STATISTICA, anno LXVIII, n. 1, 2008

THE JOINT CALIBRATION MODEL IN PROBABILISTIC WEATHER FORECASTING: SOME PRELIMINARY ISSUES

P. Agati, D.G. Calò, L. Stracqualursi

1. INTRODUCTION

Weather forecasting has been traditionally viewed as a deterministic task: given a set of input data, a single prediction is generated by sophisticated numerical models of the atmosphere. Indeed, several uncertainty sources are involved in environmental prediction (for instance, initial conditions, model formulations, unrepresented feedbacks) and only a probabilistic framework can convey these uncertainties in the prediction. Such a perspective is offered by Ensemble Prediction Systems, introduced in weather forecasting in the last decade (Gneiting and Raftery, 2005).

In this paper, we show how the Joint Calibration Model (Agati *et al.*, 2007) can provide a solution to the problem of information combining in probabilistic forecasting of weather variables.

The paper is organized as follows. An overview on ensembles and their role in weather prediction is presented in section 2. Section 3 briefly describes the Joint Calibration Model and presents a pseudocode of the procedure. In section 4 a case study on sea-level pressure data is discussed.

2. WEATHER FORECASTING AND ENSEMBLES

The atmosphere is a fluid. Information regarding the current state of a fluid can be used as input into mathematical models, which exploit dynamics and thermodynamics equations in order to predict the state of the fluid at some time in the future. Such models are known as numerical weather prediction (NWP) models.

A fundamental characteristic of a chaotic system such as the atmosphere is its sensitive dependence on initial conditions (often referred to as the "butterfly effect"): *i.e.*, very small differences in the initial state tend to be amplified over time, often quite rapidly, so that similar initial conditions may produce large variations in the long term behavior of the system, and evolve into quite different final states. This means that inevitably – no matter how accurately the current state of the

atmosphere (temperature, humidity, etc.) is measured – tiny errors in the measurements are amplified until they make any long-term weather forecast unreliable.

In order to deal with this uncertainty, operational weather centres worldwide – such as the European Centre for Medium-Range Weather Forecasts (ECMWF), the British Met Office, the National Centers for Environmental Prediction (NCEP), just to mention a few – use *multi-initial condition* Ensemble Prediction Systems (EPS). Such an EPS consists of multiple runs of a single NWP model, started with slightly different sets of initial conditions, where each set has been generated by perturbing the 'best' mathematical representation of the current state of the atmosphere. The resulting multiple predictions, supported and enhanced by the use of adequate statistical post-processing techniques (EMOS, Gneiting *et al.*, 2004; BMA, Raftery *et al.*, 2005), are now routinely used to draw a probability distribution over future events or quantities. This last one represents the most detailed and interesting form of predictive information and allows to perform risk-based assessments of (severe) weather impacts. In addition, a *deterministic-style* ensemble forecast can be obtained by 'averaging' multiple predictions, where ensemble spread is informative about the inherent reliability of the final prediction.

However, it is worth noting that *initial condition* uncertainty is not the only uncertainty source in environmental forecasting (Gneiting and Raftery, 2005; Collins, 2007). Without claiming to be exhaustive, at least two additional sources can be identified:

– *model parameter* uncertainty. Any numerical model, designed for weather (or climate change) prediction, depends on a number of adjustable, high-impact parameters. Some of them are physic, measurable quantities, subject to only measurement errors; others are simplified representations of sub-grid scale physical processes – i.e., processes occurring in the atmospheric system on spatial scales smaller than the NWP model resolution – and, as such, subject to considerable uncertainty. In any case, different choices of model parameter settings can lead to different numerical predictions regarding the future state of the atmospheric system;

– *model structure* uncertainty. It regards choices to be made in designing the mathematical model: typically, the choice of the structure of precipitation and cloud scheme as well as convective scheme. Obviously, different structures result in different numerical predictions.

In recent years, several tools have been suggested and implemented to address model uncertainty (Collins and Knight, 2007). In *perturbed-physics* ensemble systems, multiple runs are performed by varying physical parameters of a NWP model within their physically acceptable range (Collins *et al.*, 2006; Rougier and Sexton, 2007; Annan and Hargreaves, 2007). *Multi-model* ensemble systems consist of different NWP models, characterized by different dynamical and mathematical formulations, physical parameterizations, horizontal and vertical resolution (Krishnamurti *et al.*, 1999, 2000; Vitart, 2006; Tebaldi and Knutty, 2007). A stochastic representation of parameterized physical processes which are unresolved by traditional deterministic parameterization is implemented in *stochastic physics* ensembles, where randomness is introduced into the model run itself (Buizza *et al.*, 1999; Lin and Neelin, 2002).

Ensemble systems play today a fundamental role in weather prediction. They can represent and measure forecasting uncertainty, thereby acquiring scientific, socioeconomic and ethical value. Specifically, ensembles allow probabilistic – i.e., distributional – forecasting (National Research Council, 2006) as well as deterministic-style forecasts, which have been established to be generally more accurate than the predictions from the individual ensemble members (Gneiting and Raftery, 2005; Raftery *et al.*, 2005).

The next step in the evolution of predictive systems is the development of *multi-model multi-initial* ensembles (the so-called *ensembles of ensembles*, also known as *grand ensembles*; Graham, 2001), which should be able to deal with uncertainty arising from different sources as well as with interactions among them. In this framework, advanced statistical post-processing techniques are needed: here we present a promising approach to information combining, which is able to cope with multi-initial EPS and, above all, is fit for treating multi-model EPS and grand ensembles.

3. THE JOINT CALIBRATION MODEL

In this section, a brief illustration of the Joint Calibration Model (JCM) is given; for more details, see Agati *et al.* (2007). After introducing the basic ideas underlying the method, we describe the proposed combining procedure using a pseudocode, in order to provide a compact and environment-independent description of the key principles of the algorithm.

3.1. *Basic ideas*

Let's suppose that an unknown quantity y has to be forecast on the basis of the forecasts $f_1,...,f_k,...,f_K$ provided by K ensemble members $M_1,...,M_k,...,M_K$. We will use the following notation:

– let $g_k = g_k(y|f_k)$ be the predictive probability density function (pdf) associated with f_k (k = 1, 2, ..., K). For the sake of simplicity, we assume that g_k is parameterized by a location parameter, coinciding with f_k, and a scale parameter v_k: for example, g_k denotes the pdf of a Gaussian random variable $\mathcal{N}(f_k, v_k)$;

– let $L(\cdot)$ be the likelihood function of the data $\mathbf{f} = [f_k]'_{k=1,2,...,K}$ and $\mathbf{v} = [v_k]'_{k=1,2,...,K}$;

– let $p(y)$ be a prior predictive pdf (which may also be uninformative).

By recasting Morris' information combining algorithm (Morris, 1977) in a predictive context, the ensemble final pdf can be thought of as the posterior pdf $p(y|g_1, ...,g_k, ...,g_K)$ and written, via Bayes' theorem, under the assumption $L(\mathbf{v}|y) = L(\mathbf{v})$, as follows:

$$p(y|\mathbf{f},\mathbf{v}) \propto L(\mathbf{f}|\mathbf{v},y)\cdot p(y) \tag{1}$$

As the likelihood function in (1) can be shown to be proportional to the product of the individual pdfs, adjusted by a *joint* "calibration function" $C(y)$, JCM takes the following form:

$$p(y|\mathbf{f},\mathbf{v}) \propto C(y)\cdot\prod_{k=1}^{K} g_k(y|f_k,v_k)\cdot p(y) \tag{2}$$

According to the Maximum A Posteriori (MAP) principle, we suggest to take the value $\hat{y}_{JCM}$ maximizing (2) as the final deterministic forecast for y:

$$\hat{y}_{JCM} = arg\,max_y\, p(y|\mathbf{f},\mathbf{v}) \tag{3}$$

It is worth noting that, unlike most standard combining methods (Cooke, 1991), JCM neither attaches a weight to each individual predictive pdf nor yields the ensemble predictive pdf as a weighted average of individual pdfs. It works by means of the overall calibration function $C(y)$, which models the predictive performance of the ensemble members by taking into account bias, variability and mutual dependence in their assessments of y.

The functional form of $C(\cdot)$ is derived through a parametric modelization, in order to allow its assessment by means of a relatively small number of parameters. This proposal is briefly sketched in the following, after a few notes which explain the structure of $C(y)$.

Let's denote by τ_k the *performance indicator* associated with g_k. It is nothing but the Probability Integral Transform (PIT) value (Pearson, 1933; Rosenblatt, 1952), i.e. the value $0 \le \tau_k \le 1$ that the predictive cumulative distribution function (cdf) $G_k(\cdot|f_k,v_k)$ attains at the verification y_0 of y:

$$\tau_k = \tau_k(f_k,v_k,y_0) = \int_{-\infty}^{y_0} g_k(y|f_k,v_k)\,dy \;=\; G_k(y_0|f_k,v_k) \tag{4}$$

The PIT value has received much attention in the literature as a key *diagnostic* tool (in the form of PIT histogram) for evaluating the predictive performance of forecasts when they take the form of probability distributions whereas the observations are real-valued quantities (Dawid, 1984; Diebold *et al.*, 1998; Gneiting *et al.*, 2007). In JCM, the PIT value is used in the combining process itself, as a *calibration* tool. Specifically, the calibration function $C(y)$ is obtained by defining a joint *performance function* $\varphi(\boldsymbol{\tau}|\mathbf{v})$ as a conditional density on the vector $\boldsymbol{\tau} = [\tau_k]'_{k=1,2,\ldots,K} = [G_k(y)]'_{k=1,2,\ldots,K} = \mathbf{G}(y)$ given $\mathbf{v}$, and viewing it as a function of y (for fixed $\mathbf{f}$). It attaches an admissibility degree to any y value considered as the realization of the K-dimensional vector $\boldsymbol{\tau}$:

$$C(y) = \varphi[\mathbf{G}(y)|\mathbf{v}] = \varphi(\boldsymbol{\tau}|\mathbf{v}) \tag{5}$$

Under the assumption of normality of the *K*-dimensional vector of the performance *log*-odds $\tilde{\boldsymbol{\tau}} = [\tilde{\tau}_k]'_{k=1,...,K}$, with $\tilde{\tau}_k = \ln[\tau_k/(1-\tau_k)] \in \Re$, the calibration function $C(y)$ depends on the mean vector $\tilde{\mathbf{t}}$ and the covariance matrix $\mathbf{S}$ of $\tilde{\boldsymbol{\tau}}$, and can be written as follows:

$$C(y) = \varphi[\mathbf{G}(y)|\mathbf{v}] \propto$$

$$\propto \prod_{k=1}^{K} \frac{1}{G_k(y)\cdot[1-G_k(y)]} \cdot \exp\left[-\frac{1}{2}(\tilde{\mathbf{G}}(y)-\tilde{\mathbf{t}})' \mathbf{S}^{-1} (\tilde{\mathbf{G}}(y)-\tilde{\mathbf{t}})\right] \tag{6}$$

where $\tilde{\mathbf{G}}(y) = [\tilde{G}_k]' = [\ln(G_k/(1-G_k))]'$. The parameters $\tilde{\mathbf{t}}$ and $\mathbf{S}$ can be estimated by least squares on a training set consisting of *n* verifying observations $\{y_1, ..., y_i, ..., y_n\}$ and the corresponding forecasts from the ensemble members $\{f_{1k}, ..., f_{ik}, ..., f_{nk}\}$ for $k = 1, 2, ..., K$. In other words, the estimates are the solutions of the following minimization problem,

$$\min_{\tilde{\mathbf{t}},\,\mathbf{S}} \Psi(\tilde{\mathbf{t}},\mathbf{S}) = \min_{\tilde{\mathbf{t}},\,\mathbf{S}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7}$$

where $\hat{y}_i$ is obtained according to equation (3).

3.2. *A pseudocode of the procedure*

The input information consists of the following datasets:

a) the dataset containing the verifying observations $\{y_1,..., y_i,..., y_n\}$ and the corresponding forecasts from the ensemble members $\{f_{1k},..., f_{ik},..., f_{nk}\}$ for $k = 1, 2, ..., K$. In the following, it will be referred to as training dataset;

b) the dataset containing the values of the scale parameter v_k associated to the forecasts $\{f_{1k},..., f_{ik},..., f_{nk}\}$ for $k=1, 2, ..., K$. As shown in equation (5), the calibration function $C(y)$ is defined conditionally on vector **v**, hence the elements in **v** must be assessed before fitting JCM.[1]

The multiparameter θ, indexing the Joint Calibration Model and representing the variables of the objective-function Ψ in (7), varies in the set:

$$\Theta = \{\theta = (\tilde{\mathbf{t}},\mathbf{S}): \ \tilde{\mathbf{t}} \in \Re^K, \ \mathbf{S} \in \mathcal{M}^{+}_{K\times K}\} \tag{8}$$

[1] In the study illustrated in section 4, a common value v_k has been associated to all forecasts. It has been assessed by simple linear regression of f_k on y (on the training dataset) as the mean of squared residuals.

where $\mathcal{M}^{+}_{K\times K}$ is the set of (strictly) positive definite $K\times K$ matrices. The number of variables in the minimization problem is equal to $K+K(K+1)/2$.

The following sketch of the procedure refers to a situation where a non-informative prior in (2) is adopted, so that maximizing the posterior pdf $p(y\,|\,\mathbf{f},\mathbf{v})$ is equivalent to maximizing the likelihood function $L(\mathbf{f}\,|\,\mathbf{v},y)$, a problem which is simplified by resorting to the logarithmic transformation.

Phase 1. Initialization:

Choose a starting point: $\theta=(\tilde{\mathbf{t}},\mathbf{S}) \quad \leftarrow \quad \theta_0$

Phase 2. Optimization:

Do until convergence

For i = 1, ..., n

$$\hat{y}_i \quad \leftarrow \quad \arg\max_y \ln\{L\ (\mathbf{f}_i\,|\,\mathbf{v}_i, y)\}$$

where

$$\ln\{L(\mathbf{f}_i\,|\,\mathbf{v}_i, y)\} = -\sum_{k=1}^{K}\ln\,[G(y_i)\cdot(1-G(y_i))] +$$

$$-\ 0.5\cdot\left[\sum_{k=1}^{K}\frac{(y_i-f_{ik})^2}{v_{ik}}+(\mathbf{G}(y_i)-\tilde{\mathbf{t}})'\,\mathbf{S}^{-1}(\mathbf{G}(y_i)-\tilde{\mathbf{t}})\right]$$

Endfor

Evaluate the objective-function $\Psi(\theta)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

Update θ according to the optimization algorithm

Endo

In the empirical study described in section 4, the optimization problem is addressed by the L-BFGS method (Byrd *et al.*, 1995). As L-BFGS procedure is unable to guarantee convergence to global solutions, a multi-start strategy has been adopted.

4. A CASE STUDY: SEA LEVEL PRESSURE DATA

In this section we compare the performance of our method (JCM) with that of Bayesian Model Averaging (BMA) by Raftery *et al.*, 2005. Our combining procedure has been implemented in Fortran95; Raftery's one is implemented in the R package `ensembleBMA`.

The comparison is carried out on the sea level pressure dataset (SLP, available at http://www.stat.washington.edu.MURI), used by Raftery in his paper. The dataset contains 48-hour sea-level pressure forecasts and observations (in milli-

bars) in Pacific Northwest for the period from January 12, 2000 to June 30, 2000 at 00.00 hours UTC. Forecasts are provided by the University of Washington mesoscale short-range ensemble system. It is a five-member multianalysis single-model ensemble[2], driven by initial and lateral boundary conditions obtained from major operational weather centers worldwide: the ensemble members are denoted as AVN, ETA, NGM, GEM and NOGAPS.

The results of our study are presented in two subsections: the former deals with a single test day, with the aim of illustrating an example where correlation between ensemble members' performance is high; in the latter, the predictive performance of the Joint Calibration Model is compared with that of Bayesian Model Averaging on all 39 test days considered by Raftery *et al.* in their study.

4.1 *Some critical notes*

For each test day considered in Raftery *et al.* (2005), a preliminary exploratory analysis has been carried out on the training datasets corresponding to different training period lengths. In fact, as Raftery points out, there is a trade-off when choosing how many days should be used in the sliding-window training period: weather patterns change over time and thus a short training period should be preferable in order to adapt rapidly to seasonally varying model biases, changes in the performance of the ensemble member models and changes in environmental conditions; on the other hand, the longer the training period, the better the parameters are estimated.

This preliminary analysis has shown that all the training datasets are characterized by two main features:

a) performance indicators of any two members are moderately correlated, except for the pair NGM, GEM (whose Pearson correlation coefficient ranges from 0.85 to 0.89);

b) estimated variances v_k of the forecasts provided by the single members are similar: that is, the ratio between any two estimates is around 1. This feature is coeherent with the nature of the data, arising from a single-model ensemble, and makes the data ideal for BMA, since this method assumes that all the predictive pdfs from the members share a common scale parameter.

Since JCM seems to take advantage from high values of positive correlation between the performance of the ensemble members (as illustrated in Agati *et al.*, 2007), we have focused our attention on the pair NGM, GEM. In addition, we have looked for the instance where they differ most in terms of forecast variability: the test day selected according to this criterion is May 1st, and the corresponding training dataset consists of April 17th, 24th, 29th , that is the most recent 3 days in the dataset. It is worth noting that the selected example is interesting also because it may represent situations where the size of the training data is small.

Both JCM and BMA have been carried out and their final deterministic-style

[2] It is known as MM5, acronym of the fifth-generation Pennsylvania State University-National Center of Atmospheric Research Mesoscale Model.

forecasts have been compared with observed data in terms of root mean square error (RMSE). JCM is 15% better than BMA (RMSE = 3.50 and RMSE = 4.12 for JCM and BMA, respectively). We have observed that BMA attaches an approximately null weight to one of the members: this means that BMA discards one of the members, which is deemed as redundant. On the contrary, JCM is able to exploit information from correlated performance indicators (since it is designed for modelling such indicators, as shown in section 3.1), so that both the members are useful.

Moreover, for a more general comparison, BMA was run for the same test set as before with eight different training period lengths. The results are shown in the second column of table 1: BMA attains the best predictive performance only if we increase the training period up to 10 calendar days.

Finally, we tested BMA on the five-member ensemble data, for the same test day and training period lengths defined above. As shown in the third column of table 1, the experiment suggests that including information from additional members having favourable characteristcs[3] may yield some improvements. However, the least RMSE value, corresponding to a 30-days training period length, is not less than JCM one, obtained with only 2 members and 3 training days.

TABLE 1

RMSE for BMA (test day: May 1st)

Training period length (days)	2 members	5 members
3	4.12	4.42
5	4.00	3.86
10	3.80	3.90
15	4.15	3.71
20	4.16	3.61
25	4.24	3.86
30	3.88	3.60
35	4.23	3.69
40	4.23	3.67

4.2 *The empirical study*

We have explored JCM predictive effectiveness on the whole dataset according to the empirical study described by Raftery *et al* (2005), being aware that in most of SLP training datasets the conditions realizing the full potential of JCM are not fulfilled.

More specifically: all five ensemble members have been taken into account; 39 test days, ranging from April 24th to June 30th, have been considered; a 25-day sliding-window training period has been used. Due to missing data, the number of calendar days spanned by the training dataset is sometimes larger than 25.

For each test day, both JCM and BMA were carried out and their deterministic-style forecast accuracy was assessed in terms of RMSE. Summary measures are given in table 2.

[3] That is, low values of correlation between performance indicators and homogeneous forecast variability.

TABLE 2

Results of the empirical study: descriptive statistics of RMSE values for 39 test days

	Procedure	
	BMA	JCM
Mean	2.59	2.63
Inter-quartile range	0.85	0.98
Range	3.93	2.05

The analysis shows that JCM can be competitive with BMA:

– RMSE mean values are similar: BMA has RMSE only 1.7% less (corresponding to a difference of just 0.04 millibar) when compared to that of JCM;

– JCM seems to be more robust: JCM range is about 50% less when compared to BMA range. If we consider inter-quartile range, which neglects extreme situations, the procedures show similar variability in their performance.

5. CONCLUDING REMARKS

Our paper addresses the problem of information combining in probabilistic weather forecasting. Specifically, we put the Joint Calibration Model (JCM, Agati *et al.*, 2007) in the context of Ensemble Prediction Systems. JCM offers a modelization of the distribution of the well known Probability Integral Transform (PIT value), which is usually adopted as a diagnostic tool in evaluating the predictive performance of probabilistic forecasters. A case study is presented, where the potentialities of the method are explored and the accuracy of deterministic-style forecasts from JCM is compared with that from Bayesian Model Averaging (BMA, Raftery *et al.*, 2005). The empirical analysis shows that JCM can be competitive with BMA in terms of RMSE and has a more stable general behavior.

Improvements could be obtained by further investigations about the optimization algorithm and the MAP criterion implementation.

Moreover, the effectiveness of our proposal should be explored according to the diagnostic paradigm of "maximizing the sharpness of the [final] predictive distribution subject to calibration" (Gneiting *et al*, 2007).

The most promising direction of future research consists in applying JCM to combining ensembles from individual ensemble systems, in the context of environmental prediction and financial risk management.

Dipartimento di Scienze Statistiche "P. Fortunati"
Università di Bologna

PATRIZIA AGATI
DANIELA GIOVANNA CALÒ
LUISA STRACQUALURSI

ACKNOWLEDGMENTS

The authors are grateful to Paolo Arena for his assistance in Fortran programming of JCM procedure.

REFERENCES

P. AGATI, D.G. CALÒ, L. STRACQUALURSI (2007), *A Joint Calibration Model for combining predictive distributions*, "Statistica", 2007, pp. 167-182.

J.D. ANNAN, J.C. HARGREAVES (2007), *Efficient estimation and ensemble generation in climate modelling*, "Philosophical transactions of the Royal Statistical Society" A, 365, 2077-2088.

R. BUIZZA, M. MILLER, T.N. PALMER (1999), *Stochastic representation of model uncertainties in the ECMWF ensemble prediction system*, "Quarterly Journal of the Royal Meteorological Society, 125, pp. 2887-2908.

R.H. BYRD, P. LU, J. NOCEDAL, C. ZHU (1995), *A limited memory algorithm for bound constrained optimization*, "SIAM Journal of Scientific Computing", 16, pp. 1190-1208.

M. COLLINS (2007), *Ensemble and probabilities: a new era in the prediction of climate change*, "Philosophical Transactions of the Royal Society", A, 365, pp. 1957-1970.

M. COLLINS, B.B.B. BOOTH, G.R. HARRIS, J. M. MURPHY, D.M.H. SEXTON, M.J. WEBB (2006), *Towards quantifying uncertainty in transient climate change*, "Climate Dynamics", 27, pp. 127-147.

M. COLLINS, S. KNIGHT eds (2007), Theme issue *Ensembles and probabilities: a new era in the prediction of climate change*, "Philosophical Transactions of the Royal Society", A, 365, 1957-2191.

R. M. COOKE (1991), *Experts in uncertainty. Opinion and subjective probability in science*, Oxford University Press, Oxford.

A. P. DAWID (1984) *Statistical theory: the prequential approach*, "Journal of the Royal Statistical Society", A, 147, pp. 278-290.

F. X. DIEBOLD, T. A. GUNTHER, A.S. TAY (1998), *Evaluating density forecasts with applications to financial risk management*, "International Economic Review", 39, pp. 863-883.

T. GNEITING, F. BALABDAOUI, A.E. RAFTERY (2007), *Probabilistic forecasts, calibration and sharpness,* "Journal of the Royal Statistical Society" B, 69, pp. 243-268.

T. GNEITING, A. RAFTERY (2005), *Weather forecasting with ensemble methods*, "Science", 310, pp. 248-249.

T. GNEITING, A.H. WESTVELD, A.E. RAFTERY, T. GOLDMAN (2004), *Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation*, Technical report n. 449, Department of Statistics, University of Washington.

R.J. GRAHAM (2001), *The ensemble of ensembles: multiple-model forecasting for seasonal timescales*, Seasonal Modelling and Prediction Group, Met. Office, United Kingdom, available at www.wmo.ch/pages/prog/wcp/wcasp/clips/modules/documents/mod3_ensembleofensemble.ppt

T.N. KRISHNAMURTI, C.M. KISHTAWAL, T. LAROW, D. BACHIOCHI, E. WILLIFORD, S. GADGIL, S. SURENDRAN (2000), *Improved weather and seasonal climate forecasts from multimodel superensemble*, "Science", 285, pp. 1548-1550.

T.N. KRISHNAMURTI, C.M. KISHTAWAL, Z. ZHANG, T. LAROW, D. BACHIOCHI, E. WILLIFORD, S. GADGIL, S. SURENDRAN (2000), *Multimodel Ensemble Forecasts for Weather and Seasonal Climate*, "Journal of Climate", 13, pp. 4196-4216.

J.W.B. LIN, J.D. NEELIN (2002), *Considerations for stochastic convective parameterization*, "Journal of the Atmospheric Sciences", 59, pp. 959-975.

P. A. MORRIS (1977), *Combining expert judgements: a Bayesian approach*, "Management Science", 23, pp. 679-693.

NATIONAL RESEARCH COUNCIL, Committee on Estimating and Communicating Uncertainty in Weather and Climate Forecasts (2006) *Completing the forecast: characterizing and communicating uncertainty for better decisions using weather and climate forecasts*, National Academies Press, Washington, available at http://www.nap.edu/catalog. php?record_id=11699 #toc

K. PEARSON (1933), *On a method of determining weather a sample of size* n *supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random*, "Biometrika", 25, pp. 379-410.

A.E. RAFTERY, T. GNEITING, F. BALABDAOUI, M. POLAKOWSKI (2005), *Using Bayesian model averaging to calibrate forecast ensembles*, "Monthly Weather Review", 133, 1155-1174.

M. ROSENBLATT (1952), *Remarks on a multivariate transformation*, "Annals of Mathematical Statistics", 23, pp. 470-472.

J. ROUGIER, D.S. SEXTON (2007), *Inference in ensemble experiments*, "Philosophical Transactions of the Royal Society", A, 365, pp. 2133-2143.

F. VITART (2006), *Seasonal forecasting of tropical storm frequency using a multi-model ensemble*, "Quarterly Journal of Meteorological Society", 132, pp. 647-666.

SUMMARY

The Joint Calibration Model in probabilistic weather forecasting: some preliminary issues

Ensemble Prediction Systems play today a fundamental role in weather forecasting. They can represent and measure uncertainty, thereby allowing distributional forecasting as well as deterministic-style forecasts. In this context, we show how the Joint Calibration Model (Agati *et al.*, 2007) – based on a modelization of the Probability Integral Transform distribution – can provide a solution to the problem of information combining in probabilistic forecasting of continuous variables. A case study is presented, where the potentialities of the method are explored and the accuracy of deterministic-style forecasts from JCM is compared with that from Bayesian Model Averaging (Raftery *et al.*, 2005).