STATISTICA, anno LXVII, n. 2, 2007

A JOINT CALIBRATION MODEL FOR COMBINING PREDICTIVE DISTRIBUTIONS

P. Agati, D.G. Calò, L. Stracqualursi

1. INTRODUCTION

In many research fields, as for example in probabilistic weather forecasting, valuable predictive information about a future random phenomenon may come from several, possibly heterogeneous, sources. Forecast combining methods have been developed over the years in order to deal with *ensembles* of sources (Collins, 2007). The aim is to combine several predictions in such a way to improve forecast accuracy and reduce risk of bad forecasts (Winkler and Clemen, 2004): for example, the ensemble mean usually outperforms the individual ensemble members.

In this framework, we propose the use of a Bayesian approach to information combining, which consists in treating the predictive probability density functions (pdfs) from the individual ensemble members as data in a Bayesian updating problem. The likelihood function is shown to be proportional to the product of the pdfs, adjusted by a joint "calibration function" describing the predicting skill of the sources (Morris, 1977). We propose to model the calibration function in terms of bias, scale and correlation and to estimate its parameters according to the least squares criterion. The performance of our method is investigated and compared with that of Bayesian Model Averaging (Leamer, 1978; Raftery *et al.*, 2005) on simulated data.

The paper is organized as follows. In section 2 Morris' model is rephrased in a probabilistic forecasting context. Section 3 presents the proposed forecast combining model and describes our choice about the final deterministic forecast. The comparison with Bayesian Model Averaging is illustrated in section 4. Finally, in section 5 the conclusions are given.

2. MORRIS' APPROACH IN THE FRAMEWORK OF FORECAST COMBINING

In the context of information combining, according to Morris (1977), the reception of K expert answers may be viewed as the outcome of an experiment. A likelihood function may be associated with it and possibly used to update a prior judgment via Bayes theorem. In such a way, the information combining process just becomes an information updating process.

This general principle can be applied to the aggregation of any kind of information, ranging from the combination of point estimates to the combination of probability distributions. In the following, we rephrase Morris' algorithm in a predictive context.

More precisely, we suppose an unknown quantity y has to be forecast on the basis of the forecasts $f_1,\ ...,f_k,\ ...,f_K$ provided by K models (ensemble members) $M_1,\ ...,M_k,\ ...,M_K$. Denoting by $g_k = g_k(y|f_k)$ the individual ensemble member predictive pdf associated with (and parameterized by) $f_k\,(k = 1,\ 2,\ ...,\ K)$, the Bayesian algorithm can be written as:

$$p(y|\,g_1,\ ...,g_k,\ ...,g_K) \propto L(g_1,\ ...,g_k,\ ...,g_K\,|y)\cdot p(y) \tag{1}$$

where

– $p(y|\,g_1,\ ...,g_k,\ ...,g_K)$ denotes the ensemble posterior predictive pdf;
– $L(\cdot)$ denotes the likelihood function for the experimental data $\{g_1,\ ...,g_k,\ ...,g_K\}$;
– $p(y)$ is a prior predictive pdf (which may also be uninformative).

What makes the Bayesian approach rather difficult to apply is the assessment of the likelihood function: a joint probability assessment over the set of pdfs from the ensemble members, which must account for the differences in performance and the dependence among the ensemble members. Two assumptions allow to express $L(\cdot)$ in a form easier to be modelled (Morris, 1977).

Assumption (*i*). Each $g_k(\cdot)$ is parameterized by a location parameter m_k (coinciding with f_k, in the present case) and a shape parameter v_k. For example, $g_k(\cdot)$ denotes the pdf of a Gaussian random variable $\mathcal{N}(f_k,\ v_k)$. Then, equation (1) becomes:

$$p(y|\,g_1,\ ...,g_k,\ ...,g_K) \ = \ p(y|\,\mathbf{f},\mathbf{v}) \propto L(\mathbf{f}|\,\mathbf{v},y)\cdot L(\mathbf{v}|y)\cdot p(y) \tag{2}$$

where $\mathbf{f} = [f_k]'_{k=1,2,...,K}$ and $\mathbf{v} = [v_k]'_{k=1,2,...,K}$.

Assumption (*ii*). The joint probability density value assigned to the event "the shape parameters of the g_ks are $v_1,\ ...,v_k,\ ...,v_K$" does not depend on y: $L(\mathbf{v}|y) = L(\mathbf{v})$.[1] Using this assumption, equation (2) takes the form:

$$p(y|\,\mathbf{f},\mathbf{v}) \propto L(\mathbf{f}|\,\mathbf{v},y)\cdot p(y) \tag{3}$$

[1] As stochastic independence is reciprocal, assumption (*ii*) can be also expressed as *invariance to scale* about y, that is $p(y|\,\mathbf{v}) = p(y)$: vector **v** alone gives no information regarding y.

where the conditional likelihood $L(\mathbf{f}|\,\mathbf{v},y)$ has to be viewed as a function of *y*: it represents the joint probability – conditioned upon the variances vector **v** – of the event "the ensemble members $M_1,\ ...,M_k,\ ...,M_K$ will give the predictions $f_1,\ ...,f_k,\ ...,f_K$, respectively".

For the purpose of assessing $L(\mathbf{f}|\,\mathbf{v},y)$, Morris (1977) introduces the notions of *performance indicator* and *performance function.*

The performance indicator τ_k associated with $g_k(\cdot)$ is defined as the cumulative distribution function $G_k(\cdot|f_k,v_k)$ evaluated at the observed value y_0 of *y*:

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

where $0 \le \tau_k \le 1$. For example, if the observed value is the 0.3-quantile of $g_k(\cdot)$, then $\tau_k = 0.3$.

The performance function, denoted by $\varphi\,(\boldsymbol{\tau}|\,\mathbf{v},y)$, is defined as a conditional joint density on the *K*-dimensional vector $\boldsymbol{\tau} = [\tau_k]'_{k=1,2,...,K} = [G_k(y)]'_{k=1,2,...,K} = \mathbf{G}(y)$, given **v** and *y*.

Given the vector **v**, for any fixed value of *y*, a monotonic decreasing relationship exists between corresponding elements in $\boldsymbol{\tau}$ and $\mathbf{f}$. So, a change of variable allows to show that (Morris, 1977):

$$L(\mathbf{f}|\,\mathbf{v},y) \;= C(y)\cdot\prod_{k=1}^{K} g_k(y|f_k,v_k)\,, \tag{5}$$

where:

$$C(y) \;=\; \varphi\,[\mathbf{G}(y)|\,\mathbf{v},y] \;=\; \varphi(\boldsymbol{\tau}|\,\mathbf{v},y) \tag{6}$$

is called joint *calibration* function. It is nothing but the performance function $\varphi\,(\boldsymbol{\tau}|\,\mathbf{v},y)$ looked at as a function of *y* (for fixed $\mathbf{f}$): it expresses the admissibility degrees assigned to each possible *y* value looked at as the realization of the *K*-dimensional vector $\boldsymbol{\tau}$.

Therefore, equation (5) shows that the likelihood function can be obtained as the product of the pdfs from the ensemble members, adjusted by function $C(\cdot)$: it is this last function that models the predictive performance of the ensemble members and their mutual dependence in assessing *y*.

By substituting (5) into (3), the posterior predictive pdf can be written as:

$$p(y|\,\mathbf{f},\mathbf{v}) \propto C(y)\cdot\prod_{k=1}^{K} g_k(y|f_k,v_k)\cdot p(y) \tag{7}$$

which describes the structural form of what we call "Joint Calibration Model".

3. MODELLING THE PERFORMANCE FUNCTION: THE JOINT CALIBRATION MODEL

According to the Maximum A Posteriori (MAP) principle, we suggest to take the value maximizing equation (7)

$$\hat{y}_{JCM} = \arg\max_{y} p(y \mid \mathbf{f}, \mathbf{v}) \tag{8}$$

as the final deterministic forecast for y yielded by the Joint Calibration Model (JCM).

However, implementing JCM requires that function $C(y)$ is properly specified. In other words, once the scale parameters in **v** have been somehow assessed (see section 4), a conditional pdf $\varphi(\boldsymbol{\tau} \mid \mathbf{v}, y)$ on the *K*-dimensional performance indicator variate $\boldsymbol{\tau}$ should be specified.

This task is less demanding if function $\varphi(\boldsymbol{\tau} \mid \mathbf{v}, y)$ can be assumed to take the same value whatever be the observed value of y (*equivariance to shift* assumption, Morris 1977):

$$\varphi(\boldsymbol{\tau} \mid \mathbf{v}, y) \;=\; \varphi(\boldsymbol{\tau} \mid \mathbf{v}) \tag{9}$$

However, it still remains a frustratingly difficult task, especially in the absence of an adequate parametric modelling, which would allow to assess the entire function by means of a relatively small number of parameters.

There exist several suitable choices about a parametric probabilistic model for the *K*-dimensional performance variate $\boldsymbol{\tau}$. Some preliminary remarks are necessary in order to motivate our choice:

– according to definition (4), each element τ_k of vector $\boldsymbol{\tau}$ is a (cumulate) probability;

– when modeling a joint pdf $\varphi(\cdot \mid \mathbf{v})$ on the variate $\boldsymbol{\tau}$, it needs to take into account that “values [...] near 0 or 1 will ordinarily have smaller standard errors than those around ½. [...] A possibility is to suppose some transform of probability, like log-odds, has constant variance” (Lindley, 1990);

– log-odds lie in the range $-\infty$ to $+\infty$: probabilities that are less, equal or greater than 0.5 correspond to negative, zero, or positive log-odds, respectively. Therefore, modelling the performance function in terms of log-odds, instead of probabilities, is advantageous also because the range of log-odds is coherent with a Gaussian distribution, which is attractive for its good analytic properties and the clear interpretation of its parameters.

For these reasons, a reasonable choice is to assume:

$$\tilde{\boldsymbol{\tau}} \sim \mathcal{N}_K\,(\tilde{\mathbf{t}}, \mathbf{S}) \tag{10}$$

where

– $\tilde{\boldsymbol{\tau}}$ denotes the *K*-dimensional vector of log-odds $[\tilde{\tau}_k]'_{k=1,\dots,K}$, with $\tilde{\tau}_k =$ $= \ln[\tau_k/(1-\tau_k)] \in \Re$ for $k = 1, 2, \dots, K$;

– $\tilde{\boldsymbol{\tau}}$ and $\mathbf{S}$ denote the mean vector and the covariance matrix of the *K*-variate Gaussian distribution, respectively.

The analytical form of function $\varphi(\boldsymbol{\tau}|\,\mathbf{v})$ can be obtained by using a change of variable from $\tilde{\boldsymbol{\tau}}$ to $\boldsymbol{\tau}$. Denoting by $\psi(\cdot|\,\mathbf{v})$ the model in (10) for the performance function of the transform $\tilde{\boldsymbol{\tau}}$, the well-known change formula yields:

$$\varphi(\boldsymbol{\tau}|\,\mathbf{v}) = \left|\,\mathrm{J}_{\tilde{\boldsymbol{\tau}}\to\boldsymbol{\tau}}\,\right| \cdot \psi_{\tilde{\boldsymbol{\tau}}}(\boldsymbol{\tau}|\,\mathbf{v}) \tag{11}$$

As the Jacobian of the transformation $\tilde{\boldsymbol{\tau}} \to \boldsymbol{\tau}$ is:

$$\mathrm{J}_{\tilde{\boldsymbol{\tau}}\to\boldsymbol{\tau}} = \prod_{k=1}^{K} \frac{1}{\tau_k(1-\tau_k)} \tag{12}$$

the resulting performance function of the variate $\boldsymbol{\tau}$ is:

$$\varphi(\boldsymbol{\tau}|\,\mathbf{v}) = c \cdot \prod_{k=1}^{K} \frac{1}{\tau_k(1-\tau_k)} \cdot \exp\left[-\frac{1}{2}(\tilde{\boldsymbol{\tau}}-\tilde{\mathbf{t}})'\,\mathbf{S}^{-1}\,(\tilde{\boldsymbol{\tau}}-\tilde{\mathbf{t}})\right] \tag{13}$$

where *c* denotes the normalization constant.

Finally, the calibration function $C(y)$, defined in (6), can be obtained as follows.

Definition (4) implies that:

$$\tilde{\boldsymbol{\tau}} = \tilde{\mathbf{G}}(y) \tag{14}$$

where $\tilde{\mathbf{G}}(y) = [\tilde{G}_k]' = [\ln(G_k/(1-G_k))]'_{k=1,\ldots,K}$. By substituting equation (14) in (13), $C(y)$ takes the form:

$$\begin{aligned} C(y) &= \varphi(\mathbf{G}(y)|\,\mathbf{v}) = \\ &= c \cdot \prod_{k=1}^{K} \frac{1}{G_k(y)\cdot[1-G_k(y)]} \cdot \exp\left[-\frac{1}{2}(\tilde{\mathbf{G}}(y)-\tilde{\mathbf{t}})'\,\mathbf{S}^{-1}\,(\tilde{\mathbf{G}}(y)-\tilde{\mathbf{t}})\right] \end{aligned} \tag{15}$$

It's worth noting that the calibration function, as represented by (15), is univocally defined by two parameters: the mean vector and the covariance matrix of $\tilde{\boldsymbol{\tau}}$. These parameters are estimated by least squares on a training set consisting of *n* verifying observations $y_1, \ldots, y_i, \ldots, y_n$ of y and the corresponding forecasts $f_{i1}, \ldots, f_{ik}, \ldots, f_{iK}$ from the ensemble members, for *i* = 1, 2, ..., *n*. In other words, the estimates are the solutions of the following minimization problem,

$$\min_{\tilde{\mathbf{t}},\,\mathbf{S}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2\,, \tag{16}$$

where $\hat{y}_i$ is obtained according to equation (8).

4. A COMPARISON WITH BAYESIAN MODEL AVERAGING: A SIMULATION STUDY

The performance of JCM is compared with that of Bayesian Model Averaging (BMA) on a Monte Carlo study.

BMA (Leamer, 1978; Hoeting *et al.*, 1999) is a standard method for combining predictive pdfs from different sources. Section 4.1 briefly presents BMA extension to dynamical models proposed by Raftery *et al.* (2005). In section 4.2 the simulated settings explored in our study are described and the results are discussed.

4.1. *Bayeisan Model Averaging*

Let y be a quantity to be predicted on the basis of the determinisitic forecasts $f_1, ..., f_k, ..., f_K$ given by K models $M_1, ..., M_k, ..., M_K$. BMA predictive model can be written as follows (Raftery *et al.*, 2005):

$$p(y|f_1, ..., f_k, ..., f_K) = \sum_{k=1}^{K} w_k g_k(y|f_k) \tag{17}$$

where:

- $g_k(y|f_k)$ is the individual ensemble member predictive pdf associated with f_k;
- $w_k \geq 0$ represents the weight assigned to $g_k(\cdot)$ on the basis of the performance of model M_k in a training period, with $\sum_{k=1}^{K} w_k = 1$.

Raftery *et al.* (2005) restrict their attention to the situation where the conditional pdfs $g_k(y|f_k)$ are approximated by normal distributions, each centered at a linear function of the forecast, $a_k + b_k f_k$, with $a_k, b_k \in \Re$. That is,

$$y|f_k \sim \mathcal{N}(a_k + b_k f_k, \sigma^2), \quad k = 1, 2, ..., K \tag{18}$$

Therefore, according to BMA, the pdf of the quantity of interest is a weighted average of pdfs centered on the individual bias-corrected forecasts. The conditional expectation of y given the forecasts is taken as the BMA final forecast $\hat{y}_{BMA}$:

$$\hat{y}_{BMA} = E(y|f_1, ..., f_k, ..., f_K) = \sum_{k=1}^{K} w_k(a_k + b_k f_k) \tag{19}$$

Model parameters, a_k, b_k, w_k (k = 1, 2, ..., K) and σ^2, are estimated on a training dataset, consisting of forecasts from the ensemble members and verifying observations. Firstly, a_k and b_k are estimated by simple linear regression of the observations on the forecasts. Then, the estimation of w_k and σ^2 is carried out by maximum likelihood via the EM algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 1997).

4.2. *The simulation study*

The performance of JCM with respect to BMA has been evaluated on simulated ensembles. Both the methods are compared to the ensemble mean.

Two remarks about the way JCM has been implemented are needed:

- a non-informative prior in (7) has been adopted;
- the scale parameters in vector **v** should be known before fitting JCM. In absence of models M_1, ..., M_k, ..., M_K providing a variability measure together with the deterministic forecast[2], they have to be somehow assessed. In our study, the generic element v_k has been estimated on the training data by the mean of squared residuals in a simple linear regression of f_{ik} on y_i, for $i = 1, 2, ..., n$.

For the sake of simplicity in designing the simulated settings, the number K of ensemble members was taken equal to 2.

The predicted values for both the members have been simulated starting from the observed values $y_1,...,y_i,...,y_n$ of sea-level pressure y (in mb) in Pacific Northwest in the second half of June 2000 (SLP dataset available at http://www.stat.washington.edu/MURI).

The forecasts given by the ensemble members were drawn from a bivariate normal distribution with diagonal covariance matrix: the i-th value predicted by member 1 was simulated from a Normal distribution with mean equal to the observed value y_i and variance equal to 1; the i-th value predicted by member 2 was simulated from a Normal distribution with mean equal to $y_i + 1$ and variance equal to 2. With this setup, the performance indicators of the two ensemble members are mutually independent; moreover, member 2 tends to provide biased forecasts, which are more variable than those given by member 1.

In order to explore the effectiveness of the compared methods in the presence of some association between members' performance, we considered the same situation described above also with non-diagonal covariance matrices. Positive and relatively high values of Pearson's correlation coefficient r between members' performance have been examined, as this is the most problematic context where a combining algorithm can be applied (Cooke, 1991). For this reason, we reported only the results for r=0.8 and r=0.9. The situation corresponding to r=0 has been taken as a benchmark, because it represents a null situation, where the ensemble members behave independently from each other.

[2] Recently, in probabilistic weather forecasting such a variability measure can be assessed by running the model in the same starting time, using different initial conditions or physical parametrizations (see e.g. http://www.nhc.noaa.gov/modelsummary.shtml).

TABLE 1

Training period and test day (in calendar days of June 2000), for the 10 replicates of the simulation. n denotes the number of sea-level pressure observed values

Replicate	Training period (3 days)	Test day
Replicate 1	14th, 16th-17th (*n*=488)	18th (*n*=164)
Replicate 2	17th, 18th, 21th (*n*=461)	22th (*n*=168)
Replicate 3	18th, 21th, 22th (*n*= 468)	23th (*n*=164)
Replicate 4	21th – 23th (*n*=468)	24th (*n*=171)
Replicate 5	22th – 24th (*n*=503)	25th (*n*=164)
Replicate 6	23th – 25th (*n*=499)	26th (*n*=163)
Replicate 7	24th – 26th (*n*=498)	27th (*n*=160)
Replicate 8	25th – 27th (*n*=487)	28th (*n*=154)
Replicate 9	26th – 28th (*n*=477)	29th (*n*=164)
Replicate 10	27th – 29th (*n*=478)	30th (*n*=155)

We generated 10 replicates for each setting (r=0, r=0.8 and r=0.9). For each replicate, a training period of 3 calendar days has been considered and the following day has been taken as a test period (see table 1). For some days the data were missing, so that the number of calendar days spanned by the training dataset was sometimes larger than three.

On each training dataset, both BMA and JCM were carried out. The BMA procedure described in section 4.1 has been implemented by Raftery *et al.* in the R package `ensembleBMA`. We implemented JCM procedure in R code, resorting to simulated annealing (implemented in the R function `optim`) in order to address the minimization problem in (16).

The predictive performance of the compared procedures has been assessed in terms of root mean square error (RMSE).

TABLE 2

Results of the simulation study.
The values are root mean square RMSEs (with the corresponding range reported in brackets)

Predictive model	Pearson's correlation coeffcient		
	$r = 0$	$r = 0.8$	$r = 0.9$
Member 1	0.99 (0.96,1.03)	1.01 (0.95, 1.11)	0.99 (0.91, 1.02)
Member 2	2.23 (2.13, 2.41)	2.22 (2.09, 2.49)	2.20 (2.05, 2.30)
Ensemble mean	1.22 (1.16, 1.34)	1.52 (1.40, 1.70)	1.52 (1.42, 1.58)
BMA	0.94 (0.85, 1.07)	1.00 (0.93, 1.08)	0.98 (0.92, 1.02)
JCM	0.92 (0.87, 1.00)	0.90 (0.85, 0.96)	0.76 (0.70, 0.86)

As the results in table 2 show, both JCM and BMA outperform the ensemble mean for each r value. This is due to the fact that the ensemble mean is based on the assumption that the ensemble members are unbiased, mutually independent and with common variability. When r=0, only the second assumption holds, and when $r \neq 0$ no assumption holds at all, while both BMA and JCM allow for bias correction, variability differences and correlations.

Therefore, as it could be expected, the ensemble mean is negatively affected by the presence of positive correlation. On the contrary, BMA error values seem to be quite stable over different correlation levels, while JCM error values even get lower with increasing correlation, up to an improvement of 23% over BMA when r=0.9. Invariance of BMA performance over different correlation levels is probably due to the fact that, even though BMA does not include any correlation parameter, it is fitted via maximum likelihood, so allowing for correlation as an intrinsic feature of the data. On the other hand, JCM seems to take advantage from high values of positive correlation, as if the calibration parameters were able to exploit information from correlated members in order to improve the final JCM forecast.

Finally, another notable difference with respect to BMA which emerges from these results is that JCM final forecasts outperform also each individual ensemble member.

5. CONCLUDING REMARKS

A Bayesian joint calibration model (JCM) for combining predictive distributions from forecast models has been proposed, where "forecast model" is a generic term referring to any tool used to generate a prediction of a future event, such as the state of the atmosphere.

JCM, as well as Bayesian Model Averaging (BMA), is designed to produce probabilistic forecasts, and as a by-product also produces deterministic forecasts. The performance of JCM with respect to BMA has been evaluated in terms of root mean square error (RMSE) on simulated ensembles. Both BMA and JCM outperform the ensemble mean. BMA error values seem to be quite stable over different correlation levels between the predictions provided by the forecast models, while JCM error values even get lower with increasing correlation, up to an improvement of 23% over BMA. In addition, JCM deterministic forecasts have a lower RMSE than any of the individual ensemble members.

At present we are investigating issues related to the application of JCM on real data.

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

PATRIZIA AGATI
DANIELA GIOVANNA CALÒ
LUISA STRACQUALURSI

RIFERIMENTI BIBLIOGRAFICI

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.

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

A.P. DEMPSTER, N.M. LAIRD, D.B. RUBIN (1977), *Maximum likelihood from incomplete data via the EM algorithm*, "Journal of the Royal Statistical Society", B, 39, pp. 1-39.

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.

J.A. HOETING, D.M. MADIGAN, A.E. RAFTERY, C.T. VOLINSKY (1999), *Bayesian Model Averaging. A tutorial*, "Statistical Science", 14, pp. 382-401.

E.E. LEAMER (1978), *Specification Searches*, Wiley, New York.

D.V. LINDLEY (1990), *The 1988 Wald Memorial Lectures: the present position in Bayesian Statistics*, "Statistical Science", 5, pp. 44-65.

G.J. McLACHLAN, T. KRISHNAN (1997), *The EM algorithm and Extensions*, Wiley, New York.

P. MONARI, P. AGATI (2001), *Fiducial inference in combining expert judgements*, "Statistical Methods and Applications", 10, pp. 81-97.

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

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

J.R. RHOME (2007), *Technical summary of the National Hurricane Center track and intensity models*, National Oceanic and Atmospheric Administration, National Weather Service, available at http://www.nhc.noaa.gov/modelsummary.shtml

R.L. WINKLER, R.T. CLEMEN (2004), *Multiple experts vs multiple methods: combining correlation assessments*, "Decision Analysis", 1, pp. 167-176.

SUMMARY

A joint calibration model for combining predictive distributions

In many research fields, as for example in probabilistic weather forecasting, valuable predictive information about a future random phenomenon may come from several, possibly heterogeneous, sources. Forecast combining methods have been developed over the years in order to deal with *ensembles* of sources: the aim is to combine several predictions in such a way to improve forecast accuracy and reduce risk of bad forecasts.

In this context, we propose the use of a Bayesian approach to information combining, which consists in treating the predictive probability density functions (pdfs) from the individual ensemble members as data in a Bayesian updating problem. The likelihood function is shown to be proportional to the product of the pdfs, adjusted by a joint "calibration function" describing the predicting skill of the sources (Morris, 1977). In this paper, after rephrasing Morris' algorithm in a predictive context, we propose to model the calibration function in terms of bias, scale and correlation and to estimate its parameters according to the least squares criterion. The performance of our method is investigated and compared with that of Bayesian Model Averaging (Raftery, 2005) on simulated data.