Implementation of Ensemble Prediction Systems post-processing methods, for electric system management

Gogonel, Adriana
EDF R&D, OSIRIS Department
1 av Charles de Gaulle
92140 Clamart, France
E-mail: Adriana.Gogonel@edf.fr

Collet, Jérôme
EDF R&D, OSIRIS Department
1 av Charles de Gaulle
92140 Clamart, France
E-mail: Jerome.Collet@edf.fr

Bar-Hen, Avner
University of Paris Descartes, MAP5 Laboratory
45 rue des Saints Pères
75006 Paris, France
E-mail: Avner.Bar-Hen@mi.parisdescartes.fr

In this paper we show how we implement post processing methods on the ensemble prediction systems (EPS) in order to improve temperature forecasts provided by Meteo-France. The principle of the EPS is to run several scenarios of the same model with slightly different input data to simulate the uncertainty of modeling and input measures. The post-processing methods we test are the Best-Member method (proposed by Fortin) and the Bayesian method (proposed by Raftery). Their results are compared using scores verifying the skill and/or the spread of the EPS.

Introduction

This study is presenting the Ensemble Prediction Systems (EPS) of forecasting temperature and their use for the electric system management, at EDF France. The temperature is a major risk-factor in the electricity consumption in France as 30% of the buildings have electric heating.

The EPS are conceived in order to give the probability of the meteorological events and the zone of inherent uncertainty in every planned situation. The principle of the EPS is to run several scenarios of the same model with slightly different input data in order to simulate the uncertainty. Then we obtain a probability distribution function informing us about the probability of realization of a forecast.

Data description The data set we are working on corresponds to the period May 31st, 2007 - January 22nd, 2008 and contains forecasts up to 14 time-horizons corresponding to 7 days (1 horizon corresponds to 12 hours). Is is a data-set provided by Meteo-France as an ensemble of weather prediction system which contains 51 members, or 51 equiprobable scenarios obtained by running the same forecasting model with slightly different initial conditions. These scenarios are randomly named by numbers from 0 to 50. Hence, the uncertainty added to ensemble members is not related to the number of the ensemble member (except the scenario 0 as the one with no perturbation of the initial conditions added). That means that for a given day d the scenario k is not the same as the scenario k for another day j. Nevertheless for a given day d, the scenario k producing the forecast for $d+1$ is the same that produces forecasts up to $d+14$.

The temperature forecast value used currently by the electric system management is the average of the 51 values of the temperature provided by the Meteo-France EPS. To improve its prediction, we propose two statistical post-processing methods: the Best Member Method (BMM) proposed by V. Fortin (see Ref [1]) and the Bayesian Averaging Method (BMA) proposed by Raftery (see Ref [2]). After implementing it on our data-set we will compare them by the mean of criteria: skill or accuracy criteria (how close the forecasts are to the observations) and spread or variability criteria (how well the forecasts represent the uncertainty). We set the horizon-time and make a uni-variate study. Up to three days there are good deterministic forecasts, so we choose to study the 7th, 9th and 11th horizon-time. We only present in this document the results for the 7th horizon.

Methods

Best Member Method

The idea of this method is to design for each lead time in the data set, the best forecast among all 51s (in our case) and to construct an error pattern using only the errors made by those "best members" and then to "dress" all members with this error pattern. This approach fails in cases where the undressed ensemble members are already over dispersive and the solution is to dress and weight each member differently, using a different error distribution for each order statistic of the ensemble. So we can distinguish two more specialized methods: one with constant dressing, or the un-weighted members method and one with variable dressing, or the weighted members method (see Ref [1]).

The un-weighted members method The temperature prediction system provides the forecasts $\mathbf{x}_{t,k,j}$, where k is the scenarios number, t is the time and j is the time-horizon. The method is presented in univariate case so from the start j is fixed, hence $\mathbf{x}_{t,k,j}$ becomes $\mathbf{x}_{t,k}$. Let $\mathbf{y}_t$ be the unknown variable which is forecasted at the moment t, and let $X_t = \{\mathbf{x}_{t,k}, k = 1, 2, ..., K\}$ be the set of all ensemble members of the forecasting system. Given X_t the purpose is to obtain, a probabilistic forecasts i.e. $p(\mathbf{y}_t|\mathbf{x}_t)$ in order to provide many more predictive simulations $p(\mathbf{y}_m|\mathbf{x}_m)$ sampled from $p(\mathbf{y}_t|X_t)$ with $m = 1, .., M$ and $M \gg K$.
The basic idea of the method is to "dress" each ensemble member $x_{t,k}$ with a probability distribution being the error made by this member when it happened to give the best forecast. The best scenario is $\mathbf{x}_t^*$ defined as the one minimizing $\|\mathbf{y}_t - \mathbf{x}_{t,k}\|$ for a given norm $\|\cdot\|$: $\mathbf{x}_t^* = \arg_{\mathbf{x}_{t,k}} \min \|\mathbf{y}_t - \mathbf{x}_{t,k}\|$.
For a given norm and an archive of past forecasts a probability distribution is created from the realizations of $\varepsilon_t^* = \mathbf{y}_t - \mathbf{x}_{t,k}^*$: $p(\mathbf{y}_t|X_t) \approx \frac{1}{K}\sum_{k=1}^{K} p_{\varepsilon^*}(\mathbf{y}_t - \mathbf{x}_{t,k})$
Application In this first sub method we will treat all the ensemble members equally, we build only one probability distribution function for all them. We choose to simulate 10 times more values, that is $M = 10 \times K = 10 \times 51$ predictions instead of $K = 51$ that we had initially. In the first graphique of the Fig. 1 we can observed superposed on the same graphic, the three curves: initial forecasts average, simulated un-weighted forecasts average and the observations. Even if the curve of the forecasts we simulated is still not very close to the observation curve, it is closer than the curve of initial curve of forecasts. This result is encouraging but we will do some other verifications to establish its skill and spread.

The weighted members method In this second sub-method we take into account the statistical rank of the ensemble members. Hence, we have:

- $\mathbf{x}_{t,(k)}$, the kth statistical rank of the ensemble members $X_t = \{\mathbf{x}_{t,k}, k = 1, 2, ..., K\}$
- $\varepsilon_{(k)}^* = \{y_t - \mathbf{x}_t^* |\ \mathbf{x}_t^* = \mathbf{x}_{t,(k)}, t = 1, 2, , ..., T\}$ the errors of the best ensemble members for every t moment in a database of past forecasts, when the best forecast has rank k .

- p_k be the probability that $x_{t,(k)}$ be the best member, i.e. $p_k = Pr[\mathbf{x}_t^* = \mathbf{x}_{t,(k)}]$

In the same way as we did for the first sub method, for the archive of past forecasts and a given norm we create a probability distribution from the realizations of $\varepsilon_{(k)}^* = \mathbf{y}_t - \mathbf{x}_{t,k}^*$, and that is $\varepsilon_{(k)}(t) = \mu_{prev}(t) + \exp(\nu_{prev}(t))\mathcal{N}(0,1)$

Application For this sub method also we choose to simulate 10 times more values, that is 10x51 predictions instead of 51 that we had initially. But unlike the un-weighted method, here we treat the ensemble members differently, by their classes of statistical rank, so we build 51 different probability distribution functions. The curve of the weighed forecasts is closer than the initial predictions curve to the observation, but not as close as for the unweighed forecasts sub-method (Fig. 1). We will also do other verifications to established its skill and spread.

Bayesian model averaging

It is a statistical method for postprocessing Model Output which allows to provide calibrated and sharp predictive probability density functions (PDFs) even if the output itself is not calibrated (see Ref [2]).

Application Applying the bayesian method is constructing the BMA PDF as a weighted sum of normal PDFs, where the weights are reflecting the ensemble members overall performance over the training period. Let y^T be the quantity to be forecasted and $M_1, ..., M_K$ K statistical models providing forecasts. According to the law of total probability, the forecasts PDF, $p(y)$ is given by: $p(y) = \sum_{k=1}^{K} p(y|M_k)p(M_k|y^T)$ where $p(y|M_k)$ is the forecast PDF based on M_k and $p(M_k|y^T)$ is the posterior probability of model M_k being correct given the training data and tells if the model is fitting the training data. The sum of all k posterior probabilities corresponding to the k models is 1: $\sum_{k=1}^{K} p(M_k|y^T) = 1$. This allows us to use them as weights, so to define the BMA PDF as a weighted average of the conditional PDFs.

The first and an important step of this method is to choose the length of the training period. We are looking for the best compromise as the advantage of a short training period is that it is able to adapt rapidly to changes (as weather patterns and model specification change over time) and the advantage of a longer training period is that the BMA parameters are better estimated. We compare training period lengths (from 10 to 50 days, by 5 days step) by measurements as the Mean Absolute Error (MAE) and the Continuous Ranked Probability Score (CRPS). For our data-set the values of the MAE and the CRPS decrease substantially up to 30 days increase slowly from 30 to 50. Hence a length of 30 days seems be a good choice.

We can now apply the BMA on our data-set of 30 days using the R-packages EnsembleBMA (see [3]) and Verification (see [4]). Results study starts with the curves graphique. In the Fig.1 we can see the curve of the BMA Simulations is not closer to the observations curve than the curve of the initial forecasts. Other scores are calculated in the section below to decide on the spread and skill of the BMA forecasts.

Comparing methods by skill and spread criteria

We compare below three kind of scores: standard mesures, reliability scores and resolution scores for the initial forecasts (IPred), unweighted forecast (Un-F) from the best member method, weighted forecasts from the best member method (W-F) and the forecasts obtained with the bayesian method(BMA-F). We only present the main scores of every category.

Bias It is a standard mesure, Biais $= \frac{\frac{1}{N}\sum_{i=1}^{N} F_i}{\frac{1}{N}\sum_{i=1}^{N} O_i}$ (F= forecasts, O=observations) the perfect score is 1 so the ones obtained for the three methods are very good (see Fig. 2) but it is possible to get a

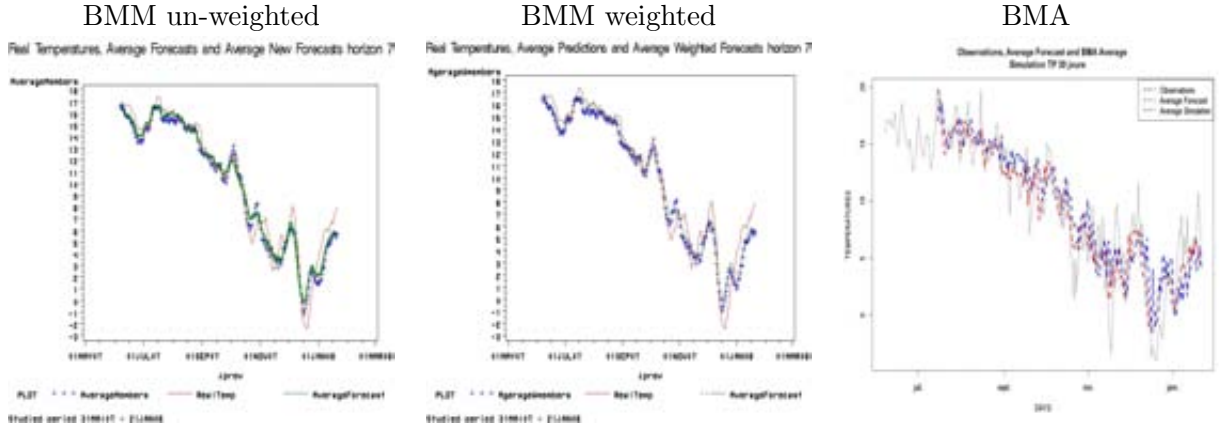

Figure 1: Comparison of the curves of the averages forecasts, for the three methods.

Forecasts	**Bias**	$\mathbf{R}^2$	**RMSE**
Initial forecasts	$0,98$	$0,89$	$2,54$
Unweighted forecasts	$0,99$	$0,89$	$2,74$
Weighted forecasts	$0,98$	$0,89$	$2,74$
Bayesian Forecasts	$1,04$	$0,88$	$2,77$

Figure 2: The values of the standard measures for the three applied methods.

perfect score for a bad forecast if there are compensating errors.

Correlation Coefficient It measures the correspondence between forecasts and observations and is given by the variance between forecasts and observations $\mathbf{R}^2 = \frac{\sum_{i=1}^{n}(y_i-\overline{y})(o_i-\overline{o})}{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2\sum_{i=1}^{n}(o_i-\overline{o})^2}}$. A perfect correlation coefficient is 1. The (R^2) we obtain for the three methods are very close, between them: $0,88$ for the bayesian forecasts and $0,89$ for the others (see Fig.2). That shows a good correlation between observation and forecasts. The degree of correlation is kept after post processing the forecasts.

Root mean square error (RMSE) It is defined as the root square of the $MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$. As for any error, the smaller RMSE the better. The RMSE's values for the three methods are low, between 2,54 for the IPrev and 2,77 for the BMA-F (see Fig.2) showing small errors of the models.

The Talagrand diagram It is a reliability criteria, more known as the PIT (Probability Integral Transform). It measures how well the ensemble spread of the forecast represents the true variability (uncertainty) of the observations. For each time instant (day) we consider the ensemble of the forecasts values (the observation value included). The values within this ensemble are ordered and the position of the observation is noted (the rank). Repeating the procedure for all the lead time (day of the period) we obtain a histogram of observations rank. By examining the shape of the Talagrand diagram, we can draw conclusions on the bias of the overall system and the adequacy of its dispersion.

In the Figure3 we have the Talagrand diagrams for the three cases. For the Un-F and W-F the Talagrand diagram have a dome-shape. This is specific for ensembles spread too large, where too many observations are falling near the center of the ensemble. The BMA-F has a "U" shape showing under dispersion with a small bias, as it is not a symmetric histogram. The better spread is that of the BMA-F, better than the IPred.

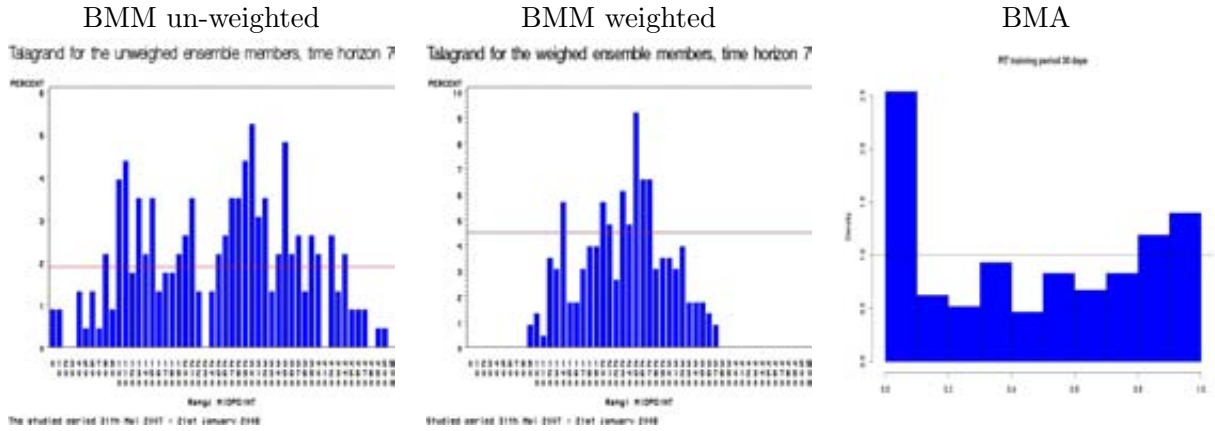

Figure 3: Comparison of the Talagrand Diagrams for the three methods.

Forecasts	**MAE**	**CRPS**
Initial forecasts	2, 0	1, 05
Unweighted forecasts	2, 1	1, 09
Weighted forecasts	2,1	1, 13
Bayesian Forecasts	2, 2	1, 49

Figure 4: The values of CRPS and MAE for the applied methods as well as for the initial forecasts.

Mean absolute error (MAE) It measures overall accuracy and is defined as $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - o_i|$. The smaller the mean absolute error, the better. When we compare the MAE for the initial forecasts and the MAE for the forecasts obtained by the three methods, we find closed values (see Fig. 4). Hence, post processing the forecasts by the three methods does not increase the MAE.

The Continuous rank probability score (CRPS) It measures the difference between the forecast and observed cumulative distribution functions (CDFs). The CRPS is related to the rank probability score, but compares a full distribution with the observation, where both are represented as CDFs. If F is the CDF of the forecast distribution and x is the observation , the CRPS is defined as:

$$(1) \quad CRPS(F,x) = \int_{-\infty}^{+\infty} [F(y) - \mathbb{1}\{y \geq x\}]^2 \, dy$$

where $\mathbb{1}\{y \geq x\}$ denotes a step function along the real line that attains the value 1 if $y \geq x$ and the value 0 otherwise. In the case of probabilistic forecasts the CRPS is a probability-weighted average of all possible absolute differences between forecasts and observations.
The CRPS provides a diagnostic of the global skill of an EPS, the perfect CRPS is 0, a higher value of the CRPS indicates a lower skill of the EPS. The values of CRPS for the three methods, calculated for the entire studied period are between 1 and 1,5 (see 4). There are good values proving a high skill of the new created EPS, nevertheless the CRPS of the BMA-F is significantly higher than the IPred's CRPS i.e. the EPS obtained by BMA has lower skill than the initial EPS.

The Reliability diagram is the plot of observed frequency against forecast probability for all probability categories. A good reliability implies a curve close to the diagonal. In the Figure 5 we can

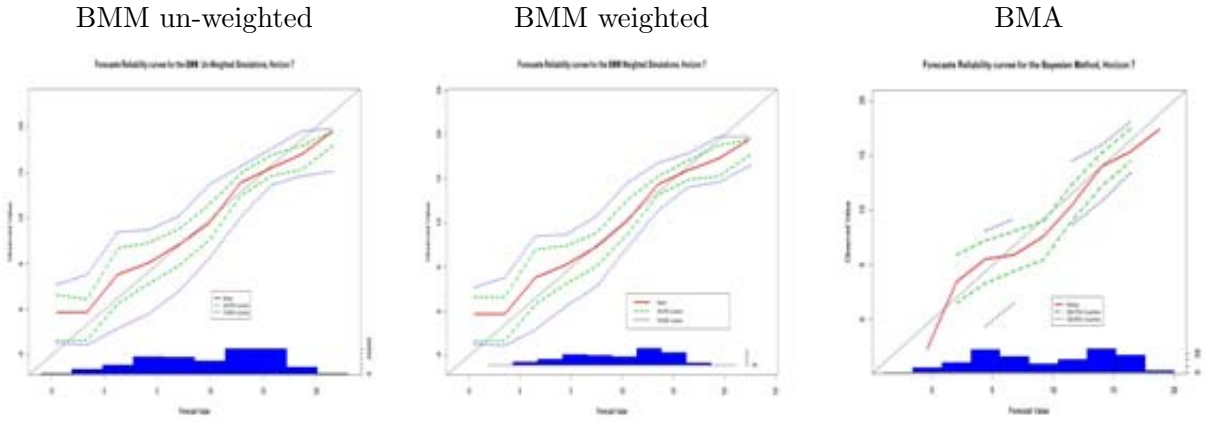

Figure 5: Comparison of the Reliability Diagrams for the three methods.

see the reliability curves for our three forecasts data-sets. All three diagrams are close to the diagonal, but with deviations from time to time, showing a conditional bias, especially for the BMA-F.

Conclusion

From all our diagnoses results are not as good as we could expect it. The BMA method improves the spread of the ensemble but not its skill. The BM Method does not improve the spread of the EPS but it improves its skill (as showed by the curves graphiques) for the extremes temperatures. The other scores verifying skill (CRPS, MAE) don't show an improvement but this is explained by the fact that they provide a diagnostic of the *global* skill of an EPS. Thus we will concentrate on extreme temperatures for future research as there is a need in electricity management, to measure and reduce blackout residues.

Figures

REFERENCES

[1] Article FORTIN, V.; FAVRE, A.; SAID, M. Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member Q. J. R. Meteorol. Soc., 2006, 132, 1349 -1369
[2] Article RAFTERY, A.; GNEITING, T.; BALABDAOUI, F.; POLAKOWSKI, M. Using Bayesian Model Averaging to Calibrate Forecast Ensembles Physical Review, 2004, 20
[3] Techreport FRALEY, C.; RAFTERY, A. E.; GNEITING, T.; SLOUGHTER, J. M. ensembleBMA: An R Package for Probabilistic Forecasting using Ensembles and Bayesian Model Averaging Department of Statistics University of Washington, 2009
[4] Techreport POCERNICH, M. Verification Package: examples using weather forecasts. National Center for Atmospheric Sciences (NCAR), 2010