



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 Nihörster

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 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.

1 Probabilistic skill in ensemble seasonal
2 forecasts*

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

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

4 February 26, 2014

5 **Abstract**

6 Operational seasonal forecasting centres employ simulation mod-
7 els to make probability forecasts of future conditions on seasonal to
8 annual lead times. Skill in such forecasts is reflected in the infor-
9 mation they add to purely empirical statistical models, or to earlier
10 versions of simulation models. An evaluation of seasonal probabil-
11 ity forecasts from the DEMETER and the ENSEMBLES multi-model
12 ensemble experiments is presented. Two particular regions are con-
13 sidered (Nino3.4 in the Pacific and Main Development Region in the
14 Atlantic); these regions were chosen before any spatial distribution
15 of skill were examined. The ENSEMBLES models are found to have
16 skill against the climatological distribution on seasonal time scales;
17 for models in ENSEMBLES which have a clearly defined predecessor
18 model in DEMETER the improvement from DEMETER to ENSEM-
19 BLES is discussed. Due to the long lead times of the forecasts and
20 the evolution of observation technology, the forecast-outcome archive
21 for seasonal forecast evaluation is small; arguably evaluation data for

*Under review for Quarterly Journal of the Royal Meteorological Society

22 seasonal forecasting will always be precious. Issues of information con-
23 tamination from in-sample evaluation are discussed, impacts (both
24 positive and negative) of variations in cross-validation protocol are
25 demonstrated. Other difficulties due to the small forecast-outcome
26 archive are identified. The claim that the multi-model ensemble pro-
27 vides a “better” probability forecast than the best single model is
28 examined and challenged. Significant forecast information beyond the
29 climatological distribution is also found in a probability forecast based
30 on persistence. On seasonal time scales, the ENSEMBLES simulation-
31 based probability forecasts add significantly more information to em-
32 pirical probability forecasts than on decadal scales. It is suggested
33 most skillful operational seasonal forecasts available would meld in-
34 formation both from simulation models and empirical models.

35 1 Introduction

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

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

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

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

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

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

¹Attention 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.

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

114 **2 The seasonal multi-model ENSEMBLES fore-** 115 **casts**

116 The ENSEMBLES multi-model ensemble experiment for seasonal-to-annual
117 forecasting comprises global coupled atmosphere-ocean climate models from
118 the UK Met Office (UKMO), Météo France (MF), the European Centre for
119 Medium-Range Weather Forecasts (ECMWF), the Leibniz Institute of Ma-
120 rine Sciences at Kiel University (IFM-GEOMAR) and the Euro-Mediterranean
121 Centre for Climate Change (CMCC-INGV) in Bologna ([9]). In each case
122 the ensemble simulations include all the major radiative forcings; none of the
123 coupled models has flux adjustments ([13, 36, 9]). A set of seasonal hindcast
124 simulations cover the 46 year period from 1960 to 2005. For each launch
125 date the atmosphere and ocean for each model were initialized using realistic

126 estimates of their observed states, providing an ensemble consisting of nine
127 initial condition ensemble members for each model. Hindcast simulations
128 were launched on the first day of February, of May, of August and of Novem-
129 ber each year over the hindcast period and run for seven months. This set
130 of 46 seasonal forecasts for each launch date is analysed below. Additionally
131 each model, with the exception of CMCC-INGV, was run for an extended
132 period up to a lead time of 14 months from the November launch.

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

144 3 Defining probabilistic forecast skill

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

153 I. J. Good's logarithmic score (Ignorance) (see [10, 25, 5]), is unique
154 among several scoring rules ([37]) designed for evaluating the skill of proba-
155 bilistic forecasts. It is the only proper and local score² for continuous vari-

²Proper meaning that it cannot be optimized by hedging the probabilistic forecasts to-
ward 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:

$$S(p(y), Y) = -\log_2(p(Y)), \quad (1)$$

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, $(p_t, Y_t, t = 1, \dots, K)$, the empirical Ignorance score is:

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

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

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

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

183 to produce seasonal probability forecasts (for a full description see [6], and
 184 Appendix A). The climatological distribution is estimated by kernel dress-
 185 ing all available historical observations under cross-validation (see Appendix
 186 B). Figure 1 shows an example of the kernel dressed and blended proba-
 187 bilistic forecast distributions for a subset (over the period 1995-2000) of the
 188 IFS(ECMWF) hindcast simulations from ENSEMBLES for the Nino3.4 in-
 189 dex, launched in November. The blue shaded regions indicate the forecast
 190 percentiles between 1-99% and the red line shows the observed outcome (from
 191 the ERA40 reanalysis) for comparison. The grey shaded bands show the per-
 192 centiles between 1-99% for the climatological distribution.

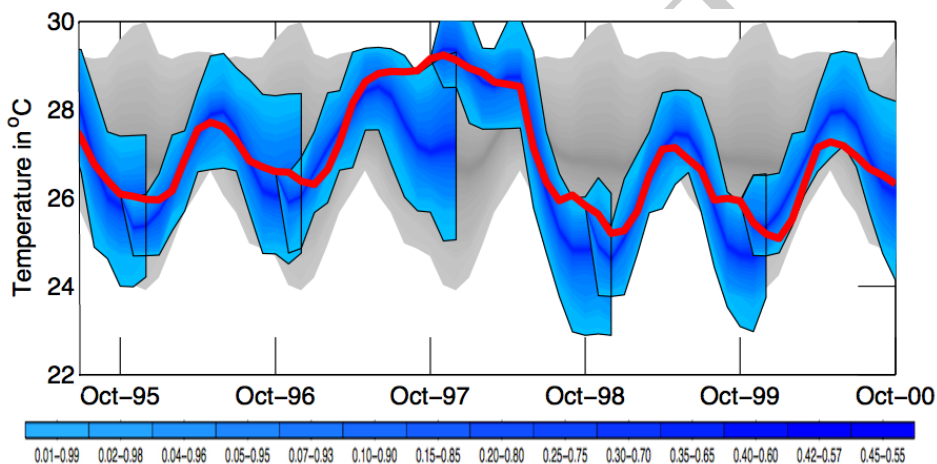


Figure 1: Probabilistic forecast distributions for the IFS(ECMWF) hindcast simulations from ENSEMBLES for the Nino3.4 index, launched in November over the period 1995-2000. The blue shaded regions indicate the forecast percentiles between 1-99% and the red line shows the observed outcome from the ERA40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

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

198 4 ENSEMBLES seasonal forecast skill

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

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

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

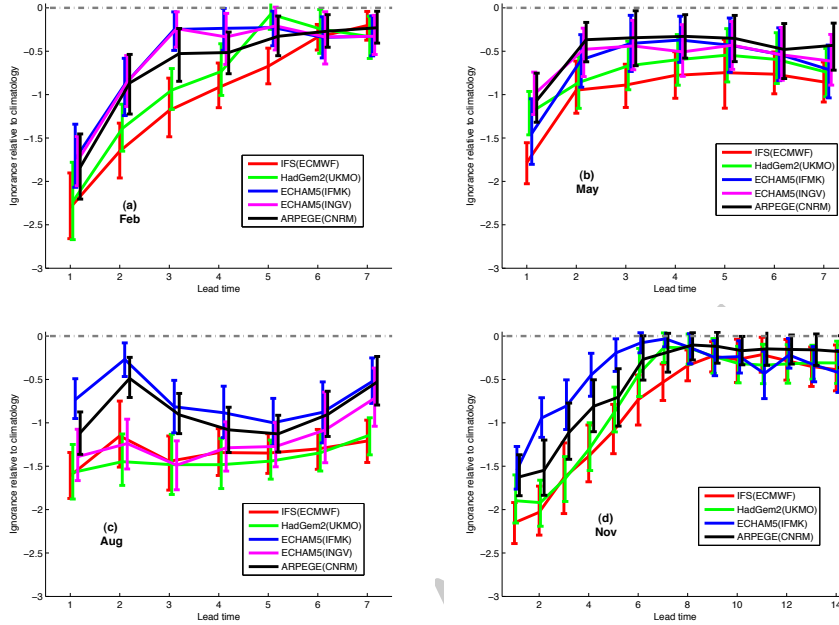


Figure 2: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to climatology as a function of lead time in months. The four different panels show the hindcasts initialized in (a) February, (b) May, (c) August and (d) November. Zero Ignorance indicates a model has no skill relative 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.

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

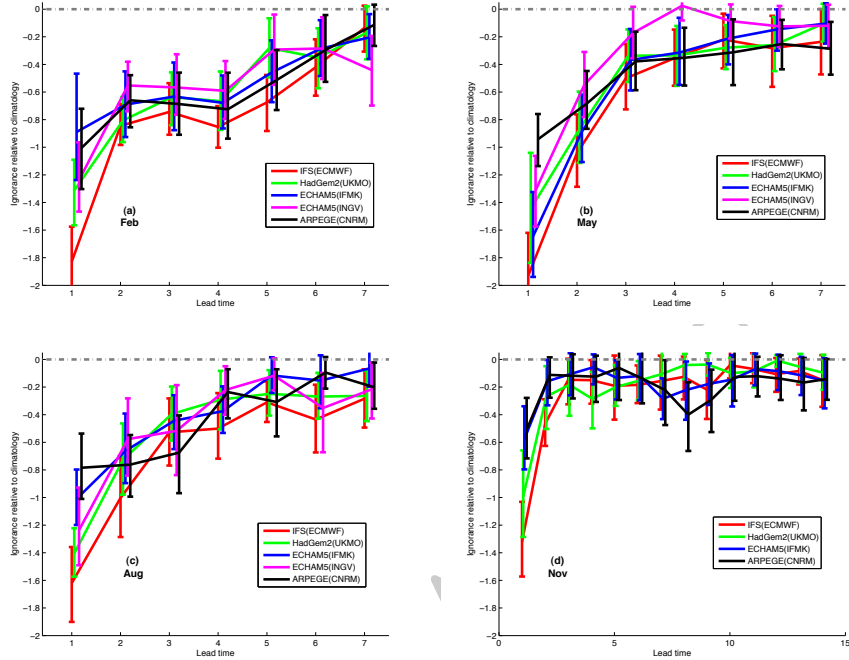


Figure 3: Ignorance score of each model from ENSEMBLES for the MDR index relative to climatology as a function of lead time in months. The four different panels show the hindcasts initialized in (a) February, (b) May, (c) August and (d) November. Zero Ignorance indicates a model has no skill relative 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.

242 Figure 4 shows the Ignorance of each of the ENSEMBLES models for the
 243 Nino3.4 index relative to the IFS(ECMWF) model. There are indeed some
 244 cases where the IFS(ECMWF) model outperforms all other models despite
 245 the overlapping bootstrap resampling intervals in Figure 2. For example,
 246 the IFS(ECMWF) model systematically outperforms the ARPEGE(CNRM),

247 ECHAM5(INGV) and ECHAM5(IFMK) models particularly at early lead
 248 times for most launch dates. In the case analysed above, there is substantial
 249 information in the forecasts from the ENSEMBLES models for the Nino3.4
 250 index even at longer lead times; the IFS(ECMWF) model shows higher skill
 251 (often exceeding 0.5 bits in the first 6 months) relative to the other seasonal
 252 forecast models used in ENSEMBLES.

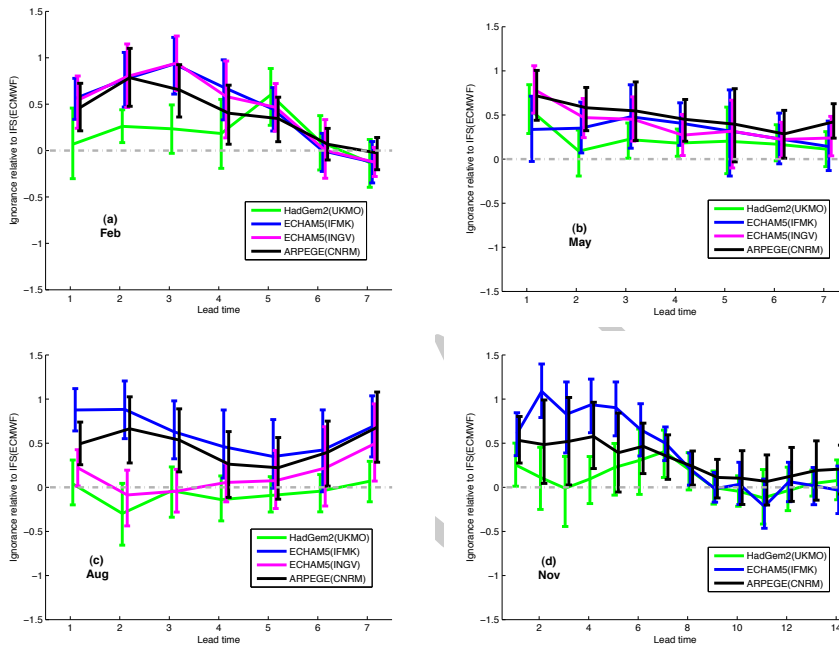


Figure 4: Ignorance score of the ENSEMBLES model forecasts for the Nino3.4 index relative to the IFS(ECMWF) model as a function of lead time in months. Zero Ignorance indicates a model has no skill relative to the IFS(ECMWF) model and negative relative Ignorance scores suggest a model is more skillful 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).

253 5 Contrasting skill of ENSEMBLES & DEME- 254 TER

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

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

274 In order to measure the improvement of forecast performance due to
275 model development from the DEMETER to the ENSEMBLES project, the
276 Ignorance of the forecast distributions derived from pairs of model simula-
277 tions from each project is compared. Although seven European simulation
278 models were used in the DEMETER project, only those models that corre-
279 spond to earlier “versions” of those used in ENSEMBLES are considered.

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

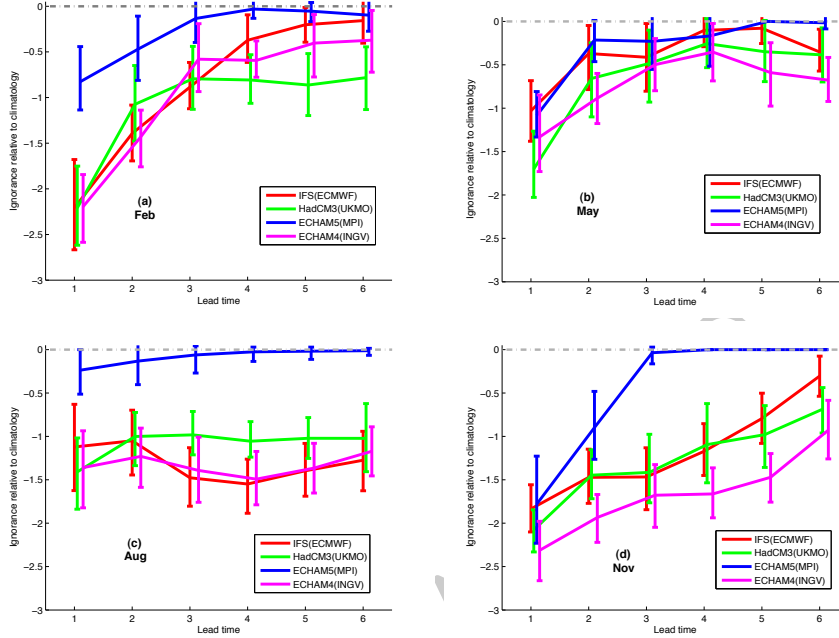


Figure 5: Ignorance score of each model from DEMETER for the Nino3.4 index relative to climatology as a function of lead time in months. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model is more skillful than climatology. Bootstrap 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.

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

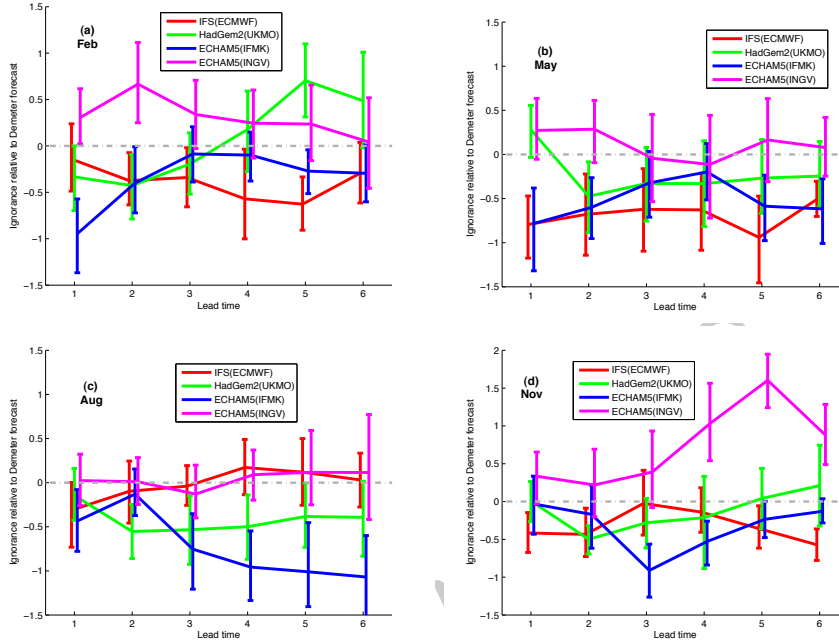


Figure 6: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to the corresponding DEMETER forecasts as a function of lead time in months. Zero Ignorance indicates an ENSEMBLES model has no added skill relative to the corresponding DEMETER model and negative relative Ignorance 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.

294 6 Contrasting ENSEMBLES seasonal skill with 295 persistence forecasts

296 In the previous sections the climatological distribution was used as a bench-
297 mark against the performance of the ENSEMBLES and the DEMETER sea-
298 sonal hindcasts. Whilst comparing skill between dynamical models and cli-

matology 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 model³. 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

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

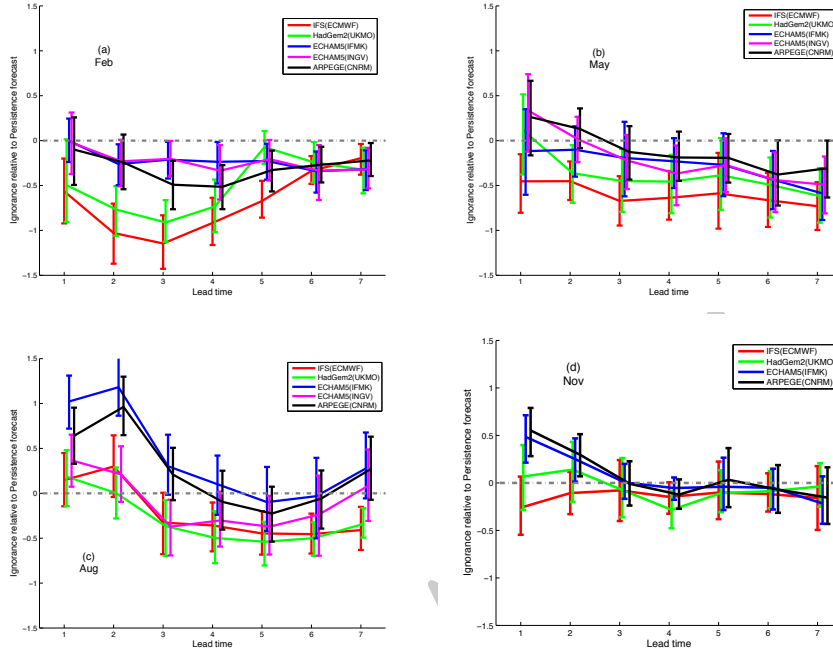


Figure 7: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to persistence forecasts as a function of lead time in months. The four different panels show the hindcasts initialized in (a) February, (b) May, (c) August and (d) November. Scores below zero indicate that an ENSEMBLES 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.

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

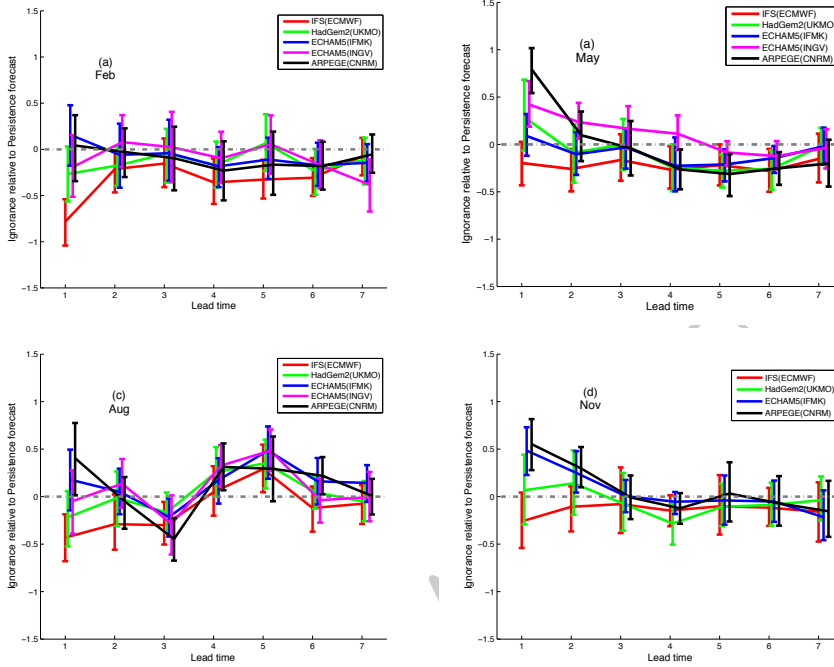


Figure 8: Ignorance score of each model from ENSEMBLES for the MDR index relative to persistence forecasts as a function of lead time in months. The four different panels show the hindcasts initialized in (a) February, (b) May, (c) August and (d) November. Scores below zero indicate that an ENSEMBLES 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.

340 members could be tested more directly.

341 Figure 9 shows the effect of decreasing the number of ensemble members
 342 on the forecast skill for the Nino3.4 index from the IFS(ECMWF) model
 343 launched in November. The skill of two-member ensembles (red) and four-
 344 member ensembles (green) are shown relative to the full nine-member en-
 345 semble (the zero line) both as a set of random draws from the nine original
 346 members without replacement (Figure 9a) and as the average Ignorance of

347 all two- or four-ensemble member combinations (Figure 9b). In Figure 9a
348 most two- and four-member combinations show less skill than the full nine-
349 member ensemble, with only a few ensemble member combinations scoring
350 better than the original ensemble now and then. Figure 9b shows that de-
351 creasing the number of ensemble members systematically decreases the av-
352 erage skill (that is, increases the Ignorance score) across all lead times. This
353 result holds both when decreasing from nine members to four members and
354 when decreasing from four to two ensemble members. At a lead time of six
355 months, where the IFS(ECMWF) model still has non-trivial skill relative to
356 climatology (Figure 2), for example, the two-member forecast places $\sim 7\%$
357 and the four-member ensemble places $\sim 3\%$ less probability density on av-
358 erage on the outcome⁴ relative to the nine-member ensemble (Figure 9b).
359 This result suggests that increasing the current ensemble size of nine would
360 further improve the forecast performance⁵.

361 A larger ensemble could be obtained either by increasing the number
362 of ensemble members from one particular model, or, alternatively, by com-
363 bining simulations from different models to form a multi-model ensemble
364 (see [21, 35]). Of course developing a new, ideally independent model is
365 more costly than increasing the number of ensemble members from an exist-
366 ing model. Combining the output of different (independent) models might,
367 however, have the added advantage of reducing the systematic bias of any
368 single model⁶. One may therefore expect to obtain significantly more in-
369 formation by using multi-model outputs than by increasing the number of
370 ensemble members from a single model.

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

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

⁵Operational 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.

⁶In 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”).

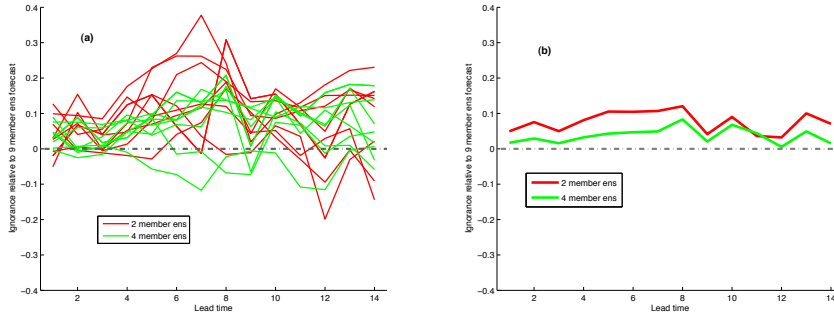


Figure 9: (a) Ignorance of the IFS(ECMWF) model as a function of lead time in months for the Nino3.4 index. The green (red) lines represent the skill of a subset of four-member (two-member) ensemble forecasts relative to the full nine-member ensemble forecast. Each four-member and two-member ensemble 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).

376 ensembles containing four members randomly drawn from the 36 available
 377 ensemble members (nine members from each of four models) without re-
 378 placement. Similarly, Figure 10b shows the skill of multi-model ensembles
 379 containing nine randomly drawn members. The blue line in each case shows
 380 the skill of the full multi-model ensemble, containing 36 members from sim-
 381 ulations of the IFS(ECMWF), HadGem2(UKMO), ECHAM5(IFMK) and
 382 ARPEGE(CNRM) models. The four-member multi-model forecasts are shown
 383 to perform substantially worse than the nine-member IFS(ECMWF) ensem-
 384 ble (indicated by positive Ignorance scores), particularly over short lead times
 385 (up to eight months). The skill of the nine-member multi-model forecasts
 386 are generally increased compared to the four-member forecasts, however,
 387 the single-model, IFS(ECMWF), forecast is still shown to be more skillful⁷

⁷As 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

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

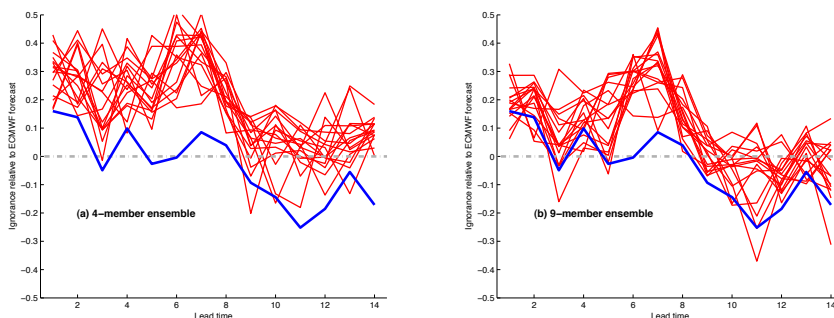


Figure 10: Ignorance of multi-model forecasts as a function of lead time in months for the Nino3.4 index, launched in November, relative to the nine-member IFS(ECMWF) forecast. The blue line represents the multi-model forecast using all 36 ensemble members from the four ENSEMBLES models, equally weighted. 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.

394 8 The importance of being proper

395 It is sometimes said that a multi-model ensemble forecast is more skillful
396 than any of its constituent single-model ensemble forecasts. This may be the
397 case in terms of reducing root-mean-square (RMS) like scores (see [21, 11,
398 4, 35, 36, 2]). For probability forecasts, the definition of skill should reflect
399 the characteristics of the forecast problem. While RMS scores are effectively
400 optimal in linear stochastic systems, they are misleading in evaluating non-
401 linear forecast systems, even when the data is not precious. Indeed RMS
402 scores can be misleading even in the limit of an infinite forecast-verification
403 archive (see [19]). Improvements in RMS skill when using multi-model ensem-
404 bles may be due to error cancellation from independent model contributions
405 (see [11, 15, 4]). For example, if some of the single-model ensembles lie below
406 the observations and some lie above then the ensemble mean could lie closer
407 to the observed outcome than any single ensemble member. While such an
408 error cancellation would reduce the RMS score, rewarding the multi-model
409 forecast more than any single model contribution, a proper skill score ([5])
410 would not credit this “false” skill. Similarly, combining ensemble members
411 from different models may serve to reduce the variance of ensemble mean
412 statistics, which in turn may lead to a lower RMS score. Indeed, if the en-
413 semble variance is large, adding “information free” ensemble members at the
414 mean value will reduce the RMS error, but need not improve a probabilistic
415 score.

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

427 9 Multiple models Ensembles when data are 428 precious

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

$$p_{mm} = \sum_i \omega_i p_i, \quad (4)$$

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

446 To avoid complications with fitting model weights a simple iterative method
447 to combine models is used below: First, a reference forecast distribution is
448 derived from the ensemble members of one particular candidate model, in
449 this case the IFS(ECMWF) forecasts, which were argued to provide the most
450 skillful seasonal forecasts for the Nino3.4 index back in Section 4. Each of the
451 other candidate models, in turn, is then combined with the IFS(ECMWF)
452 model by deriving a forecast distribution from the ensemble members of
453 both models, equally weighted. The skill of each two-model combination is
454 computed in terms of Ignorance relative to the IFS(ECMWF) reference fore-
455 cast and shown in Table 1 for the November launch forecasts of the Nino3.4
456 index. Each model combination shows the average relative Ignorance (neg-
457 ative scores indicate an improvement over simply using the IFS(ECMWF)
458 forecast). Positive values in the 5th, 8th and 11th columns of Table 1 show
459 that there is no clear improvement in skill for any two-model combination in

460 this case, particularly at lead times less than eight months. Arguably beyond
461 eight months the improvements in skill are not significant; the bootstrap re-
462 sampling intervals overlap with zero relative skill in each case. Table 2 shows
463 the corresponding results when other models are combined with the UKMO
464 model. In this case combining with ECMWF tends to improve the average
465 Ignorance at all lead times (negative values in 4th and 5th columns of Table
466 2), but no other combination does this. Starting with ECMWF, combining
467 UKMO has a much smaller effect. In cases where significant improvements
468 are found from such a model combination then further models could be in-
469 cluded into the multi-model forecast by choosing those models which yield
470 the biggest improvement in skill and adding them into the forecast one by
471 one with equal weight until no further skill can be added. In this case, how-
472 ever, results suggest that the most skillful seasonal forecasts are provided by
473 using ensemble members from a single model.

474 **10 Establishing skill when data are precious**

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

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

493 Figure 11 shows the skill of forecasts from the ENSEMBLES models using
494 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.

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

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

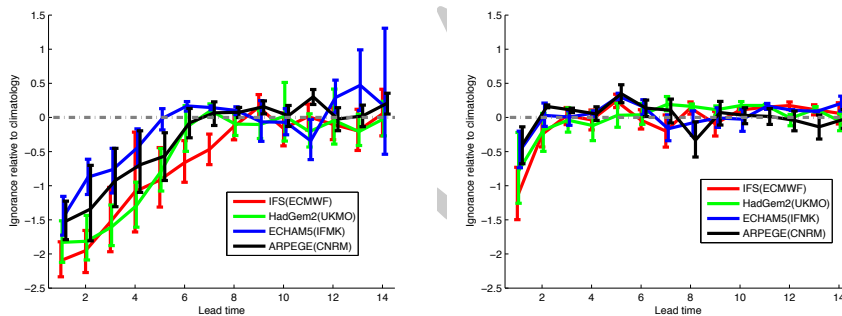


Figure 11: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true cross-validation for, (a) forecasts of the Nino3.4 index and (b) forecasts of the MDR index launched in November. Zero Ignorance indicates a model has no skill relative to climatology 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.

510 Figure 12 illustrates the effect of the different cross-validation protocols

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

525 **11 Conclusions**

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

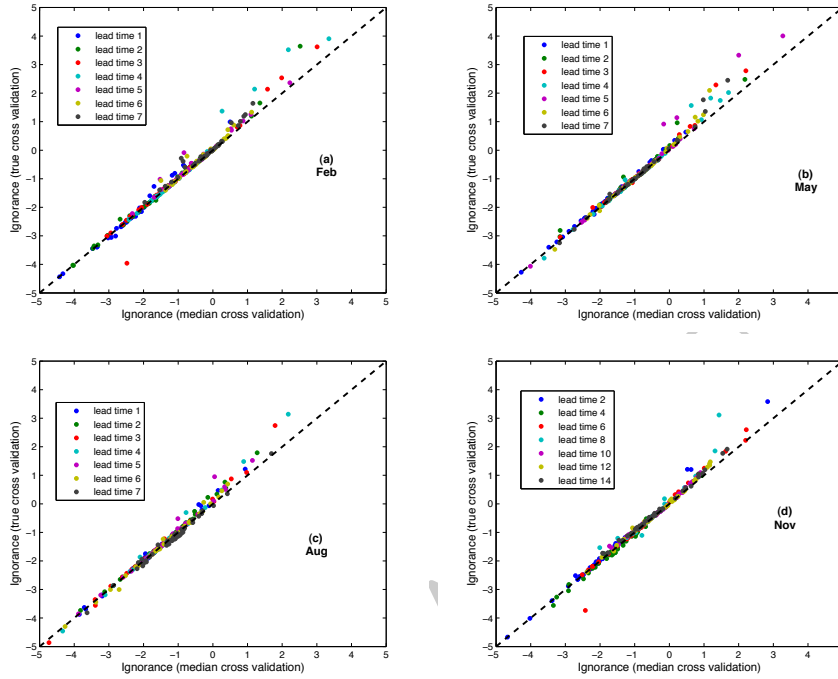


Figure 12: Comparison of Ignorance scores for the IFS(ECMWF) model from ENSEMBLES relative to climatology using the median and true cross-validation protocols for forecasts of the Nino3.4 index, launched in the months as indicated. On average the true cross-validation protocol shows a reduction in skill (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).

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

551 A From Simulation to a PDF

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

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

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

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

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

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

$$p(\cdot) = \alpha p_m(\cdot) + (1 - \alpha) p_c(\cdot), \quad (7)$$

578 where p_m is the density function generated by dressing the model ensemble
579 and p_c is the estimate of climatological density. The blending parameter α
580 determines how much weight is placed in the model. Specifying the three
581 values (kernel width σ , kernel offset μ and weight α) at each lead time de-
582 fines the forecast distribution. These parameters are fitted simultaneously by
583 optimising the empirical Ignorance score, using a cross-validation protocol⁸
584 as described in Appendix B.

585 B Information Contamination and Cross-validation

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

601 First, define the forecast probability distribution to be $p(x, X_t, \Theta)$, $t =$
602 $1, \dots, N$, where X represents the ensemble forecast at time t , Θ represents
603 a vector of parameters (including the kernel width σ , offset μ and blending
604 parameter α) to be fitted and N is the number of forecasts. The correspond-
605 ing outcomes are defined to be s_t . For each forecast at time $j = 1, \dots, N$,
606 leave out one pair of forecast-outcome data (X_j, s_j) and use the remaining

⁸As 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.

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

611 The leave-one-out protocol described in the previous paragraph is not
612 pure cross-validation as $\bar{\Theta}$ arguably contains information from every (X_j, s_j)
613 when the median is taken. To achieve pure cross-validation, the following
614 protocol is adopted. For each forecast at time j , first leave out (X_j, s_j) , then
615 for the remaining set apply the median cross-validation protocol described
616 above to obtain N parameter values Θ_j . The value Θ_j at each time j is then
617 independent of (X_j, s_j) . The forecast empirical Ignorance is then given by
618 $\sum_{j=1}^N -\log_2 p(s_j, X_j, \bar{\Theta}_j)$. This protocol ensures that the parameters Θ_j have
619 no explicit dependence on the datum used to evaluate them at the cost of a
620 smaller learning set(s). Even in this case, the datum was known to the ana-
621 lyst. Indeed, use of a common archive in DEMETER and in ENSEMBLES
622 (Stream Two) clouds the possibility of assigning clear statistical significance
623 to estimates of expected skill.

624 Acknowledgements

625 This research was supported by the EU Framework 6 ENSEMBLES project;
626 it was also supported both by the LSE’s Grantham Research Institute on
627 Climate Change and the Environment and the ESRC Centre for Climate
628 Change Economics and Policy, funded by the Economic and Social Research
629 Council and Munich Re. L.A.S. gratefully acknowledges support from Pem-
630 broke College, Oxford.

631 References

- 632 [1] A. Alessandri, A. Borrelli, S. Masina, P. D. Pietro, A. Carril, A. Cher-
633 chi, S. Gualdi, and A. Navarra. The INGV-CMCC seasonal prediction
634 system: Improved ocean initial conditions. *Monthly Weather Review*,
635 138, 29302952 (2010).
- 636 [2] A. Alessandri, A. Borrelli, A. Navarra, A. Arribas, M. Déqué, P. Rogel,
637 and A. Weisheimer. Evaluation of probabilistic quality and value of the

- 638 ensembles multimodel seasonal forecasts: Comparison with DEMETER.
639 Monthly Weather Review, 139, 2 (2011).
- 640 [3] J. M. Bernardo. Expected information as expected utility. Annals of
641 Statistics, 7 (7):686C690 (1979).
- 642 [4] N. E. Bowler, A. Arribas, K. R. Mylne. The benefits of multi-analysis
643 and poor-mans ensembles. Monthly Weather Review, 136, 41134129
644 (2008).
- 645 [5] J. Bröcker, L. A. Smith. Scoring Probabilistic Forecasts: On the Impor-
646 tance of Being Proper. Weather and Forecasting, 22 (2), 382-388 (2006).
- 647 [6] J. Bröcker and L. A. Smith. From ensemble forecasts to predictive dis-
648 tribution functions. Tellus A, 60, 663-678 (2007).
- 649 [7] C. A. S. Coelho, D. B. Stephenson, M. Balmaseda, F. J. Doblas-Reyes
650 and G. J. van Oldenborgh. Towards an integrated seasonal forecasting
651 system for South America. Journal of Climate, 19, 3704-3721 (2006).
- 652 [8] F. J. Doblas-Reyes, R. Hagedorn and T. N. Palmer. The rationale behind
653 the success of multi-model ensembles in seasonal forecasting. Part II:
654 Calibration and combination. Tellus A, 57 (2005).
- 655 [9] F. J. Doblas-Reyes, A. Weisheimer, T. N. Palmer, J. M. Murphy and
656 D. Smith. Forecast quality assessment of the ENSEMBLES seasonal-to-
657 decadal Stream 2 hindcasts. Technical Memorandum (ECMWF), 621
658 (2010).
- 659 [10] I. J. Good. Rational decisions. Journal of the Royal Statistical Society,
660 XIV(1):107-114 (1952).
- 661 [11] R. Hagedorn, F. J. Doblas-Reyes, and T. N. Palmer. The rationale be-
662 hind the success of multi-model ensembles in seasonal forecasting. Part
663 I: Basic concept. Tellus A, 57, 219233 (2005).
- 664 [12] R. Hagedorn and L. A. Smith. Communicating the value of probabilistic
665 forecasts with weather roulette. Meteorological Applications, 16(2):143-
666 155 (2009).

- 667 [13] C. D. Hewitt and D. J. Griggs. Ensembles-based Predictions of Climate
668 Changes and their Impacts. *Eos, Transactions American Geophysical*
669 *Union*, 85, p566 (2004).
- 670 [14] B. Ingleby and M. Huddleston. Quality control of ocean temperature
671 and salinity profiles - historical and real-time data. *Journal of Marine*
672 *Systems*, 65:158-175 (2007).
- 673 [15] I.-S. Kang and J. Yoo. Examination of multi-model ensemble seasonal
674 prediction methods using a simple climate system. *Climate Dynamics*,
675 26:285294 (2006).
- 676 [16] J. L. Kelly, Jr. A New Interpretation of Information Rate. *Bell System*
677 *Technical Journal*, 35:917-926 (1956).
- 678 [17] B. P. Kirtman, D. Min, J. M. Infanti, J. L. Kinter III, D. A. Paolino, Q.
679 Zhang, H. van den Dool, S. Saha, M. P. Mendez, E. Becker, P. Peng, P.
680 Tripp, J. Huang, D. G. DeWitt, M. K. Tippett, A. G. Barnston, S. Li, A.
681 Rosati, S. D. Schubert, M. Rienecker, M. Suarez, Z. E. Li, J. Marshak,
682 Y.-K. Lim, J. Tribbia, K. Pegion, W. J. Merryfield, B. Denis, and E. F.
683 Wood. The North American Multi-Model Ensemble (NMME): Phase-1
684 Seasonal to Interannual Prediction, Phase-2 Toward Developing Intra-
685 Seasonal Prediction. *Bulletin of the American Meteorological Society*,
686 (2013).
- 687 [18] T. N. Krishnamurti, C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z.
688 Zhang, C. E. Williford, S. Gadgil, and S. Surendran. Improved Weather
689 and Seasonal Climate Forecasts from Multimodel Superensemble. *Sci-*
690 *ence*, 285(5433):15481550 (1999).
- 691 [19] P. E. McSharry and L. A. Smith. Better nonlinear models from noisy
692 data: Attractors with maximum likelihood. *Physical Review Letters*, 83,
693 (21):4285-4288 (1999).
- 694 [20] G. J. van Oldenborgh, M. A. Balmaseda, L. Ferranti, T. N. Stockdale
695 and D. L. T. Anderson. Did the ECMWF seasonal forecast model out-
696 perform statistical ENSO forecast models over the last 15 years? *Journal*
697 *of Climate*, 18, 3240-3249 (2005).
- 698 [21] T. N. Palmer, A. Alessandri, U. Andersen, P. Cantelaube, M. Davey,
699 P. Délecluse, M. Déqué, E. Diez, F. J. Doblas-Reyes, H. Feddersen, R.

- 700 Graham, S. Gualdi, J.-F. Gu er emy, R. Hagedorn, M. Hoshen, N. Keenly-
701 side, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B.
702 Orfila, P. Rogel, J.-M. Terres, M. C. Thomson. Development of a Euro-
703 pean multimodel ensemble system for seasonal-to-interannual prediction
704 (DEMETER). *Bulletin of the American Meteorological Society*, 85, 853-
705 872 (2004).
- 706 [22] P. Peng, A. Kumar, H. van den Dool and A. G. Barnston. An analysis of
707 multimodel ensemble predictions for seasonal climate anomalies. *Journal*
708 *of Geophysical Research*, 107(23):4710 (2002).
- 709 [23] A. E. Raftery, T. Gneiting, F. Balabdaoui and M. Polakowski. Us-
710 ing Bayesian model averaging to calibrate forecast ensembles. *Monthly*
711 *Weather Review*, 131, 1155-1174 (2005).
- 712 [24] B. Rajagopalan, U. Lall and S. E. Zebiak. Categorical climate forecasts
713 through regularization and optimal combination of multiple GCM en-
714 sembles. *Monthly Weather Review*, 130:1792-1811 (2002).
- 715 [25] M. S. Roulston and L. A. Smith. Evaluating Probabilistic Forecasts
716 Using Information Theory. *Monthly Weather Review*, 130(6):1653-1660
717 (2002).
- 718 [26] D. M. Smith, R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H.
719 Pohlmann, and A. A. Scaife. Skilful multi-year predictions of atlantic
720 hurricane frequency. *Nature Geoscience*, 3(12):846-849 (2010).
- 721 [27] L.A. Smith. Identification and Prediction of Low-Dimensional Dynam-
722 ics. *Physica D*, 58 (1-4): 50-76, (1992).
- 723 [28] L. A. Smith. Proceedings International School of Physics “Enrico
724 Fermi”. CXXXIII, 177, Bologna, Italy (1997).
- 725 [29] L. A. Smith, H. L. Du, and S. Higgins. Necessary Conditions for as-
726 signing Sensible Model Weights in Seasonal and Decadal Forecasting. in
727 preparation for *Tellus* (2013).
- 728 [30] E. B. Suckling and L. A. Smith. An evaluation of decadal probability
729 forecasts from state-of-the-art climate models. *Journal of Climate*, 26,
730 23 (2013).

- 731 [31] The Met Office. 3-month outlook for contingency planning:
732 User guidance. HM Government document, The Met Of-
733 fice, Devon UK, [http://www.metoffice.gov.uk/media/pdf/g/o/3-](http://www.metoffice.gov.uk/media/pdf/g/o/3-month_Outlook_User_Guidance-150.pdf)
734 [month_Outlook_User_Guidance-150.pdf](http://www.metoffice.gov.uk/media/pdf/g/o/3-month_Outlook_User_Guidance-150.pdf) (accessed: December 2013).
- 735 [32] H. M. Van Den Dool. Empirical methods in short-term climate predic-
736 tion, Oxford University Press, (2007).
- 737 [33] F. Vitart, M. R. Huddleston, M. Déqué, D. Peake, T. N. Palmer, T.
738 N. Stockdale, M. K. Davey, S. Ineson and A. Weisheimer. Dynamically-
739 based seasonal forecast of Atlantic tropical storm activity issued in June
740 by EUROSIP. *Geophysical Research Letters*, 34, L16815 (2007).
- 741 [34] B. Wang, J.-Y. Lee, I.-S. Kang, J. Shukla, C. K. Park, A. Kumar, J.
742 Schemm, S. Cocks, J. S. Kug, J. J. Luo, T. Zhou, B. Wang, X. Fu, W.
743 T. Yun, O. Alves, E. Jin, J. Kinter, B. Kirtman, T. Krishnamurti, N.
744 Lau, W. Lau, P. Liu, P. Pegion, T. Rosati, S. Schubert, W. Stern, M.
745 Suarez, and T. Yamagata. Advance and prospectus of seasonal predic-
746 tion: assessment of the APCC/CliPAS 14-model ensemble retrospective
747 seasonal prediction (1980-2004). *Climate Dynamics*, 33(1):93117 (2009).
- 748 [35] A. P. Weigel, M. A. Liniger, and C. Appenzeller. Can multi-model
749 combination really enhance the prediction skill of probabilistic ensem-
750 ble forecasts? *Quarterly Journal of the Royal Meteorological Society*,
751 134(630):241C260 (2008).
- 752 [36] A. Weisheimer, F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Ar-
753 ribas, M. Déqué, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel.
754 ENSEMBLES: A new multi-model ensemble for seasonal-to-annual
755 predictions and Skill and progress beyond DEMETER in forecasting
756 tropical Pacific SSTs. *Geophysical Research Letters*, 36(21) (2009).
- 757 [37] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*, Volume
758 91, Second Edition (International Geophysics). Academic Press, 2 edi-
759 tion (2005).

1 Supplemental Material
2 Probabilistic skill in ensemble seasonal
3 forecasts

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

6 February 26, 2014

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

- 9 • Details of the ENSEMBLES & DEMETER simulation models used in
10 the seasonal forecast evaluation are given in Table 1 and Table 2. Ta-
11 ble 1 contains a description of the simulation models that constitute
12 the ENSEMBLES seasonal hindcast experiment. There were seven
13 comprehensive European global coupled atmosphere-ocean models de-
14 veloped in the DEMETER project. Table 2 lists a subset of simulation
15 models from the DEMETER project, which are directly comparable to
16 the models used for the ENSEMBLES hindcasts.
- 17 • Ignorance scores of both the DEMETER and the ENSEMBLES fore-
18 casts using the true cross-validation protocol (as described in Appendix
19 B of the main manuscript) are presented in Figure 1-4, which can be
20 compared with those generated using median cross-validation (See Fig-
21 ure 2, 3, 5, 6 of the main manuscript).
- 22 • Figure 5a and 5b illustrate the results shown in Table 1 and Table 2
23 of the main manuscript in terms of the additional information gained
24 from multi-model combination with the best and second best ranked
25 models, respectively.
- 26 • Table 3 shows the statistics of each of the four simulation models'
27 forecast performance in rank ordered according to Ignorance score for
each forecast of Nino3.4 index at November launch. It appears the

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

30 • Ignorance score of each model from ENSEMBLES relative to persis-
31 tence forecasts as a function of lead time at November launch for both
32 Nino3.4 index and MDR index is shown in Figure 6. In Figure 7d &
33 8d of the main manuscript, results after lead time 7 were not presented
34 in order to conveniently compare with the rest of the panels of Figure
35 7 & 8. Results for the whole range of lead time is presented here in
36 Figure 6.

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

DRAFT

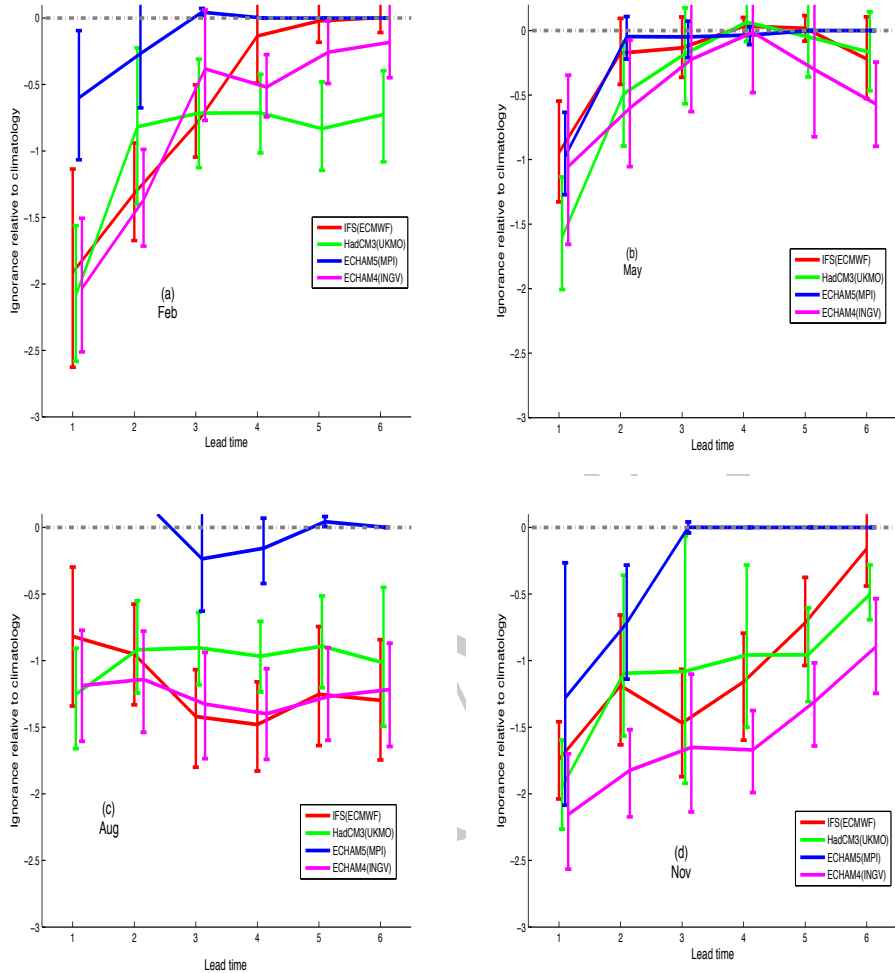


Figure 1: Ignorance score of each model from DEMETER for the Nino3.4 index relative to climatology as a function of lead time in months using true leave-one-out cross-validation. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model 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.

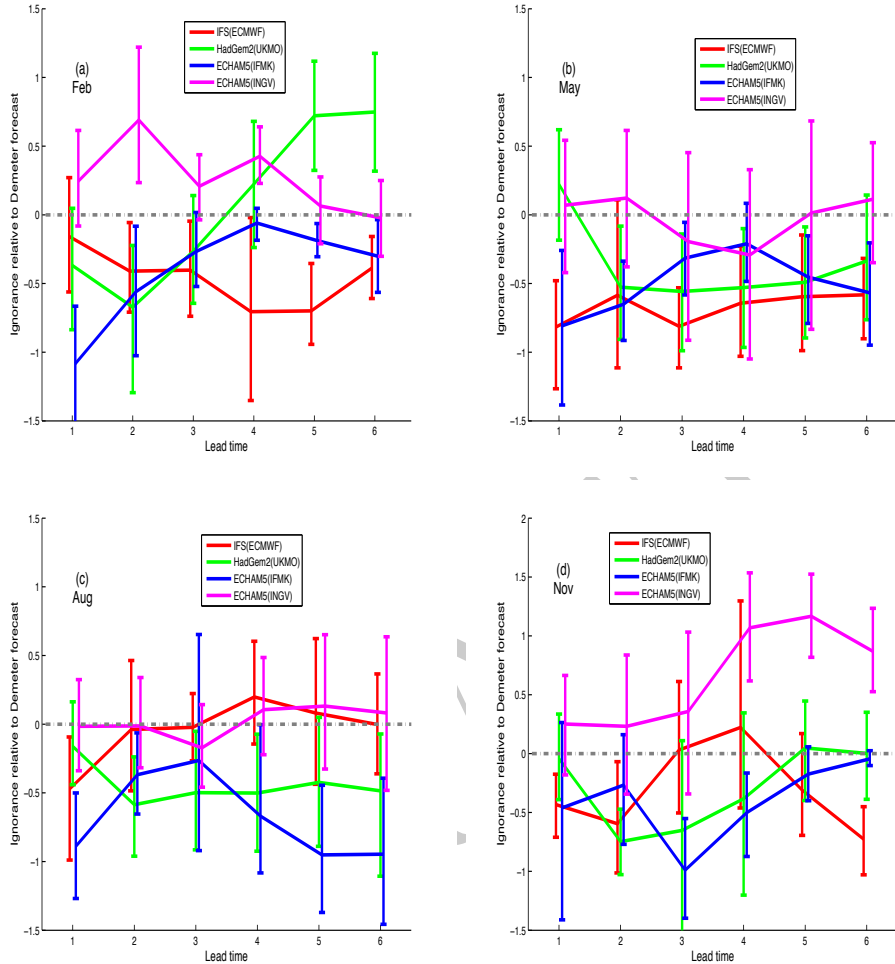


Figure 2: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to the equivalent DEMETER forecasts as a function of lead time in months using true cross-validation. Zero Ignorance indicates an ENSEMBLES model has no skill relative to the corresponding DEMETER model and negative 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.

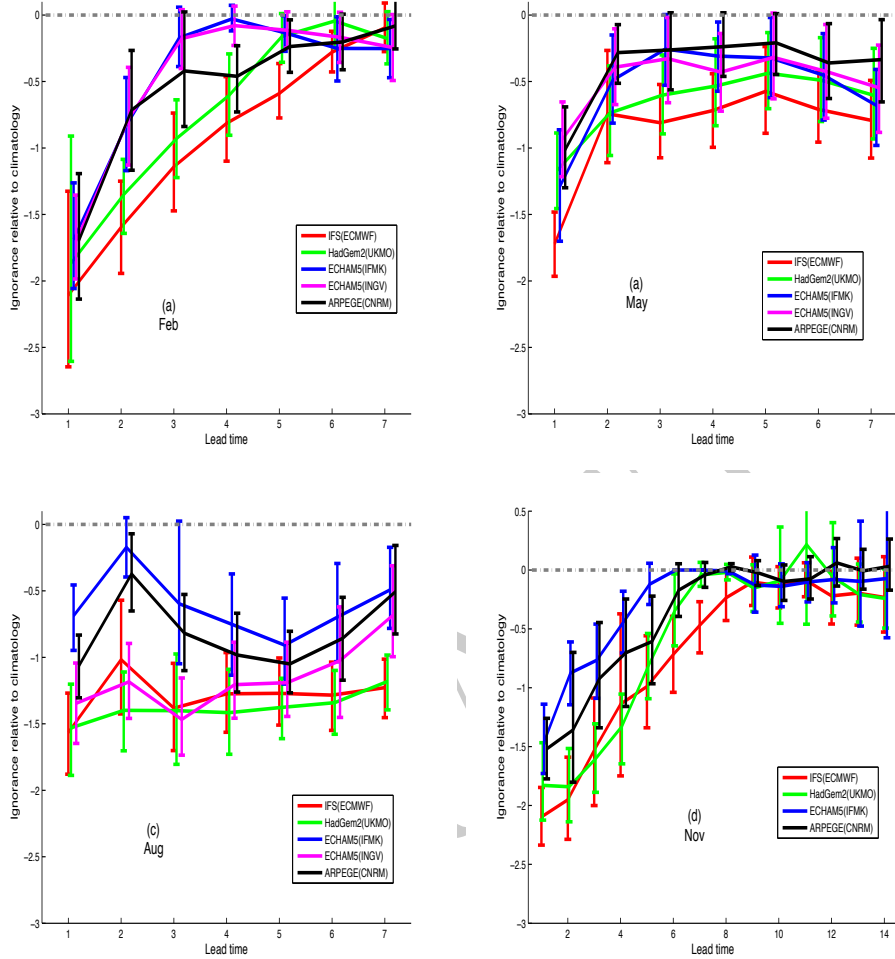


Figure 3: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true leave-one-out cross-validation for forecasts of the Nino3.4 index. The four different panels show the hindcasts initialized in February, May, August and November. Zero Ignorance 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.

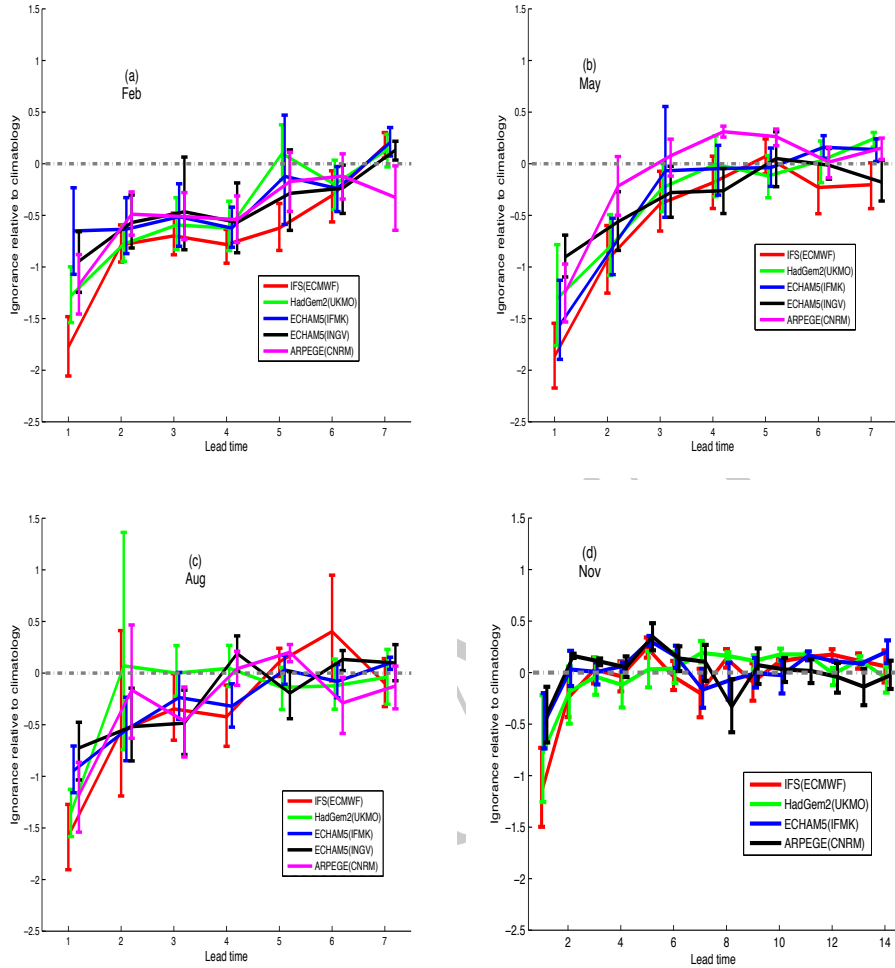


Figure 4: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true leave-one-out cross-validation for forecasts at Main Development Region. The four different panels show the hindcasts initialized in February, May, August and November. Zero 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.

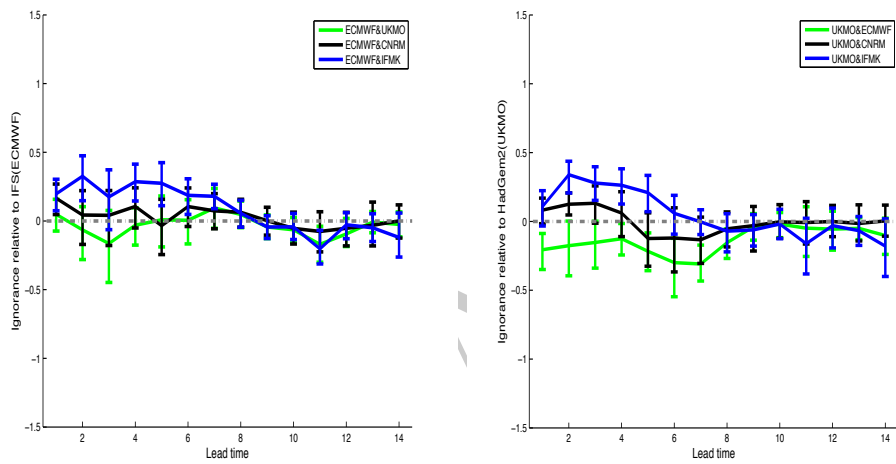


Figure 5: Ignorance score of each two-model forecast combination, as labelled, relative to a) the IFS(ECMWF) forecast; b) the HadGem2(UKMO) forecast, at each lead time for forecasts of the Nino3.4 index, launched in November. In each case the individual models are also blended with the climatological 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.

Institute	Atmosphere		Ocean		References
	Model	Resolution	Model	Resolution	
ECMWF	IFS CY31R1	T1159/L62	HOPE	0.3-1.4/L29	Stockdale et al., [2009]; Balmaseda et al., [2008]
UKMO	HadGEM2-A	N96/L38	HadGEM2-O	0.33-1/L20	Collins et al., [2008]
CNRM	ARPEGE4.6	T63	OPA8.2	2/L31	Daget et al., [2009]; Salas y Melia, [2002]
IFMK	ECHAM5	T63/L31	MPI-OM1	1.5/L40	Keenlyside et al., [2005]; Jungclaus et al., [2006]
INGV	ECHAM5	T63/L19	OPA8.2	2/L31	Alessandri et al., [2009]; Di Pietro and Masina, [2009]

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

Institute	Atmosphere		Ocean		References
	Model	Resolution	Model	Resolution	
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%
	$p(x \geq No.Rank1)$	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%
	$p(x \geq No.Rank1)$	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%
	$p(x \geq No.Rank1)$	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%
	$p(x \geq No.Rank1)$	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. $p(x \geq No.Rank1)$ 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.

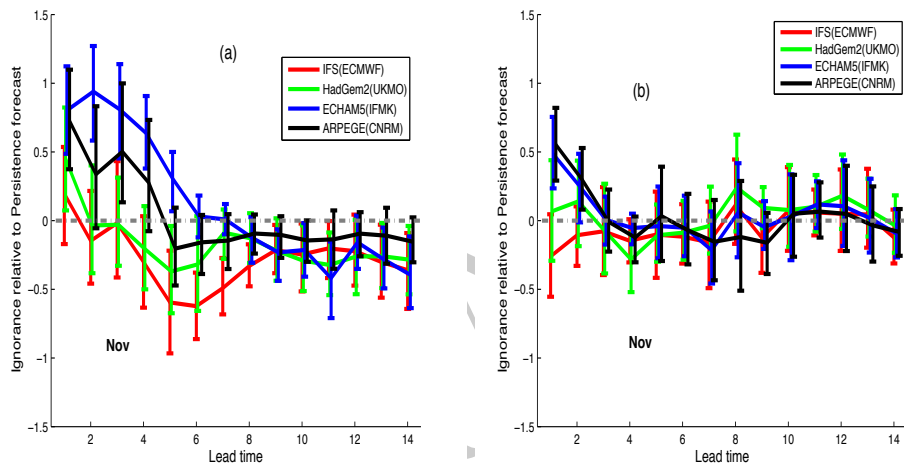


Figure 6: Ignorance score of each model from ENSEMBLES for a) the Nino3.4 index; b) the MDR index, relative to persistence forecast as a function of lead time in months for November launch. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

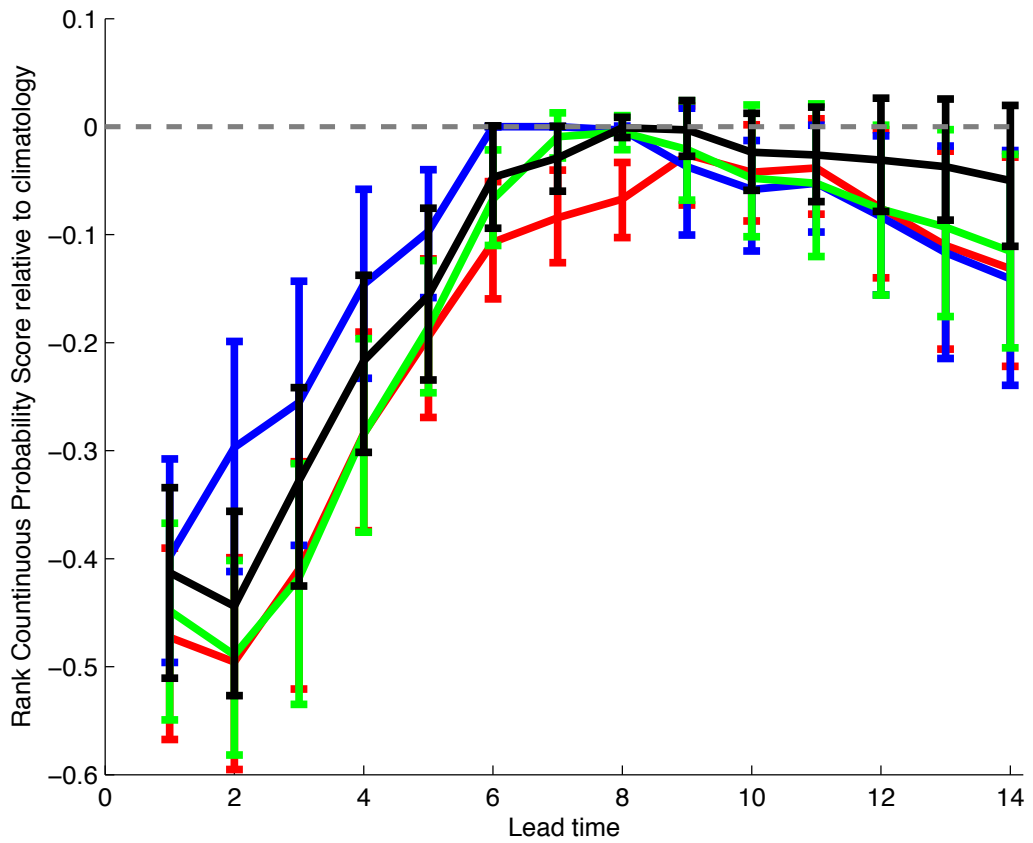


Figure 7: Rank continuous probability score of each model from ENSEMBLES for the Nino3.4 index relative to climatological forecast as a function of lead time in months for November launch. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

1 Supplemental Material
2 Probabilistic skill in ensemble seasonal
3 forecasts

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

5 February 26, 2014

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

- 8 • Details of the ENSEMBLES & DEMETER simulation models used in
9 the seasonal forecast evaluation are given in Table 1 and Table 2. Ta-
10 ble 1 contains a description of the simulation models that constitute
11 the ENSEMBLES seasonal hindcast experiment. There were seven
12 comprehensive European global coupled atmosphere-ocean models de-
13 veloped in the DEMETER project. Table 2 lists a subset of simulation
14 models from the DEMETER project, which are directly comparable to
15 the models used for the ENSEMBLES hindcasts.
- 16 • Ignorance scores of both the DEMETER and the ENSEMBLES fore-
17 casts using the true cross-validation protocol (as described in Appendix
18 B of the main manuscript) are presented in Figure 1-4, which can be
19 compared with those generated using median cross-validation (See Fig-
20 ure 2, 3, 5, 6 of the main manuscript).
- 21 • Figure 5a and 5b illustrate the results shown in Table 1 and Table 2
22 of the main manuscript in terms of the additional information gained
23 from multi-model combination with the best and second best ranked
24 models, respectively.
- 25 • Table 3 shows the statistics of each of the four simulation models'
26 forecast performance in rank ordered according to Ignorance score for
27 each forecast of Nino3.4 index at November launch. It appears the

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

30 • Ignorance score of each model from ENSEMBLES relative to persis-
31 tence forecasts as a function of lead time at November launch for both
32 Nino3.4 index and MDR index is shown in Figure 6. In Figure 7d &
33 8d of the main manuscript, results after lead time 7 were not presented
34 in order to conveniently compare with the rest of the panels of Figure
35 7 & 8. Results for the whole range of lead time is presented here in
36 Figure 6.

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

DRAFT

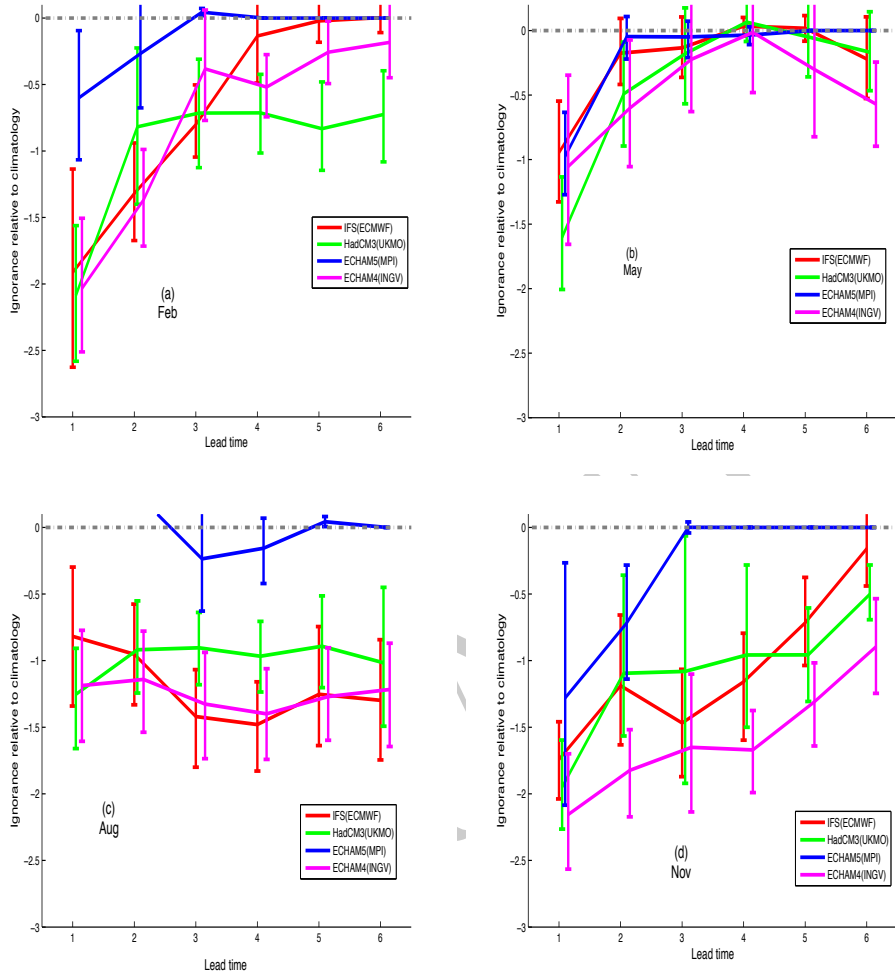


Figure 1: Ignorance score of each model from DEMETER for the Nino3.4 index relative to climatology as a function of lead time in months using true leave-one-out cross-validation. Zero Ignorance indicates a model has no skill relative to climatology and negative relative Ignorance scores suggest a model 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.

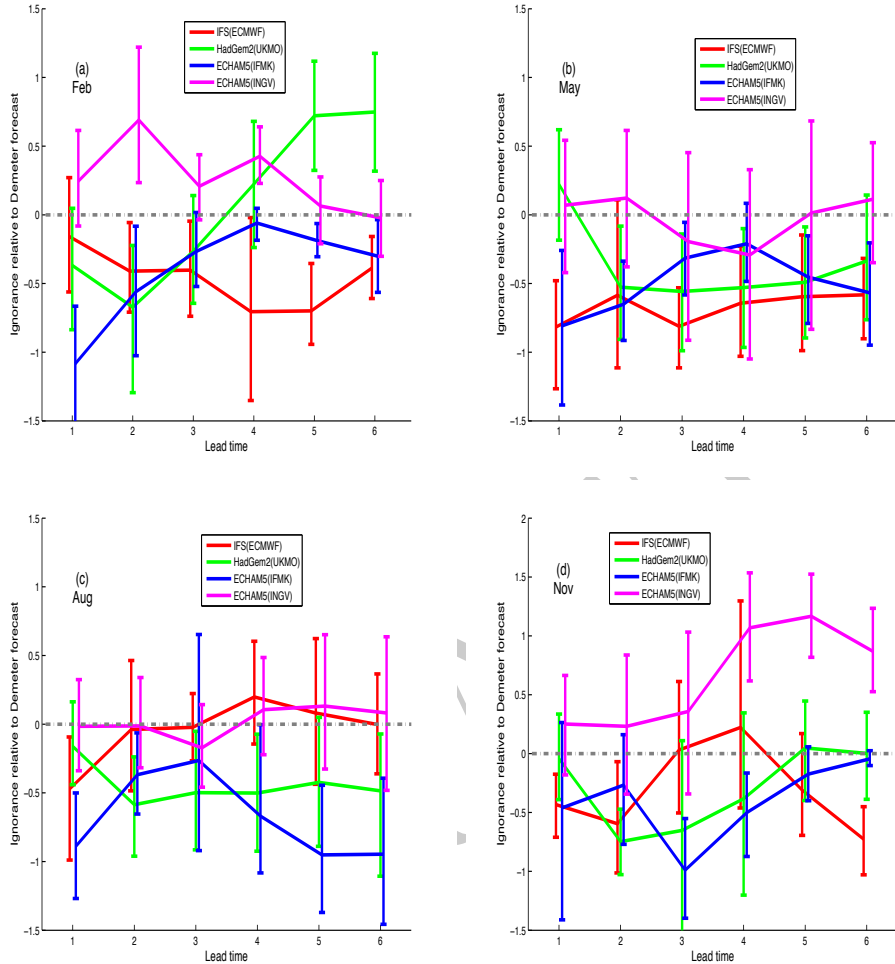


Figure 2: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to the equivalent DEMETER forecasts as a function of lead time in months using true cross-validation. Zero Ignorance indicates an ENSEMBLES model has no skill relative to the corresponding DEMETER model and negative 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.

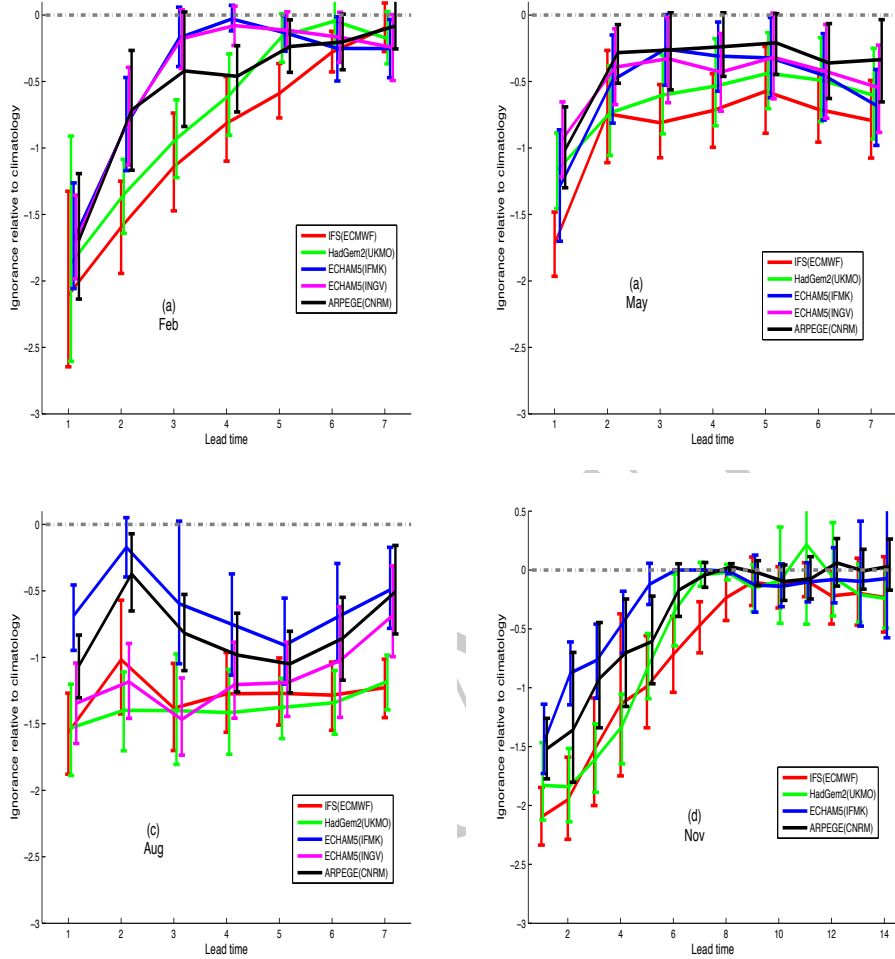


Figure 3: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true leave-one-out cross-validation for forecasts of the Nino3.4 index. The four different panels show the hindcasts initialized in February, May, August and November. Zero Ignorance 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.

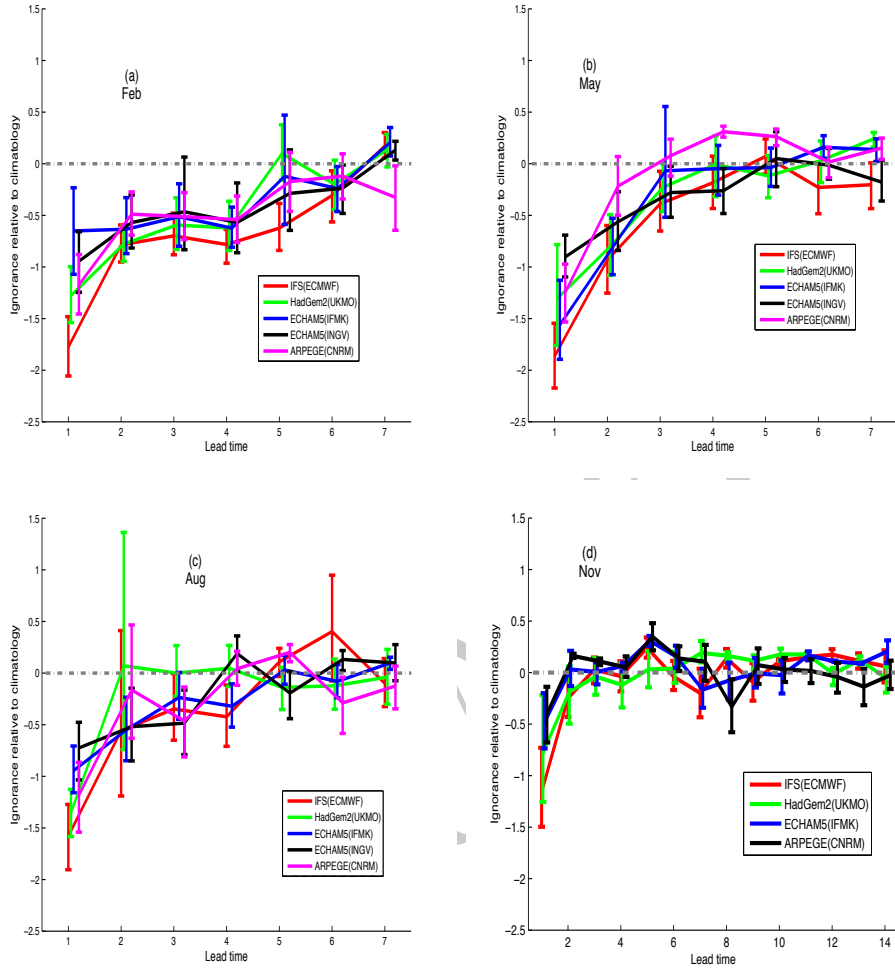


Figure 4: Ignorance score of each model from ENSEMBLES relative to climatology as a function of lead time in months using true leave-one-out cross-validation for forecasts at Main Development Region. The four different panels show the hindcasts initialized in February, May, August and November. Zero 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.

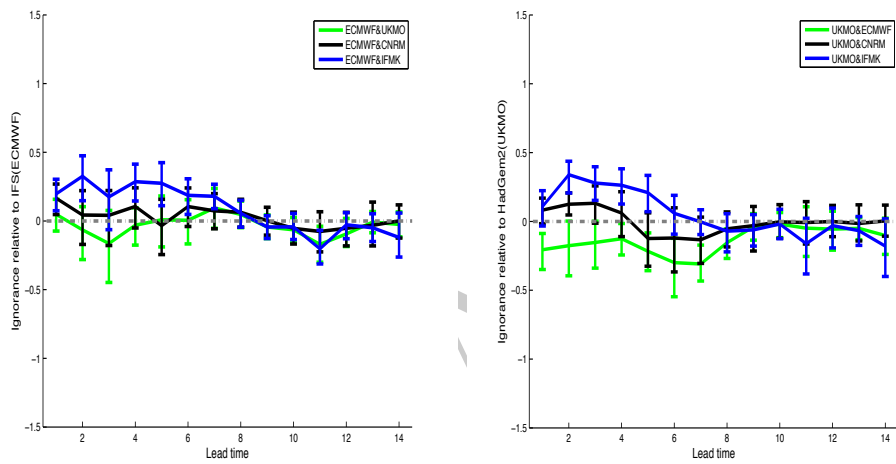


Figure 5: Ignorance score of each two-model forecast combination, as labelled, relative to a) the IFS(ECMWF) forecast; b) the HadGem2(UKMO) forecast, at each lead time for forecasts of the Nino3.4 index, launched in November. In each case the individual models are also blended with the climatological 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.

Institute	Atmosphere		Ocean		References
	Model	Resolution	Model	Resolution	
ECMWF	IFS CY31R1	T1159/L62	HOPE	0.3-1.4/L29	Stockdale et al., [2009]; Balmaseda et al., [2008]
UKMO	HadGEM2-A	N96/L38	HadGEM2-O	0.33-1/L20	Collins et al., [2008]
CNRM	ARPEGE4.6	T63	OPA8.2	2/L31	Daget et al., [2009]; Salas y Melia, [2002]
IFMK	ECHAM5	T63/L31	MPI-OM1	1.5/L40	Keenlyside et al., [2005]; Jungclaus et al., [2006]
INGV	ECHAM5	T63/L19	OPA8.2	2/L31	Alessandri et al., [2009]; Di Pietro and Masina, [2009]

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

Institute	Atmosphere		Ocean		References
	Model	Resolution	Model	Resolution	
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%
	$p(x \geq No.Rank1)$	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%
	$p(x \geq No.Rank1)$	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%
	$p(x \geq No.Rank1)$	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%
	$p(x \geq No.Rank1)$	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. $p(x \geq No.Rank1)$ 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.

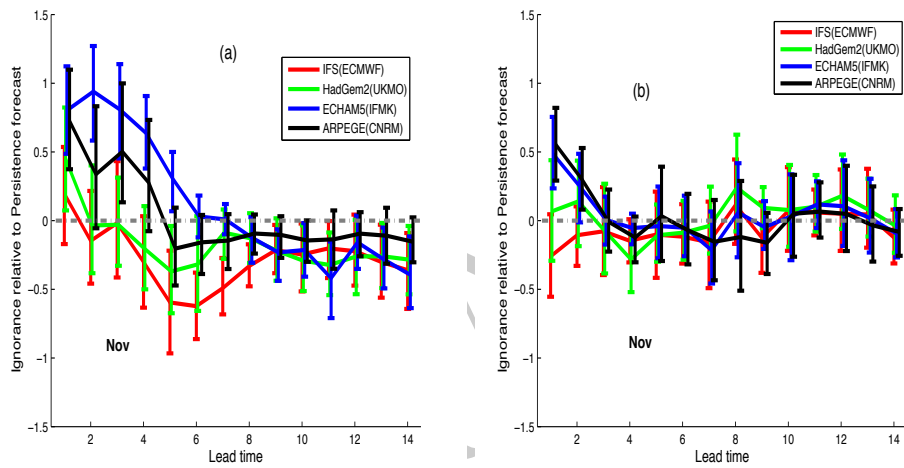


Figure 6: Ignorance score of each model from ENSEMBLES for a) the Nino3.4 index; b) the MDR index, relative to persistence forecast as a function of lead time in months for November launch. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

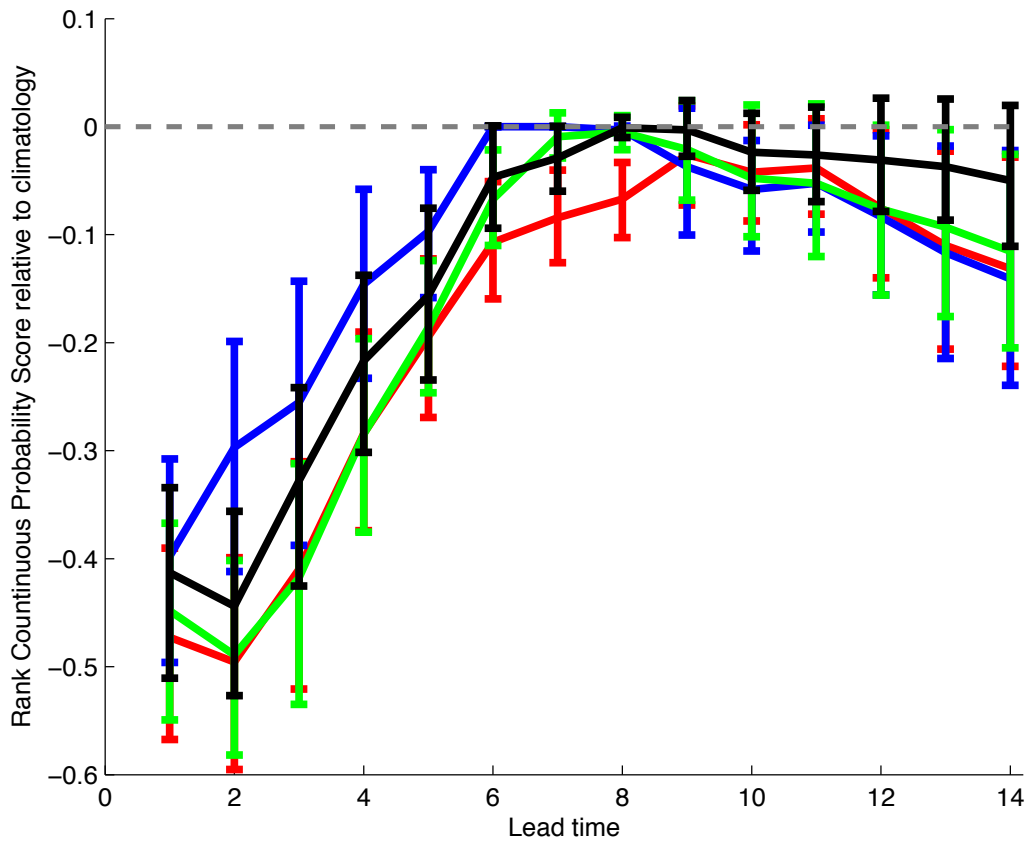


Figure 7: Rank continuous probability score of each model from ENSEMBLES for the Nino3.4 index relative to climatological forecast as a function of lead time in months for November launch. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.