CS-BIGS 5(2) : 74-87
http://www.bentley.edu/csbigs

Implementation of two statistical methods for Ensemble Prediction Systems in the management of electrical systems

Adriana Gogonel
Electricité de France and University of Paris Descartes, France

Jérôme Collet
Electricité de France, France

Avner Bar-Hen
University of Paris Descartes (MAP5), France

This paper presents a study of two statistical post-processing methods implemented on forecasts of Meteo-France temperatures provided by the ensemble prediction system (EPS) The results are useful in the management of electricity consumption at EDF France. Those methods are the Best-Member Method (BMM) proposed by Fortin (2006), and the Bayesian Model Averaging method (BMA) proposed by Raftery (2004). The idea behind the BMM is to design for each lead time in the data set the best forecast among all k forecasts provided by the temperature prediction system, to construct an error pattern using only the errors made by those "best members" and to then "dress" all the members of the initial prediction system with this error pattern. The BMA method is a statistical method which combines predictive distributions from different sources. The BMA predictive probability density function (PDF) of the quantity of interest is a weighted average of PDFs centered on the bias-corrected forecasts, where the weights are equal to posterior probabilities of the models generating the forecasts and reflect the accuracy (skill) of the models over the training period. The resulting forecasts implemented on our data set are compared with one another and compared to the initial forecasts, using scores which measure the accuracy and/or the spread of the EPS: the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), the Ignorance Score, the Continuous Rank Probability Score (CRPS), the Talagrand Diagram, the Bias and the Mean. The purpose is to improve the probability density function of the forecasts, preserving at the same time the quality of the mean forecasts. The presentation is accessible to readers with an intermediate level of statistics.

Keywords: forecasting; ensemble prediction systems; energy; Bayesian analysis

1. Introduction

1.1 Context

The energy sector is highly weather-dependent, hence it needs accurate forecasts to guaranty and optimize its activities. Predictions of production are needed to optimize the selling and distribution of electricity. Of course needs inelectricity depend on meteorological conditions.

Numerous specialists in physics (for example meteorologists) build sophisticated deterministic numerical models, with uncertainty on the input data. To take into account this uncertainty, they run the same model several times, with slight but non random perturbation of the data. It seems obvious that the results of physical models contain irreplaceable information. Nevertheless we notice that the probability distribution obtained from the model is typically not a perfect representation of the risk factor, thus we need to submit it to a statistical processing before using it.

Correlation between temperature and electricity consumption

Temperature is the main risk factor for EDF as an electricity producer in France, a country where electric heating is well developed. If we take into account the variability in temperature, the power consumed for heating on a winter given day can vary by about 20GW, which represents 40 % of the average consumption. Regarding energy, the climatic risk factor is quantitatively less important, because the difference in energy consumed between the warmest and the coldest winters represents approximately 5 % of the energy over the year.

To explain the correlation between temperature and electricity consumption we note that the French electrical load is very sensitive to temperature because of the development of electric heating since the 70's. The influence of temperature on the French load is mostly known, except for the impact of air conditioning whose trend remains difficult to estimate. Electric heating is used to maintain a temperature close to 20° C inside buildings.

Taking into account the "free" contributions to heat (sun, human heat), it is considered that electric heating turns on at approximately below 16°C. Beyond that temperature, the heat loss being proportional to the temperature difference between inside and outside, the consumption increases approximately linearly. Moreover, since buildings take a certain amount of time to warm up or to cool down, the reaction to outside temperature variations is delayed.To take into account this delay, one uses a smoothed temperature (based on an "average" temperature) as a predictor of the consumption (Bruhns et al., 2005). The situation is similar for air conditioning.

1.2 Purpose of the work

This study has for objective to improve the probabilistic distribution of forecasts provided by the Ensemble Prediction Systems (EPS) of Meteo-France, while preserving the accuracy (skill) of the mean forecasts. The initial EPS contains $k = 51$ members - scenarios of the same model - one starting with unperturbed initial weather conditions (the control forecasts) and 50 from perturbed initial conditions defined by adding small dynamically active perturbations to the operational analysis for the day. Each one of the 51 members of the studied EPS provides trajectories of temperatures for 14 time-horizons (1 horizon corresponds to 1 day). We implement statistical post-processing methods to improve its use for the management of the electric system at EDF France.

The use of the EPS method allows on the one hand to extend the horizon where we have good forecasts and on the other hand to give a measure of forecast uncertainty. Unlike deterministic solutions the probability forecast is better adapted to the analysis of risk and decision-making.

First, we study the Meteo-France temperature forecasts and the temperature realizations in retrospective mode in order to establish the statistical link between these two variables. We then examine two statistical processing methods of the pattern's outputs. From state of the art existing methods and from the results obtained by the examination of the probability forecasts, a post-processing module will be developed and tested. The goal is to achieve a robust method for calibrating statistical forecasts. This method should thus take into account the uncertainties of the inputs (represented by the 51 different initial conditions added to the pattern).

The first method is the Best-Member Method (BMM) and it has been proposed by Fortin et al. (2006). The idea is to design for each lead time in the data set the best forecast among all k forecasts provided by the temperature prediction system, to construct an error pattern using only the errors made by those "best members" and then to "dress" all the members of the initial prediction system with this error pattern. This approach fails in cases where the initial prediction systems are already over dispersed. This is the reason why a second method was introduced which allows dressing and weighting each member differently by classes of its statistical order. We present in this paper the second method, that we call W-BMM.

The second method we implement is the Bayesian Model Averaging (BMA) method proposed by Raftery et al. (2004). It is a statistical method for post processing model outputs which allows to provide calibrated and sharp predictive Probability Distribution Functions (PDFs) even if the output itself is not calibrated (forecasting are well calibrated if for a forecast with probability p, the predicted event is observed p times). The method allows for using a sliding-window training period to estimate new models parameters, instead of using the whole database of past forecasts and observations.

Results will be compared using scores which measure the skill and/or the spread of the EPS: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Ignorance Score, Continuous Rank Probability Score (CRPS), Talagrand Diagram, Reliability diagram, Bias, Mean.

2. Ensemble prediction systems (EPS)

Ensemble prediction systems are a rather new tool in operational forecasting which allows for faster and scientifically justified comparisons of several forecast models. The EPS is conceived in order to yield the probability of meteorological events and the zone of inherent uncertainty in every planned situation. It is a technique to predict the probability distribution of forecast states, given a probability distribution of random analysis error and model error.

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. In the current system Meteo-France is using, each EPS perturbation is a linear combination of singular vectors with maximum growth computed using a total energy norm. The assumption underlying the linear combination is that initial error is normally distributed in the space spanned by the singular vectors. A Gaussian sampling technique is used to sample realizations from this distribution (IFS Documentation 2006).

The EPS is based on the notion that forecast uncertainty is dominated by error or uncertainty in initial conditions. This is consistent with studies that show that, when two operational forecasts differ, the differences tend to originate in the analyses rather than in model formulation see (IFS Documentation 2002).

We note some of the primary objectives of EPS (Mallet 2008):

1. Allows for estimating the uncertainty, obtaining a representative spread of the typical uncertainty, ensuring an empirical standard deviation of the forecasts comparable with the standard deviation of the observations;

2. Provides a good estimation of the probability of an event;

3. Allows for a convenient use of linear combinations of models in forecast.

3. Verification methods for EPS

Meteorologists have been using EPS for several years now and at the same time many methods for evaluating their performances have been developed (Stanski et al. 1989). A proper scoring rule maximizes the expected reward (or minimizes the expected penalty) for forecasting one's true beliefs, thereby discouraging hedging or cheating (Jolliffe and Stephenson 2007). One can distinguish two types of methods: the ones permitting to evaluate the quality of the spread and the ones giving a score (a numerical result) permitting to evaluate the performance of the forecasts (Petit 2008).

Therefore, we need to examine the *skill* or accuracy (how close the forecasts are to the observations) and *spread* or variability (how well the forecasts represent the uncertainty). If model errors played no role, and if initial uncertainties were fully included in the EPS initial perturbations, a small spread among the EPS members would be an indication of a very predictable situation i.e. whatever small errors there might be in the initial conditions, they would not seriously affect the deterministic forecast. By contrast, a large spread indicates a large uncertainty of the deterministic forecast (Persson 2003). As for the skill, it indicates the correspondence between a given probability, and the observed frequency of an event. Statistical considerations suggest that even for a perfect ensemble (one in which all sources of forecast error are sampled correctly) there may not be a high correlation between spread and skill (Whitaker and Loughe 1998).

3.1 Standard Statistical Measures

Let y be the vector of model outputs and let o be the vector of the corresponding observations. These vectors both have n components. Their means are respectively $\overline{y}$ and $\overline{o}$.

The *Bias* is given by:

$$Bias_m = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{1}{n}\sum_{i=1}^{n} o_i} \qquad (1)$$

The *Correlation Coefficient* is given by:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(o_i-\bar{o})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(o_i-\bar{o})^2}} \qquad (2)$$

The *Mean Absolute Error (MAE)* measures overall accuracy and is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - o_i\right| \qquad (3)$$

The *root mean square error (RMSE)* has the advantage of being recorded in the same unit as the observations and it is the root square of the MSE where MSE is given by $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i-o_i)^2$.

3.2 Reliability

The reliability (or spread) measures how well the predicted probability of an event corresponds to its observed probability of occurrence. For a p probability forecast, the predicted event should be observed $round(p)$ times.

Talagrand Diagram. It is a type of bar chart in which categories are represented by bars of varying ranks rather than specific values - a histogram of ranks. The Talagrand diagram has its origins in the Probability Integral Transform diagram (PIT, see Dordonnat and Collet 2010). It measures how well the spread of the ensemble forecast represents the true variability (uncertainty) of the observations. For each period (day) we consider the ensemble of the forecasts values (including the observation value). The values within this ensemble are ordered and the position of the observation is noted (the rank). For example the rank will be 0 if the observation is below all the forecasts and N if the observation is above all the forecasts. Repeating the procedure for all the forecasts we obtain a histogram of observation ranks. 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:

- A flat histogram: the ensemble spread correctly represents forecast uncertainty. It does not necessarily indicate a skilled forecast but only measures whether the observed probability distribution is well represented by the ensemble.

- A U-shaped histogram: the ensemble spread too small, so that many observations fall outside the extremes of the ensemble

- A Dome-shaped histogram: the ensemble spread is too large, so that too many observations fall near the center of the ensemble

- Asymmetric histogram: the ensemble contains bias.

3.3 Resolution (sharpness)

The resolution (or accuracy, or skill) is the measure of the accuracy of the forecasts.

Continuous Rank Probability Score (CRPS). The CRPS measures the difference between the forecast and observed cumulative distribution functions (CDFs). The CRPS compares the 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:

$$CRPS(F,x) = \int_{-\infty}^{\infty}\left[F(y) - 1\{y \geq x\}\right]^2 dy \qquad (4)$$

where $1\{y \geq x\}$ 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 tends to increase with forecast bias and be reduced by the effects of the correlation between forecasts and observations (Schaake et al. 2007).

One of its advantages is that it has the same units as the predicted variable (so is comparable to the MAE) and does not depend on predefined classes. It is the generalization of the Brier score for the case of the continuous variables. 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.

4. Post-processing methods

4.1 The Best-Member method

The Best-Member Method was proposed by V.Fortin et al. (2006) and improves on studies previously led by Roulston and Smith (2002) then by Wang and Bishop (2005). The idea is to design for each lead time in the data set the best forecast among all 51 forecasts (in our case), to construct an error pattern using only errors made by those "best members" and to then "dress" all members with this error pattern. This approach does not work in cases where the undressed ensemble members are already over- or under-dispersed; the solution is then to weigh and dress each member differently, that is using a different error distribution for each order statistic of the ensemble. So we can distinguish two specialized methods: the one with constant dressing, or the “un-weighted members” method and the one with variable dressing, or the “weighted members” method. We implement and present in this paper the Weighted Members Method (W-BMM).

4.1.1 The Weighted Members Method

Fortin applied this method on a synthetic EPS (where the EPS was built under certain conditions - ensemble members are independent and identically distributed - and where we can vary the parameters we want in order to test different hypotheses or methods). He observed that this method failed in the case of over-dispersed or under-dispersed EPS. The explanation is that when an EPS is under-dispersed, the outcome often lies outside the spread of the ensemble. Hence, an extreme forecast has a much more chance of giving the best prediction than a forecast close to the ensemble mean. Conversely, when an ensemble is over-dispersed members close to the ensemble mean have a much higher chance to be best members than extreme forecasts. Hence, the probability that an ensemble member gives the best forecast as well as the error distribution of the best member depends on the distance to the ensemble mean. For univariate forecasts we can sort ensemble members from the smallest to the biggest, note theirs ranks and consider the rank of a member at the dressing sequence.

Let $x_{t,k,j}$ be the temperature predictions provided by a given EPS, where k is the scenario’s number, t is the time for which the forecast is made and j is the time-horizon. The method is presented in the univariate case so j is fixed, hence $x_{t,k,j}$ becomes $x_{t,k}$.

Let y_t be the unknown variable which is forecasted at time t, and let $\mathrm{X}_\mathrm{t} = \{\mathbf{x}_{\mathrm{t},k}, k = 1, 2, \ldots, \mathrm{K}\}$ be the set of all ensemble members of the forecasting system. Given X_t the purpose is to obtain a probabilistic forecast i.e. $p(y_t \mid X_t)$ in order to provide many more predictive simulations sampled from $p(y_t \mid X_t)$ where $\mathrm{X}_\mathrm{t} = \{\mathbf{x}_{\mathrm{t,m}}, \mathrm{m} = 1, 2, \ldots, \mathrm{M}\}$ with $\mathrm{M} \gg \mathrm{K}$.

The concept of conditional probability allows for incorporating additional information into the forecasts (in this case it will be the forecasts given by Meteo France).

The basic idea of the method is to "dress" each ensemble member $x_{t,k}$ with a probability distribution equal to that of the error made by this member when it happened to give the best forecast. The best scenario denoted by $\mathbf{x}^*$ is defined as the one minimizing $\| y_t - x_{t.k} \|$ for a given norm $\| . \|$. As we are working in a univariate space, the norm is the absolute value so that:

$$\mathbf{x}_t^* = \arg_{\mathbf{x}_{t,k}} \min | \mathbf{y}_\mathrm{t} - \mathbf{x}_{\mathrm{t,k}} | .$$

Let:

- $x_{t,(k)}$ be the k th member of the ensemble $\mathrm{X}_t = \{\mathbf{x}_{\mathrm{t,k}}, \mathrm{k} = 1, 2, \ldots, \mathrm{K}\}$ ordered by rank.
- $\varepsilon^*_{(k)} = \left\{ \mathbf{y}_t - \mathbf{x}_t^* \middle| \mathbf{x}_t^* = \mathbf{x}_{t,(k)}, t = 1, 2, .., T \right\}$ be the errors of the best ensemble members for every time t in a database of past forecasts, where the best forecast has rank k.
- p_k be the probability that $\mathbf{x}_{\mathrm{t,(k)}}$ is the best member, i.e $p_k = Pr[\mathbf{x}_t^* = \mathbf{x}_{t(k)}]$.

To dress each ensemble member differently, instead of resampling from the archive of all best-member errors, one resamples from $\varepsilon^*_{(k)}$ to obtain dressed ensemble members. Hence, the simulated forecasts are obtained as:

$$\mathbf{y}_{t,k,n} = \mathbf{x}_{t,(k)} + \omega \cdot \varepsilon_{t,(k),n} \tag{5}$$

where:

- the $\varepsilon_{t,(k),n}$ are drawn at random from $\varepsilon^*_{(k)}$,
- n is the number of simulations per time step,
- $\omega = \sqrt{\frac{s^2}{s^2_{\varepsilon^*}}}$ and $s^2_{\varepsilon^*} = \frac{1}{T-1}\sum_{t=1}^{T}(\varepsilon^*_{t,(k)})^2$ is the estimated variance of the best-member error.

This computation fails in the case of EPSs where the uncertainty is already over estimated (s^2 negative).

4.2 Bayesian model averaging

The Bayesian approach is based on the fact that the probability of realization of an event does not depend only on its frequency of appearance but also on the knowledge and experience of the researcher.

We base our study on the Bayesian Model Averaging method (BMA) (Raftery et al. 2004). The BMA predictive probability density function (PDF) of the quantity of interest is a weighted average of PDFs centered on the bias-corrected forecasts, where the weights are equal to the posterior probabilities of the models generating the forecasts, reflecting the models skill over the training period.

An original idea in this approach is to use a moving training period (sliding-window) to estimate new models parameters, instead of using the whole database of past forecasts and observations. This implies a choice of length for this sliding-window training period; the principle guiding this choice is that probabilistic forecasting methods should be designed to maximize sharpness subject to calibration. It is an advantage to use a short training period in order to be able to adapt rapidly to changes (since weather patterns and model specification change over time) but the longer the training period, the better the BMA parameters are estimated (Raftery et al. 2004). After comparing measurements such as the RMSE, the MAE, the CRPS for various training period lengths (from 10 to 60) Raftery et al. conclude that there are substantial gains in increasing the training period to up to 30 days, and that beyond that there is little gain. The main difference between their case and ours is that they have 5 models and we have 51 (scenarios).

Let y^T be the quantity to be forecasted and $M_1,\ldots,M_K$ be K statistical models providing forecasts. According to the law of total probability, the PDF of the forecasts $p(y)$ is given by:

$$p(y) = \sum_{k=1}^{K} p(y\,|\,M_k)\,p(M_k\,|\,y^T) \tag{6}$$

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 which a measure of whether the model fits 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. This approach uses the idea that there is a best "model" for each prediction ensemble but it is unknown. Let f_k the bias-corrected forecast provided by the model M_k which yields the best prediction, corresponding to a PDF $g_k(y\,|\,f_k)$. The BMA predictive model then is given by:

$$p(y\,|\,f_1,..,f_k) = \sum_{k=1}^{K} w_k g_k(y\,|\,f_k) \tag{7}$$

where w_k is the posterior probability of forecast k being the best one and is based on forecast k 's performance in the training period and where $\sum_{k=1}^{K} w_k = 1$. For temperature and sea-level pressure, the conditional PDF can be fitted reasonably well using a normal distribution centered at a bias-corrected forecast $a_k + b_k f_k$ as shown by Raftery et al. (2004):

$$y\,|\,f_k \sim N(a_k + b_k f_k, \sigma^2)$$

The parameters a_k, b_k as well as the w_k are to be estimated on the basis of the training data set: a_k and b_k by simple linear regression of y_t on f_{kt} for the training data and w_k , $k = 1, .., K$, and σ by maximum likelihood (Aldrich 1997) from the training data. For algebraic simplicity and numerical stability reasons it is more convenient to maximize the logarithm of the likelihood function rather than the likelihood function itself; the expectation-maximization (EM) algorithm (Dempster et al. 1977) is used.

Finally the BMA PDF is a weighted sum of normal PDFs and the weights w_k reflect the overall performance of the ensemble members over the training period, relative to other members.

5. Application

5.1 Data description

We are working on temperature forecasts provided by Meteo-France as an ensemble of weather prediction systems which contains 51 members, or 51 equiprobable scenarios obtained by running the same forecasting model with slightly different initial conditions.

The data set corresponds to the period between March 30 2007 and 20 of April 20 2011 and contains forecasts up to 14 time-horizons corresponding to 14 days (1 horizon corresponds to 24 hours). Currently the value used to predict the consumption is the mean of the 51 forecasts. In Figure 1, on top we represent for three fixed time-steps the curves of the prediction errors for the 51 scenarios as a function of time-horizon (from 1 to 14-days ahead).

The errors typically increase with the time-horizon; they are particularly small up to the 4th time-horizon. Hence, we consider that up to the 4th time-horizon the deterministic forecasts give high quality forecasts and we implement our improvement methods starting with the 5th horizon. In the same figure, on the bottom, we can see the prediction errors for all periods but for three different time-horizons; we notice the same typical correlation between the errors and the time-horizon.

As mentioned above every scenario, among the 51, gives forecasts up to 14-days ahead. The difference between the scenarios comes from the small dynamically active perturbation added to their initial conditions.

Hence this perturbation is not related to the name of the scenario (numbers between 0 and 50) and is not the same from one day of forecasting start to another [1].

The temperature measurements are performed by 26 different French stations, of which we take a weighted average to obtain a single temperature for France. The weights are defined so as to best explain electricity consumption for different French regions.

We start by setting the time-horizon. Once the horizon is fixed, we study the forecasts, starting with 5-days ahead horizons since up to 4-days ahead the determinist forecasts are very good (the Meteo-France pattern is purposely built to be under- dispersed up to 3 days). In this paper we present the 5-days ahead results. We can see in Figure 2 superposed on the same graph the curve of the realizations and the curve of the average predicted temperatures.

5.2 Implementation of the weighted Best-Member method

Let $\mathbf{y}_t$ be the temperature variable we are forecasting at time t and let $X_t = \{x_{t,k}, k = 1,2,...,K\}$ be the set of all ensemble members of the Meteo-France forecasting system. We would like to obtain a probabilistic forecast i.e. $p(y_t \mid X_t)$. The conditional probability allows for taking into account additional information in a forecast, in our case the forecasts given by Meteo-France. The best scenario $\mathbf{x}_t^*$ is the one minimizing $|\mathbf{y}_t - \mathbf{x}_{t,k}|$.

To compute the W-BMM method we use the SAS software. We use a cross-validation method to build and test our models: we divide the four years in our data set into two equal parts: the first part serves as a model building period for the model we will validate on the second part and vice versa.

As mentioned in the presentation of the method, the statistical rank of the ensemble members is taken into account. We denote the kth forecast by $\mathbf{x}_{t,(k)}$ and recall that $\varepsilon_{(k)}^* = \{y_t - \mathbf{x}_t^* \mid \mathbf{x}_t^* = \mathbf{x}_{t,(k)},\ t = 1,2,\ ,\ ...,\ T\}$ as defined above (see 4.1.1).

[1] For example: forecasts given by scenario 15 computed on July 1st for the period July 1st-July 7th take into account from the beginning a certain perturbation. That perturbation will not be the same as the one taken into account by scenario 15 when on July 2nd it provides forecasts for the period July 2nd-July 8th.

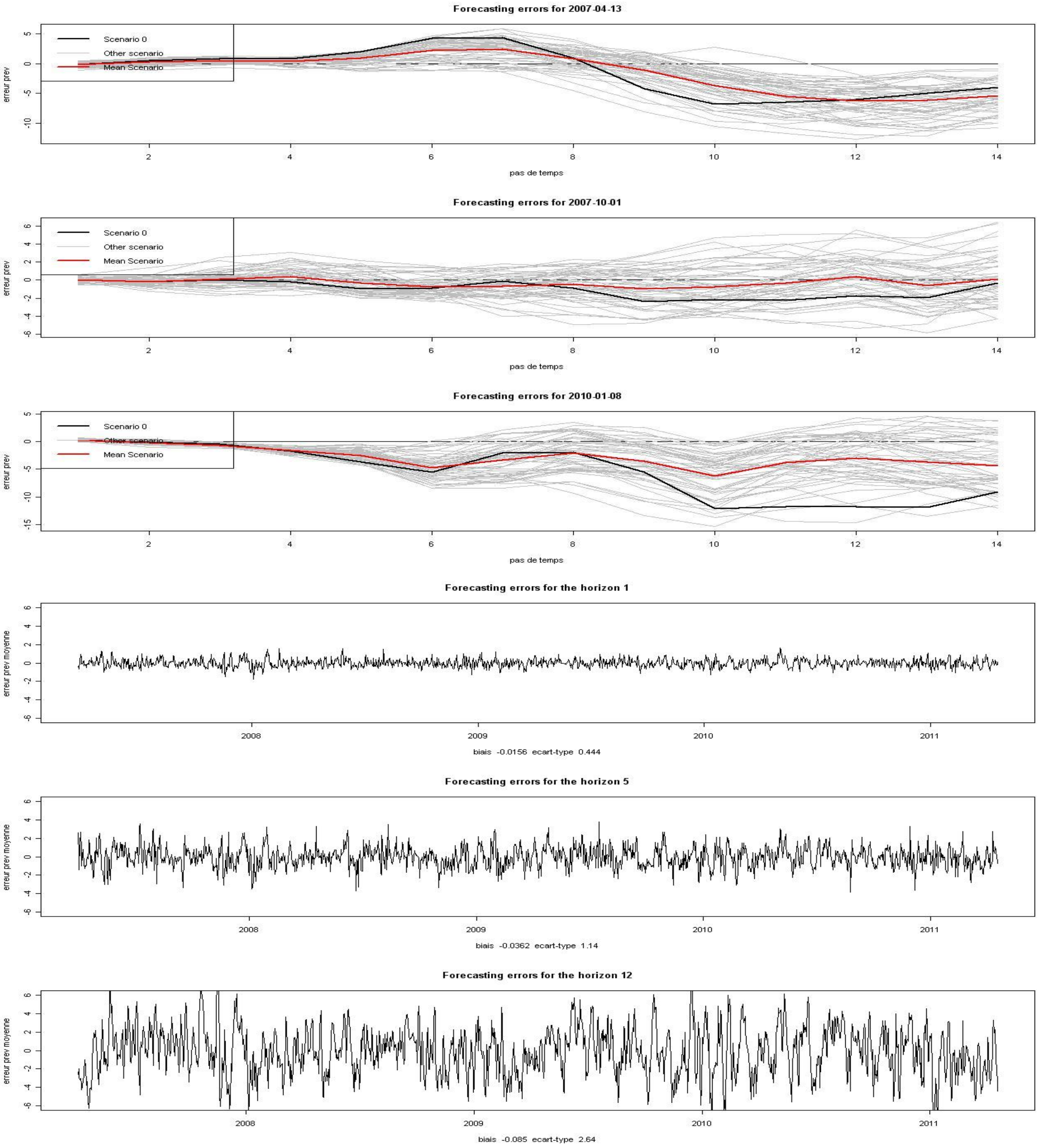


Figure 1. Figures corresponding to initial predictions. On top, prediction error curves for the 51 scenarios for three fixed days, for all 14 time-horizons (in gray are scenario errors from 1 to 51, in black scenario 0 - the one with no perturbed initial conditions - and in red the mean of the 51 scenarios). At the bottom are the forecasting errors for three different time-horizons; we can see the errors become larger with the time-horizon.

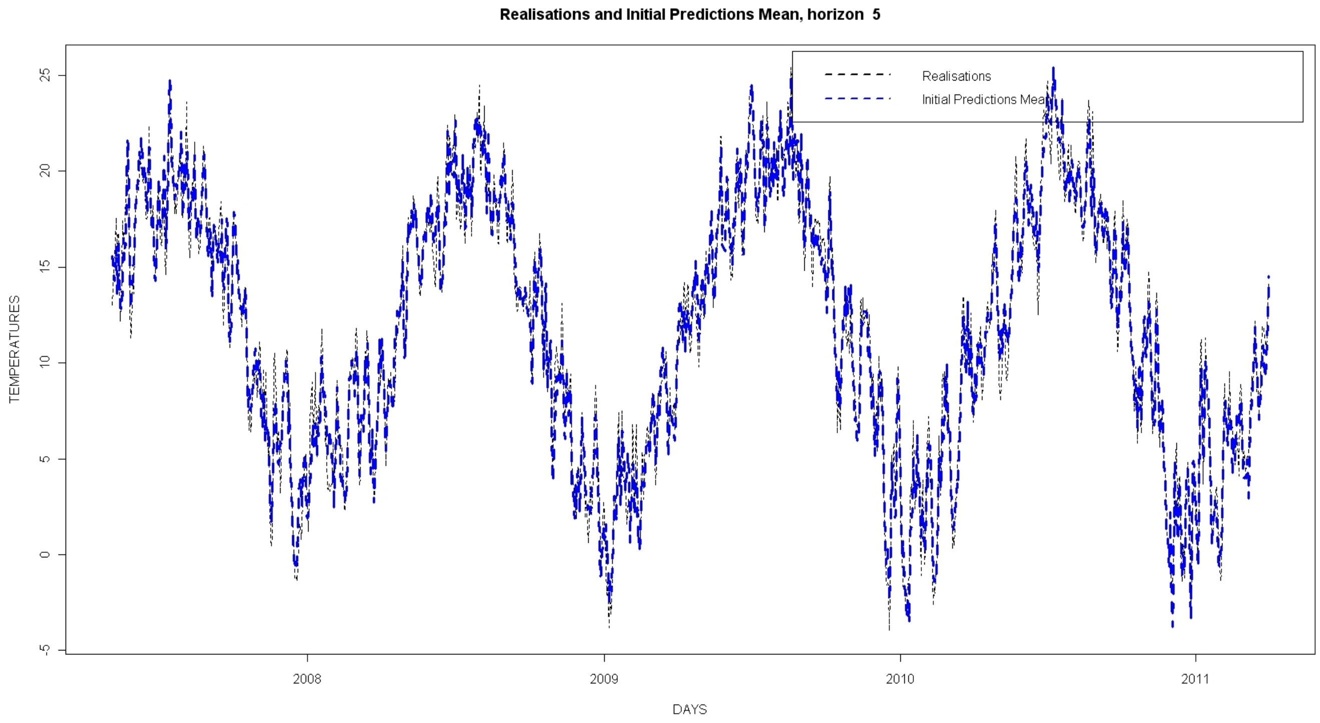


Figure 2. Graphs corresponding to initial predictions for 5-days ahead. In black is the observed temperatures curve, in blue the initial prediction means. We notice a good precision of the mean forecasts except for extreme temperatures.

Given the archive of past forecasts and a norm we create a probability distribution from the realizations of $\varepsilon^*_{(k)} = y_t - x_{t,(k)}$:

$$\varepsilon_{(k)}(t) = \mu_{\text{prev}}(t) + \exp(\nu_{\text{prev}}(t))N(0, 1) \qquad (8)$$

where:

- *t* is the time-step,
- the μ_{prev} are the values of the errors, predicted by a linear regression model M_1 as described below,
- the ν_{prev} are the logarithms of the absolute values of the residuals from the M_2 pattern, predicted by a linear regression model M_2 as described below,
- the statistical rank does not interfere directly at this stage of the study but interferes indirectly in the creation of the M_1 and M_2 patterns.

The M_1 pattern explains the prediction error with the initial forecast, the day-position within the year and the statistical rank τ_t:

$$\mu_{\text{prev}} = \alpha_1 \cdot \mathbf{x}_t + \sum_{i=1}^{3}[\alpha_{2,i} \cdot \text{a}(i) + \alpha_{3,i} \cdot \text{b}(i)] + \alpha_4 . \tau_t \qquad (9)$$

The M_2 pattern explains the logarithm of the absolute value of the residuals from the M_1 pattern i.e. ν_{prev}, with the temperature, day-position within the year and the statistical rank τ_t:

$$\nu_{\text{prev}} = \beta_1 \cdot \mathbf{x}_t + \sum_{i=1}^{3}[\beta_{2,i} . \text{a}(i) + \beta_{3,i} \cdot \text{b}(i)] + \beta_4 \cdot \tau_t$$

(10)

We therefore also have two parameters μ_{prev} (predicted by the M_1 pattern) and ν_{prev} (predicted by the M_2 pattern). Both are generated by 7 parameters: $\alpha_1, \alpha_{2,\text{i}}, \alpha_{3,\text{i}}$ for μ_{prev} (i = 1, 2, 3)

and $\beta_1, \beta_{2,\text{i}}, \beta_{3,\text{i}}$ for ν_{prev} (*i* = 1, 2, 3). Both have the same length as the studied period – 1,459. As mentioned above we use them as parameters of the normal distribution to simulate our new forecasts. We want to obtain M=10 × K=10 **x** 51=510 simulations so we will draw $N_k = p_k$ × M dressed ensemble members from each $x_{t,(k)}$. This way classes with posterior probabilities of giving better forecasts will be simulated more than classes with small such probabilities:

$$y_{t,k,n} = \mathbf{x}_{t,(k)} + \omega \times \varepsilon_{t,(k),n} \qquad (11)$$

where $\omega = p_k = Pr[\mathbf{x}_t^* = \mathbf{x}_{t(k)}]$ is the probability that $x_{t,(k)}$ gives the best forecasts among the K=51 scenarios.

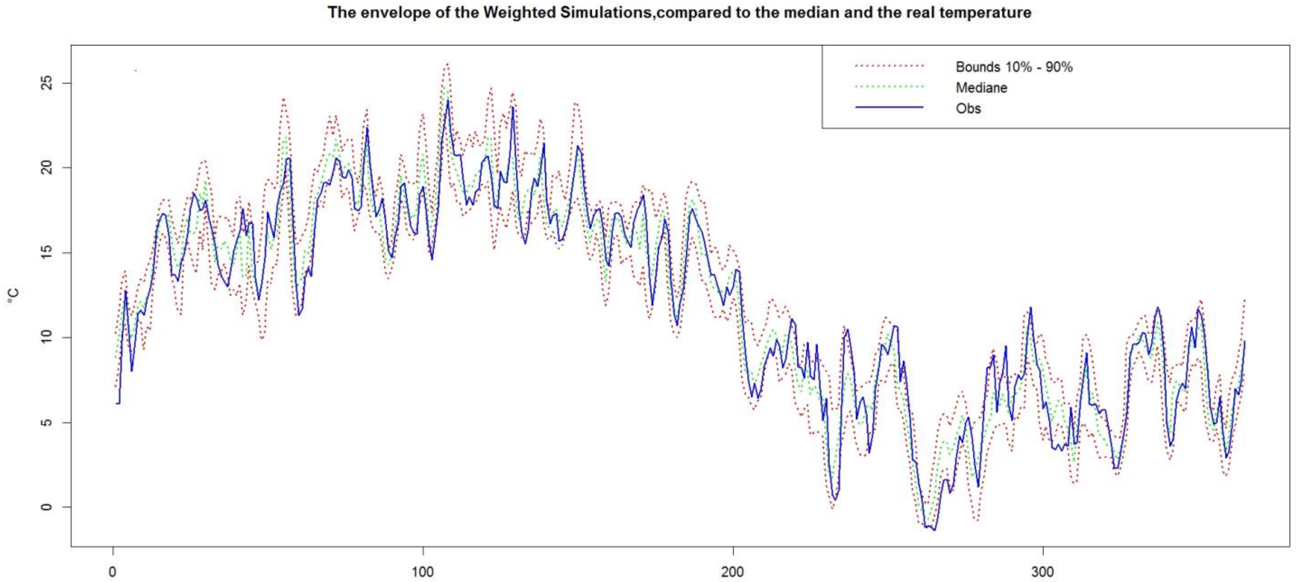


Figure 3. Probability intervals (10%-90%) for the 510 daily simulations (in red), their median (in green) and the curve with realizations (in blue) for a one-year period. We note on this graph that the median of the simulations lies in the [10%, 90%] interval.

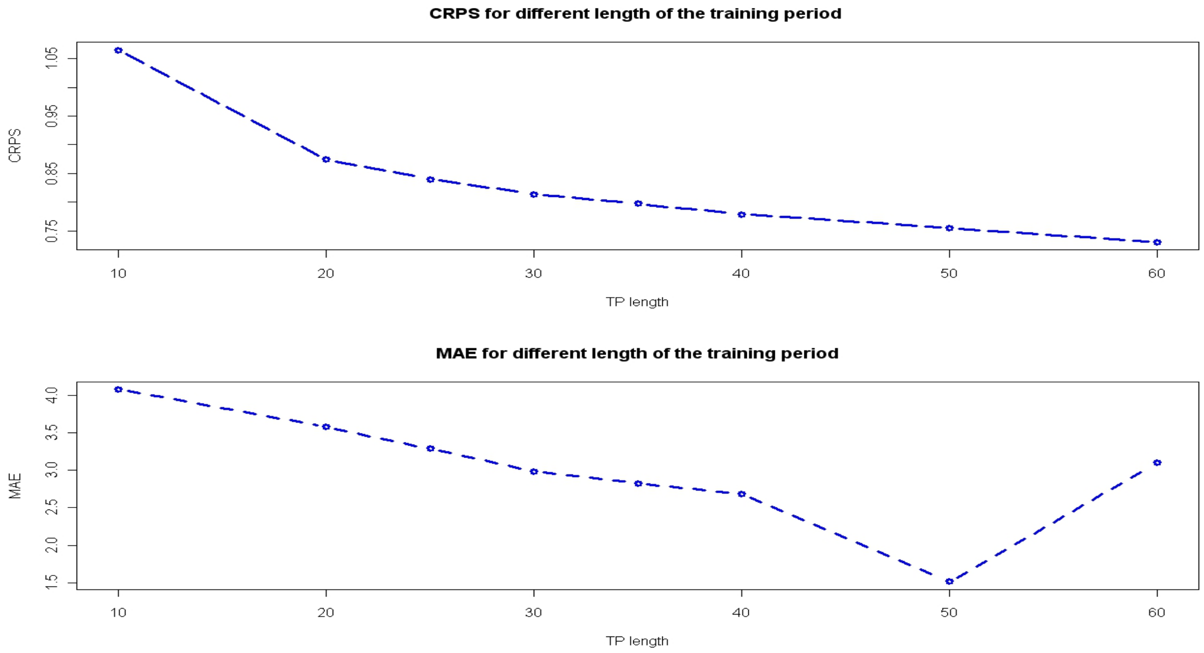


Figure 4. BMA method. CRPS and MAE for 5-days ahead for different lengths of the training period, from 10 to 50, by 5-day steps.

In Figure 3 we can observe the median of simulated forecasts, the real temperatures curve and the 10% - 90% probability interval for the simulated forecasts. The curve of the forecasts we simulated is still not perfectly close to the curve of observations. It is interesting to observe on that graph whether the real temperatures curve always lies within the 10% - 90% interval. Other results of the tests verifying skill and spread are presented in Section 6.

5.3 Implementation of the Bayesian Model Averaging method

Applying the Bayesian Model Averaging method consists in constructing the BMA PDF as a weighted sum of normal PDFs, where the weights reflect the overall performance of the ensemble members over the training period. To implement this method we use an R package for probabilistic forecasting, *ensemble BMA* created by Fraley et al. (2009) using ensemble post processing via Bayesian Model Averaging to provide functions for modeling and forecasting data. When we construct the Bayesian model we consider that forecasts ensemble members are interchangeable (because of the independence of the forecast scenarios, see section 2) that is, their forecasts can be assumed to come from the same distribution.

The first and an important step in this method is to choose the length of the training period. We are looking for a good compromise. The advantage of a short training period is that it is able to adapt rapidly to changes (since weather patterns and model specification change over time). The advantage of a longer training period is that the BMA parameters are better estimated. We compare measurements such as the Mean Absolute Error (MAE) and the Continuous Ranked Probability Score (CRPS) for different training period lengths (from 10 to 60 days, by 5 or 10 day steps).

The values of CRPS and the values of the MAE corresponding to different lengths of the training period are given in Figure 4. We notice that the two curves are alike; the CRPS decreases from 1.0 (10 days) to 0.73 (60 days) and the MAE decreases from 4.2 (10 days) to 1.5 (50 days) and then increases again to 3.1 at 60 days. A 50-days training period is chosen.

Once we have decided on the length period we construct the pattern that fit those data, so that we can obtain the new forecasts system and the corresponding probabilities. Scores are calculated in the next section to decide on the spread and skill of the BMA forecasts.

6. Comparison of the methods by means of the criteria

To compare the quality of the forecasts provided by the statistical post processing methods considered here, we use some of the criteria presented earlier in this paper. We compare three types of scores: standard measures, reliability scores and resolution scores for the initial forecasts, un-weighted forecasts from the best member method, weighted forecasts from the best member method and the forecasts obtained with the Bayesian Model Averaging method.

6.1 Standard Measures

Bias: We compare the bias for the initial forecasts and the bias for the forecasts obtained by the three methods (see Table 1). A perfect score is 1. Scores obtained for all three methods are 1, showing good forecasts but we note that it is possible to get a perfect score for a bad forecast if there are compensating errors.

Correlation Coefficient: The R^2 we obtain for two of the methods has values very close to 1: 0.96 for the Bayesian forecasts and 0.97 for the W-BMM forecasts (see Table 1). Since a perfect correlation coefficient is 1 our scores show a good correlation between observations and forecasts. The correlation coefficient for the initial predictions is 0.99 so the degree of correlation is not lost after post processing the forecasts.

Root mean square error (RMSE): The RMSE's values for two of the methods show small model errors. Nevertheless the RMSE for the initial forecasts is smaller than the RMSE for the forecasts we simulated by W-BMM (see Table 1), and the BMA RMSE is even larger. One possible explanation would be that the RMSE is influenced more by large errors rather than small errors.

From the point of view of standard measures, the forecasts we created have predictive qualities almost as good as the initial predictions.

Mean absolute error (MAE): The smaller the MAE, the better. When we compare the MAE for the initial forecasts and the MAE for the forecasts obtained by the other two methods, we find a larger value for the W-

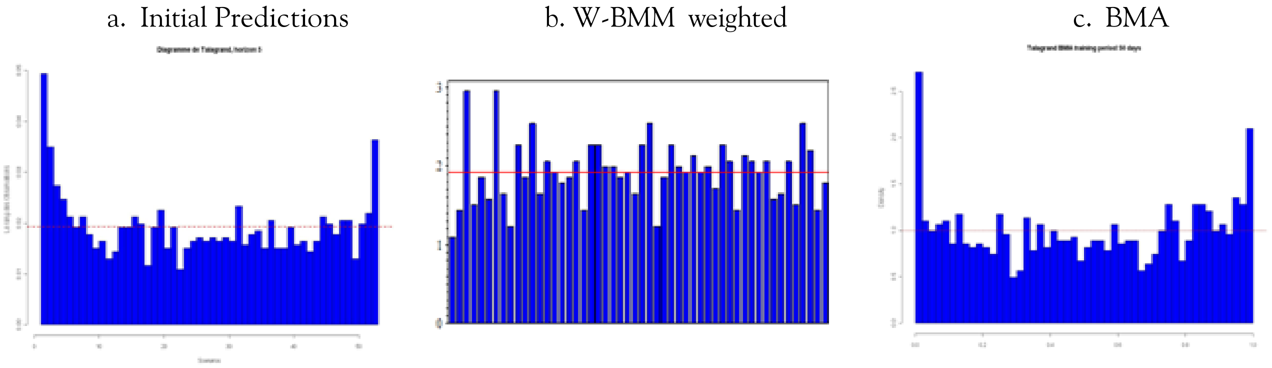
BMM: 1.30 and even larger for the BMA 1.53 (see Table 1). Hence, post processing the forecasts by these two methods slowly increases the MAE, as well as the RMSE. Nevertheless those are good values of the MAE.

Table 1. Values of the standard measures for the three methods.

Forecasts	Bias	R^2	RMSE(°C)	MAE	CRPS
Initial	1	0.99	1.14	0.88	0.63
W-BMM	1	0.97	1.74	1.18	0.63
Bayesian	1	0.96	1.90	1.52	0.75

6.2 Reliability criteria

Talagrand diagram: For the initial system of forecasts for 5-days ahead, the rank histogram is given in Figure 5a. We notice an asymmetric U-shaped histogram meaning that the ensemble spread is too small (under-dispersive) with many observations falling outside the extremes of the ensemble. The EPS is under-dispersive, so the uncertainty is under estimated. The rank histogram of the ensemble obtained by the Best Member Weighted Method has a rather flat shape – the ensemble spread correctly represents forecast uncertainty (see Figure 5b) but we notice that the extremes ranks are not so well represented. The rank histogram of the ensemble obtained by the BMA Method is given in Figure 5c. We still notice a U-diagram, but more symmetrical than the BMM one.


a. Initial Predictions

b. W-BMM weighted

c. BMA

Figure 5. Comparison of the Talagrand Rank Diagrams of the two methods and with the initial predictions

6.3 Resolution Criteria

Continuous rank probability score (CRPS): The CRPS measures the difference between the forecast and observed CDFs. The values of CRPS for the two methods calculated for the entire studied period are given in Table 1. Those are good values, knowing that the perfect CRPS is 0, proving a high skill of the new created EPS. The CRPS of the W-BMM forecasts is as good as the one of the initial predictions.

7. Conclusion

The objective of this paper was to extend the number of simulated forecasts of temperature (the 51 per day) provided by Meteo-France and still have a forecasting system with a good quality (spread and skill) that will be useful for the management of the electric system at EDF France.

Up to the 4th time horizon (1 horizon corresponds to 1 day) the deterministic forecasts give high quality forecasts so we tried to improve forecasts beyond this time-horizon. Therefore we examined two methods of statistical processing of the pattern's outputs which take into account the uncertainties of the inputs (represented by the 51 different initial conditions added to the pattern). We studied their implementation on the data-set provided by Meteo-France. It contains forecasts for the March 30 2007 - April 20 2011 period. There are 51 values of forecasts for 14 time-horizons. We studied separately several horizons-time, starting with the 5-days ahead horizon (the study of the 5th horizon presented in the current article). We want to improve the probability density function of the forecasts, preserving at the same time the quality of the mean forecasts.

The first method is the Best-Member method proposed by Fortin et al. (2006). The idea is to design for each lead time in the data set the best forecast among all k forecasts provided by the temperature prediction system, to construct an error pattern using only the errors made by those "best members" and to then "dress" all the members of the initial prediction system with this error pattern. This method allows us to extend the number of simulated temperatures. We presented here the case where the ensemble members are dressed and weighed differently by classes of its statistical order.

The second method we have implemented is the Bayesian Model Averaging method proposed by Raftery et al. (2004). It is a statistical method for post processing model outputs which allows for providing calibrated and sharp predictive PDFs even if the output itself is not calibrated. The method uses a sliding-window training period to estimate new models parameters, instead of using the whole database of past forecasts and observations.

Reliability and resolution are the attributes that determine the quality of a probabilistic prediction system. Therefore, comparing EPS (three in our case, including the initial system) involves comparing scores which measure the skill and the spread.

From the spread point of view there is significant improvement of the distribution when using the method W-BMM. The Ranks Histogram of the initial EPS shows under-dispersion and a cold and hot bias (see Figure 5) and the Ranks Histogram of the BMA maintains the same shape, but the Ranks Diagram for the W-BMM is rather flat, although there is a small effect on the extreme ranks that might be corrected by using a different modeling approach for the extreme values of the forecasts.

From the skill point of view, the Bayesian Method gives less good results than the initial predictions (see the CRPS, RMSE, MAE values in Table 1). The W-BMM method has a better (smaller) CRPS but the RMSE and MAE are larger. So from the spread point of view we can say that the quality of the initial prediction system is preserved but not improved.

The results we obtained are convenient, considering the objective: increasing the number of forecasts for improving the distribution of the Ensemble Prediction System, without losing the precision of its mean forecasts. The next step is to build a mixture model using the W-BMM for the center of the distribution and Generalized Extreme Value (GEV) models for the tails of the distribution.

Correspondence: adriana.gogonel@gmail.com

References

Aldrich, J. 2007. R. a. fisher and the making of maximum likelihood 1912 - 1922, *Statistical Science* 12 (1997), 162 –176.

Bruhns, A. Deurveilher, G. and Roy, J-S. 2005. A non-linear regression model for mid-term load forecasting and improvements in seasonality, *15th PSCC, Liege*,, 2005.

Dordonnat, V. and Collet, J. 2010. Méthodes de prévision en loi : état de l'art, *Tech. report, EDF R&D*, 2010.

Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* 39 (1977), 1–38.

IFS Documentation, 2002. The ensemble prediction system, Tech. report, *ECMWF*, 2002.

IFS Documentation, 2006 Part v: The ensemble prediction systems, Tech. report, ECMWF, 2006. V. FORTIN, A.C. FAVRE, and M. SAID, 2006. Probabilistic forecasting from ensemble pre-diction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member, Q. J. R. Meteorol. Soc. 132 (2006), 1349 –1369.

Farley, C., Raftery, A. E. Gneiting, T., 2009. and J. M. SLOUGHTER, EnsembleBMA: An r package for probabilistic forecasting using ensembles and bayesian model averaging, Tech. report, Department of Statistics University of Washington, 2009.

Jolliffe, I. T. and Stephenson, D. B., 2007. Proper scores for probability forecasts can never be equitable, *American Meteorological Society* (2007).

Mallet, V., 2008. Prévision d'ensemble, Tech. report, *INRIA Roquencourt*, 2008.

Persson, A., 2003. User guide to ecmwf forecast products, Tech. report, *ECMWF*, 2003.

Petit, T., 2008. Evaluation de la performance de prévisions hydrologiques logiques d'ensemble issues de prévisions météorologiques d'ensemble, Ph.D. thesis, Faculté Des Sciences Et De Génie Université Laval Québec, 2008.

Raftery, A.E., Gneiting, T., Balabdaoui, F., and Polakowski, M., 2004. Using Bayesian model averaging to calibrate forecast ensembles, *Physical Review* (2004), 20.

Roulston and Smith, 2002. Combining dynamical and statistical ensembles, Tellus 55A(2002), 16–30.

Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., and Seo, D. J., 2007. Precipitation and temperature ensemble forecasts from single-value forecasts, *Hydrology and Earth System Sciences Discussions 4* (2007),655 – 717.

Stanski, H.R., Wilson, L.H., and Burrows, W. R., Survey of common verification methods in meteorology, Tech. report, Environement Canada, *Service de l'environement atmosphérique.*, 1989.

Wang, X. and Bishop, C.H., Improvement of ensemble reliability with a new dressing kernel., Q. J. R. Meteorol. Soc. 131 (2005), 965–986.

Whitaker, J.S. and Loughe A. F., 1998. The relationship between ensemble spread and ensemble mean skil l, *Monthly Weather Review* 126 (1998), 3292 – 3302.