statistics, which in turn may lead to a lower RMS score. Indeed, if the ensemble variance is large, adding "information free" ensemble members at the mean value will reduce the RMS error, but need not improve a probabilistic score.

It has also been suggested that the multi-model ensemble forecast outperforms any of the single-model ensemble forecasts by reducing an apparent overconfidence in any one model (see [35, 36, 2]). Such "improvements" can be easily over-interpreted, however; merely doubling the ensemble size under the same model may significantly increase the spread of the forecast distribution. Another way to widen the ensemble spread is simply to blend ([6]) the model forecast distribution with an estimate of the climatological distribution, based on the historical observations (see Appendix A for details). Two single-model forecasts may be ranked differently before and after blending with the climatological distribution. The effect of multi-model combination on seasonal forecast skill is investigated below.

9 Multiple models Ensembles when data are precious

There are many ways in which forecast distributions, generated from ensembles of individual model runs can be combined to produce a single probabilistic multi-model forecast distribution. One approach may be to assign equal weight to each model and simply sum the distributions generated from each model to obtain a single probabilistic distribution (see [11]). When different forecast models do not provide equal amounts of information, one may want to weight the models according to some measure of past performance, see for example [18, 24, 8]. The combined multi-model forecast is the weighted linear sum of the constituent distributions,

$$p_{mm} = \sum_i \omega_i p_i, \tag{4}$$

where the p_i is the forecast distribution from model i and ω_i its weight, with $\sum_i \omega_i = 1$. The weighting parameters may be chosen by minimizing the Ignorance score for example, although fitting ω_i in this way can be costly and is typically complicated by different models sharing information. And, of course, the weights of individual models are expected to vary as a function of lead time. Another, perhaps more fundamental problem of such a weighting procedure is that ω_i are likely to be over- or under-fitted when the forecast-outcome archive is small ([22, 29]).

To avoid complications with fitting model weights a simple iterative method to combine models is used below: First, a reference forecast distribution is derived from the ensemble members of one particular candidate model, in this case the IFS(ECMWF) forecasts, which were argued to provide the most skillful seasonal forecasts for the Nino3.4 index back in Section 4. Each of the other candidate models, in turn, is then combined with the IFS(ECMWF) model by deriving a forecast distribution from the ensemble members of both models, equally weighted. The skill of each two-model combination is computed in terms of Ignorance relative to the IFS(ECMWF) reference forecast and shown in Table 1 for the November launch forecasts of the Nino3.4 index. Each model combination shows the average relative Ignorance (negative scores indicate an improvement over simply using the IFS(ECMWF) forecast). Positive values in the 5th, 8th and 11th columns of Table 1 show that there is no clear improvement in skill for any two-model combination in this case, particularly at lead times less than eight months. Arguably beyond eight months the improvements in skill are not significant; the bootstrap resampling intervals overlap with zero relative skill in each case. Table 2 shows the corresponding results when other models are combined with the UKMO model. In this case combining with ECMWF tends to improve the average Ignorance at all lead times (negative values in 4th and 5th columns of Table 2), but no other combination does this. Starting with ECMWF, combining UKMO has a much smaller effect. In cases where significant improvements are found from such a model combination then further models could be included into the multi-model forecast by choosing those models which yield the biggest improvement in skill and adding them into the forecast one by one with equal weight until no further skill can be added. In this case, however, results suggest that the most skillful seasonal forecasts are provided by using ensemble members from a single model.

10 Establishing skill when data are precious

The DEMETER and the ENSEMBLES seasonal hindcast archive contains merely 46 independent forecast-outcome pairs for each launch date. At seasonal forecast timescales and longer, no true out-of-sample evaluation can be achieved on human timescales; evaluations today must necessarily be in-sample. In this case, it is desirable to strike a balance between using as much of the available data as possible to obtain the best results and holding back enough data so as to avoid information contamination (overfitting) which would lead to poor estimates of real-time operational skill.

The results shown in the previous sections used median cross-validation protocol as described in Appendix B; no additional data is held back in the evaluation of probabilistic forecast distributions beyond that excluded when determining the kernel parameters. While using median values for u, σ and α seems unlikely to allow significant information contamination, this median leave-one-out protocol is not "true" cross-validation. In a true cross-validation protocol, more than one segment of data at a time must be removed from the fitting protocol. This reduces chance of information contamination, it also reduces true quality of the estimation when data are precious. Appendix B details both protocols.

Figure 11 shows the skill of forecasts from the ENSEMBLES models using true cross-validation. Figure 11a shows the Ignorance score for forecasts

LT	ECMWF	ECMWF&UKMO			ECMWF&CNRM			ECMWF&IFMK		
		5%	mean	95%	5%	mean	95%	5%	mean	95%
1	-2.15	-0.08	0.05	0.16	0.05	0.17	0.28	0.07	0.20	0.30
2	-2.03	-0.29	-0.07	0.10	-0.17	0.04	0.24	0.15	0.33	0.47
3	-1.63	-0.44	-0.16	0.08	-0.21	0.04	0.23	-0.09	0.18	0.37
4	-1.36	-0.17	-0.03	0.10	-0.05	0.11	0.26	0.13	0.29	0.41
5	-1.10	-0.19	0.01	0.16	-0.25	-0.04	0.16	0.09	0.28	0.42
6	-0.73	-0.16	0.01	0.17	-0.04	0.11	0.25	0.03	0.19	0.31
7	-0.53	-0.05	0.09	0.22	-0.07	0.07	0.20	0.09	0.18	0.26
8	-0.34	-0.06	0.05	0.15	-0.04	0.06	0.16	-0.04	0.06	0.15
9	-0.23	-0.14	-0.04	0.05	-0.10	0.00	0.11	-0.14	-0.04	0.04
10	-0.27	-0.16	-0.06	0.03	-0.17	-0.05	0.06	-0.14	-0.04	0.05
11	-0.22	-0.32	-0.17	-0.02	-0.22	-0.08	0.06	-0.33	-0.20	-0.08
12	-0.28	-0.20	-0.09	0.01	-0.17	-0.05	0.07	-0.13	-0.03	0.07
13	-0.35	-0.08	-0.01	0.06	-0.20	-0.03	0.11	-0.14	-0.05	0.05
14	-0.39	-0.12	-0.03	0.07	-0.12	0.00	0.13	-0.31	-0.12	0.03

Table 1: Ignorance of each two-model forecast combination, as labeled, relative to the IFS(ECMWF) forecast for each (monthly) lead time for seasonal forecasts of the Nino3.4 index, launched in November. In each case the individual models are also blended with the climatological distribution using blending parameters that minimize the Ignorance score. Each two-model combination shows the average relative Ignorance and the $5-95\%$ bootstrap resampling intervals, which provide an estimate of sampling uncertainty of the relative skill score. For comparison, the second column shows the skill of the (single) ECMWF model relative to climatology.

LT	UKMO	UKMO&ECMWF			UKMO&CNRM			UKMO&IFMK		
		5%	mean	95%	5%	mean	95%	5%	mean	95%
1	-1.90	-0.35	-0.21	-0.08	-0.02	0.08	0.17	-0.01	0.11	0.22
2	-1.92	-0.41	-0.18	0.01	0.03	0.12	0.21	0.22	0.34	0.44
3	-1.64	-0.33	-0.15	-0.01	0.00	0.13	0.26	0.14	0.28	0.40
4	-1.29	-0.24	-0.13	0.00	-0.09	0.06	0.20	0.13	0.26	0.38
5	-0.87	-0.37	-0.22	-0.09	-0.34	-0.12	0.07	0.06	0.21	0.33
6	-0.43	-0.49	-0.30	-0.11	-0.38	-0.12	0.09	-0.11	0.06	0.20
7	-0.13	-0.45	-0.31	-0.16	-0.30	-0.13	0.02	-0.09	0.00	0.08
8	-0.14	-0.26	-0.15	-0.06	-0.20	-0.05	0.06	-0.24	-0.07	0.06
9	-0.24	-0.15	-0.04	0.05	-0.21	-0.03	0.12	-0.18	-0.06	0.05
10	-0.32	-0.12	-0.02	0.08	-0.10	0.00	0.10	-0.12	-0.02	0.08
11	-0.33	-0.24	-0.05	0.12	-0.15	-0.01	0.13	-0.40	-0.16	0.03
12	-0.32	-0.22	-0.06	0.09	-0.11	0.00	0.10	-0.17	-0.03	0.11
13	-0.31	-0.13	-0.05	0.03	-0.14	-0.02	0.12	-0.17	-0.07	0.03
14	-0.31	-0.24	-0.10	0.03	-0.11	0.00	0.10	-0.39	-0.18	0.01

Table 2: Ignorance of each two-model forecast combination, as labeled, relative to the HadGem2(UKMO) forecast for each (monthly) lead time for seasonal forecasts of the Nino3.4 index, launched in November. In each case the individual models are also blended with the climatological distribution using blending parameters that minimize the Ignorance score. Each two-model combination shows the average relative Ignorance and the $5-95\%$ bootstrap resampling intervals, which provide an estimate of sampling uncertainty of the relative skill score. For comparison, the second column shows the skill of the (single) UKMO model relative to climatology.

of the Nino3.4 index, launched in November. Comparing Figure 11a with Figure 2d shows clearly a reduction in skill at longer lead times under the true cross-validation protocol, as well as a widening of the bootstrap resampling intervals in some cases. Significant skill above climatology is demonstrated only up to a lead time of four months. Similarly Figure 11b shows the skill of the ENSEMBLES model forecasts for the MDR index. In this case significant skill above climatology is shown to vanish beyond a lead time of two months.

The preferred cross-validation protocol when the data archive is small is unclear. The approach taken here is to consider more than one protocol. The true cross-validation protocol employed in this section (Figure 11) reflects the expected reduction in the skill of models simply because less data is used to calibrate the forecasts. The median cross-validation protocol (Figure 2 and 3) runs the risk of overfitting the dressing parameters for in-sample evaluation, however. Only out of sample evaluation could establish which effect dominates in this case.

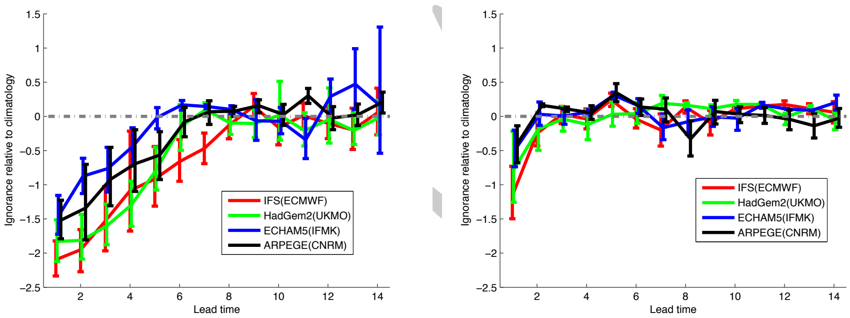

Figure 11: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true cross-validation for, (a) forecasts of the Nino3.4 index and (b) forecasts of the MDR index launched in November. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. Skill is typically reduced compared to the median cross-validation protocol (Figures 2d and 3d), particularly at very early lead times over the MDR. The bootstrap resampling intervals are also widened in some cases.

Figure 12 illustrates the effect of the different cross-validation protocols on the calculated skill of the seasonal forecasts. The figure shows Ignorance scores for the IFS(ECMWF) model from ENSEMBLES relative to climatology using the median (x-axis) and true (y-axis) cross-validation protocols for forecasts of the Nino3.4 index. Each of the four panels corresponds to a different forecast launch month (as indicated). As expected, on average the true cross-validation protocol suggests less skill (that is, larger Ignorance scores) relative to median cross-validation. This improvement on average is not systematic across individual forecasts. The reduction of skill under true cross-validation protocol is small in most cases, giving increased confidence to results using median cross-validation. The most prominent differences are at the highest values of Ignorance where the forecasts have little skill under either protocol. For the November launch this typically occurs at longer lead times (beyond seven months). The argument here is merely that it is important to consider questions of cross-validation when data are precious.

11 Conclusions

The current generation of seasonal forecasts will retire before the forecast-outcome archive grows significantly larger: seasonal verification data are precious! This complicates forecast calibration and evaluation must be performed using cross-validation with only a small sample. Nevertheless probabilistic seasonal forecasts based on the ENSEMBLES stream II experiment demonstrate increased skill in forecasting sea surface temperatures in the Nino3.4 region over that of the DEMETER model simulations. Further analysis suggests that increasing the ensemble size could potentially improve forecast skill further. Such evaluations of skill, on the other hand, should be analysed with care. RMS-based skill scores can obscure skill in nonlinear systems. The statistical characteristics reflected in RMS scores differ from those using proper scoring rules, which are recommended for evaluations of such nonlinear systems as in weather and climate dynamics. The evidence of skill presented, particularly at moderate lead times, is shown to be robust to different choices of appropriate (proper) scores (see Supplementary Material), and may prove to have nontrivial value in application. Simulation based forecasts clearly outperform climatological probability forecasts in many cases. The fact that empirical persistence-based probability forecasts provide a significantly stronger challenge suggests that, in practice, the skill of operational forecast systems can be enhanced with information from

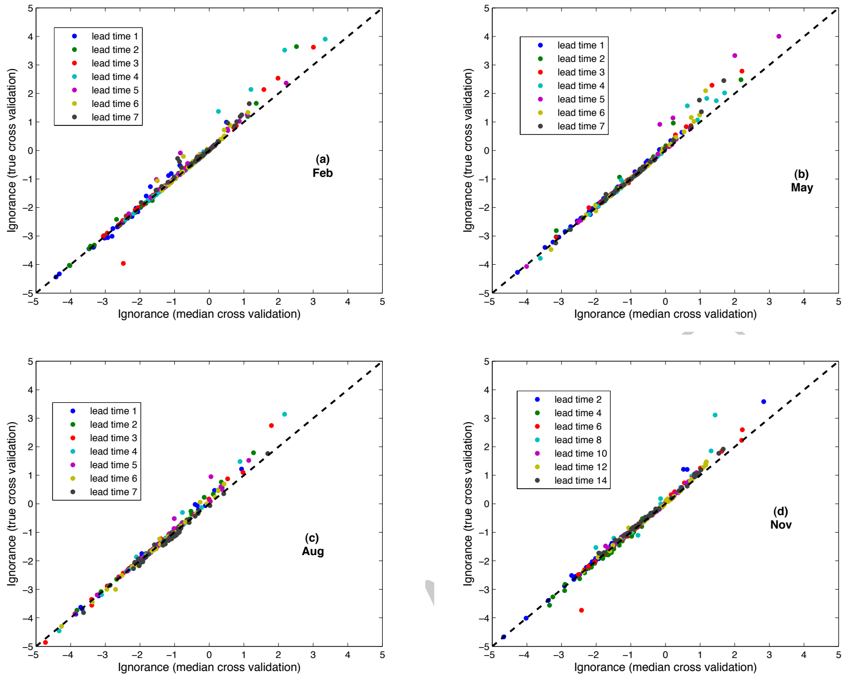

Figure 12: Comparison of Ignorance scores for the IFS(ECMWF) model from ENSEMBLES relative to climatology using the median and true cross-validation protocols for forecasts of the Nino3.4 index, launched in the months as indicated. On average the true cross-validation protocol shows a reduction in skill (larger Ignorance scores) compared to median cross-validation, although individual forecasts can score better. The reduction of skill when using the true cross-validation protocol is most prominent at higher values of Ignorance (when the forecasts are already demonstrating poor skill under the median cross-validation protocol), which for the November launch typically occurs at longer lead times (beyond seven months).

the richer empirical models. Distinguishing the limitations of this level of skill for decision-making from the limitations of our current skill scores and evaluation methodologies will also prove of great value, both in terms of informing future experimental designs for multi-model ensemble projects and for determining the value of these forecast systems to decision-makers.

A From Simulation to a PDF

An ensemble of simulations is transformed into a probabilistic distribution function by a combination of kernel dressing and blending with climatology (see [6]). An N-member ensemble at time t is given as $X_t = [x_t^1, ..., x_t^N]$, where x_t^i is the value of a physical quantity (for example the SST in the MDR region) for the ith ensemble member. For simplicity, all ensemble members under given a model are treated as exchangeable. In other words, the ensemble interpretation does not depend on the ordering of the ensemble members as long as they are generated by the same model ([6]). Kernel dressing defines the model-based component of the density as:

$$p(y : X, \sigma) = \frac{1}{N\sigma} \sum_{i}^{N} K\left(\frac{y - (x^i - \mu)}{\sigma}\right), \tag{5}$$

where y is a random variable corresponding to the density function p and K is the kernel, taken here to be

$$K(\zeta) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}\zeta^2). \tag{6}$$

Thus each ensemble member contributes a Gaussian kernel centred at $x^i - \mu$. Here μ is an offset, which accounts for any systematic "bias". For a Gaussian kernel, the kernel width σ is simply the standard deviation determined empirically as discussed below.

For any finite ensemble, there remains the chance of $\sim \frac{2}{N}$ that the outcome lies outside the range of the ensemble even when the outcome is selected from the same distribution as the ensemble itself. Given the nonlinearity of the model, such outcomes can be very far outside the range of the ensemble members. In addition to N being finite, in practice, of course, the simulations are not drawn from the same distribution as the outcome as the ensemble simulation system is not perfect. To improve the skill of the probabilistic forecasts, the kernel dressed ensemble may be blended with an estimate of the climatological distribution of the system (see [6] for more details, and [23] for a Bayesian approach). The blended forecast distribution is then written as

$$p(\cdot) = \alpha p_m(\cdot) + (1 - \alpha)p_c(\cdot), \tag{7}$$

where p_m is the density function generated by dressing the model ensemble and p_c is the estimate of climatological density. The blending parameter α determines how much weight is placed in the model. Specifying the three values (kernel width σ, kernel offset μ and weight α) at each lead time defines the forecast distribution. These parameters are fitted simultaneously by optimising the empirical Ignorance score, using a cross-validation protocol[8] as described in Appendix B.

B Information Contamination and Cross-validation

Ideally, forecast performance is evaluated "out-of-sample", with new data unknown at the time the model parameters where determined (much less data seen by the analyst). Given a large forecast-outcome archive, cross-validation reduces information contamination and over-fitting when working in-sample (that is, when evaluating a model on the sample sample used to fit the apramers of that model) by dividing the archive into two sets. A training set, used to build the forecast model and fit the parameters, and a testing set, used to get an estimate the skill and likely performance of the model. The process can be repeated to examine the robustness of the results, but information from the test set(s) must not be used to improve the forecast model. When the archive is small and will increase only slowly, one does not have the luxury of this approach. Calibration and evaluation are at best performed under more complex cross-validation; the ideal protocol is not clear and the results can be expected to change with the protocol. A median protocol and a true leave-one-out protocol are defined below.

First, define the forecast probability distribution to be $p(x, X_t, \Theta)$, $t = 1, ..., N$, where X represents the ensemble forecast at time t, Θ represents a vector of parameters (including the kernel width σ, offset μ and blending parameter α) to be fitted and N is the number of forecasts. The corresponding outcomes are defined to be s_t. For each forecast at time $j = 1, ..., N$, leave out one pair of forecast-outcome data (X_j, s_j) and use the remaining

[8]As only 46 years of data are used in this case, any estimation of the two parameters lacks robustness. If one has 4000 years of data, one could draw multiple 46-year data sets from them and estimate the parameters for each sample set. In experiments with simple systems, it turns out that the variation of such estimates is large (see [29]). Note that a 46 year hindcast archive of the full ensemble system may not be available to aid the construction of operational forecast systems.

forecast-outcome data pairs to determine the parameter Θ_j by minimizing the empirical score (in this paper Ignorance is used). The median value, $\bar{\Theta}$, of the set of N Θ_j is then used in the forecast model. This "median protocol" maintains a large learning set with only slight information contamination.

The leave-one-out protocol described in the previous paragraph is not pure cross-validation as $\bar{\Theta}$ arguably contains information from every (X_j, s_j) when the median is taken. To achieve pure cross-validation, the following protocol is adopted. For each forecast at time j, first leave out (X_j, s_j), then for the remaining set apply the median cross-validation protocol described above to obtain N parameter values $\bar{\Theta}_j$. The value $\bar{\Theta}_j$ at each time j is then independent of (X_j, s_j). The forecast empirical Ignorance is then given by $\sum_{j=1}^{N} -\log_2 p(s_j, X_j, \bar{\Theta}_j)$. This protocol ensures that the parameters Θ_j have no explicit dependence on the datum used to evaluate them at the cost of a smaller learning set(s). Even in this case, the datum was known to the analyst. Indeed, use of a common archive in DEMETER and in ENSEMBLES (Stream Two) clouds the possibility of assigning clear statistical significance to estimates of expected skill.

Acknowledgements

This research was supported by the EU Framework 6 ENSEMBLES project; it was also supported both by the LSE's Grantham Research Institute on Climate Change and the Environment and the ESRC Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council and Munich Re. L.A.S. gratefully acknowledges support from Pembroke College, Oxford.

References

[1] A. Alessandri, A. Borrelli, S. Masina, P. D. Pietro, A. Carril, A. Cherchi, S. Gualdi, and A. Navarra. The INGV-CMCC seasonal prediction system: Improved ocean initial conditions. Monthly Weather Review, 138, 29302952 (2010).

[2] A. Alessandri, A. Borrelli, A. Navarra, A. Arribas, M. Déqué, P. Rogel, and A. Weisheimer. Evaluation of probabilistic quality and value of the

ensembles multimodel seasonal forecasts: Comparison with DEMETER. Monthly Weather Review, 139, 2 (2011).

[3] J. M. Bernardo. Expected information as expected utility. Annals of Statistics, 7 (7):686C690 (1979).

[4] N. E. Bowler, A. Arribas, K. R. Mylne. The benefits of multi-analysis and poor-mans ensembles. Monthly Weather Review, 136, 41134129 (2008).

[5] J. Bröcker, L. A. Smith. Scoring Probabilistic Forecasts: On the Importance of Being Proper. Weather and Forecasting, 22 (2), 382-388 (2006).

[6] J. Bröcker and L. A. Smith. From ensemble forecasts to predictive distribution functions. Tellus A, 60, 663-678 (2007).

[7] C. A. S. Coelho, D. B. Stephenson, M. Balmaseda, F. J. Doblas-Reyes and G. J. van Oldenborgh. Towards an integrated seasonal forecasting system for South America. Journal of Climate, 19, 3704-3721 (2006).

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

[9] F. J. Doblas-Reyes, A. Weisheimer, T. N. Palmer, J. M. Murphy and D. Smith. Forecast quality assessment of the ENSEMBLES seasonal-to-decadal Stream 2 hindcasts. Technical Memorandum (ECMWF), 621 (2010).

[10] I. J. Good. Rational decisions. Journal of the Royal Statistical Society, XIV(1):107-114 (1952).

[11] R. Hagedorn, F. J. Doblas-Reyes, and T. N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. Tellus A, 57, 219233 (2005).

[12] R. Hagedorn and L. A. Smith. Communicating the value of probabilistic forecasts with weather roulette. Meteorological Applications, 16(2):143-155 (2009).

[13] C. D. Hewitt and D. J. Griggs. Ensembles-based Predictions of Climate Changes and their Impacts. Eos, Transactions American Geophysical Union, 85, p566 (2004).

[14] B. Ingleby and M. Huddleston. Quality control of ocean temperature and salinity profiles - historical and real-time data. Journal of Marine Systems, 65:158-175 (2007).

[15] I.-S. Kang and J. Yoo. Examination of multi-model ensemble seasonal prediction methods using a simple climate system. Climate Dynamics, 26:285294 (2006).

[16] J. L. Kelly, Jr. A New Interpretation of Information Rate. Bell System Technical Journal, 35:917-926 (1956).

[17] B. P. Kirtman, D. Min, J. M. Infanti, J. L. Kinter III, D. A. Paolino, Q. Zhang, H. van den Dool, S. Saha, M. P. Mendez, E. Becker, P. Peng, P. Tripp, J. Huang, D. G. DeWitt, M. K. Tippett, A, G. Barnston, S. Li, A. Rosati, S. D. Schubert, M. Rienecker, M. Suarez, Z. E. Li, J. Marshak, Y.-K. Lim, J. Tribbia, K. Pegion, W. J. Merryfield, B. Denis, and E. F. Wood. The North American Multi-Model Ensemble (NMME): Phase-1 Seasonal to Interannual Prediction, Phase-2 Toward Developing Intra-Seasonal Prediction. Bulletin of the American Meteorological Society, (2013).

[18] T. N. Krishnamurti, C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran. Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble. Science, 285(5433):15481550 (1999).

[19] P. E. McSharry and L. A. Smith. Better nonlinear models from noisy data: Attractors with maximum likelihood. Physical Review Letters, 83, (21):4285-4288 (1999).

[20] G. J. van Oldenborgh, M. A. Balmaseda, L. Ferranti, T. N. Stockdale and D. L. T. Anderson. Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over the last 15 years? Journal of Climate, 18, 3240-3249 (2005).

[21] T. N. Palmer, A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Délécluse, M. Déqué, E. Diez, 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. Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bulletin of the American Meteorological Society, 85, 853-872 (2004).

[22] P. Peng, A. Kumar, H. van den Dool and A. G. Barnston. An analysis of multimodel ensemble predictions for seasonal climate anomalies. Journal of Geophysical Research, 107(23):4710 (2002).

[23] A. E. Raftery, T. Gneiting, F. Balabdaoui and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review, 131, 1155-1174 (2005).

[24] B. Rajagopalan, U. Lall and S. E. Zebiak. Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. Monthly Weather Review, 130:17921811 (2002).

[25] M. S. Roulston and L. A. Smith. Evaluating Probabilistic Forecasts Using Information Theory. Monthly Weather Review, 130(6):1653-1660 (2002).

[26] D. M. Smith, R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife. Skilful multi-year predictions of atlantic hurricane frequency. Nature Geoscience, 3(12):846-849 (2010).

[27] L.A. Smith. Identification and Prediction of Low-Dimensional Dynamics. Physica D, 58 (1-4): 50-76, (1992).

[28] L. A. Smith. Proceedings International School of Physics “Enrico Fermi”. CXXXIII, 177, Bologna, Italy (1997).

[29] L. A. Smith, H. L. Du, and S. Higgins. Necessary Conditions for assigning Sensible Model Weights in Seasonal and Decadal Forecasting. in preparation for Tellus (2013).

[30] E. B. Suckling and L. A. Smith. An evaluation of decadal probability forecasts from state-of-the-art climate models. Journal of Climate, 26, 23 (2013).

[31] The Met Office. 3-month outlook for contingence planning: User guidance. HM Government document, The Met Office, Devon UK, http://www.metoffice.gov.uk/media/pdf/g/o/3-month_Outlook_User_Guidance-150.pdf (accessed: December 2013).

[32] H. M. Van Den Dool. Empirical methods in short-term climate prediction, Oxford University Press, (2007).

[33] F. Vitart, M. R. Huddleston, M. Déqué, D. Peake, T. N. Palmer, T. N. Stockdale, M. K. Davey, S. Ineson and A. Weisheimer. Dynamically-based seasonal forecast of Atlantic tropical storm actvity issued in June by EUROSIP. Geophysical Research Letters, 34, L16815 (2007).

[34] B. Wang, J.-Y. Lee, I.-S. Kang, J. Shukla, C. K. Park, A. Kumar, J. Schemm, S. Cocke, J. S. Kug, J. J. Luo, T. Zhou, B.Wang, X. Fu, W. T. Yun, O. Alves, E. Jin, J. Kinter, B. Kirtman, T. Krishnamurti, N. Lau, W. Lau, P. Liu, P. Pegion, T. Rosati, S. Schubert, W. Stern, M. Suarez, and T. Yamagata. Advance and prospectus of seasonal prediction: assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (19802004). Climate Dynamics, 33(1):93117 (2009).

[35] A. P. Weigel, M. A. Liniger, and C. Appenzeller. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Quarterly Journal of the Royal Meteorological Society, 134(630):241C260 (2008).

[36] A. Weisheimer, F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Déqué, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel. ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions and Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. Geophysical Research Letters, 36(21) (2009).

[37] D. S. Wilks. Statistical Methods in the Atmospheric Sciences, Volume 91, Second Edition (International Geophysics). Academic Press, 2 edition (2005).

Supplemental Material
Probabilistic skill in ensemble seasonal forecasts

Leonard A. Smith, Hailiang Du, Emma B. Suckling and Falk Niehörster
Centre for the Analysis of Time Series, LSE, UK

February 26, 2014

This document provides supplementary material for the manuscript (Smith et al. Probabilistic skill in ensemble seasonal forecasts).

- Details of the ENSEMBLES & DEMETER simulation models used in the seasonal forecast evaluation are given in Table 1 and Table 2. Table 1 contains a description of the simulation models that constitute the ENSEMBLES seasonal hindcast experiment. There were seven comprehensive European global coupled atmosphere-ocean models developed in the DEMETER project. Table 2 lists a subset of simulation models from the DEMETER project, which are directly comparable to the models used for the ENSEMBLES hindcasts.

- Ignorance scores of both the DEMETER and the ENSEMBLES forecasts using the true cross-validation protocol (as described in Appendix B of the main manuscript) are presented in Figure 1-4, which can be compared with those generated using median cross-validation (See Figure 2, 3, 5, 6 of the main manuscript).

- Figure 5a and 5b illustrate the results shown in Table 1 and Table 2 of the main manuscript in terms of the additional information gained from multi-model combination with the best and second best ranked models, respectively.

- Table 3 shows the statistics of each of the four simulation models' forecast performance in rank ordered according to Ignorance score for each forecast of Nino3.4 index at November launch. It appears the

IFS(ECMWF) and HadGEM2(UKMO) comes first or second much more often than the other two models.

- Ignorance score of each model from ENSEMBLES relative to persistence forecasts as a function of lead time at November launch for both Nino3.4 index and MDR index is shown in Figure 6. In Figure 7d & 8d of the main manuscript, results after lead time 7 were not presented in order to conveniently compare with the rest of the panels of Figure 7 & 8. Results for the whole range of lead time is presented here in Figure 6.

- Rank continuous probability score of each model from ENSEMBLES for the Nino3.4 index relative to climatological forecast at November launch is illustrated in Figure 7. The results are consistent with the evaluation using Ignorance score (See Figure 2d of the main manuscript).

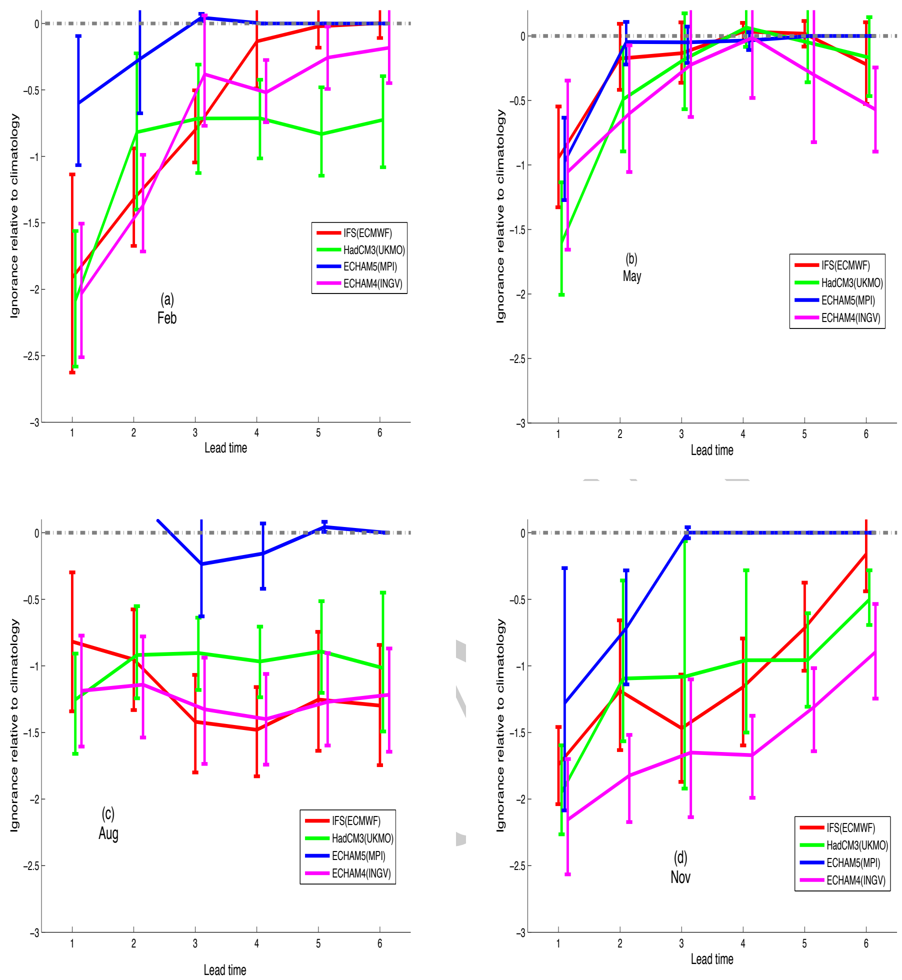

Figure 1: Ignorance score of each model from DEMETER for the Nino3.4 index relative to climatology as a function of lead time in months using true leave-one-out cross-validation. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. All models, with the exception of ECHAM5(MPI) are significantly more skillful than climatology at most lead times, particularly for forecasts launched in August and November. Note that ECHAM5(MPI) significantly under perform climatology at short lead for forecasts launched in August.

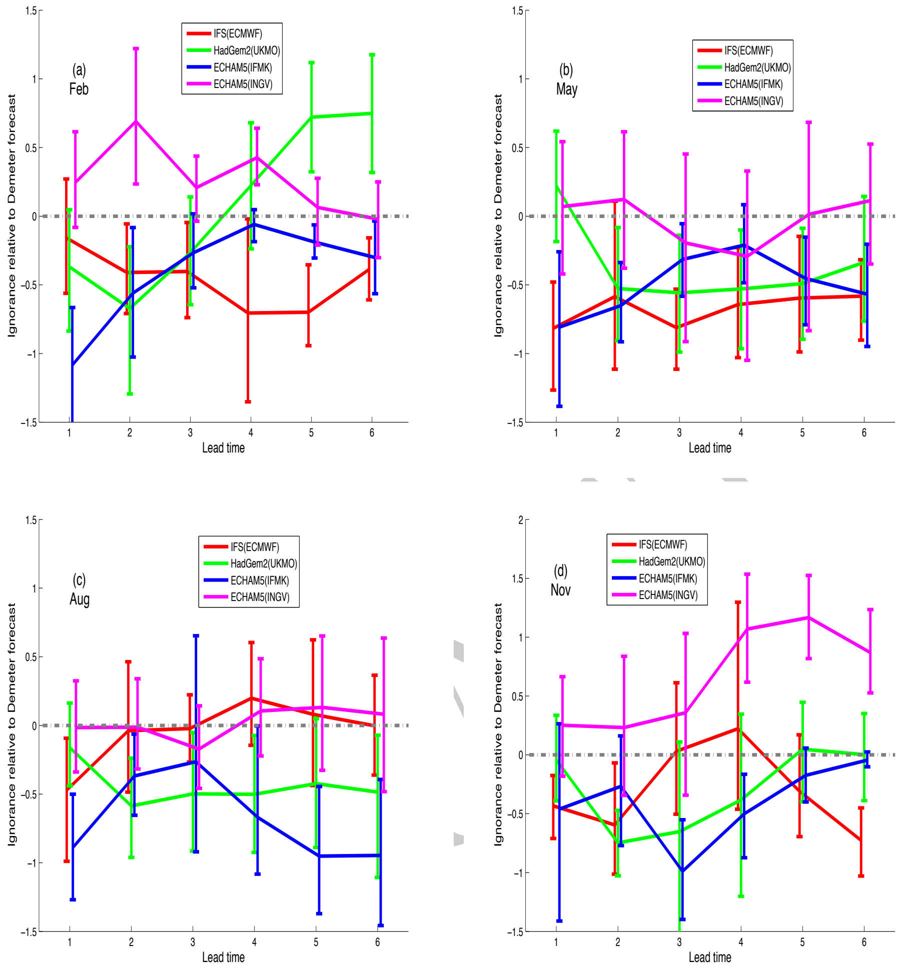

Figure 2: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to the equivalent DEMETER forecasts as a function of lead time in months using true cross-validation. Zero Ignorance indicates an ENSEMBLES model has no skill relative to the corresponding DEMETER model and negative relative Ignorance scores suggest the ENSEMBLES model is more skillful than that of the corresponding DEMETER model. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. ENSEMBLES models typically demonstrate improvements, of up to one bit in some cases, over their corresponding DEMETER models. ECHAM5(INGV) is an exception to this improvement and is shown to perform worse in ENSEMBLES than its DEMETER version.

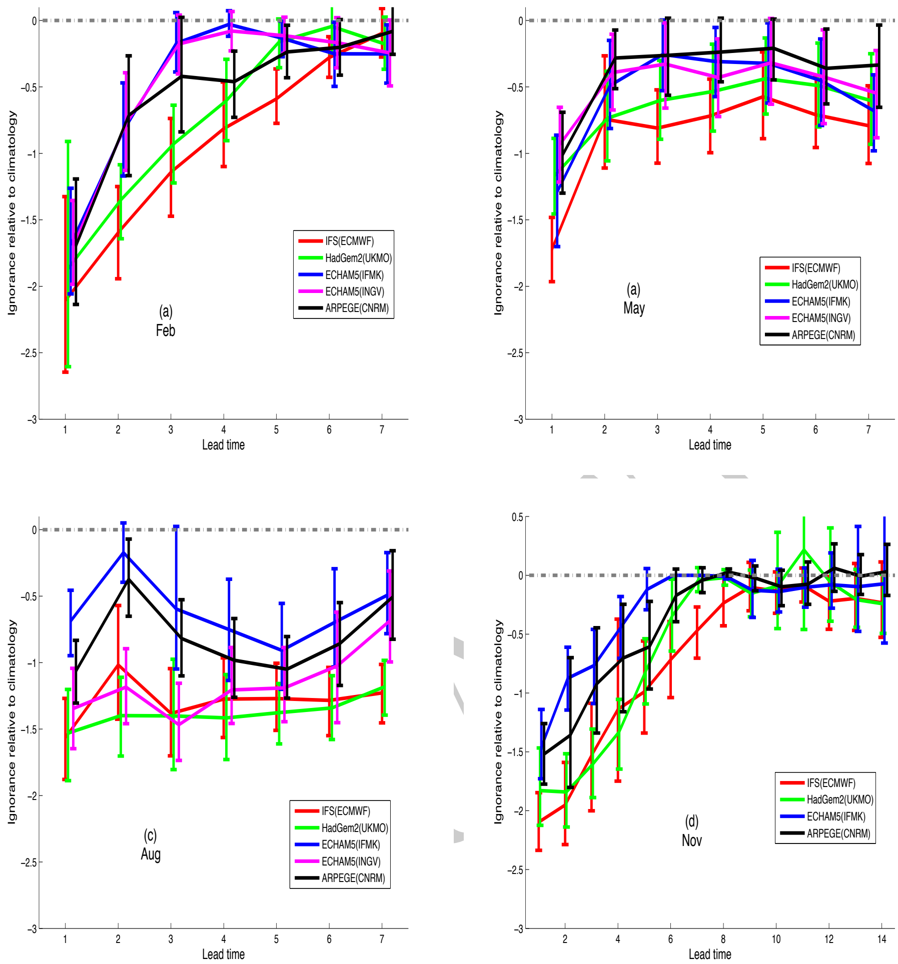

Figure 3: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true leave-one-out cross-validation for forecasts of the Nino3.4 index. The four different panels show the hindcasts initialized in February, May, August and November. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. Skill is generally reduced compared to the median cross-validation procedure (Figure 2. in the manuscript). The bootstrap resampling intervals are also widened in some cases.

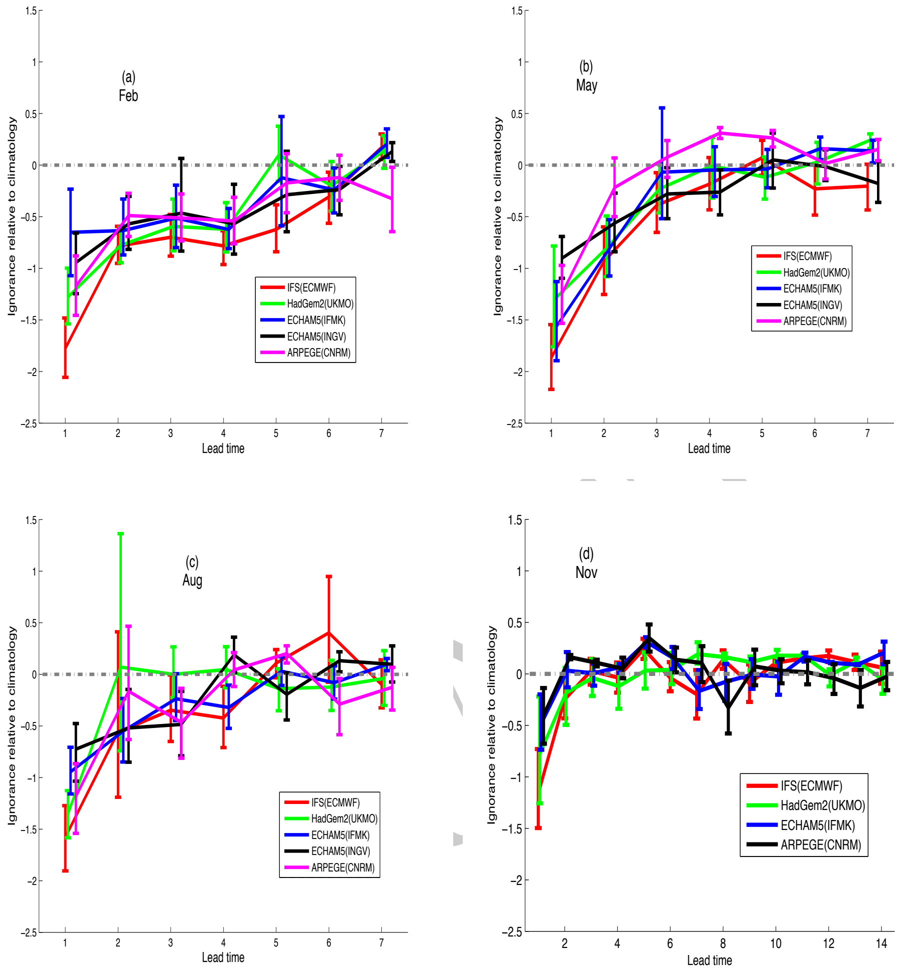

Figure 4: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true leave-one-out cross-validation for forecasts at Main Development Region. The four different panels show the hindcasts initialized in February, May, August and November. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. Skill is generally reduced compared to the median cross-validation procedure (Figure 3. in the manuscript). The bootstrap resampling intervals are also widened in some cases.

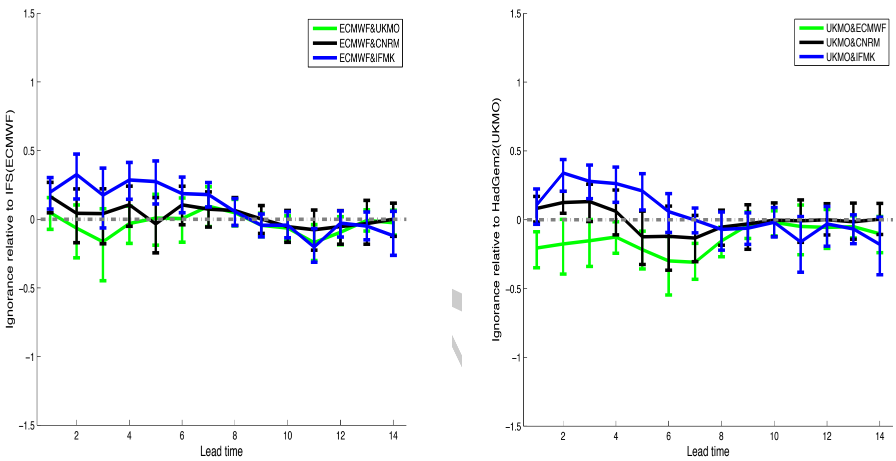

Figure 5: Ignorance score of each two-model forecast combination, as labelled, relative to a) the IFS(ECMWF) forecast; b) the HadGem2(UKMO) forecast, at each lead time for forecasts of the Nino3.4 index, launched in November. In each case the inividual models are also blended with the climatological distribution (dressing and blending parameter values are fitted using median cross-validation protocol). Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

	Atmosphere		Ocean			
Institute	Model	Resolution	Model	Resolution	Initialization	References
ECMWF	IFS CY31R1	T159/L62	HOPE	0.3-1.4/L29	ERA-40 and ECMWF operational analysis for atmosphere and land, ensemble of ocean reanalyses + SST perturbations, singular vectors in atmosphere	Stockdale et al.,[2009]; Balmaseda et al., [2008]
UKMO	HadGEM2-A	N96/L38	HadGEM2-O	0.33-1/L20	As for ECMWF plus improved soil moisture anomaly assimilation	Collins et al., [2008]
CNRM	ARPEGE4.6	T63	OPA8.2	2/L31	As for ECMWF	Daget et al., [2009]; Salas y Melia, [2002]
IFMK	ECHAM5	T63/L31	MPI-OM1	1.5/L40	Permutations of 20th century coupled simulations with restored SSTs	Keenlyside et al., [2005]; Jungclaus et al., [2006]
INGV	ECHAM5	T63/L19	OPA8.2	2/L31	AMIP-type simulations for atmosphere, ensemble of ocean reanalyses + SST perturbations	Alessandri et al, [2009]; Di Pietro and Masina, [2009]

Table 1: Description of the simulation models that constitute the ENSEMBLES seasonal hindcast experiment.

	Atmosphere		Ocean			
Institute	Model	Resolution	Model	Resolution	Initialization	References
ECMWF	IFS	T95/L40	HOPE-E	0.3-1.4X1.4 L29	Wind stress and SST perturbations	Gregory et al., [2000], Wolff et al. [1997]
UKMO	HadAM3	N96/L19	GloSea OGCM	0.3-1.25X2.5 L23	Wind stress and SST perturbations	Pope et al., [2000], Gordon et al., [2000]
MPI	ECHAM5	T42/L19	MPI-OM1	0.52.5X2.5 L23	Atmospheric conditions from the coupled initialization run (lagged method)	Roeckner et al., [1996], Marsland et al., [2003]
INGV	ECHAM4	T42/L19	OPA8.1	0.5-1.5X2.0 L31	Wind stress and SST perturbations	Roeckner et al., [1996]; Madec et al., [1998]

Table 2: The subset of simulation models from the DEMETER project, which are directly comparable to the model used for the ENSEMBLES hindcasts.

Model	Lead time	1	2	3	4	5	6	7
ECMWF	No. Rank 1	18	14	14	12	20	22	16
	% of Rank 1	43.9%	34.2%	34.2%	29.3%	48.8%	53.7%	39.0%
	% of Rank 1 or 2	75.6%	75.6%	68.3%	56.1%	63.4%	70.7%	73.2%
	$p(x >= No.Rank1)$	0.006	0.122	0.122	0.318	0.001	0.001	0.033
UKMO	No. Rank 1	17	24	23	22	11	9	18
	% of Rank 1	41.5%	58.5%	56.1%	53.7%	26.8%	21.9%	43.9%
	% of Rank 1 or 2	70.7%	82.9%	82.9%	73.1%	63.4%	78.1%	75.6%
	$p(x >= No.Rank1)$	0.015	0.000	0.000	0.000	0.452	0.730	0.006
CNRM	No. Rank 1	5	2	2	6	7	6	4
	% of Rank 1	12.2%	4.9%	4.9%	14.6%	17.1%	14.6%	9.8%
	% of Rank 1 or 2	39.0%	22.0%	22.0%	39.0%	39.0%	26.8%	31.7%
	$p(x >= No.Rank1)$	0.987	1.000	1.000	0.964	0.917	0.964	0.996
IFMK	No. Rank 1	1	1	2	1	3	4	3
	% of Rank 1	2.4%	2.4%	4.9%	2.4%	7.3%	9.8%	7.3%
	% of Rank 1 or 2	14.6%	19.5%	26.8%	31.7%	34.2%	24.4%	19.5%
	$p(x >= No.Rank1)$	1.000	1.000	1.000	1.000	0.999	0.996	0.999

Table 3: Four simulation models' forecast performance is rank ordered according to Ignorance score for each forecast of Nino3.4 index at November launch. The number of times each model rank the first, the percentage of each model rank the first and the percentage of each model rank the first or second. $p(x >= No.Rank1)$ is the probability that the number of times a model rank the first no less than the observed No. Rank 1 assuming all four models are equally good.

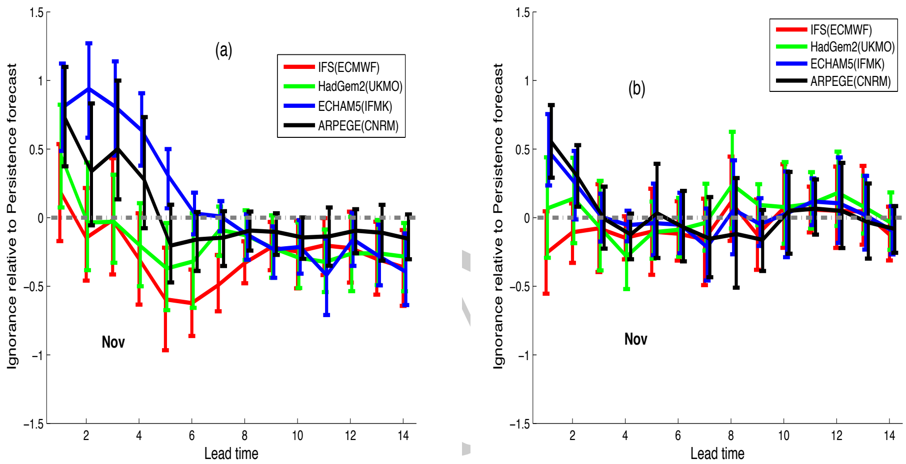

Figure 6: Ignorance score of each model from ENSEMBLES for a) the Nino3.4 index; b) the MDR index, relative to persistence forecast as a function of lead time in months for November launch. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

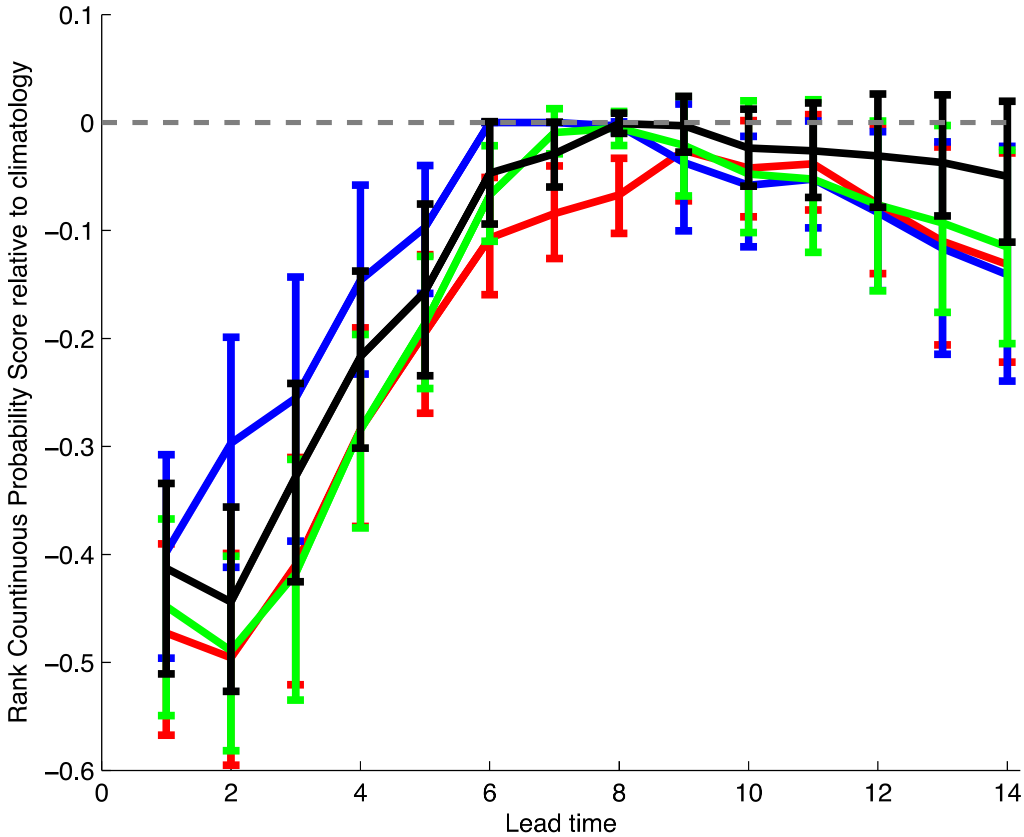

Figure 7: Rank continuous probability score of each model from ENSEMBLES for the Nino3.4 index relative to climatological forecast as a function of lead time in months for November launch. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

Supplemental Material
Probabilistic skill in ensemble seasonal forecasts

Leonard A. Smith, Hailiang Du, Emma B. Suckling and Falk Niehörster
Centre for the Analysis of Time Series, LSE, UK

February 26, 2014

This document provides supplementary material for the manuscript (Smith et al. Probabilistic skill in ensemble seasonal forecasts).

- Details of the ENSEMBLES & DEMETER simulation models used in the seasonal forecast evaluation are given in Table 1 and Table 2. Table 1 contains a description of the simulation models that constitute the ENSEMBLES seasonal hindcast experiment. There were seven comprehensive European global coupled atmosphere-ocean models developed in the DEMETER project. Table 2 lists a subset of simulation models from the DEMETER project, which are directly comparable to the models used for the ENSEMBLES hindcasts.

- Ignorance scores of both the DEMETER and the ENSEMBLES forecasts using the true cross-validation protocol (as described in Appendix B of the main manuscript) are presented in Figure 1-4, which can be compared with those generated using median cross-validation (See Figure 2, 3, 5, 6 of the main manuscript).

- Figure 5a and 5b illustrate the results shown in Table 1 and Table 2 of the main manuscript in terms of the additional information gained from multi-model combination with the best and second best ranked models, respectively.

- Table 3 shows the statistics of each of the four simulation models' forecast performance in rank ordered according to Ignorance score for each forecast of Nino3.4 index at November launch. It appears the

IFS(ECMWF) and HadGEM2(UKMO) comes first or second much more often than the other two models.

- Ignorance score of each model from ENSEMBLES relative to persistence forecasts as a function of lead time at November launch for both Nino3.4 index and MDR index is shown in Figure 6. In Figure 7d & 8d of the main manuscript, results after lead time 7 were not presented in order to conveniently compare with the rest of the panels of Figure 7 & 8. Results for the whole range of lead time is presented here in Figure 6.

- Rank continuous probability score of each model from ENSEMBLES for the Nino3.4 index relative to climatological forecast at November launch is illustrated in Figure 7. The results are consistent with the evaluation using Ignorance score (See Figure 2d of the main manuscript).

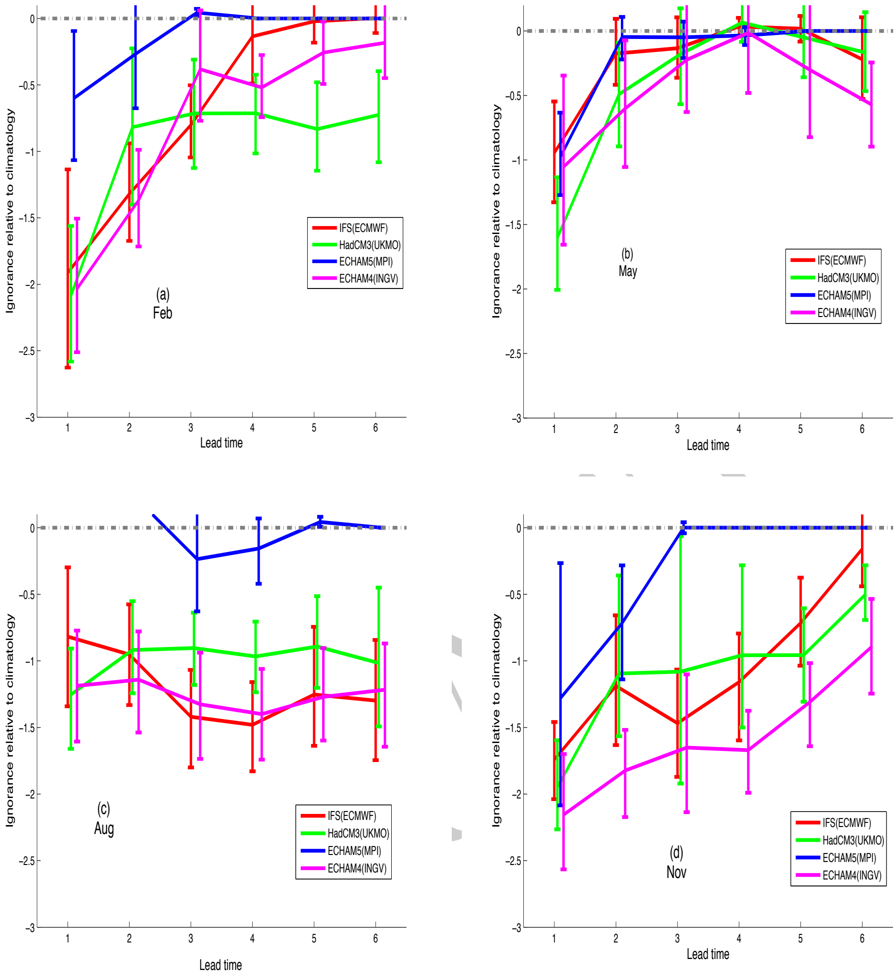

Figure 1: Ignorance score of each model from DEMETER for the Nino3.4 index relative to climatology as a function of lead time in months using true leave-one-out cross-validation. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. All models, with the exception of ECHAM5(MPI) are significantly more skillful than climatology at most lead times, particularly for forecasts launched in August and November. Note that ECHAM5(MPI) significantly under perform climatology at short lead for forecasts launched in August.

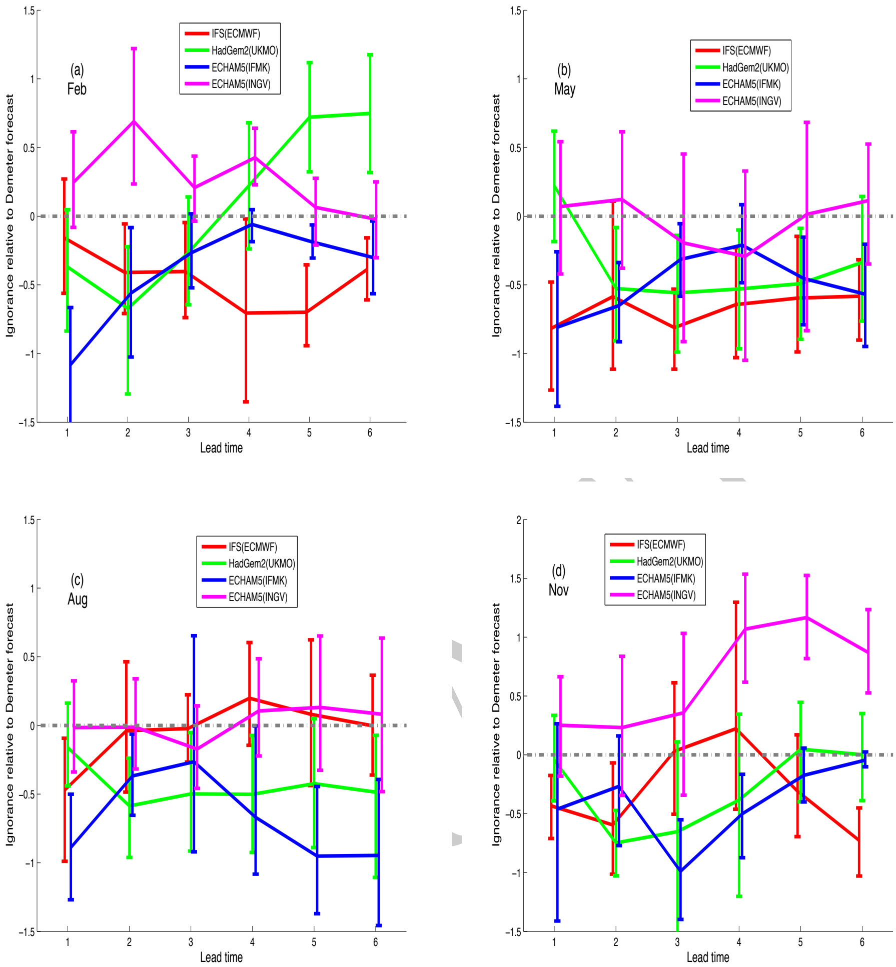

Figure 2: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to the equivalent DEMETER forecasts as a function of lead time in months using true cross-validation. Zero Ignorance indicates an ENSEMBLES model has no skill relative to the corresponding DEMETER model and negative relative Ignorance scores suggest the ENSEMBLES model is more skillful than that of the corresponding DEMETER model. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. ENSEMBLES models typically demonstrate improvements, of up to one bit in some cases, over their corresponding DEMETER models. ECHAM5(INGV) is an exception to this improvement and is shown to perform worse in ENSEMBLES than its DEMETER version.

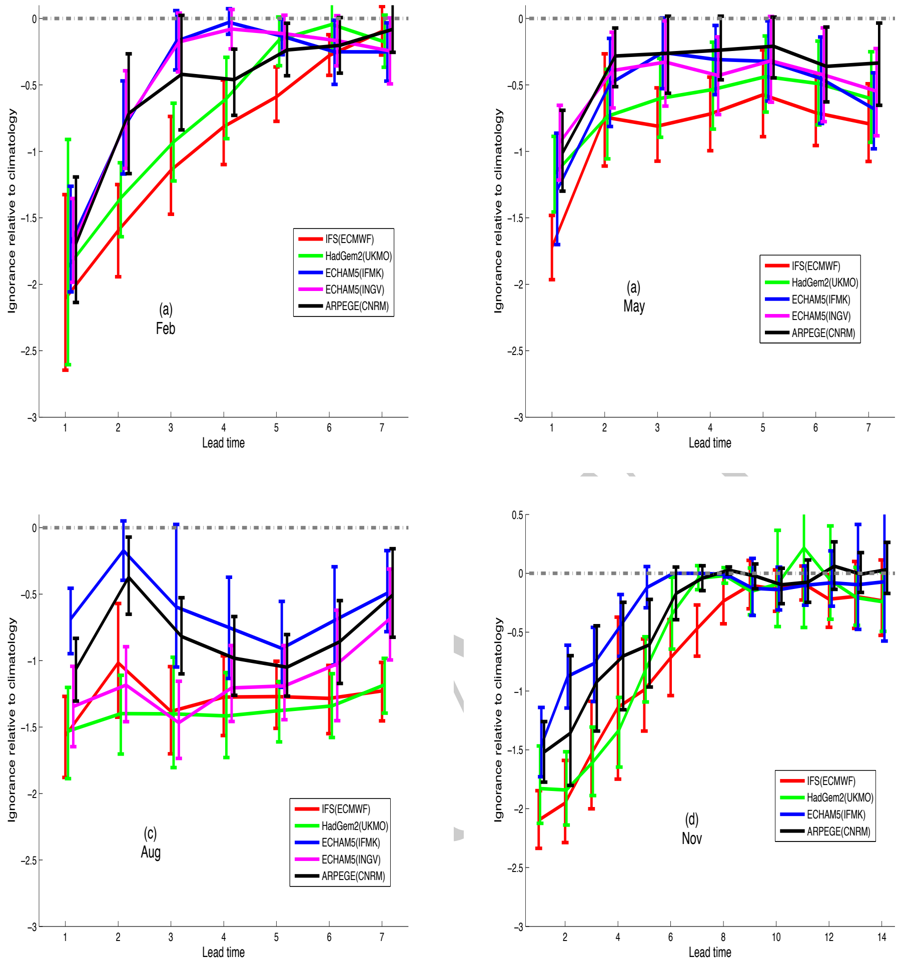

Figure 3: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true leave-one-out cross-validation for forecasts of the Nino3.4 index. The four different panels show the hindcasts initialized in February, May, August and November. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. Skill is generally reduced compared to the median cross-validation procedure (Figure 2. in the manuscript). The bootstrap resampling intervals are also widened in some cases.

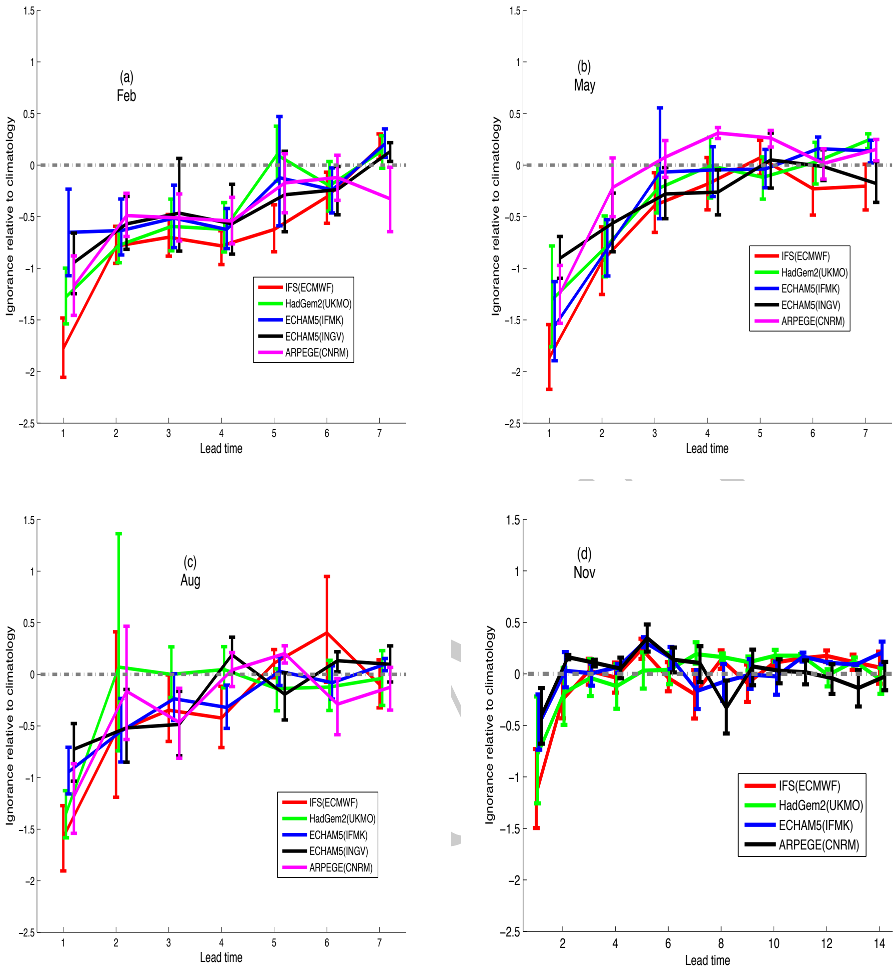

Figure 4: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true leave-one-out cross-validation for forecasts at Main Development Region. The four different panels show the hindcasts initialized in February, May, August and November. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples. Skill is generally reduced compared to the median cross-validation procedure (Figure 3. in the manuscript). The bootstrap resampling intervals are also widened in some cases.

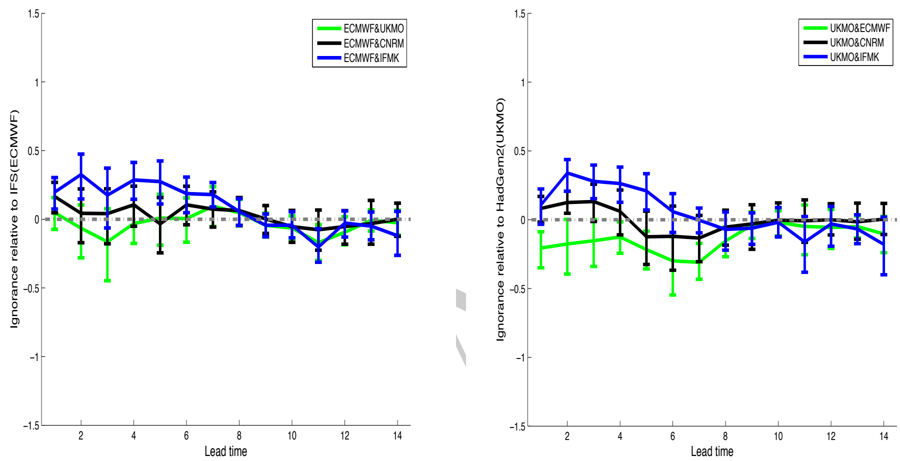

Figure 5: Ignorance score of each two-model forecast combination, as labelled, relative to a) the IFS(ECMWF) forecast; b) the HadGem2(UKMO) forecast, at each lead time for forecasts of the Nino3.4 index, launched in November. In each case the inividual models are also blended with the climatological distribution (dressing and blending parameter values are fitted using median cross-validation protocol). Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

	Atmosphere		Ocean			
Institute	Model	Resolution	Model	Resolution	Initialization	References
ECMWF	IFS CY31R1	T159/L62	HOPE	0.3-1.4/L29	ERA-40 and ECMWF operational analysis for atmosphere and land, ensemble of ocean reanalyses + SST perturbations, singular vectors in atmosphere	Stockdale et al.,[2009]; Balmaseda et al., [2008]
UKMO	HadGEM2-A	N96/L38	HadGEM2-O	0.33-1/L20	As for ECMWF plus improved soil moisture anomaly assimilation	Collins et al., [2008]
CNRM	ARPEGE4.6	T63	OPA8.2	2/L31	As for ECMWF	Daget et al., [2009]; Salas y Melia, [2002]
IFMK	ECHAM5	T63/L31	MPI-OM1	1.5/L40	Permutations of 20th century coupled simulations with restored SSTs	Keenlyside et al., [2005]; Jungclaus et al., [2006]
INGV	ECHAM5	T63/L19	OPA8.2	2/L31	AMIP-type simulations for atmosphere, ensemble of ocean reanalyses + SST perturbations	Alessandri et al, [2009]; Di Pietro and Masina, [2009]

Table 1: Description of the simulation models that constitute the ENSEMBLES seasonal hindcast experiment.

	Atmosphere		Ocean			
Institute	Model	Resolution	Model	Resolution	Initialization	References
ECMWF	IFS	T95/L40	HOPE-E	0.3-1.4X1.4 L29	Wind stress and SST perturbations	Gregory et al., [2000], Wolff et al. [1997]
UKMO	HadAM3	N96/L19	GloSea OGCM	0.3-1.25X2.5 L23	Wind stress and SST perturbations	Pope et al., [2000], Gordon et al., [2000]
MPI	ECHAM5	T42/L19	MPI-OM1	0.52.5X2.5 L23	Atmospheric conditions from the coupled initialization run (lagged method)	Roeckner et al., [1996], Marsland et al., [2003]
INGV	ECHAM4	T42/L19	OPA8.1	0.5-1.5X2.0 L31	Wind stress and SST perturbations	Roeckner et al., [1996]; Madec et al., [1998]

Table 2: The subset of simulation models from the DEMETER project, which are directly comparable to the model used for the ENSEMBLES hindcasts.

Model	Lead time	1	2	3	4	5	6	7
ECMWF	No. Rank 1	18	14	14	12	20	22	16
	% of Rank 1	43.9%	34.2%	34.2%	29.3%	48.8%	53.7%	39.0%
	% of Rank 1 or 2	75.6%	75.6%	68.3%	56.1%	63.4%	70.7%	73.2%
	$p(x >= No.Rank1)$	0.006	0.122	0.122	0.318	0.001	0.001	0.033
UKMO	No. Rank 1	17	24	23	22	11	9	18
	% of Rank 1	41.5%	58.5%	56.1%	53.7%	26.8%	21.9%	43.9%
	% of Rank 1 or 2	70.7%	82.9%	82.9%	73.1%	63.4%	78.1%	75.6%
	$p(x >= No.Rank1)$	0.015	0.000	0.000	0.000	0.452	0.730	0.006
CNRM	No. Rank 1	5	2	2	6	7	6	4
	% of Rank 1	12.2%	4.9%	4.9%	14.6%	17.1%	14.6%	9.8%
	% of Rank 1 or 2	39.0%	22.0%	22.0%	39.0%	39.0%	26.8%	31.7%
	$p(x >= No.Rank1)$	0.987	1.000	1.000	0.964	0.917	0.964	0.996
IFMK	No. Rank 1	1	1	2	1	3	4	3
	% of Rank 1	2.4%	2.4%	4.9%	2.4%	7.3%	9.8%	7.3%
	% of Rank 1 or 2	14.6%	19.5%	26.8%	31.7%	34.2%	24.4%	19.5%
	$p(x >= No.Rank1)$	1.000	1.000	1.000	1.000	0.999	0.996	0.999

Table 3: Four simulation models' forecast performance is rank ordered according to Ignorance score for each forecast of Nino3.4 index at November launch. The number of times each model rank the first, the percentage of each model rank the first and the percentage of each model rank the first or second. $p(x >= No.Rank1)$ is the probability that the number of times a model rank the first no less than the observed No. Rank 1 assuming all four models are equally good.

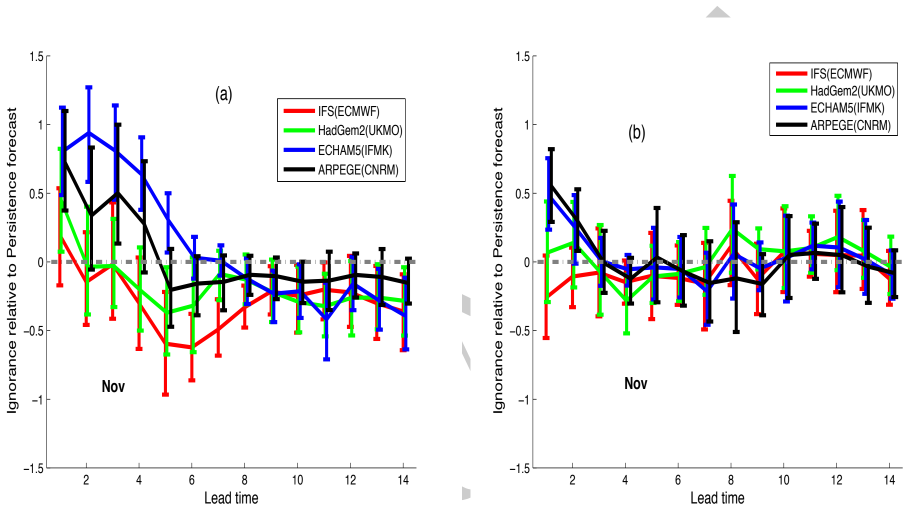

Figure 6: Ignorance score of each model from ENSEMBLES for a) the Nino3.4 index; b) the MDR index, relative to persistence forecast as a function of lead time in months for November launch. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

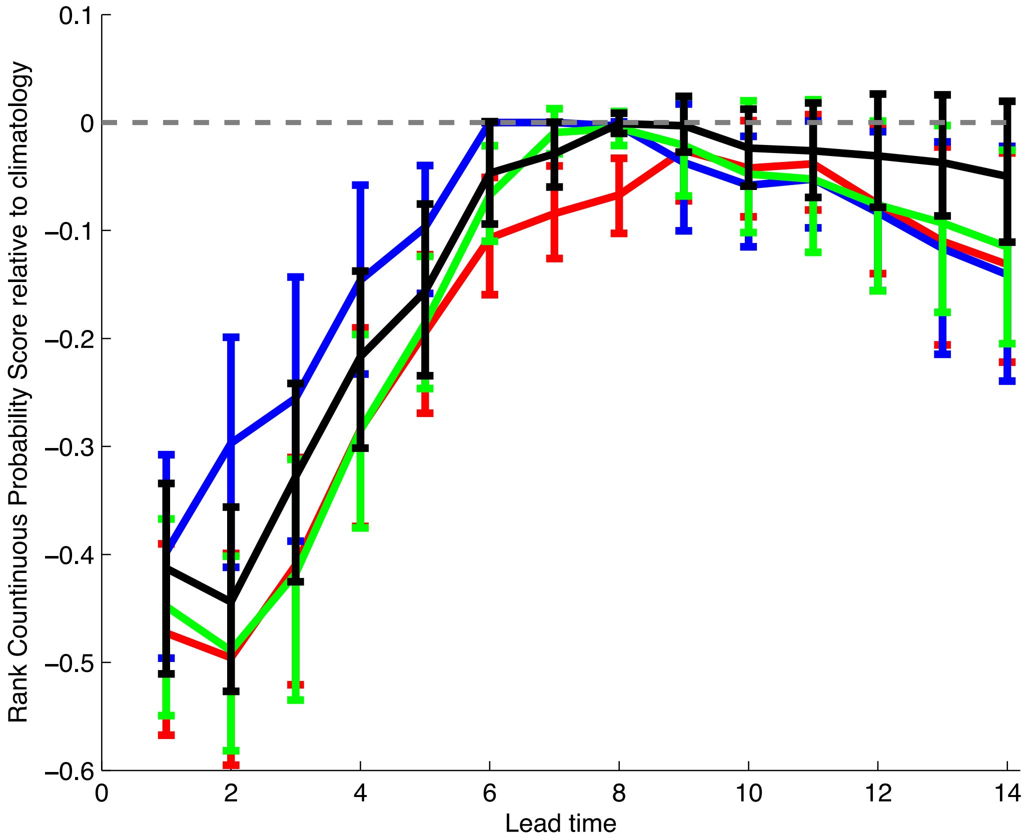

Figure 7: Rank continuous probability score of each model from ENSEMBLES for the Nino3.4 index relative to climatological forecast as a function of lead time in months for November launch. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.