Project no. GOCE-CT-2003-505539

Project acronym: ENSEMBLES

Project title: ENSEMBLE-based Predictions of Climate Changes and their Impacts

Instrument: Integrated Project

Thematic Priority: Global Change and Ecosystems

Deliverable 5.13: Report illustrating the challenges of interpretation and application of an ensemble of imperfect models on seasonal timescales

Due date of deliverable: Feb 28, 2007
Actual submission date: Jul 20, 2007

Start date of project: 1 September 2004 Duration: 60 Months

Organisation name of lead contractor for this deliverable LSE

Revision [final]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the Consortium (including the Commission Services)	

D5.13 Report illustrating the challenges of interpreting and the application of an ensemble of imperfect models on seasonal time scales

Centre for the Analysis of Time Series

July 20, 2007

1 Introduction

This report illustrates the challenges faced when combining ensemble forecast information from a number of imperfect models. In particular we consider the difficulties faced when weighting models with different skill given a limited forecast-verification archive, as is almost always the case in multi-model seasonal forecasting, and show that the relative weights assigned to different models vary considerably with the size of the forecast archive. The work follows a similar methodology to that set out by Doblas-Reyes in [5] but using imperfect models and combining the ensembles probabilistically. We begin in Section 2 by discussing the various types of ensemble information available along with methods to interpret this information into probabilis-

tic forecast distributions. Section 3 reports an illustrative experiment using a non-linear chaotic Map. The issues highlighted using the simple map are then explored in Section 4 using the DEMETER data set. The model weights are shown to strongly depend on the size of the forecast archive.

2 Ensembles

Forecast uncertainty has, in recent years, been addressed primarily through ensemble techniques. In particular, ensembles over initial conditions have been used to address observational uncertainty; ensembles over model parameters have been used to address parameter uncertainty and multi-model ensembles have been used to address model inadequacy. The ENSEMBLES project[1] aims to "develop an ensemble prediction system for climate change based on the principal state-of-the-art, high resolution, global and regional Earth System models developed in Europe, validated against quality controlled, high resolution gridded datasets for Europe, to produce for the first time, an objective probabilistic estimate of uncertainty in future climate at the seasonal to decadal and longer timescales". Combining information from multiple models provides an opportunity to produce better forecasts, whereby the inadequacies of one model can hopefully be addressed by the strengths of another. The path to an objective probability density function (PDF) is not yet clear.

[1] `http://www.ensembles-eu.org/`

In order to make best use of these different forecasts one would like their relative skills to be reflected in the relative weightings of the models in the combined forecast. Different individual models will likely have very

	Leadtime					
Name	1	2	3	4	5	6
`ecmwf`	-1.67	-1.11	-0.90	-0.60	-0.49	-0.26
`ingv`	-1.71	-1.05	-0.62	-0.41	-0.28	-0.21
`lodyc`	-1.84	-1.22	-0.82	-0.75	-0.62	-0.30
`maxplanck`	-1.05	-0.52	-0.37	-0.22	0.03	0.18
`ukmet`	-1.46	-0.91	-0.66	-0.63	-0.52	-0.31

Table 1: Ignorance of kernel dressed institutional hindcasts, blended with a climatology conditioned on the month, for Nino 3.4 index. Negative ignorance scores indicate skill. Zero ignorance corresponds to the average climatological performance.

different performances; this has been documented in [8] for the DEMETER data set. Table 1 extends this comparison to probabilistic skill, where the institutional models have been kernel dressed using the method discussed below. Moreover, the ranking of the model skill may well depend on the leadtime and the variable of interest. It has been argued, for example in [5] that combining forecasting information and taking account of each model's relative skill produces better forecasts, but when the forecast verification archive is small it is not clear how accurately one can judge each model's

skill. How are we to tell if a given archive is "small" ?

An ensemble is designed to express information regarding the uncertainty in that forecast. This uncertainty information is best communicated by issuing a probabilistic forecast distribution, or a density for continuous forecasts. A number of methods have been put forward for turning ensemble forecasts into probabilistic forecasts. This process, here termed *ensemble interpretation* can be straightforward, i.e. by counting the number of ensemble members or more sophisticated such as: best member dressing [13], plotting positions [10], Rank Histogram Recallibration [9], Multiple implementation of single-integration MOS equations [6], Bayesian model averaging [11] and Kernel Dressing [2]. Regardless of the individual interpretation method it is important that the combination method takes into account the probabilistic information and that the weightings are determined by minimising some proper skill score [7, 4].

The weighting of different models can be viewed as an extension of the individual ensemble interpretation methods. As such, additional models increase the number of parameters to be determined. Determining the interpretation parameters and weights is a statistical problem and the quality of the estimates depends greatly on the size of the forecast verification archive.

An unavoidable feature of imperfect models is the existence of "model busts", situations where the model is unable to shadow [1] (i.e. stay close to the observations) regardless of the initial condition provided. The ensemble

interpretation scheme and model weights can be unduly influenced by the presence of model busts in the forecast-verification archive. It has been suggested [2] that a climatological forecast, based on historical observations, be included in the multi-model combination in order to better account for model busts.

In the work that follows, the ensemble forecast for each model is turned into a continuous forecast distribution by kernel dressing each of the ensemble member with a Gaussian distribution. The Gaussian kernel is centred on the ensemble member and the spread set so as to minimise the average Ignorance score over the historical archive. The resultant probabilistic forecast distribution P is then is blended with climatology P_{clim} to produce a single probabilistic forecast C.

$$C(x) \quad = \quad \alpha P + (1 - \alpha) P_{clim}, \tag{1}$$

where α is the weight assigned to the model forecast. The weighting parameter can be determined simultaneously with the ensemble interpretation schemes or seperately. Given sufficient data additional forecast distributions can be included in the blend.

3 Moran-Ricker Map Illustration

To demonstrate the significance played by the relatively small forecast-verification archive we consider four imperfect dynamical models of the

Moran-Ricker Map.

The Moran-Ricker Map [12] $F(x) : x_i \rightarrow x_{i+1}$ is defined to be:

$$x_{i+1} = x_i e^{\gamma(1-x_i)}, \tag{2}$$

where $\gamma = 2.9$. The imperfect models of the Moran-Ricker system are taken to be truncated expansions of equation 2. The models are as follows:

- P_1 - 10^{th} order polynomial truncation;
- P_2 - 12^{th} order polynomial truncation;
- P_3 - exponentiated 8^{th} order logarithmic expansion;
- P_4 - 12^{th} order Fourier expansion.

In addition we take P_{clim} to be the climatological distribution of historical observations, unconditioned on the current state.

The initial condition ensembles are formed by taking perturbations restricted to the model attractor. The magnitude of the perturbations are such that the error in the ensemble mean after two iterations is commensurate with the ensemble spread at lead time 2. Each ensemble forecast consists of 9 ensemble members, as in the DEMETER project.

To investigate the impact of small forecast-verification archives we first estimate the model weights using a large archive of 2000 data points. We then sample 22 data points without replacement and estimate α using the smaller data set. Here, 22 was chosen to be commensurate with the data

sizes of the DEMETER data set. This sub-sampling is repeated 128 times. Figure 1 shows the distribution of estimates for α for P_1 based on a sub-sample of 22 forecast-verification points. The dashed line in each plot shows the α value estimated using 2000 forecast-verification points. Note that the distribution of α becomes wider as the leadtime increases. In addition a number of sub-samples assigned unit weight to the model, thereby placing absolute confidence in the model output. This over confidence is likely due to the limited number of model busts present in the training data. Similar results are seen for models P_2, P_3 and P_4 in Figures 5,6 and 7 respectively.

4 DEMETER Hindcasts

A probabilistic forecast has skill if its Ignorance score is less than[2] that of a climatological distribution not conditioned on the current state. The climatological distribution depends on the month of the year. Figure 2 shows the annual cycle of the climatological distribution of temperature for Niño 3.4 index along with the dressed ECMWF hindcasts. Comparing the two forecast distributions on probability paper (as in [3]) in Figure 3, shows the ECMWF dressed forecast assigning relatively high probabilities in climatological low probability regions. This ECMWF forecast clearly has skill over the month-conditioned climatology as can be seen in Figure 4, which shows the performance of the January and July start-date DEMETER hindcasts

[2]Here skill scores are treated like cost functions where smaller scores are more desirable.

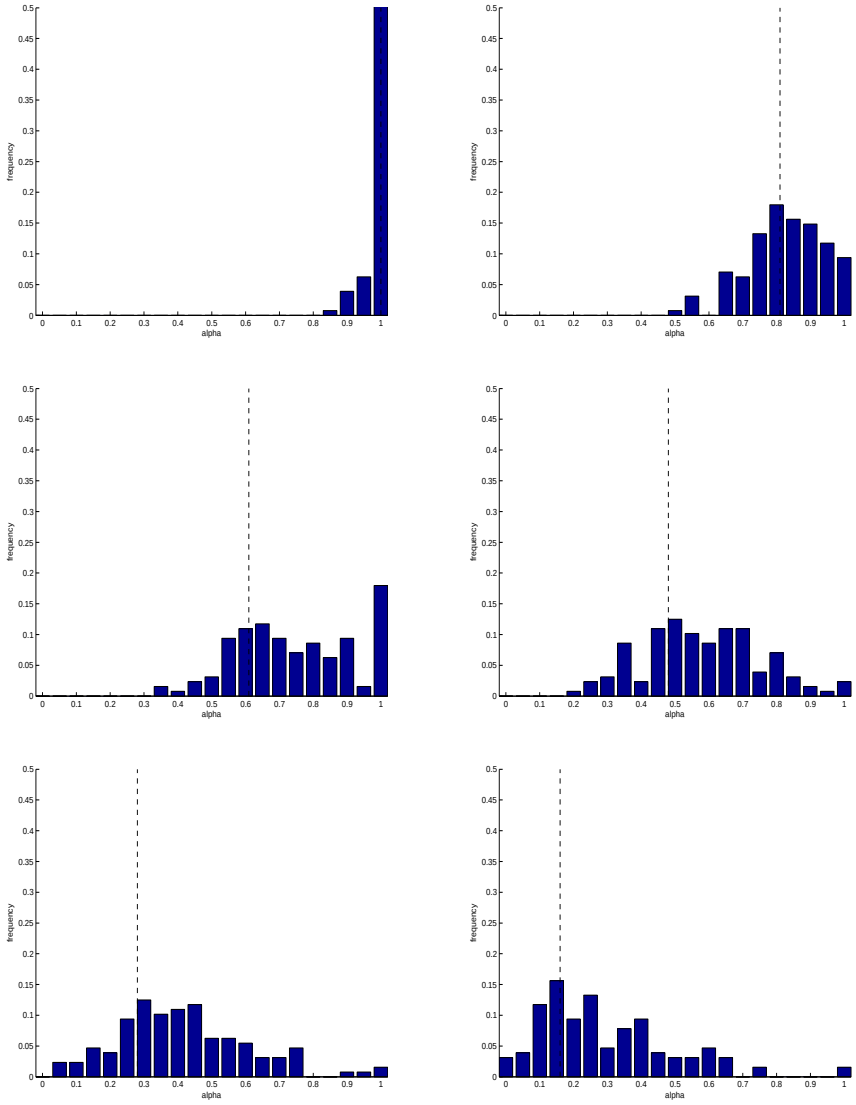

Figure 1: Distribution of model blending parameter α for P_1 over 128 samples of 22 data points for leadtime 1 (top left), 2 (top right), 3 (middle left), 4 (middle right), 5 (bottom left), 6 (bottom right). The dashed line indicates the α value computed using 2000 forecast verification pairs.

based on the ECMWF model, dressed and blended with climatology for Niño 3.4.

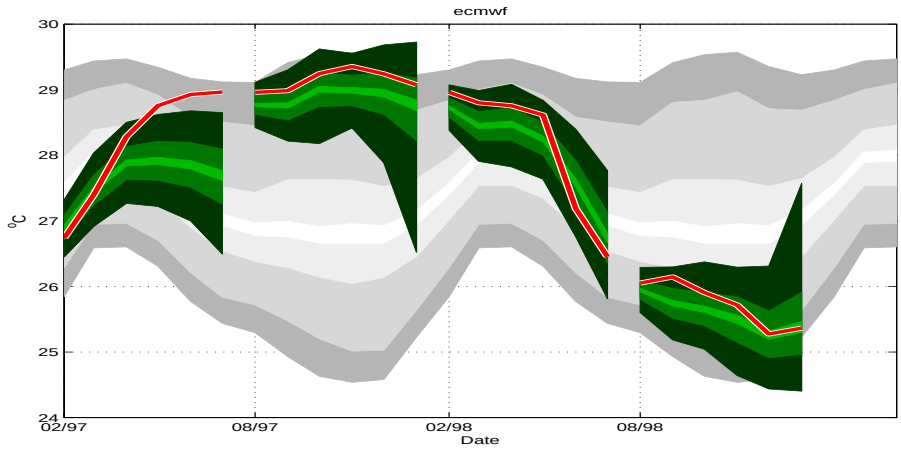

Figure 2: Left - Dressed hindcasts of ECMWF model (green) and conditioned climatology (grey). Shaded regions show the 5-95% (dark green), 25-75% (green), 45-55% (light green) for the EMCWF forecast, grey regions are 1-99% (dark grey), 5-95% (grey), 25-75% (light grey), 45-55% (white). The red line is the verification.

5 Discussion

The fact that seasonal forecasts will have small forecast-verification archives for the foreseeable future suggests there is value in learning how to train multi-model, multi-IC ensembles with a small archive.

At present, the most robust results come from dressing each model's ensemble and blending it with climatology; and then blending these prob-

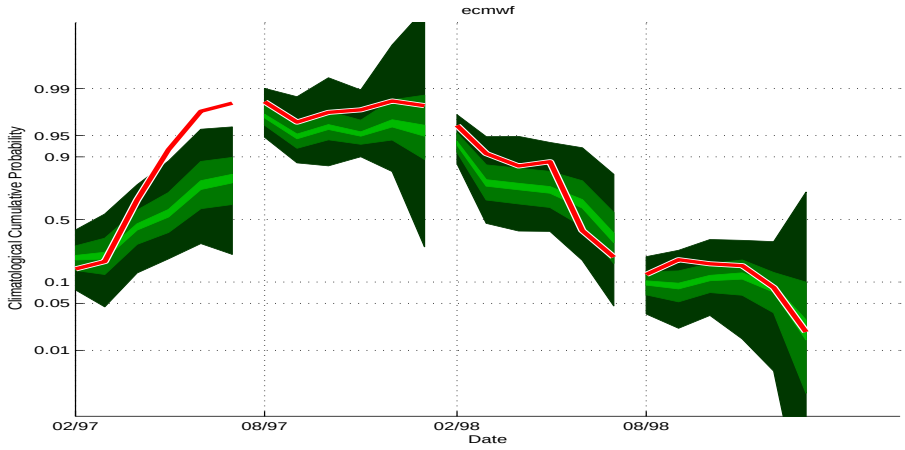

Figure 3: Dressed hindcasts of ECMWF model (green) on probability paper defined by the climatological probabilities. Shaded regions show the 5-95% (dark green), 25-75% (green), 45-55% (light green) for the EMCWF forecast.

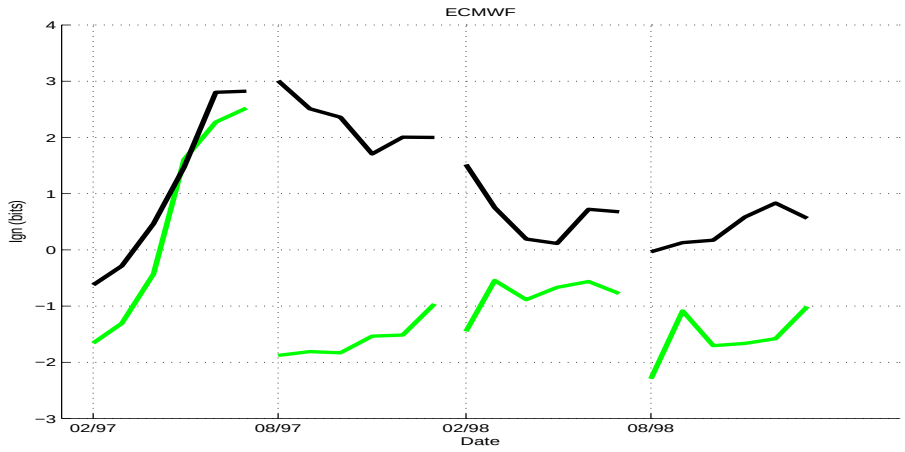

Figure 4: Ignorance scores for dressed ECMWF model (green) and climatology (black), zero Ignorance corresponds to average climatological performance over each month.

ability distributions, only accepting the inclusion of a new model when it yields a significant improvement, ideally an improvement well spread across the archive. Archives on this order are much too small to allow for the simultaneous estimation of model kernels and blend parameters; relatively poor models which by chance contribute an apparently skilful member to a forecast bust are assigned too large a weight. Alternative approaches, for example assigning equal weights and the same kernel parameter to every model, and then blending the entire ensemble with climatology, tend to suffer from similar effects. Fitting models separately will yield kernels that are wider than necessary and some hybrid approach is likely to be preferred. No general solution is expected to be optimal; variations in relative model skill and the particular members of the archive will impact each application. We are currently developing more robust methods for determining these parameters.

References

[1] *Nonlinear Dynamics and Statistics*, chapter Disentangling Uncertainty and Error: On the predictability of nonlinear systems, pages 31–64. Boston:Birkhauser, 2000.

[2] J. Bröcker and L.A. Smith. From ensemble forecasts to predictive distribution functions. submitted to Tellus.

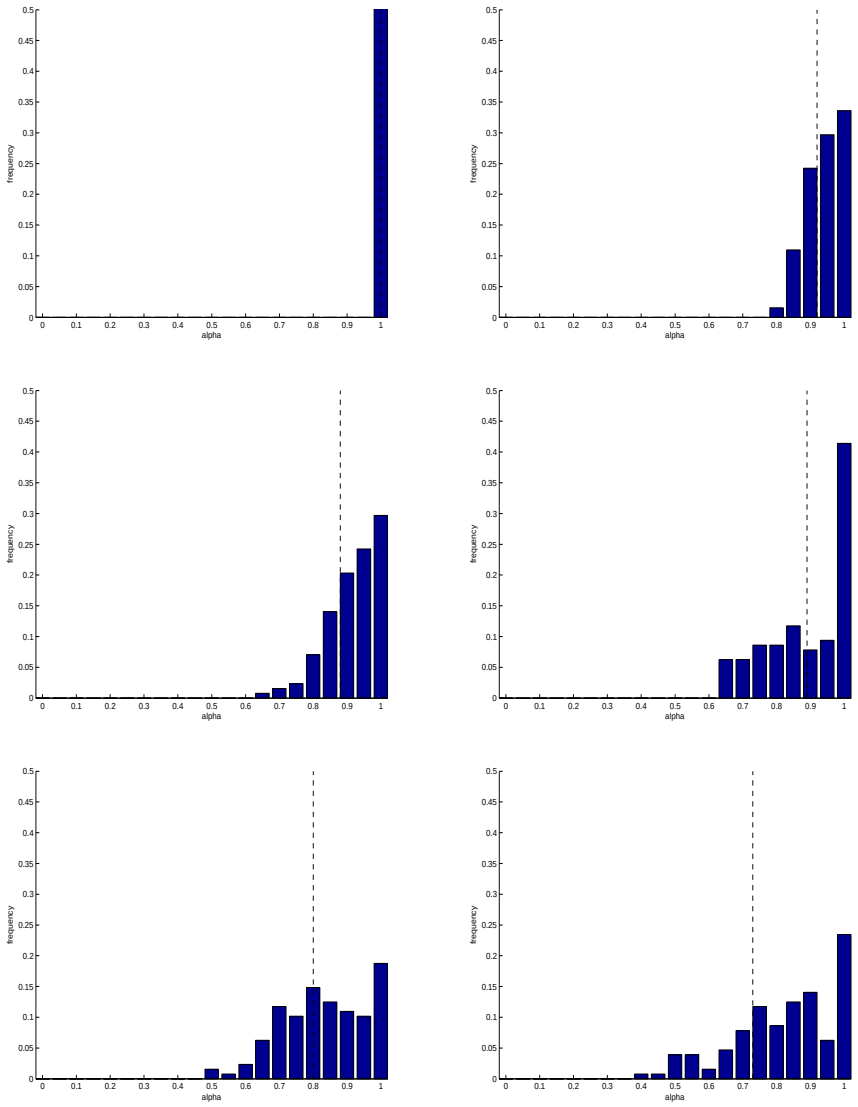

Figure 5: Distribution of model blending parameter α for P_2 over 128 samples of 22 data points for leadtimes 1 (top left), 2 (top right), 3 (middle left), 4 (middle right), 5 (bottom left), 6 (bottom right). The dashed line indicates the α value computed using 2000 forecast verification pairs.

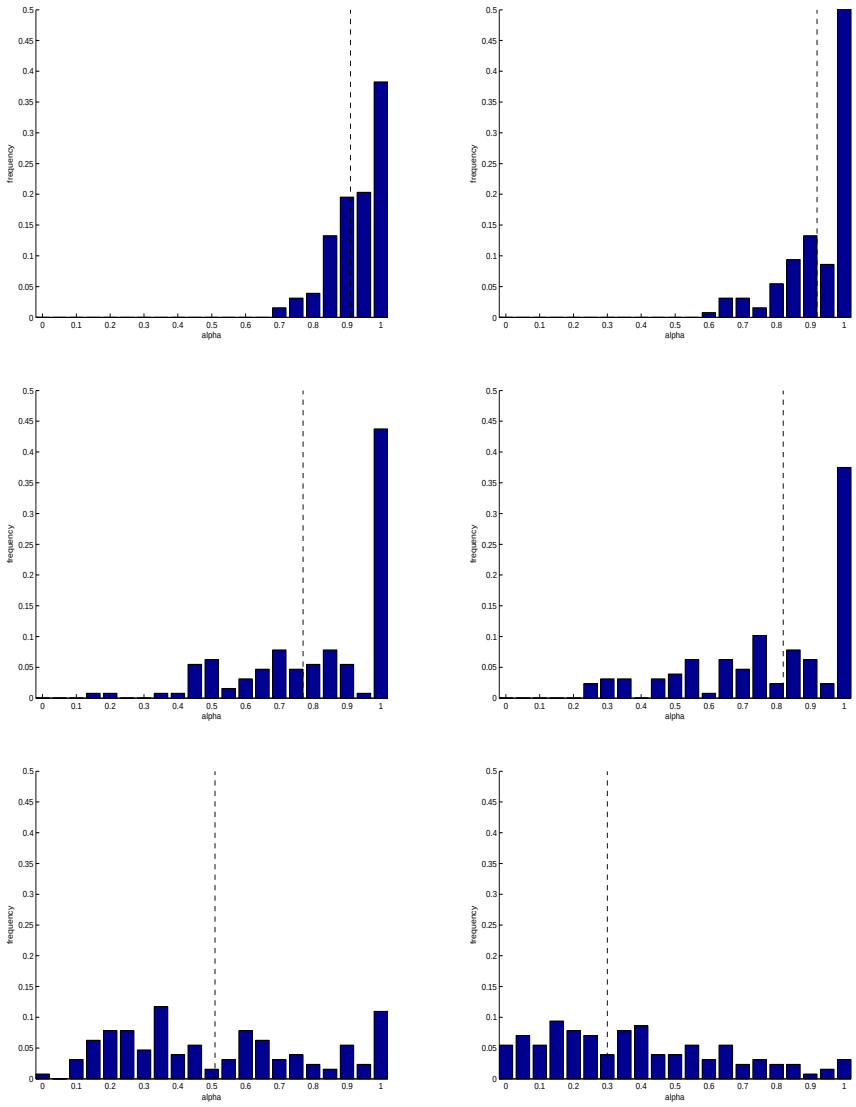

Figure 6: Distribution of model blending parameter α for P_3 over 128 samples of 22 data points for leadtimes 1 (top left), 2 (top right), 3 (middle left), 4 (middle right), 5 (bottom left), 6 (bottom right). The dashed line indicates the α value computed using 2000 forecast verification pairs.

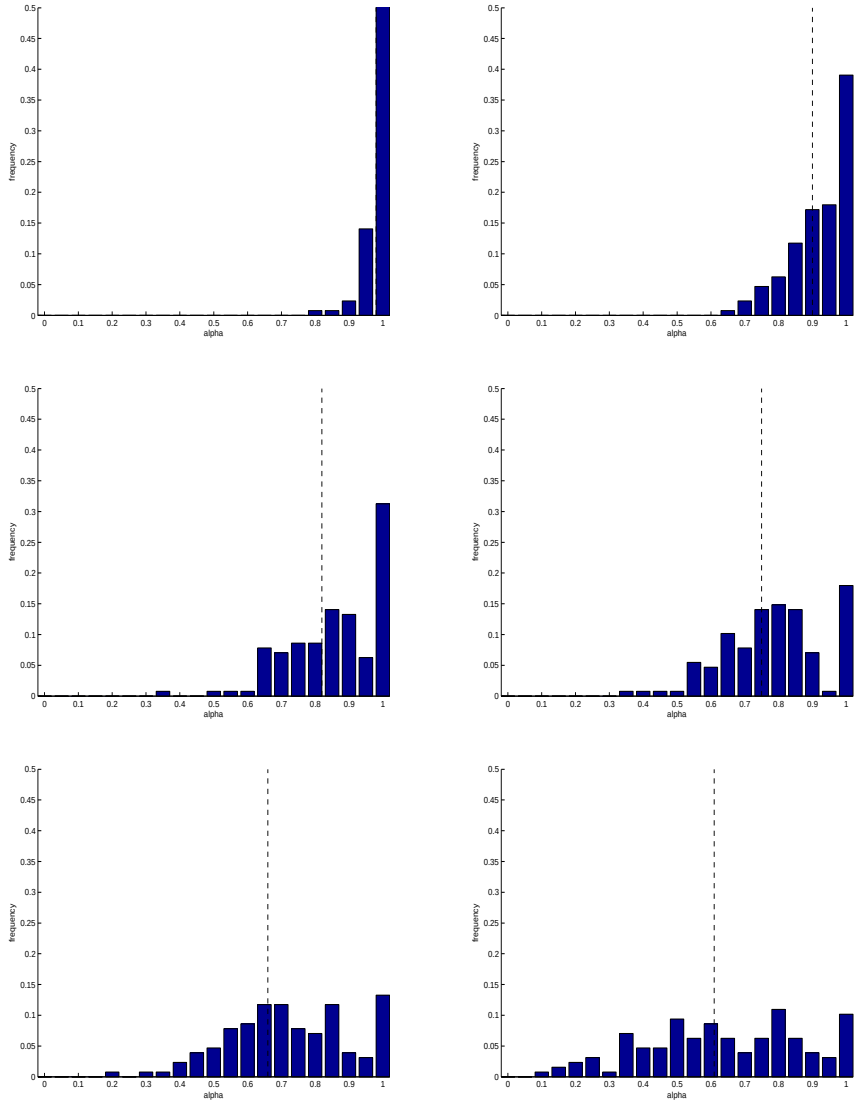

Figure 7: Distribution of model blending parameter α for P_4 over 128 samples of 22 data points for leadtimes 1 (top left), 2 (top right), 3 (middle left), 4 (middle right), 5 (bottom left), 6 (bottom right). The dashed line indicates the α value computed using 2000 forecast verification pairs.

[3] J. Bröcker and L.A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661, 2007.

[4] J. Bröcker and L.A. Smith. Scoring probabilistic forecasts: On the importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

[5] Francisco J. Doblas-Reyes, Renate Hagedorn, and T.N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting. Part II: Calibration and combination. *Tellus*, 57(3):234–252, May 2005.

[6] M.C. Erickson. Medium-range prediction of PoP and max/min in the era of ensemble model output. In *Preprints, 15th Conference on Weather and Forecasting*, pages J35–J38. Am. Meteorol. Soc.

[7] T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. Technical Report 463, Department of Statistics, University of Washington, September 2004.

[8] Renate Hagedorn, Francisco J. Doblas-Reyes, and T.N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting - Part I: Basic concept. *Tellus A*, 57(3):219–233, May 2005.

[9] T.M. Hamill and S.J. Colucci. Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Weather Rev.*, 126:711–724, 1998.

[10] H.L. Harter. Another look at plotting positions. *Commun. Stat. A-Theor.*, 13:1613–1633, 1984.

[11] A.E. Raftery, F. Balabdaoui, T. Gneiting, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.*, 133:1155–1174, 2005.

[12] W.E. Ricker. Stock and recruitment. *Journal of the Fisheries Research Board of Canada*, 11:559–623, 1954.

[13] M. Roulston and L. Smith. Combining dynamical and statistical ensembles. *Tellus A*, 55:16–30, 2003.