Ensemble forecast of photovoltaic power with online CRPS learning

Jean Thorey, Christophe Chaussin, Vivien Mallet

Ensemble forecast of photovoltaic power with online CRPS learning

J. Thorey[a,c], C. Chaussin[b], V. Mallet[c]

[a] CEREA, joint research laboratory ENPC ParisTech – EDF R&D, Université Paris-Est, Marne-La-Vallée, France
[b] EDF R&D, Chatou, France
[c] INRIA, Paris, France

Abstract

We provide probabilistic forecasts of photovoltaic (PV) production, for several PV plants located in France up to 6 days of lead time, with a 30-min timestep. First, we derive multiple forecasts from numerical weather predictions (ECMWF and Météo France), including ensemble forecasts. Second, our parameter-free online learning technique generates a weighted combination of the production forecasts for each PV plant. The weights are computed sequentially before each forecast using only past information. Our strategy is to minimize the Continuous Ranked Probability Score (CRPS). We show that our technique provides forecast improvements for both deterministic and probabilistic evaluation tools.

Introduction

Improved photovoltaic power integration needs better power forecasts. Forecasters may pursue efforts to improve meteorological models, weather-based power models or statistical post-processing methods. For our part, we focus on the following case: a forecaster, willing to provide probabilistic PV power forecasts, retrieves multiple meteorological forecasts (possibly from various sources). In this general setting, numerous state-of-the-art methods can be tested and combined.

Meteorological forecasts can either be deterministic single forecasts or an ensemble of forecasts, usually at coarser resolution. Inman et al. (2013) provide a review of PV forecasting methods with deterministic forecasts. Ensemble forecasting and more generally probabilistic forecasting has been widely covered in the meteorological community (Gneiting and Katzfuss, 2014). Only recently, ensemble-based forecasting techniques are tested for PV (Zamo et al., 2014), while these techniques are more common for wind and wind power forecasting (Ren et al., 2015).

A recent benchmark of deterministic and probabilistic PV forecasts is analyzed in Sperati et al. (2015), along with classical diagnostic tools. Probabilistic forecasts rely on the estimation of quantiles of the predicted probability density function (PDF). Quantile regression (Almeida et al., 2015) and analogs (Alessandrini et al., 2015; Huang and Perry, 2015) are amongst the most popular techniques for quantile estimation in PV. These techniques do not require an ensemble of forecasts as they can rely only on the historical variability of the forecasts and production data. The main drawback of most of the previously cited methods is that they use a single method and not a combination of several methods.

A forecaster having multiple forecasts hopefully wishes to combine them. In our case, we combine deterministic forecasts, quantile forecasts and ensemble forecasts, which is seldom the case. We combine these different types of forecasts to take advantage of their diversity. On the one hand, ensemble members describe several meteorological situations. On the other

hand, quantile forecasts are built from the errors of a deterministic forecast, which describes a single meteorological situation with a finer resolution than the ensemble forecasts. Quantile forecasts estimate the inability of the forecaster to provide a perfect deterministic forecast.

The forecasts combination can be carried out in an optimal way. A batch process would not produce an estimation based on all available data but only on a limited learning data set. A batch process can be updated but it will only mimic an online learning technique. On the contrary, online learning techniques provide rules for combining forecasts, see the monograph Cesa-Bianchi and Lugosi (2006). The combination rules stemming from online learning depend only on the available past information at each forecast step and come with theoretical performance guarantee under essentially no assumptions (concerning prior weights, underlying stochastic process or distributions). The theoretical guarantee of the online learning algorithm can be seen as a long term performance guarantee without stationarity or ergodicity assumptions. Online learning techniques have been tested for several applications: electricity consumption, ozone concentration, wind and geopotential fields, and solar irradiance (Stoltz, 2010; Mallet et al., 2009; Mallet, 2010; Devaine et al., 2013; Baudin, 2015; Thorey et al., 2015).

This paper presents application results with our innovative approach (Thorey et al., 2016), whose purpose is to combine multiple forecasters in a linear opinion pool (Genest and McConway, 1990; Geweke and Amisano, 2011). The originality of our technique is to use combination rules deriving from online learning techniques in order to minimize the CRPS of the weighted empirical distribution function. We stress here the fact that our method provides theoretical guarantee and that it does not rely on distribution assumptions. Besides, the algorithm has a low computational cost and is parameter-free. Our framework is inspired from the work of Gaillard et al. (2016), which focuses on quantile scoring functions.

Minimizing the CRPS is a common strategy in the meteorological literature to obtain calibrated probabilistic forecasts. However, standard techniques do not offer theoretical guarantees of robustness and usually resort to strong assumptions on the distributions. For example, Bayesian model averaging (BMA) techniques provide a mixture of parametric distributions, usually a Gaussian sum (Raftery et al., 2005) or gamma distributions sum for wind and precipitation applications (Sloughter et al., 2010, 2007). Non-homogeneous regression fits the parameters of a parameterized distribution using characteristics of the ensemble of forecasts (Gneiting et al., 2005; Wilks, 2009; Thorarinsdottir and Gneiting, 2010). For instance, a Gaussian distribution is fitted using a linear model between the mean of the distribution and the mean of the forecasts. Besides, likelihood maximization with the logarithm loss is not an appropriate tool in our setting since it fails to produce satisfactory scores for a discrete probability distribution. A discussion on local scores such as the logarithm loss is addressed by Bröcker and Smith (2007b).

The main contributions of this manuscript are twofold:

- We show probabilistic forecasts performance on a large data set comprising 219 PV power plants with deterministic, quantile and deterministic forecasts from two meteorological centers (ECMWF and Météo France). We evaluate PV forecasts that are used operationally at a country-scale.

- Our statistical postprocessing technique creates a weighted empirical distribution by CRPS minimization with theoretical guarantees under essentially no assumptions.

In Section 1, we introduce the production data sets and the forecasts from ECMWF and Météo France. We detail our method to generate deterministic, quantile and ensemble PV forecasts from ensemble and deterministic weather forecasts. We finish this section by describing linear opinion pools, or in other words, by describing how we build a probabilistic forecast from multiple pointwise forecasts. The evaluation tools are described in Section 2, with a focus on the CRPS. Our statistical post-processing method is explained in Section 3. We detail how the weights of the linear opinion pool are updated. Numerical results and discussions are developed in Section 4.

The deterministic and probabilistic predictive skills of the present forecasts are computed. In particular, we highlight the benefits of using our online learning algorithm compared to simply using uniform weights.

1 Methods

1.1 Production and meteorological data

The production data cover 219 PV power plants in metropolitan France with 21 consecutive months (January 2012 to October 2013). The total power of the plants is referred to as France production. We wish to provide production forecasts for each power plant and for France production. The data are shown as load factor, i.e. scaled by the installed capacity. France production forecasts are the weighted sums of the plant forecasts w.r.t. the installed capacity of each plant.

Forecast data are summarized in Table 1 and 2. We use data from two meteorological centers (ECMWF and Météo France), both deterministic forecasts and ensembles of forecasts: HRES and ENS for ECMWF, and ARPEGE and PEARP for Météo France (Courtier et al., 1991; Descamps et al., 2015; Palmer et al., 2009), up to a lead time of 6 days. Note that the deterministic forecasts are not the unperturbed members of the ensembles of forecasts but different forecasts, with better resolution.

We are interested in predicting the 30-min average power output of the plants. We only show results for the following hours of the day 06:00, 09:00, 12:00, 15:00, and 18:00 UTC in order to save computation time and to avoid the issue of temporal interpolation of our forecast data. Ensemble forecasts are only solar irradiance forecasts while deterministic forecasts also include total cloud cover Tcc and 2-m temperature T_{2m}. The ensemble PEARP is available for longer lead times but only with a time step of 6 (and not 3) hours. Consequently, for our application we restricted the use of PEARP up to 2 days.

1.2 Conversion of meteorological forecasts to production forecasts

Our regression technique is inspired from Bacher et al. (2009) and Lorenz et al. (2009). This regression technique has been successfully applied in the benchmark of Sperati et al. (2015), where the technique ranked first in the deterministic PV forecasting competition.

The training set ranges from early 2012 to February 2013 (nearly 400 days). The testing set with the remaining days of 2013 is about 240-day long. For a given weather forecast, the following technique is applied for each time of the day and for each power plant independently.

First, clear sky indices τ_P and τ_I are generated from the production P and the solar forecasts I:

$$\tau_P = \frac{P}{P_{cc}} \quad \text{and} \quad \tau_I = \frac{I}{I_{cc}}\,, \tag{1}$$

where clear sky production P_{cc} and clear sky solar radiation I_{cc} are the production and solar radiation in clear sky conditions. The clear sky profiles P_{cc} and I_{cc} are respectively estimated from the production P and the solar forecasts I thanks to quantile regression introduced below.

Quantile regression uses a piecewise linear asymmetric loss function $\mathrm{QS}_\alpha(x, y)$ called the quantile score (or pinball loss) of level α (Koenker and Hallock, 2001):

$$\mathrm{QS}_\alpha(x, y) = \alpha(y - x)_+ + (1 - \alpha)(x - y)_+\,, \tag{2}$$

where $(\cdot)_+ = \max(\cdot, 0)$. The expectation (over y) of $\mathrm{QS}_\alpha(x, y)$ is minimized if x correctly estimates the quantile of level α of y.

The clear sky profiles P_{cc} and I_{cc} are based on quantile regressions with only trigonometric polynomials of time as inputs, see Lorenz et al. (2009). For P_{cc}, we have

$$P_{cc} = \sum_{j=1}^{4} a_j \sin(j\lambda s) + a'_j \cos(j\lambda s)\,, \tag{3}$$

where λ is the wavenumber corresponding to a year and s is the number of time elapsed since the beginning of the year. The quantile regression determines

D	D + 1	D + 2	D + 3	D + 4	D + 5
PEARP	PEARP	x	x	x	x
Det	Det	Det	Det	x	x
ENS	ENS	ENS	ENS	ENS	ENS

Table 1: Forecast availability with lead time. PEARP is the Météo France ensemble, Det defines the deterministic forecasts Arpège and HRES, and ENS is the ECMWF ensemble.

Label	Nature	Origin	Timestep	Resolution	Base time	Number of forecasts
HRES	deterministic	ECMWF	3 h	0.13°	0 h	1
ARPEGE	deterministic	Météo France	1 h	0.10°	0 h	1
ENS	ensemble	ECMWF	3 h	0.25°	0 h	50
PEARP	ensemble	Météo France	3 h	0.20°	18 h	34

Table 2: Forecast weather data. The indicated resolutions may change for further lead times than those of the present article.

the coefficients a_j and a'_j for a level of quantile close to 90%. A similar procedure is applied to build I_{cc}.

1.2.1 Deterministic PV forecasts

The main idea of the statistical modelling is to use both meteorological and temporal information to provides estimates of the production index τ_P. The model can be seen as a sum of weather-related terms and trigonometric polynomials of time as in Equation 3. The whole model is decomposed into several parts with: (1) a weather-related model, (2) an analysis of the residuals with trigonometric polynomials of time and (3) a multiplicative bias correction.

The first part of the statistical analysis is a linear regression between the production index τ_P and the meteorological variables (the clear sky index τ_I, the total cloud cover Tcc, and the temperature T_{2m}). Non-linear dependencies are taken into account by introducing several terms such as the squared clear sky index τ_I^2 and cross terms between variables $\tau_I(T_{2m} - \overline{T}_{2m})$. The quantity $T_{2m} - \overline{T}_{2m}$ is the deviation of the temperature T_{2m} from its local seasonal average value $\overline{T}_{2m}$. The linear regression estimates the coefficients b_i to produce

$$\widehat{\tau_P} = b_0 + b_1\tau_I + b_2\tau_I^2 + b_3 Tcc + b_4\tau_I(T_{2m} - \overline{T}_{2m})\,. \quad (4)$$

A secondary statistical model is then fitted on the residuals $\widehat{\tau_P} - \tau_P$. The objective of this secondary model is to reduce the seasonal biases and other remaining errors of the first model. We use the elapsed time s and the production forecasts $\widehat{\tau_P}$ as inputs to build

$$\widehat{\widehat{\tau_P}} = c_0\widehat{\tau_P} + \sum_{j=1}^{4} c_j \sin(j\lambda s) + c'_j \cos(j\lambda s)\,. \quad (5)$$

The model parameters of the first two steps are set with the forecasts with lead times less than 24 h. The motivation behind is that the forecasts with short lead times are presumably the most accurate forecasts to fit the main parameters. Still, forecasting long lead times may necessitate slight corrections compared to short lead times, hence we introduce a third step, which takes into account the lead time of the forecasts. The third step is a multiplicative correction γ applied to $\widehat{\widehat{\tau_P}}$, such that $\gamma\widehat{\widehat{\tau_P}}/\tau_P$ is equal to 1 on average. In this third step, both γ and $\widehat{\widehat{\tau_P}}$ are lead time dependent.

The statistical regression scheme is slightly different for ensemble and deterministic forecasts. For ensemble forecasts, the input variable of the linear regression is simply the solar irradiance of the unperturbed member without other weather variables. The same conversion model is used for all the members of a given ensemble.

1.2.2 Quantile PV forecasts

For each deterministic production forecast, we build 19 quantile forecasts (of order 5 to 95) for a total of 38 additional forecasts. They are referred to as deterministic quantiles as opposed to the ensemble members. The idea is to train PV quantile forecasts based on the value of the deterministic PV forecast, see (Nielsen et al., 2006). We follow the idea that we should first precisely estimate the mean of the distribution and only then estimate the quantiles. The quantile regressions are carried out independently for each lead time.

We apply quantile regressions on the residuals of the deterministic forecast obtained at the end of Section 1.2.1. The inputs are the deterministic forecast and trigonometric polynomials of s, similarly to Equation 5.

Concerning France production deterministic quantiles, they are not set to a weighted sum of quantiles of the plants, but they are determined from the deterministic forecast of France production. In other words, the deterministic forecasts of the plants are summed to generate France production forecast, and this forecast is used to generate quantile forecasts for France production.

At this point, the forecaster has a total of $50 + 34 + 2 \times 19 + 2 = 124$ forecasts up to the lead time of 48 hours, 90 forecasts up to the lead time of 96 hours, and 50 forecasts up to the lead time of 138 hours.

1.3 Linear opinion pools

Using a discrete Cumulative Distribution Function (CDF) based on several forecasts allows us to model any CDF without distribution assumption.

Let the x_m be M forecasts (or members) and the u_m be M weights given to the forecasts. The forecaster's CDF

$$\mathrm{G} = \sum_m u_m \mathrm{H}_m \tag{6}$$

is designed as a weighted combination of unit step functions, where $\mathrm{H}_m(x) = \mathrm{H}(x - x_m)$ equals 0 before x_m and 1 otherwise. The mth step of G is centered on x_m and its height equals the weight u_m. The weights

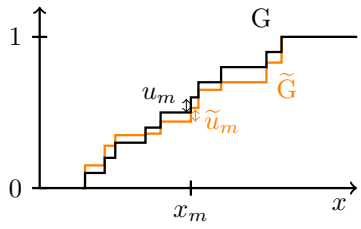

Figure 1: Illustration of weighted CDFs. The CDFs G and $\widetilde{\mathrm{G}}$ are built with the same locations x_m. However the weights u_m and $\widetilde{u}_m$ given to a member are respectively different for G and $\widetilde{\mathrm{G}}$.

u_m are non-negative and sum to one ($\boldsymbol{u} \in \mathcal{P}_M$ the simplex of R^M). This weighted CDF is also known as model mixture or linear opinion pool.

The impact of the weights u_m are illustrated in Figure 1 and 2. Two CDFs G and $\widetilde{\mathrm{G}}$ using the same locations x_m are shown in Figure 1. The CDF G is built with uniform weights $u_m = 1/M$, while the weights $\widetilde{u}_m$ of $\widetilde{\mathrm{G}}$ are not uniform. We show in Figure 2 an illustration of probabilistic forecasts in two different cases: with equal weights for all members and with possibly different weights given by our online learning algorithm. A visual inspection indicates that the online learning algorithm provides a better estimation of the median and a larger spread of the distribution. We emphasize that methods involving weighted empirical distribution functions necessitate that the forecasts x_m are sufficiently dispersed. Since ensemble forecasts are usually under-dispersed, we do not expect such methods to provide satisfactory results on ensemble forecasts from a single ensemble without postprocessing. For our part, we use two ensembles of forecasts and quantile forecasts in order to improve the spread of the weighted forecast.

2 Evaluation

In the following we describe classical diagnostic tools used in Section 4, see for example the monograph of Jolliffe and Stephenson (2012) for further references.

We begin by describing the CRPS as it is at the heart of our learning method.

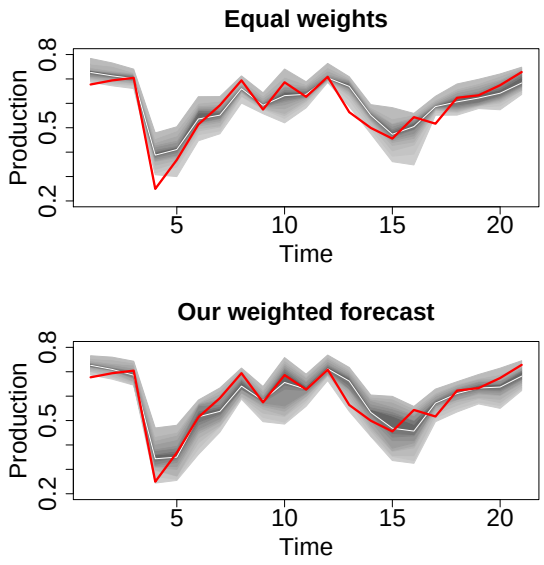

Figure 2: Time series of France production forecasts scaled by the installed capacity (12 hours of lead time, for several consecutive days). Top: equal weights for all members, (b): our forecast with online learning of the weights. Real production is in red and the median of the forecasted distribution is in white.

2.1 The CRPS

The CRPS is a classical scoring function in meteorology (Hersbach, 2000; Candille and Talagrand, 2005). The CRPS is the generalization over all thresholds of the Brier score (Brier, 1950). Let G be the cumulative distribution function of a forecaster describing the i.i.d. random variables X and X$'$, and y be the observation revealed to the forecaster. The CRPS is defined as

$$\mathrm{CRPS}(\mathrm{G}, y) = \int (\mathrm{G} - \mathrm{H}_y)^2 = \mathrm{E}(|\mathrm{X}-y|) - \frac{1}{2}\mathrm{E}(|\mathrm{X}-\mathrm{X}'|)\,, \tag{7}$$

where $\mathrm{H}_y(x) = \mathrm{H}(x-y)$ is the CDF assigned to y, the unit step function H centered on y. The CRPS reduces to the absolute error for deterministic forecasts.

Assuming that y is a random variable, described by the CDF F, the averaged quantity $\mathrm{E}_y(\mathrm{CRPS}(\mathrm{G}, y))$ (on the observation) is minimized only for F = G. This property makes the CRPS a strictly proper scoring rule (Gneiting and Raftery, 2007), and as such it explains why the CRPS is a classical evaluation tool for probabilistic forecasts.

We highlight the fact the CRPS can also be written as a sum of quantile scores (Gneiting and Ranjan, 2011):

$$\mathrm{CRPS}(\mathrm{G}, y) = 2\int_0^1 \mathrm{QS}_\alpha(\mathrm{G}^{-1}(\alpha), y)\mathrm{d}\alpha\,. \tag{8}$$

The strategies of minimizing the CRPS or minimizing several quantile losses are therefore closely related.

For a CDF step function, the corresponding CRPS is computed as:

$$\mathrm{CRPS}\left(\sum_{m=1}^{M} u_m \mathrm{H}_m, y\right) = \sum_{m=1}^{M} u_m |x_m - y| - \frac{1}{2}\sum_{m,k=1}^{M} u_m u_k |x_m - x_k|\,. \tag{9}$$

which is also concisely noted $\ell(\boldsymbol{u})$ in Section 3.

2.2 Other diagnostic tools

The scores are all presented only for the test period, usually averaged over time. Besides the CRPS, we also show results for the celebrated RMSE and MAE for which our forecast is the weighted average $\sum_{m=1}^{M} u_m x_m$. The RMSE of the predictions $\hat{y}$ with respect to the observations y is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}\,, \tag{10}$$

and for the MAE:

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T}|\hat{y}_t - y_t|\,. \tag{11}$$

We use daily scores to show the deterioration of the scores with the increasing lead time. To keep the range of the daily score consistent, the daily score is weighted by the average production of the related hour of the day $\bar{y}_h$. For a score S_h depending on the lead time h, the daily score

$$S^{(d)} = \frac{\sum_h S_h \times \bar{y}_h}{\sum_h \bar{y}_h} \tag{12}$$

is computed with summation over the available lead times h corresponding to the same daily lead time.

Skill scores are useful to compare prediction performance. In this paper, the reference prediction chosen for skill scores is our weighted forecast. Skill scores for a given score S are written

$$S_{pred}^{skill} = \frac{S_{ref} - S_{pred}}{S_{ref}}\,, \tag{13}$$

so that our forecast shows better scores when the skill scores of the other forecasts are negative.

3 Online learning with the CRPS

3.1 Background

3.1.1 Regret bounds

An online learning algorithm determines the weights $u_{m,t}$ using only the available past information. In other words, such algorithm is an update rule indicating the value of the weights $u_{m,t}$ and relying only on the values of the past forecasts and observations $x_{m,t'}$ and $y_{t'}$ with $t' < t$ as described in Section 3.2.

The regret of the algorithm

$$\sum_{t=1}^{T}\ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{P}_M}\sum_{t=1}^{T}\ell_t(\boldsymbol{u}) \tag{14}$$

is an indication of the predictive skill of the weights $u_{m,t}$ for a given set of forecasts and observations. In our case, the notation ℓ_t refers to the CRPS as in Equation 9 with a highlighted dependency on the weight. The regret compares the loss of the algorithm generating the weights $u_{m,t}$ (left term) with the loss of the best combination with weights constant in time (right term). The forecast with best weights constant in time can be known only at the end of the experiment and is called the oracle. By definition, the oracle has a better score than individual forecasts (MAE of each forecast), and than any subset ensemble with uniform weights.

Online learning algorithms come with a theoretical guarantee on the long term performance based on a bound on the regret such as:

$$\sum_{t=1}^{T}\ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{P}_M}\sum_{t=1}^{T}\ell_t(\boldsymbol{u}) \leq o(T)\,. \tag{15}$$

In the sense of the theoretical guarantee 15, the algorithm competes against the best combination with weights constant in time and also with any subset ensemble with uniform weights.

Note that other definitions of regret bounds exist. They may include a supremum taken over all possible values of $x_{m,t}$ and y_t in the left term of Equation 15 so that the regret is maximized. This case considers

the worst scenario of forecasts and observations for the algorithm.

3.1.2 Linearized losses

It is common practice in online learning to use linearized losses, by computing the loss gradients w.r.t. the weights. For the CRPS, the loss gradient $\widetilde{\ell}_{m,t}$ of the mth forecaster can be written as

$$\widetilde{\ell}_{m,t} = \frac{\partial \ell_t}{\partial u_m}(\boldsymbol{u}_t) = |x_{m,t} - y_t| - \sum_{k=1}^{M} u_{k,t}|x_{m,t} - x_{k,t}| + y_t - \sum_{k=1}^{M} u_{k,t} x_{k,t}\,. \tag{16}$$

The last two terms are identical for all forecasters and appear due to terms $1 - \sum_{m=1}^{M} u_m$ hidden in the expression of the CRPS, see Appendix B of Thorey et al. (2016). The loss gradient has two main terms: the distance of $x_{m,t}$ to y_t and the weighted distance of $x_{m,t}$ to the ensemble members. A very good member is therefore close to the observation and far from the other members. A neutral member is equally distant to the observation and to the other members.

The interest of the loss linearization can be seen in terms of regret bounds, as we now detail (see also Devaine et al. (2013)). An algorithm formulated for linear losses $\sum_{m=1}^{M} u_m \widetilde{\ell}_m$ comes with theoretical guarantee against the best member. By applying this algorithm on the gradient losses of any convex differentiable loss, we obtain a theoretical guarantee on the nonlinear loss against the best fixed combination of members.

Indeed, the convexity and the differentiability of ℓ_t gives

$$\ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}) \leq (\boldsymbol{u}_t - \boldsymbol{u})^\top \nabla \boldsymbol{\ell}_t(\boldsymbol{u}_t) = \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \boldsymbol{u}^\top \widetilde{\boldsymbol{\ell}}_t\,. \tag{17}$$

for any two vectors $\boldsymbol{u}_t, \boldsymbol{u} \in \mathcal{P}_M$. Summing over time, we get the following regret bound inequalities:

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{P}} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) = \sup_{\boldsymbol{u}\in\mathcal{P}} \left(\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}) \right) \tag{18}$$

$$\leq \sup_{\boldsymbol{u}\in\mathcal{P}} \left(\sum_{t=1}^{T} \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \boldsymbol{u}^\top \widetilde{\boldsymbol{\ell}}_t \right) \tag{19}$$

$$= \sum_{t=1}^{T} \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \min_{\text{expert k}} \sum_{t=1}^{T} \widetilde{\ell}_{k,t}\,. \tag{20}$$

As a consequence, a theoretical guarantee for the nonlinear losses $\ell_t(\boldsymbol{u})$ is obtained. In other words, knowing a regret bound for expression 20 provides a regret bound for expression 18.

3.2 Example of general algorithm

Initialization: $\boldsymbol{u}_1$;

For each time index t = 1, 2, ..., T

1. get the vector of predictions data $\boldsymbol{x}_t$,
2. compute the forecaster's choice G_t with $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$,
3. get the verification y_t and compute $\boldsymbol{u}_{t+1}$, based on the update rule.

The initial weight vector $\boldsymbol{u}_1$ is arbitrarily set, e.g., to $[1/M, \ldots, 1/M]^\top$.

3.3 ML-Poly

In this article we use a learning algorithm from Gaillard et al. (2014) called ML-Poly for Polynomially weighted averages with multiple learning rates. The algorithm ML-Poly, described in Table 3, has no parameters and is not computationally costly. The algorithm relies on terms $\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t}$ that compare the performance of each member to the performance of the weighted ensemble.

The algorithm gives higher weights to the members performing better than the weighted ensemble with member-dependent learning rates. The regret of the mth member $R_{m,t}$ quantifies the regret for

update the regret of each member	$R_{m,t} = R_{m,t-1} + \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t}$
update the learning rate of each member	$\eta_{m,t} = 1/\left(1 + \sum_{t'=1}^{t} (\boldsymbol{u}_{t'}^\top \widetilde{\boldsymbol{\ell}}_{t'} - \widetilde{\ell}_{m,t'})^2\right)$
compute the weights	$u_{m,t+1} = \eta_{m,t}(R_{m,t})_+ \;/\; \boldsymbol{\eta_t}^\top (\boldsymbol{R}_t)_+$

Table 3: ML-Poly algorithm, at time t after y_t is given. The vectors $\boldsymbol{\eta_t}$ and $\boldsymbol{R}_t$ have M coordinates, respectively $\eta_{m,t}$ and $R_{m,t}$. The functions $(\cdot)_+$ are applied to a vector by applying it component-wise.

not having given higher weights to a forecaster. The learning rate $\eta_{m,t}$ checks whether the mth member's performance is in average close to the performance of the weighted forecast and can be seen as a confidence term. The members with more confidence can see their weights vary more quickly. The ideas behind ML-Poly are on the one hand an adaptation of the algorithm Prod of Cesa-Bianchi et al. (2005) to multiple learning rates, and on the other hand the introduction of the polynomial potential described in Cesa-Bianchi and Lugosi (2003) and giving the terms $(R_{m,t})_+$.

The regret bound of ML-Poly is expressed against the best member for the linearized losses. For all sequences of losses $\widetilde{\ell}_{m,t} \in [0,1]$, the cumulated loss of ML-Poly is bounded:

$$\sum_{t=1}^{T} \boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t \leq \min_{1\leq m\leq M} \left\{ \sum_{t=1}^{T} \widetilde{\ell}_{m,t} + \sqrt{M(1+\ln(1+T))\left(1 + \sum_{t=1}^{T} (\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t})^2\right)} \right\}. \tag{21}$$

As opposed to the bound of Equation 15, the bound of ML-Poly is of second order due to the term $\sum_{t=1}^{T} (\boldsymbol{u}_t^\top \widetilde{\boldsymbol{\ell}}_t - \widetilde{\ell}_{m,t})^2$. The worst case scenario gives a bound $\mathcal{O}(\sqrt{MT \ln T})$, indicating that even in the worst case, the weighted forecast will perform at least as well as the best forecast. In the case of i.i.d. sequences of losses, the regret bound is practically constant. A detailed analysis of second-order bounds can be found in (Gaillard et al., 2014). Besides, other algorithms showing second order bounds are described in Koolen and Van Erven (2015); Luo and Schapire (2015); Wintenberger (2017).

4 Application

4.1 Experiment setup

Local production data may be unfortunately unavailable for given days and plants. In such cases we removed the related data. However we did not modify France production capacity factor to account for local unavailability, because in our opinion, a challenging task for online learning technique is to reduce biases which may be caused by local null production.

The algorithm is run independently for each lead time and production site (including France production). We run the algorithm as if production data is available at the end of each day. For long lead times where several observations arrive between the delivery of a forecast and the reception of the corresponding observation, we use shifted weights in order to use the latest available observation. For example with the shorthand notation $\boldsymbol{u}_{\text{lead time, day}}$ and with a lead time of 36 h, the weights $\boldsymbol{u}_{36\,\text{h},d}$ were delivered at the beginning of $d-1$ to forecast $y_{12:00,d}$. After $y_{12:00,d-1}$ is known, the weights $\boldsymbol{u}_{36\,\text{h},d+1}$ are computed. The key point is that the weight update uses $\boldsymbol{u}_{36\,\text{h},d}$ instead of $\boldsymbol{u}_{36\,\text{h},d-1}$ to check the combination performance against $y_{12:00,d-1}$ to generate $\boldsymbol{u}_{36\,\text{h},d+1}$.

The production forecasts from PEARP and ENS are sorted by rank in order to associate clearly a weight with an ensemble member. As a result, all the members belong to one of the four sorted subensembles, except for the two deterministic forecasts.

We define a climatological reference for diagnostic purposes, called climatology forecast. For time t, we use 2 months of production data centered on t to estimate a so-called climatological mean and 19 quantiles of climatological production. The climatological mean is used for deterministic evaluations

(bias, RMSE, MAE) and the quantiles are used for the CRPS. This method produces a rather 'skilled' reference because the climatology is not only evaluated on the training period but on a rolling period.

We define the raw forecast as the forecast with uniform weights. We use this forecast to assess the gain brought by our online learning algorithm.

The results are shown for PV production forecasts only, and not meteorological variables.

4.2 Results

In this section we only show the results for the individual plants. The results obtained for France production are quite similar to those obtained for the plants and are shown in Appendix A.

4.2.1 Scores and skill scores

First we show the classical scores RMSE, MAE, CRPS and bias in Figure 3 on a daily average, see Equation 12. The confidence intervals indicate the variability of the scores obtained for the plants. The scores are shown for our weighted forecast as well as the raw forecast, the ECMWF deterministic forecast and the climatology forecast. Our weighted forecast gets the best scores up to a lead time of 4 days. Note that our forecast has a quite low bias. For days 5 and 6 the ENS members are the only members available in our study, hence a change of slope in the daily scores. Even for a lead time of 6 days, the climatological forecasts is the worst forecast. Therefore numerical weather predictions may be used to forecast plants productions for a lead time of several days at the 30-min timestep. It is noticeable that our regression scheme minimizes the CRPS and also achieves improvements on the other scores (RMSE, MAE and bias). We tried to identify situations where our algorithm provides a particular improvement over the raw ensemble, but we did not find discriminatory criteria. For example, the installed capacity of the plant or the CRPS of the raw ensemble are not explanatory statistics of the CRPS skill scores of the raw ensemble.

The CRPS skill scores of 5 ensembles (with uniform weights) are shown in Figure 4. The skill scores are

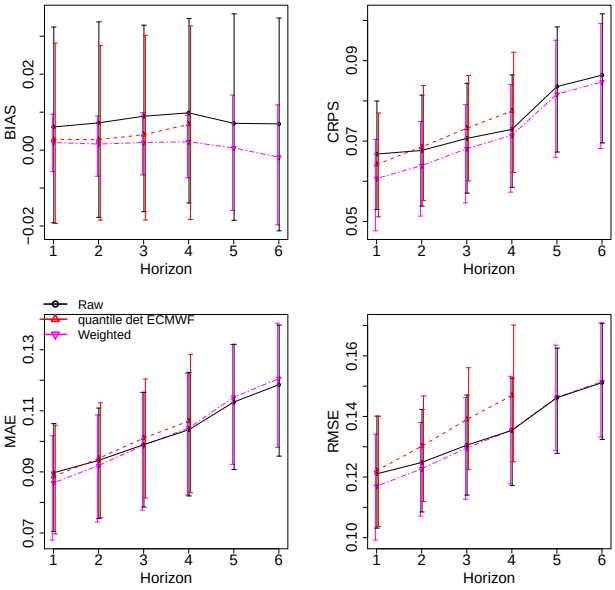

Figure 3: RMSE, MAE, CRPS and bias for the daily scores, for all sites. The results are shown for 3 forecasts: our weighted forecast, the raw forecast (all members with uniform weights), the deterministic forecast of the ECMWF (and its quantiles for the CRPS). The climatology scores are the following : bias $= -0.001$, CRPS $= 0.089$, MAE $= 0.139$, RMSE $= 0.167$. The confidence intervals are derived from the scores of all sites.

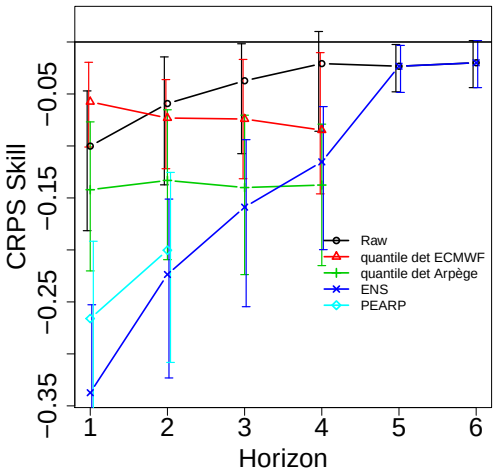

Figure 4: CRPS skill scores for all sites based on Equation 13. Our weighted forecast is the reference prediction for the skill scores. The confidence intervals correspond to the scores of all sites.

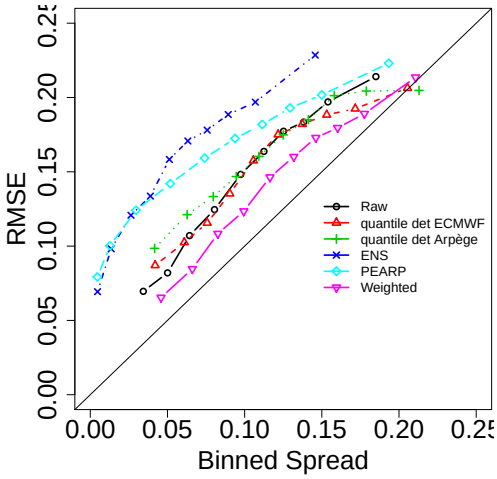

Figure 5: Spread skill diagrams for 36 h of lead times for all sites.

assessed against our weighted forecast. The 5 ensembles are the 4 subensembles of our complete ensemble and the complete ensemble as well. Our weighted forecast performs better than any of the 5 ensembles. The best ensemble with uniform weights is (in average) the complete ensemble. This may be due to the variety of the forecasts in the complete ensemble. Although the quantile ensemble from HRES (quantile det ECMWF) performs well before 24 hours of lead time, it is beaten by the complete ensemble afterwards and its skill decreases with time. The skill of the ECMWF ensemble (ENS) increases notably with time, from the worst skill for day 1 to a satisfactory skill for day 4.

4.2.2 Diagnostic tools

Improvements are also shown for several other diagnostic tools but only for a lead time of 36 h (12:00, D+1) for the sake of brevity. Better results are obtained for shorter lead times and conversely worst results are obtained for longer lead times. By better we mean improvement of our weighted forecast over the raw ensemble.

The spread-skill diagram checks whether the spread of an ensemble (binned into categories) is consistent with the error of the ensemble mean. The squared spread $\sum u_m(x_m - \boldsymbol{u}^\top \boldsymbol{x})^2$ and the square error $(\boldsymbol{u}^\top \boldsymbol{x} - y)^2$ are averaged in each bin and their rooted square are plotted against each other. The spread and the error should be ideally equal (Fortin et al., 2014). On the graph, the curves should match the first diagonal. The spread-skill diagram of the ensembles of our study is shown on Figure 5. We see that our weighted forecasts are closer to the first diagonal than any other subensemble with uniform weights. Our weighted forecasts for plants are still under-dispersive, while the correction is better in the case of France production as shown in Appendix A. The weights provided by the online learning algorithm are larger for the outer members of the ensemble, and especially the lowest members. Consequently the spread of the weighted ensemble is larger than the spread of the raw ensemble and the positive bias of the raw ensemble is mitigated. Besides, the ECMWF ensemble shows the lowest spread and the ensemble PEARP presents very large and very small spreads. However, when the ensemble PEARP shows a small spread, the error is quite larger than the spread.

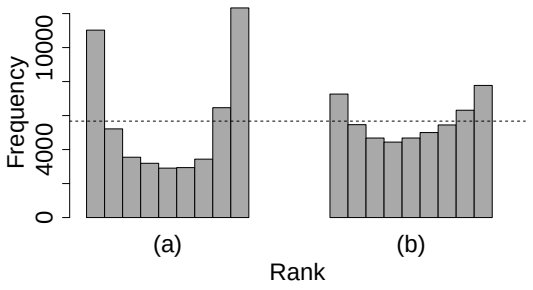

Figure 6: Rank histograms for 36 h of lead times for all sites; (a) the raw ensemble; (b) our weighted forecast. The dotted line illustrates the ideal case of a flat rank histogram.

The rank histogram (Anderson, 1996; Talagrand et al., 1999; Hamill and Colucci, 1997) or Probability Integral Transform (PIT) is built with the values of the CDFs of the forecaster reached by the verifications along an experiment. The ideal rank histogram is flat. The rank histogram of our weighted forecast and the raw ensemble are shown in Figure 6. The rank histogram of our weighted forecast is closer to the ideal rank histogram than the rank diagram of the raw ensemble. The raw ensemble is under-dispersive, since it presents a U-shape. This is consistent with the results shown on Figure 5.

For a given binary event, the reliability diagram checks whether the observed frequency and the forecasted frequency of the event match (Atger, 2004; Bröcker and Smith, 2007a). The forecasted probabilities of the event are binned into categories. The observed frequency of the event for each category is the share of occurence of the event. The ideal reliability diagram shows a curve along the first diagonal. We also show the number of occurences of the event in each binned category. We use the following event "the production level is lower than the average production", where we use the climatological production defined above as local average production. We show the reliability diagrams of our weighted forecast and the raw ensemble in Figure 7. We see that our weighted forecast is very well calibrated for event

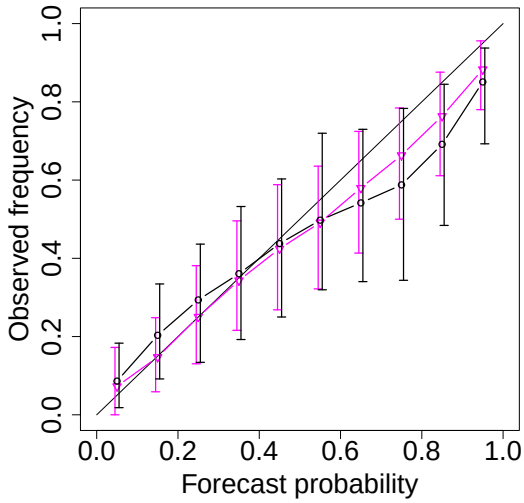

Figure 7: Reliability diagrams for lead times 36 hours for all sites; black circle: the raw ensemble; magenta triangles: our weighted forecast.

with low probability, but tends to overpredict the occurence of the event when the event is highly likely.

Conclusion

We have applied the algorithm ML-Poly for the minimization of the CRPS, in order to provide probabilistic forecasts. The algorithm does not depend on any parameter or assumptions on distributions such as Gaussianity, and comes with theoretical guarantee of performance. The regret bound ensures our forecast to perform at least as well as the best forecast in the ensemble.

Our case study investigated the PV production of several power plants in France and the total production of the plants. We have shown that our weighted forecast improves on the raw ensemble, which is the best ensemble with uniform weights. Interestingly, we show that CRPS minimization brings improvement on classical scores for the ensemble mean and probabilistic diagnostic tools. Indeed, the forecast-

ing capability measured by classical scores (RMSE, MAE, CRPS and bias) are improved by our online learning algorithm up to a lead time of 4 days. Besides, the online learning algorithm provides a spread correction as shown on the spread-skill diagrams and on the rank histograms. The results obtained for France production forecasts and plants forecasts are quite similar.

Future work should investigate the generation of specialized experts on meteorological regimes. For example, an expert specialized in clear sky production could improve the forecasting capability of the ensemble. The quantiles are already specialized, but the ensemble members from ENS and PEARP are converted to production using the same model as for the control member. The investigation of weights prior may also be of interest. The update rule ML-Poly does not use the value of the upcoming forecasts $x_{m,t}$ for computation of the weights $u_{m,t}$, while weights prior may take this additional information into account.

Acknowledgements

The EDF company (Électricité de France) and ANRT (Association nationale de la recherche et de la technologie) are sincerely acknowledged for providing thesis fellowship and supporting this work.

References

Alessandrini, S., Delle Monache, L., Sperati, S., Cervone, G., 2015. An analog ensemble for short-term probabilistic solar power forecast. Applied Energy 157, 95–110.

Almeida, M.P., Perpiñán, O., Narvarte, L., 2015. PV power forecast using a nonparametric PV model. Solar Energy 115, 354 – 368.

Anderson, J.L., 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations 9, 1518–1530.

Atger, F., 2004. Estimation of the reliability of ensemble-based probabilistic forecasts. Quarterly Journal of the Royal Meteorological Society 130, 627–646.

Bacher, P., Madsen, H., Nielsen, H.A., 2009. Online short-term solar power forecasting. Solar Energy 83, 1772–1783.

Baudin, P., 2015. Prévision séquentielle par agrégation d'ensemble: application à des prévisions météorologiques assorties d'incertitudes. Ph.D. thesis. Paris-Sud XI.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability 78, 1–3.

Bröcker, J., Smith, L.A., 2007a. Increasing the reliability of reliability diagrams. Weather and forecasting 22, 651–661.

Bröcker, J., Smith, L.A., 2007b. Scoring probabilistic forecasts: The importance of being proper. Weather and Forecasting 22, 382–388.

Candille, G., Talagrand, O., 2005. Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society 131, 2131–2150.

Cesa-Bianchi, N., Lugosi, G., 2003. Potential-based algorithms in on-line prediction and game theory. Machine Learning 51, 239–261.

Cesa-Bianchi, N., Lugosi, G., 2006. Prediction, learning, and games. Cambridge University Press.

Cesa-Bianchi, N., Mansour, Y., Stoltz, G., 2005. Improved second-order bounds for prediction with expert advice, in: International Conference on Computational Learning Theory, Springer. pp. 217–232.

Courtier, P., Freydier, C., Geleyn, J., Rabier, F., Rochas, M., 1991. The Arpège project at Météo-France, in: ECMWF Seminar Proceedings, pp. 193–231.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., Cébron, P., 2015. PEARP, the Météo-France short-range ensemble prediction system. Quarterly Journal of the Royal Meteorological Society 141, 1671–1685.

Devaine, M., Gaillard, P., Goude, Y., Stoltz, G., 2013. Forecasting electricity consumption by aggregating specialized experts. Machine Learning 90, 231–260.

Fortin, V., Abaza, M., Anctil, F., Turcotte, R., 2014. Why should ensemble spread match the rmse of the ensemble mean? Journal of Hydrometeorology 15, 1708–1713.

Gaillard, P., Goude, Y., Nedellec, R., 2016. Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. International Journal of Forecasting 32, 1038–1050.

Gaillard, P., Stoltz, G., van Erven, T., 2014. A second-order bound with excess losses, in: Proceedings of COLT'14, JMLR: Workshop and Conference Proceedings. pp. 176–196.

Genest, C., McConway, K.J., 1990. Allocating the weights in the linear opinion pool. Journal of Forecasting 9, 53–73. URL: http://dx.doi.org/10.1002/for.3980090106, doi:10.1002/for.3980090106.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. Journal of Econometrics 164, 130–141.

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annual Review of Statistics and Its Application 1, 125–151.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378.

Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. Monthly Weather Review 133, 1098–1118.

Gneiting, T., Ranjan, R., 2011. Comparing density forecasts using threshold-and quantile-weighted scoring rules. Journal of Business & Economic Statistics 29.

Hamill, T.M., Colucci, S.J., 1997. Verification of Eta/RSM short-range ensemble forecasts 125, 1312–1327.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting 15, 559–570.

Huang, J., Perry, M., 2015. A semi-empirical approach using gradient boosting and -nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. International Journal of Forecasting , –.

Inman, R.H., Pedro, H.T., Coimbra, C.F., 2013. Solar forecasting methods for renewable energy integration. Progress in Energy and Combustion Science 39, 535–576.

Jolliffe, I., Stephenson, D., 2012. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley.

Koenker, R., Hallock, K., 2001. Quantile regression: An introduction. Journal of Economic Perspectives 15, 43–56.

Koolen, W.M., Van Erven, T., 2015. Second-order quantile methods for experts and combinatorial games., in: COLT, pp. 1155–1175.

Lorenz, E., Hurka, J., Heinemann, D., Beyer, H.G., 2009. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of 2, 2–10.

Luo, H., Schapire, R.E., 2015. Achieving all with no parameters: Adanormalhedge, in: Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015, pp. 1286–1304.

Mallet, V., 2010. Ensemble forecast of analyses: Coupling data assimilation and sequential aggregation. J. Geophys. Res. 115.

Mallet, V., Stoltz, G., Mauricette, B., 2009. Ozone ensemble forecast with machine learning algorithms. J. Geophys. Res. 114.

Nielsen, H.A., Madsen, H., Nielsen, T.S., 2006. Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. Wind Energy 9, 95–108.

Palmer, T., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., Steinheimer, M., Weisheimer, A., 2009. Stochastic parametrization and model uncertainty. European Centre for Medium-Range Weather Forecasts.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles 133, 1,155–1,174.

Ren, Y., Suganthan, P., Srikanth, N., 2015. Ensemble methods for wind and solar power forecasting—a state-of-the-art review. Renewable and Sustainable Energy Reviews 50, 82 – 91.

Sloughter, J.M., Gneiting, T., Raftery, A.E., 2010. Probabilistic wind speed forecasting using ensembles and bayesian model averaging. Journal of the American Statistical Association 105, 25–35.

Sloughter, J.M.L., Raftery, A.E., Gneiting, T., Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using bayesian model averaging. Monthly Weather Review 135, 3209–3220.

Sperati, S., Alessandrini, S., Pinson, P., Kariniotakis, G., 2015. The "weather intelligence for renewable energies" benchmarking exercise on short-term forecasting of wind and solar power generation. Energies 8, 9594–9619.

Stoltz, G., 2010. Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique 151, 66–106.

Talagrand, O., Vautard, R., Strauss, B., 1999. Evaluation of probabilistic prediction system. Proceedings of the ECMWF Workshop on Predictability.

Thorarinsdottir, T.L., Gneiting, T., 2010. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. Journal of the Royal Statistical Society: Series A (Statistics in Society) 173, 371–388.

Thorey, J., Mallet, V., Baudin, P., 2016. Online learning with the CRPS for ensemble forecasting. Quarterly Journal of the Royal Meteorological Society .

Thorey, J., Mallet, V., Chaussin, C., Descamps, L., Blanc, P., 2015. Ensemble forecast of solar radiation using tigge weather forecasts and helioclim database. Solar Energy 120, 232–243.

Wilks, D.S., 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. Meteorological Applications 16, 361–368.

Wintenberger, O., 2017. Optimal learning with bernstein online aggregation. Machine Learning 106, 119–141.

Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., 2014. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. part ii: Probabilistic forecast of daily production. Solar Energy 105, 804–816.

Appendix A Results for France production

In this Appendix, we show the results for France production, while the results for the individual sites are shown in Section 4. The results of France production forecasts and plants forecasts are roughly similar. Our online learning algorithm provides improvements over the raw ensemble up to a lead time of a few days. Because it is easier to forecast the power output of the total production, the forecast quality is better than for individual sites. This statement is verified for all diagnostic tools shown below.

We show in Figure 8 the average bias, CRPS, MAE and RMSE for France production. We see the scores

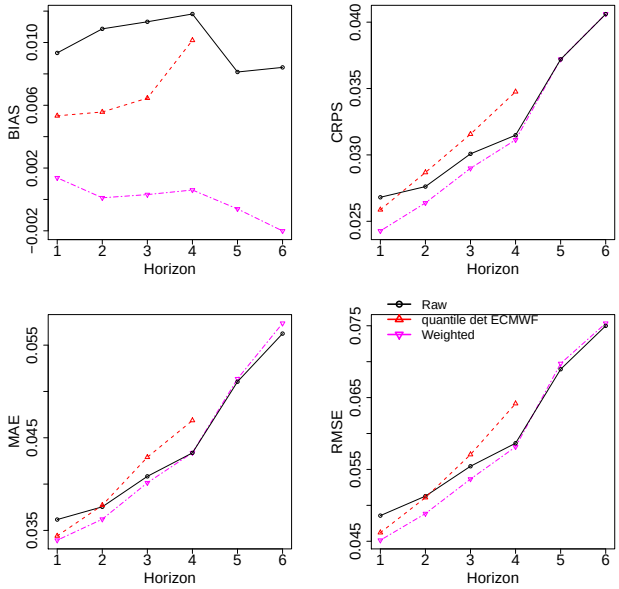

Figure 8: RMSE, MAE, CRPS and bias for the daily scores of France production. The results are shown for 3 forecasts: our weighted forecast, the raw forecast (all members with uniform weights), the deterministic forecast of the ECMWF (and its quantiles for the CRPS). The climatology scores are the following : bias $= -0.001$, CRPS $= 0.055$, MAE $= 0.081$, RMSE $= 0.101$.

of the sites are more than twice as large as the score of France production, but for the bias. Our online learning algorithm provides improvement for bias, CRPS, MAE and RMSE up to a lead time of 4 days.

The CRPS skill scores are shown in Figure 9. Once again, the score trends are mostly equivalent to those obtained for the sites. The quantile ensemble from HRES (quantile det ECMWF) has good scores for short lead times and the raw ensemble is the best ensemble with uniform weights after 24 h of lead time. Our online learning algorithm provides an improvement of roughly 10% over the raw ensemble for the first 24 h of lead time. This improvement decreases with time quickly than for the sites. It is remarkable that the CRPS skill score of the quantile en-

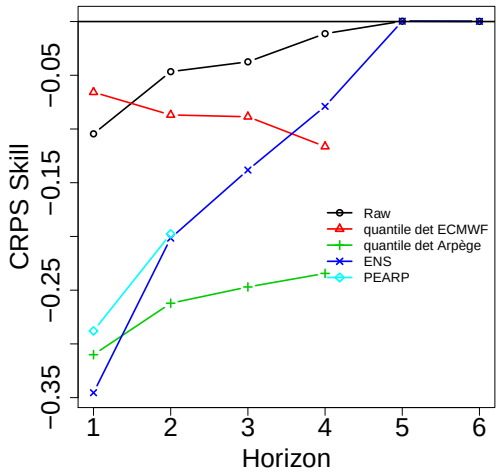

Figure 9: CRPS Skill score.

semble from Arpège ("quantile det Arpège") shows much better results for the plants than for France production. Indeed the skill score of "quantile det Arpège" is around -15% for the plants and is stable, while it is at least below -24% for France production. For days 5 and 6, the weights brought by our algorithm do not vary much from the uniform distribution. Consequently, the skill scores are close to one.

The following probabilistic diagnostic tools are only for 09:00, 12:00, and 15:00 of day 2 (lead times 33, 36, 39 hours). In Figure 10, we compare the rank histogram for the raw ensemble and our weighted forecast. The raw ensemble is largely underdispersive with a positive bias (over-estimation). Our online learning algorithm manages to reduce the under-dispersion of the raw ensemble. This statement is verified on the spread skill diagram in Figure 11. We see that the spread and errors of our weighted forecasts match approximately, while the other ensembles (with uniform weights) are underdispersive with respect to their errors.

A correction of the forecast reliability is illustrated in Figure 12. The event "the production level is lower than the average production" is used (same as for the sites). A visual inspection shows that the raw ensemble tends to underpredict the occurence of low pro-

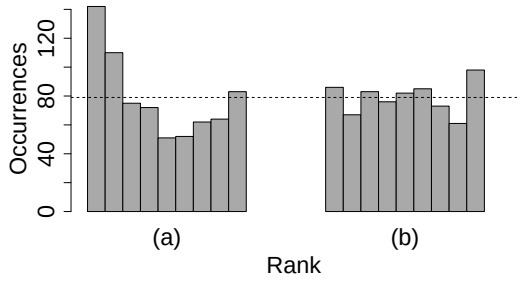

Figure 10: Rank histograms for lead times 33, 36, 39 hours; (a) the raw ensemble; (b) our weighted forecast. The dotted line illustrates the ideal case of a flat rank histogram.

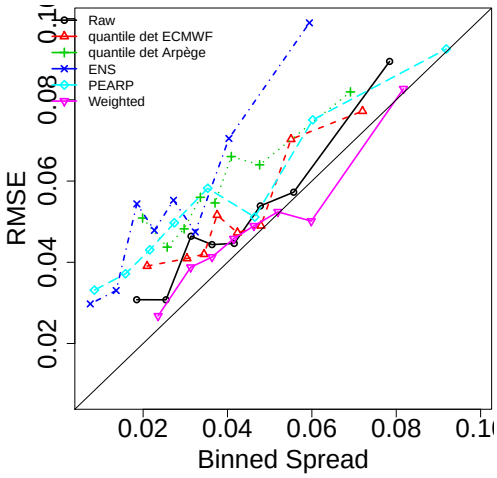

Figure 11: Spread skill for lead times 33, 36, 39 hours.

duction for a forecasted frequency between 0.3 and 0.7, when the event is likely to occur. Our weighted forecast does not show this tendency and is symmetrical with respect to the first diagonal although not perfectly aligned.

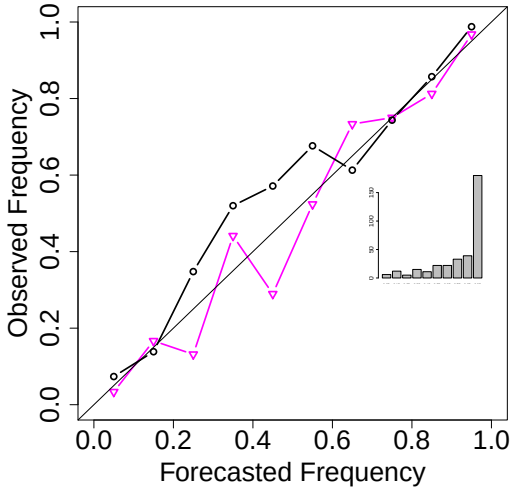

Figure 12: Reliability diagrams for lead times 33, 36, 39 hours; black round: the raw ensemble; magenta triangles: our weighted forecast.