Available online at www.sciencedirect.com

ScienceDirect

Solar Energy xxx (2014) xxx–xxx

SOLAR ENERGY

www.elsevier.com/locate/solener

A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production

M. Zamo [a,1], O. Mestre [a,*], P. Arbogast [b], O. Pannekoucke [b]

[a] *Météo-France, Direction de la Production, 42 av. Coriolis, 31057 Toulouse cedex, France*
[b] *Météo France, CNRM/GMAP/RECYF, 42 av. Coriolis, 31057 Toulouse cedex, France*

Received 7 October 2013; received in revised form 14 March 2014; accepted 17 March 2014

Communicated by: Associate Editor David Renne

Abstract

This pair of articles presents the results of a study about forecasting photovoltaic (PV) electricity production for some power plants in mainland France. Forecasts are built with statistical methods exploiting outputs from numerical weather prediction (NWP) models. Contrary to most other studies, forecasts are built without using technical information on the power plants. In each article, several statistical methods are used to build forecast models and their performance is compared by means of adequate scores. When a best forecast emerges, its characteristics are then further assessed in order to get a deeper insight of its merits and flaws. The robustness of the results are evaluated with an intense use of cross-validation.

This article deals with probabilistic forecasts of daily production 2 days ahead. By "probabilistic" we mean that our forecast models yield some quantiles of the expected production's probability distribution. Whereas probabilistic forecasts are quite usual in wind power forecasting, they are more unusual in the field of PV production forecasting. We show that our eight forecasts always perform better than a simple climatological reference forecast. Nevertheless, no forecast model clearly dominates the other, whatever the studied power plant. A post-processing technique aiming at improving forecasts may in fact decrease their performance. Finally, we use outputs from an ensemble NWP model in order to add information about the meteorological forecast uncertainty. Sometimes this ensemble improves the performance of forecasts, but this is not always true.

Keywords: Photovoltaic power forecasting; Benchmark; Statistical methods; Numerical weather prediction; Quantile regression; Ensemble numerical weather prediction

1. Introduction

In Zamo et al. (in press), we described how we forecast hourly electricity photovoltaic (PV) production one day ahead to answer needs of electricity grid managers, energy traders and producers. The forecast consisted in the mean hourly PV production at each day-time hour at some power plants in two French counties.

* Corresponding author. Permanent address: Météo France, DP/DPrévi/COMPAS, 42 av. Coriolis, 31057 Toulouse cedex 01, France. Tel.: +33 5 61 07 86 33.

E-mail addresses: michael.zamo@meteo.fr (M. Zamo), olivier.mestre@meteo.fr (O. Mestre), philippe.arbogast@meteo.fr (P. Arbogast), olivier.pannekoucke@meteo.fr (O. Pannekoucke).

[1] Principal corresponding author. Tel.: +33 5 61 07 86 36.

http://dx.doi.org/10.1016/j.solener.2014.03.026
0038-092X/

This current article deals with a farther lead time and a different purpose. We seek to forecast daily electricity PV production with an anticipation of two days. The purpose is to provide electricity producers with an accurate forecast to manage maintenance operations of their facilities in a cost-effective way. Indeed a skillful forecast should allow producers to efficiently plan those operations only when electricity production is expected to be low, in order to reduce the financial loss caused by stopping production during maintenance.

Since maintenance operations are likely to last the entire day, we forecast the expected daily production at chosen PV power plants. Furthermore, since meteorological predictions tend to be less precise with increasing lead times, the production forecast takes a probabilistic form. By "probabilistic", we mean the forecast is composed of several quantiles of the forecast production's probability distribution. One final element of our forecasting strategy is to use predictors from Météo France's ensemble numerical weather prediction (NWP) system, PEARP. PEARP is an ensemble NWP system composed of a set of 35 forecasts computed from different atmospheric initial states and with different physics. Ensemble forecasts aim at evaluating the forecast uncertainty originating from several error sources, such as initial conditions or uncertainty from unresolved physics in the case of PEARP. By taking predictors from PEARP, our goal is to assess the usefulness of incorporating such informations on meteorological uncertainty in daily production forecasts.

As stated in Zamo et al. (in press), since Statistics offers many methods to our purpose, we compare several statistical methods with appropriate scores and graphs. Two quantile regression methods are used (linear model in quantile regression and quantile regression forest) optionally taking into account PEARP information. Finally, we correct each forecast with a post-processing technique or let the forecast uncorrected. This makes a total of 8 different forecasting strategies to compare. A climatological forecast is also built, as a reference forecast. Lastly, three sets of quantiles are forecast, with a different number of quantiles. This aims at assessing the influence of the number of forecast quantiles on the predictive performance. During this benchmark, a systematic use of cross-validation is done in order to evaluate the robustness of our results.

The article proceeds as follow. Section 2 describes our main goal and the data. Section 3 details our modeling strategy. Relevant scores and assessment tools for probabilistic forecast are also described here. Section 4 presents the results of the comparative test. We conclude in Section 5 with a reminder of those main results along with some perspectives of possible improvements.

2. Purpose of the study and available data

We aim at building a forecast model of the daily production of photovoltaic electricity 2 days ahead at some power plants in mainland France. Most studies in the field of PV production forecasting rely on technical information about PV panels and/or modeling of direct and diffuse solar irradiation. This approach is presented in Heimo et al. (2012) and used in Bracale et al. (2013) for a probabilistic PV production forecast. Since we have no access to such information and models and in order to explore different solutions, we choose a direct modeling of the production with statistical methods. That is, we try to learn the statistical link between some carefully chosen meteorological predictors and the observed electricity production from a production archive. We briefly describe hereafter these two sets of data. This approach may be interesting when power plants are poorly documented and the usual approach cannot thus be employed. On the other hand, this requires a sufficient amount of past measured productions.

2.1. PV production data

The original PV production data already used in Zamo et al. (in press) are at a 10 min temporal resolution, for 28 power plants in two French counties covering about 7000 km^2 each. We will follow the same convention as in the first article by naming those two counties CountyA and CountyB. Individual power plants for CountyX are named CountyXxx with xx ranging from 1 to 18 for CountyA and from 1 to 10 for CountyB.

The forecasts we compare are numerous and require a lot of computation time. Furthermore the robustness of our results is assessed through cross-validation, which requires much more time. Thus only 5 power plants in each county are used for this part of the study. We give at the end of Section 3.1 some figures about the required computation times. The chosen power plants are evenly spread over each county. For CountyB, one power plant is specifically chosen because of its location by the seaside. For the quite hilly CountyA, some power plants are located on the reliefs, others in valleys.

For our purpose, the 10-min PV production were summed up on a daily basis between 6 and 18 UTC. This lapse roughly corresponds to day-time and working-time in mainland France. It is also compatible with the availability of predictors from PEARP (see the following subsection).

Further information about the pre-processing applied to the PV production data and the two counties can be found in Zamo et al. (in press).

2.2. Predictors

As stated above, most of our predictors come from Météo France's ensemble NWP system, named PEARP (Prévision d'Ensemble ARPege). An ensemble forecast is a set of several forecasts (or "members") built in such a way as to represent part of the uncertainty in meteorological predictions. Leutbecher and Palmer (2008) gives a good introduction to ensemble forecasting. PEARP is a

Please cite this article in press as: Zamo, M., et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Sol. Energy (2014), http://dx.doi.org/10.1016/j.solener.2014.03.026

global ensemble NWP system of 35 members computed with Météo France's deterministic model ARPEGE at a lower resolution. Two main uncertainty sources are accounted for in ensemble forecasting: uncertainty in initial conditions and model errors. In PEARP, members are run from different initial conditions in order to represent uncertainties in the analysis of the initial state of the atmosphere. Also, ten physical packages are randomly associated to members to describe the model uncertainties.

A set of 34 perturbations of initial conditions around the unperturbed member (or control member) are built to obtain 34 perturbed members. Those perturbations are based on an ensemble data assimilation with 6 members as described in Berre et al. (2007), combined with perturbations based on singular vectors computed over different areas. Details about the singular vector perturbation method can be found in Leutbecher and Palmer (2008). In a nutshell, it consists in adding to the initial state of the model a perturbation in the direction of state space along which initial errors increase the most according to a linear approximation of the NWP numerical model.

PEARP handles model uncertainty through a multiphysics approach. Indeed, it uses a set of ten coherent combinations of parametrization of unresolved physical processes at PEARP's spatial resolution. Those packages vary in their diffusion scheme, shallow convection scheme, deep convection scheme closure conditions and surface-atmosphere exchanges. Each of these ten physical packages is randomly associated to perturbed members of PEARP, with the constraint of using each package an equal number of times. More details about these physical packages can be found in Descamps et al. (2011). Also Descamps et al. (2011) and Zadra et al. (2013) contain more information about PEARP.

PEARP's global grid is stretched with an horizontal spectral truncation of T538 and a stretching factor of 2.4. This results in an horizontal spatial resolution varying from about 15 km over mainland France up to 86 km over New Zealand. Vertically, PEARP extends up to 50 km above ground level, with 65 levels. It is run twice a day at 6UTC and 18UTC with a farthest lead time of respectively 72 and 108 h. Its output frequency is 3 h for lead-times lower than 54 h, and 6 h afterward.

In this study, we use the run of 18UTC and lead times of 66 and 72 h, for which the time step is 6 h. Hence, for our purpose PEARP produces information between 6 and 18UTC for day $D+3$. The chosen predictors are bilinearly interpolated from PEARP's 4 nearest grid points to the location of each power plant. The 6-hourly predictors from PEARP are summed up or averaged to get daily predictors. Among the many outputs from PEARP, the choice of predictors is guided by the physics of PV panels. Indeed, their production is sensitive to global and diffuse irradiation and temperature, and other parameters influencing these last two parameters, such as wind speed and cloud cover. An additional constraint is to minimize the correlation between predictors, since correlated predictors may decrease the performance of some statistical regression methods. In Table 1 the eight chosen predictors are listed. The maximum sun height, i.e. the maximum daily solar elevation angle, is chosen as a proxy for the day in the year.

3. Benchmarking of several quantile regression forecast models

We detail here our benchmarking strategy to forecast quantiles of daily PV production and choose the best forecast methods. Four main points are detailed. First, the principle of the statistical methods used to forecast quantiles are introduced. Second, we specify how exactly they are used in conjunction with the 35 members of the PEARP to build production forecasts. Third, the specific tools allowing to compare probabilistic forecasts are described. Among these tools, the rank histogram can be used to post-process the statistical forecast methods in order to try and improve them. We describe how. Finally, since we have too few data, we are faced with the problem of correctly cross-validating our results. How we proceed is explained at the end of this section.

3.1. Statistical quantile regression methods

Many statistical methods exist to model the probability distribution of one predictand knowing some predictors. Here we proceed with quantile regression (QR) methods. These methods are non-parametric: they do not require any assumption about the shape of the underlying distribution. On the other hand, they provide information in the form of a discrete (albeit very large) subset of quantiles. If we note $Q_\tau(Y)$ the quantile of order $\tau \in [0;1]$ of some real random variable Y, then the probability of Y to be lower than $Q_\tau(Y)$ is exactly τ: $\mathbb{P}(Y \leqslant Q_\tau(Y)) = \tau$. Another useful definition of $Q_\tau(Y)$ in terms of $F(Y)$, the cumulative distribution function (cdf) of Y, is $Q_\tau(Y) = \inf(y \in \mathbb{R} | F(y) \geqslant \tau)$. When $Q_\tau(Y)$ is considered as a function of τ, it is called the quantile function, and is simply the inverse of the cdf $F(Y)$.

In order to get an estimate of the impact of the number of estimated quantiles on the performance of our forecast, we model three subsets of all the possible quantiles. These three subsets will be called "coarse resolution subset", "medium resolution subset" and "fine resolution subset". They contain respectively the quantiles with the regularly-spaced orders from 0.10 to 0.90 with a constant step $\Delta\tau = 0.10$, from 0.05 to 0.95 with a constant step $\Delta\tau = 0.05$ and from 0.01 to 0.99 with a constant step $\Delta\tau = 0.01$.

Amongst the many quantile regression methods, some rely on neural networks as in Cannon (2011), binary trees as in Chaudhuri and Loh (2002) or an analog to support vector machines as in Takeuchi et al. (2006). Here we use and compare the linear model in quantile regression (LMQR) presented in Koenker and Hallock (2001) and Cade and Noon (2003), and the quantile regression forest (QRF) developed by Meinshausen (2006). Their descriptions follow.

Please cite this article in press as: Zamo, M., et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Sol. Energy (2014), http://dx.doi.org/10.1016/j.solener.2014.03.026

Table 1
Table explaining the abbreviations used for the predictors' name used to forecast daily PV production.

Abbreviation	Explanation
DWSFlow	Total downward solar irradiation flow for short wavelengths, summed up between 6 and 18UTC
ThermFl	Total irradiation in the infra-red wavelengths, summed up between 6 and 18UTC
WSGrd	Horizontal wind speed, 10 m above the ground level, averaged between 6 and 18UTC
T2M	Air temperature 2 m above the ground level, averaged between 6 and 18UTC
HU2M	Air relative humidity 2 m above the ground level, averaged between 6 and 18UTC
TotCover	Total cloud cover at the vertical of a given location, averaged between 6 and 18UTC
SLP	Sea level pressure, averaged between 6 and 18UTC
SunHeight	Maximum solar elevation angle for the given day and location

3.1.1. Linear model for quantile regression

Koenker and Bassett (1978) proposed a method to estimate the quantiles of the cdf of some predictand Y under a linear modelization with a vector of predictors X. This linear model for quantile regression (LMQR), or simply quantile regression in the literature, is built as follow.

A linear relationship is supposed between the random variable Y (here daily PV production) and the vector X of p predictors: $Y = < \beta, X > + \epsilon$ where ϵ is a random error term, β is a vector of p coefficients we want to optimize and $< \cdot, \cdot >$ is the dot product. While in classical regression the loss function is the square error, in quantile regression the loss function is the dissymmetric absolute error also called check function: $\rho_\tau(u) = u \cdot (\tau - \mathbb{1}_{u<0})$, where $\tau \in (0;1)$ and $\mathbb{1}_{u<0}$ is the indicator function of the sign of $u \in \mathbb{R}$. The estimator $\hat{\beta}$ of β minimizes the sum of the loss function on the data: $\hat{\beta} = argmin_\beta \sum_{i=1}^{N} \rho_\tau(y_i - < \beta, x_i >)$, where (x_i, y_i) are the N pairs of vector of predictors x_i and corresponding observed predictand y_i. The quantity $\widehat{Y} = < \hat{\beta}, X >$ is actually an estimator of the τ-order conditional quantile of Y knowing the vector of predictors X, noted $Q_\tau(Y|X)$. This estimator will be noted $\widehat{Q}_\tau(Y|X)$.

Linear programming algorithms, such as presented in Koenker and d'Orey (1987), minimize the loss function on the data, for each order τ separately. Because of this separated estimation of β for each order τ, nothing constrains the estimator of the quantile function to be increasing with τ as it should be. In mathematical notation, we can have $\widehat{Q}_{\tau_1}(Y|X) > \widehat{Q}_{\tau_2}(Y|X)$ for $\tau_1 < \tau_2$. This problem, called "quantile crossing", can be solved by rearranging the raw estimated quantiles $\widehat{Q}_\tau(Y|X)$, as explained in Chernozhukov et al. (2010). The idea of this rearrangement method is the following. Take an estimate $\hat{f} : [0;1] \mapsto \mathbb{R}$ of some monotonic increasing function $f : [0;1] \mapsto \mathbb{R}$. This estimate is not monotonic. First compute $F_{\hat{f}}(z) = \int_0^1 \mathbb{1}_{\hat{f}(u) \leqslant z} du$ for $z \in \mathbb{R}$. Since $\mathbb{1}_{\hat{f}(u) \leqslant z}$ is positive whatever z and u, $F_{\hat{f}}$ is necessarily increasing in z. Actually, $F_{\hat{f}}$ can be interpreted as the cdf of the random variable $\hat{f}(U)$ where U follows the uniform distribution on $[0;1]$. The rearranged estimator $\hat{f}^\star(\tau)$ is then just the quantile function associated with $F_{\hat{f}}$, that is: $\hat{f}^\star(\tau) = inf\{y \in \mathbb{R} | F_{\hat{f}}(y) \geqslant \tau\}$ for $\tau \in [0;1]$, which is increasing with τ. Chernozhukov et al. (2009) shows it is also a better estimator of f than $\hat{f}$. Here $f(\tau)$ is obviously the quantile function $Q_\tau(Y|X)$ and $\widehat{Q}_\tau(Y|X)$ is its not-necessarily monotonic estimator $\hat{f}(\tau)$ obtained with LMQR. Hence our forecast for the LMQR method is the rearranged estimator $\widehat{Q}^\star_\tau(Y|X)$.

Since LMQR is a linear model, it is not bounded. Thus, negative forecast quantile are not unlikely, which is meaningless for a positive quantity such as a daily electricity production. We handle this problem by replacing every negative quantiles with zero, before rearranging the forecast.

3.1.2. Quantile regression forest

Quantile regression forest (QRF) was proposed by Meinshausen (2006) as an adaptation of random forest to quantile regression. Built exactly in the same way as random forests described in Zamo et al. (in press), a quantile regression forest is a set of binary regression trees. But for each final leaf of each tree, one does not compute the mean of the predictand's values, but instead their empirical cdf. Once the random forest is built, one determines for a new vector of predictors its associated leaf in each tree by following the binary splitting according to the predictors' values. Then the final forecast is the cdf computed by averaging the cdf from all the trees, noted $F(Y|X)$. From this averaged cdf it is straightforward to compute an estimator of whatever quantile is needed, with the definition $\widehat{Q}_\tau(Y|X) = inf\{y \in \mathbb{R} | F(y|X) \geqslant \tau\}$, where $\tau \in [0;1]$ is the desired order of the quantile.

By construction, quantile regression forest is free from the quantile crossing problem. Indeed, each tree yields a cdf that is valid, i.e. increasing (albeit not *strictly* increasing) with the quantile order τ. Averaging individual cdfs over all the trees results in a cdf that is still increasing with τ. The final cdf is also bounded between the lowest and highest value of the learning sample. It is then not possible to forecast a negative quantile for PV production. On the other hand, QRF is unable to forecast a quantile higher than the maximum measured production in the training sample.

3.2. Exploiting PEARP's information

Now we know how to estimate quantiles from a vector of predictors, we must decide how to combine the 35 members of the PEARP, that is 35 vectors of predictors. In

PEARP it is easy to follow the control, i.e. "unperturbed", member from one run to another whereas it is not for the 34 perturbed members whose physics are randomly chosen from one run to another. Consequently, we actually build two forecasts with each quantile regression method.

First each QR method is trained on the control member only, to build a "control model". With this control model and the control member's predictors only, we produce a forecast called the "control forecast". Therefore this forecast does not exploit the information in the entire PEARP ensemble but only the control member. We thus have a control LMQR forecast and a control QRF forecast.

The second forecast we build is called the "averaged forecast" and is obtained as follow from all the members of the PEARP. From the control model trained previously on the control member, we forecast one set of quantiles for each of the 35 PEARP's members. With each of these 35 quantile forecasts an empirical cdf is computed. Those 35 cdfs are averaged, thus assuming the PEARP members are equally likely. Starting from this averaged distribution we compute back our set of quantiles to get our second final forecast, the "averaged forecast". We thus have an averaged LMQR forecast and an averaged QRF forecast. By comparing the control forecast and the averaged forecast, we aim at evaluating the usefulness of using the ensemble NWP model instead of one NWP model only.

3.3. Tools for probabilistic forecasts comparison

Quantifying the performances of probabilistic forecasts requires specific tools we introduce now.

3.3.1. Continuous ranked probability score

The most common score for evaluating the global accuracy of some probabilistic forecast is the continuous ranked probability score (CRPS) presented in Hersbach (2000) for example.

For some forecast cdf F and its corresponding observation y, this score is defined as $crps(F,y) = \int_{-\infty}^{\infty} (F(x) - H(x-y))^2 dx$, where the Heaviside function $H(x)$ is equal to 0 for strictly negative x and 1 for positive x. In this definition, $H(x-y)$ is the empirical cdf of the observation y and so $crps(F,y)$ is a quadratic measure of the distance between the forecast cdf and the observed cdf. This score is negatively oriented, that is the lower the better. The mean of $crps(F,y)$ over several forecast/observation pairs (F,y) will be noted CRPS, in upper case.

As shown in Hersbach (2000), CRPS, the mean of crps over several forecast/observation pairs, can be decomposed in a reliability and a potential CRPS terms. This decomposition is much alike the decomposition of the better-known Brier score, introduced in Murphy (1973). For both terms, the lower is the better. The reliability term is a measure of the statistical consistency of the forecast distribution with the observed distribution. When a specific cdf is forecast, if the associated observations follow the forecast cdf, the forecasting system is said to be reliable. The potential CRPS is the CRPS of a perfectly reliable model. It includes the intrinsic variability of the predictand (and hence the difficulty to predict it) and the so-called resolution of the forecasting system. The resolution of a forecasting system is its ability to issue very different forecasts for different corresponding observations. Hersbach's decomposition requires the hypothesis that members in the ensemble NWP model are equiprobable. This is generally not true for a forecast in the form of quantiles. Nevertheless, since our estimated orders were regularly spaced purposedly, the resulting forecasts can be assimilated to an ensemble with equivalent members. Hersbach's decomposition is therefore relevant.

3.3.2. Rank histogram and calibration

A graphical tool to assess the reliability of some probabilistic forecast in the form of an ensemble forecast is the rank histogram, designed simultaneously but separately by several researchers such as Anderson (1996), Hamill and Colucci (1996) and Hamill (1997) or Harrison et al. (1995) and Talagrand et al. (1997). The basic idea is to rank each observation among the corresponding forecast members. Then we draw the histogram of these ranks computed on several pairs of forecast and observation. If the forecast is reliable, observations are drawn from the same distribution as the members, hence observations should have equal probability to get every possible rank. This translates graphically in a flat rank histogram, up to variations due to a limited sample. Even though our forecast quantiles are not strictly speaking an ensemble, their orders are regularly spaced. Hence if our forecasts are reliable, the observations should be evenly distributed among the intervals between each pair of successive quantiles. One drawback of the rank histogram is that flatness is a necessary but not sufficient condition for the reliability of the underlying forecasts. Indeed, a reliable forecast gives a flat histogram but a flat histogram does not necessarily indicates a reliable forecast, as discussed in Hamill (2001).

The rank histogram can be used as a means to try and improve the reliability of a forecast. This so-called "calibration technique" is described in Hamill and Colucci (1998) (page 713, starting from the last paragraph of the left column). Briefly, it consists in estimating and correcting the complete forecast cdf with information from the rank histogram built on the training sample, from the forecast quantiles $\widehat{Q}_\tau(Y|X)$ and with some linear interpolation. This aims at correcting the whole forecast distribution (its mean, spread, every moment). Fig. 1 illustrates this method on simple synthetic data. Here we want to calibrate the three forecast quartiles $\widehat{Q}_{\frac{1}{4}}(Y|X) = 35$, $\widehat{Q}_{\frac{2}{4}}(Y|X) = 60$ and $\widehat{Q}_{\frac{3}{4}}(Y|X) = 90$, in percent of the maximum observed production. The method proceeds as follow. First, build the rank histogram on the *training* sample. This gives an estimate of the proportion $\hat{p}_\tau$ of observations below each corresponding quantile $\widehat{Q}_\tau(Y|X)$. Ideally, for a perfectly reliable forecast $\hat{p}_\tau = \tau$ for every chosen τ. This is usually not true, as in our example where $\hat{p}_{\frac{1}{4}} = 0.55, \hat{p}_{\frac{2}{4}} = 0.75$ and $\hat{p}_{\frac{3}{4}} = 0.90$. Then for one forecast in the *test* sample,

Please cite this article in press as: Zamo, M., et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Sol. Energy (2014), http://dx.doi.org/10.1016/j.solener.2014.03.026

we associate to each forecast quantile $\widehat{Q}_\tau(Y|X)$ the associated proportion $\hat{p}_\tau$ computed on the training sample. The pairs $(\widehat{Q}_\tau(Y|X), \hat{p}_\tau)$ for all the quantile orders give some points of the forecast cdf (empty circles in Fig. 1). Then we linearly interpolate this estimated cdf between those points (the dashed lines in Fig. 1). Outside those available points, some assumption is required. Since our predictand is bounded, we add the points $(0, 0)$ and $(Q_1, 1)$ to our estimated cdf, where Q_1 is the maximum of forecast and observed productions in the training sample. Q_1 can be considered as an estimate of the maximum possible production (in the example, $Q_1 = 110$). Then we complete the estimated cdf with further linear interpolation thanks to those new points (dotted lines in Fig. 1). From this estimated cdf, we can finally compute the corrected forecast quantiles $\widehat{Q}^c_\tau(Y|X)$ (in the example, $\widehat{Q}^c_{\frac{1}{4}}(Y|X) = 16$, quite different from the uncorrected 35). Using this procedure we correct each forecast model with this method. Thus we will have "uncorrected" and "corrected" forecast quantiles.

3.3.3. Twofold cross-validation

In order to assess the robustness of our results, we have recourse to k-fold cross-validation, a method we explained in Zamo et al. (in press). However, our data contain only 544 dates for each power plant. Since the finest estimated quantile set contains 99 quantiles and we want to have enough potential observations between successive quantiles, we cannot choose a too high number k of training samples. Actually we choose $k = 2$, that is our data are randomly split into two samples of 272 dates. Each sample serves alternatively as a training sample and a test sample. In order to get more than 2 estimates for each statistics, we randomly repeat this 2-fold cross-validation 10 times. As a consequence each score or graph is estimated 20 times. Getting more estimates in such a way is not optimal because our estimates are of course correlated and thus do not represent exactly the true dispersion of our scores or graphs. But, to the best of our knowledge, no technique can completely get rid of this correlation problem while keeping enough data. Bootstrapping the data may be an alternative but does not allow to control the test sample size.

To sum it up, our benchmark compares 8 forecast models designated by the following abbreviations: u.c.LMQR, u.a.LMQR, u.c.QRF, u.a.QRF, c.c.LMQR, c.a.LMQR, c.c.QRF, c.a.QRF. Those abbreviations must be read as follow: we first indicate whether the forecast is corrected with the rank histogram calibration technique or left uncorrected ("c" or "u"), then whether it is a control or averaged forecast ("c" or "a") and finally which QR method is used ("LMQR" or "QRF"). For example, "c.a.QRF" is the corrected averaged forecast built from the quantile regression forest. Those eight forecasts will collectively be named "QR-based forecasts" when required. Indeed, a ninth forecast model is built, as a reference. It consists in a monthly climatology computed from the training sample. That is, for each date in the test sample, the climatological forecast is the quantile sets computed with the measured PV production in the training data from dates with the same month (not necessarily the same year). In other words, for each month, the climatological forecast is built on about 22 or 23 measurements of PV production in the training sample. This forecast is simply assigned the name "clim". Those 9 forecasts are built and compared by repeating 10 times a 2-fold cross-validation.

We use the free statistical software R available at R Development Core Team (2011) and associated packages implementing the aforementioned methods. The code was partly parallelized and ran on an operational server with 24 processing units (6 real 2.8 GHz quad-core CPUs) and a RAM of 64 GB. For one power plant and one training sample, computations require about 2 min for each QR-based methods. With adequate parallelization, this is small enough for training tens or hundreds of power plants with one training sample, as required for operational purposes.

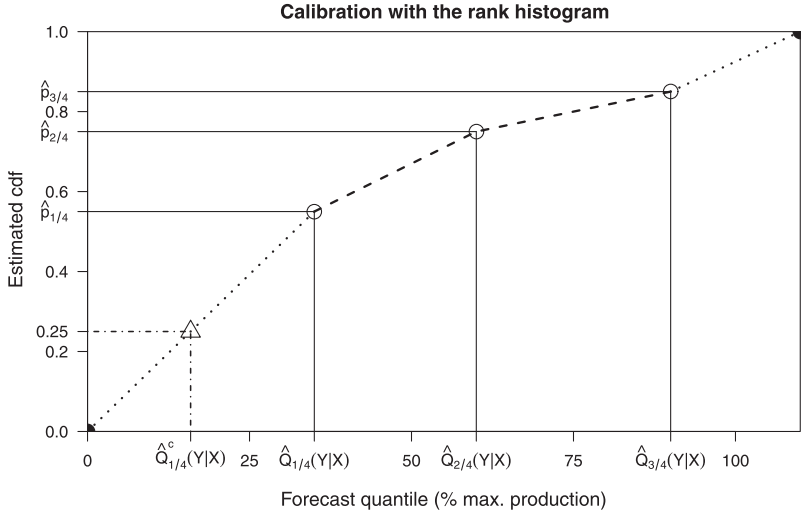

Fig. 1. Calibration with the rank histogram. Illustration with a simple synthetic example.

Please cite this article in press as: Zamo, M., et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Sol. Energy (2014), http://dx.doi.org/10.1016/j.solener.2014.03.026

But in a study, cross-validation is recommended and many forecast approaches have to be tried. Thus, with ten power plants, 20 cross-validated training samples and eight QR-based forecasts as in our study, the total computation time quickly rises to 48 h. This is without taking into account computations on the test samples, 48 h more. Consequently, in an exploratory perspective, choices have to be made in order to balance comprehensiveness of the study and acceptable computation times. We chose only ten power plants and did not study many quantile regression methods.

4. Results of the benchmark

In this section we first present the conclusions of the benchmark on the coarse resolution subset of quantile orders. Then we study the variability of those results with the resolution in the forecast quantile orders.

4.1. Coarse resolution subset of quantile orders

Figs. 2 shows the boxplots of the CRPS, the potential CRPS and the reliability term of the nine forecasts at power plant CountyB05, for the coarse resolution subset of quantile orders. Those boxplots are computed from the 20 estimates of each statistics obtained with ten 2-fold cross-validations.

According to the CRPS boxplots (top of Fig. 2), the climatological forecast is always the worst with a CRPS nearly 50% higher than the other forecasts. This is statistically significant, since the climatological CRPS boxplot does not overlap the CRPS boxplots for the QR-based forecasts. The improvement of performance compared to the climatological forecast comes from reduced potential CRPS and (with one exception, c.c.QRF explained later) reliability terms as shown in the middle and bottom parts of Fig. 2. Among the 8 QR-based forecasts, none is always better than the others, but c.c.QRF gets worse performances.

The other power plants get similar results albeit with slightly different figures. For CountyA, climatological forecasts' CRPS is about 42–72% higher than QR-based forecasts' CRPS depending on the power plant. For CountyB, climatological forecasts' CRPS is about 33–50% higher than QR-based forecasts' CRPS depending on the power plant. This means that the improvement of QR-based forecasts compared to the climatology is much more important for power plants in CountyA than in CountyB. This probably results from the more variable weather in CountyA, where using climatology as a forecast is a bad idea, than in CountyB, where the less fluctuating weather makes a climatological forecast not too bad. The median CRPS stays always between 6 and 8% of the maximum observed daily production, depending on the power plant.

Boxplots in Fig. 2 do not show the relative ordering of the CRPS for the nine forecasts and each test sample separately. Since they partially overlaps each other for QR-based forecasts, we cannot conclude whether one of these 8 methods is systematically better than the others. Table 2 allows to see this for the coarse resolution order set. It contains for each power plant the percent of test samples for which each method gets the lowest CRPS (that is, the best performance). The first result is that the climatological forecast is never the best, in agreement with the conclusion based on the CRPS boxplots. Knowing this, if the 8 QR-based forecasts performed equally, each should be the best about 12.5% of the test samples for each power plant, up to sampling errors. This does not seem to be true. As an example, c.c.QRF is never the best forecast, whatever the power plant. And for some power plants, one QR-based forecast clearly dominates the other, such as u.c.QRF for power plant CountyA02 (80% the best) or u.a.QRF for CountyB08 (65% the best). Nevertheless, no QR-based forecast dominates the other for all or most of the power plants. For this coarse resolution order set, a corrected QR-based forecast is the best for CountyA10 and CountyA11, only two power plants against eight for the uncorrected QR-based forecasts. Thus calibration does not allow to build best forecasts with so few quantiles, that is so few information on the underlying cdf. With the figures in Table 2, we can test the null hypothesis that every QR-based forecast is equally likely to be the best, vs the alternative hypothesis that at least one QR-based forecast has a probability higher than 12.5% to perform the best. A chi-square test is applied. The significance level is the usual 0.05. Table 3 shows the p-values of this test for each power plant and each order resolution. It clearly shows that for the coarse resolution order set, for most power plants we can statistically claim that at least one power plant performs better than others, in terms of CRPS. Only power plants CountyB03 and CountyB07 get a p-value near (but not very different from) the significance level. For these two power plants, one QR-based forecast can be claimed better than the others, but with less certainty. As an example, for CountyB07, u.a.LMQR, u.a.QRF and c.a.QRF are obvious candidates as best forecast according to their ranking in Table 2.

As for the reliability term, c.c.QRF is the only QR-based forecast that is less reliable than the climatological forecast, as shown in the bottom graph of Fig. 2. This worst reliability of c.c.QRF compared to clim is present whatever the power plant. From this figure, we conclude that calibration does not improve and sometimes worsens the reliability of a corrected QR-based forecast compared to the corresponding uncorrected QR-based forecast (u.c.QRF vs c.c.QRF for example). This is true whatever the power plant, for this coarse resolution subset. How can the calibration technique with the rank histogram decrease the reliability instead of improving it as we might expect? Two sources of errors can be introduced by this technique. First, the linear interpolation between and outside available quantiles may be too gross an hypothesis if not enough quantiles are known. Second, dissimilarities between the rank histograms for the training and the test

Please cite this article in press as: Zamo, M., et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Sol. Energy (2014), http://dx.doi.org/10.1016/j.solener.2014.03.026

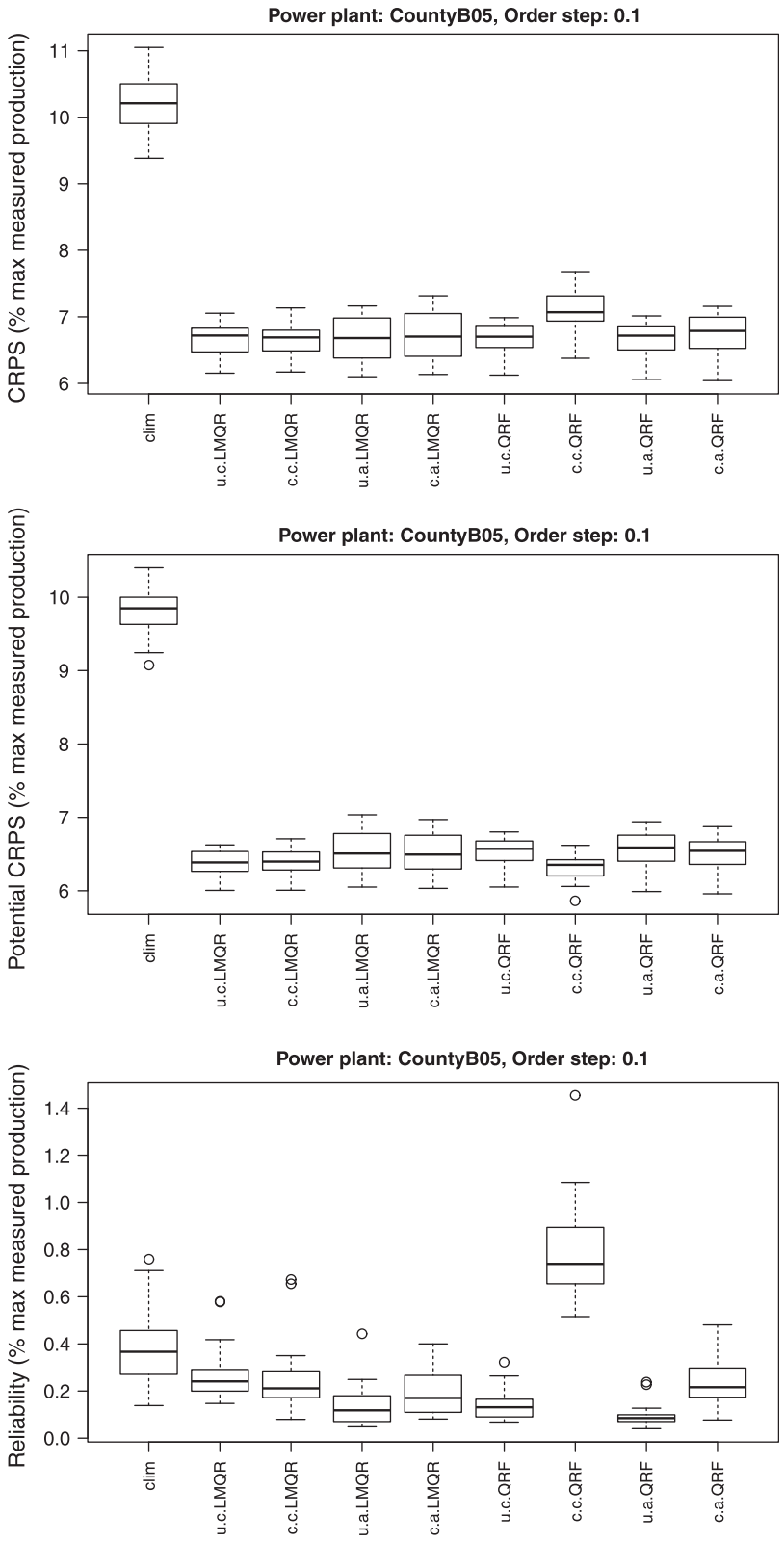

Fig. 2. CRPS and its decomposition (potential CRPS and reliability term) of all forecast methods, for power plant CountyB05 and the coarse resolution order set. Scores are given in percent of the maximum observed production. Boxplots are obtained with 10 repetitions of 2-fold cross-validation.

sample can result in an inappropriate correction. This last claim requires a comparison of the rank histograms from the training and test samples of the uncorrected forecasts and the rank histogram for the test sample of the corrected forecasts. They are drawn in Fig. 3, for CountyB05 and the control QRF forecasts. This example exhibits the most

Table 2
Percent of the times when each forecast gets the best CRPS out of the 20 cross-validation test samples, for each power plant separately (columnwise sums are 100). Bold figures designate the best forecast method(s) for the corresponding power plant. For the sake of readability, when a forecast is never the best for a specific power plant, a dot is printed instead of zero. Forecasts are for the coarse resolution order set ($\tau = 0.1, 0.2, \ldots$).

	Power plant									
	CountyA					CountyB				
	02	06	10	11	16	03	05	06	07	08
clim	–	–	–	–	–	–	–	–	–	–
u.c.LMQR	15	5	5	25	**35**	10	5	–	5	–
c.c.LMQR	5	10	**35**	**30**	15	10	10	5	–	–
u.a.LMQR	–	5	20	20	5	20	15	–	**30**	–
c.a.LMQR	–	–	10	10	–	–	5	–	10	5
u.c.QRF	**80**	**50**	25	15	25	15	**40**	5	15	5
c.c.QRF	–	–	–	–	–	–	–	–	–	–
u.a.QRF	–	30	–	–	20	**30**	10	**50**	20	**65**
c.a.QRF	–	–	5	–	–	15	15	40	20	25

striking features. The rank histogram for u.c.QRF computed on the training sample is n-shaped and not flat at all. This means that the forecast quantiles in the training samples tend to be too widespread. Therefore the calibration technique will make forecast quantiles in the test sample closer, thus narrowing each forecast probability density function. But as can be seen in the middle part of Fig. 3, the flatness of the rank histogram for u.c.QRF on the test sample is very satisfactory. Indeed, nearly every interval between successive quantiles contains roughly the expected proportion of observations, that is 0.1. Thus, the calibration technique will make the corrected forecast quantiles too close from one another. This results in the u-shaped rank histogram of c.c.QRF (bottom part of Fig. 3)): too many observations fall in the most extreme quantile intervals and not enough in the center of the distribution. That is, the corrected forecasts tends to be under-dispersive. This narrowing of the forecast probability density function after correction is illustrated in Fig. 4 for a unique forecast. The same holds true for every forecast in the test samples, but only for the forecasts based on QRF. For LMQR-based forecasts, training and test samples exhibit much more similar rank histograms before correction. Fig. 5 shows boxplots of the rank histograms for u.c.LMQR at CountyB05. It is indeed very flat and the expected frequency of observations in each order interval stands inside each boxplot. This implies we can suppose the forecasts are reliable. This holds true for every LMQR-based forecast, whatever the power plant.

Table 3
p-Values of the chi-square test to test whether all QR-based forecasts perform equally vs whether at least one QR-based method performs better than the others. The test has been applied to each power plant for each resolution order set. Bold figures indicate *p*-values near or higher than the 0.05-significance level.

Power plant	Quantile order set		
	Fine	Medium	Coarse
CountyA02	4e−07	2e−17	5e−18
CountyA06	9e−06	3e−04	6e−07
CountyA10	3e−04	6e−03	3e−03
CountyA11	8e−08	5e−04	9e−03
CountyA16	6e−03	**2e−02**	2e−03
CountyB03	6e−03	**1e−01**	**6e−02**
CountyB05	2e−03	**4e−02**	6e−03
CountyB06	5e−18	7e−12	5e−09
CountyB07	**4e−02**	**4e−02**	**3e−02**
CountyB08	6e−13	1e−06	1e−11

4.2. Variability of performance with the order resolution

Do the results of the previous subsection also hold for the two other resolution order subsets? Even when the order resolution is increased, the climatological forecast is still always beaten in terms of CRPS, whatever the power station (not shown here). This results from a reduced potential CRPS and a reduced reliability term (except for c.c.QRF). For the 8 QR-based forecasts, some boxplots of CRPS generally overlap for many power plants, whatever the order resolution. For most of the power stations, this prevents from determining at first sight from those boxplots a best forecast. But, as can be seen from Tables 4 and 5 showing the percent of test samples for which each forecast gets the lowest CRPS, there are three unchanged results whatever the order resolution. First, the u.c.QRF forecast stays the best for CountyA02 and CountyA06 whatever the order resolution. Second, c.c.QRF is never the best model. Third, no QR-based forecast dominates all the others whatever the power plant for each resolution order subset. Furthermore the best forecast for one power plant can differ for different order resolutions. Those best performances are statistically significant at the 0.05 significance level with a chi-square test, as deduced from the *p*-values in Table 3. This implies that the choice of the best model strongly depends on the specific power plant and the specific subset of quantile orders.

There is an important change in best models with the order resolution. Whereas for the coarse resolution subset, corrected QR-based forecasts perform the best only for two power plants out of ten, they get the upper hand 5 times out of 11 (with one tie) for the medium or fine resolution subsets (see Tables 2, 4 and 5). This comes from an improved reliability after calibrating with enough quantile orders, as shown in Fig. 6. In this figure, a calibrated QR-based forecast gets an equivalent or lower reliability term than the associated uncalibrated QR-based forecast, except for c.c.QRF. This last exception comes from the same differences in rank histograms between the training and test samples as explained previously for the coarse resolution subset. For the other QR-based forecasts, the improved reliability with more estimated quantiles is explained by the fact that the linear interpolation used in calibrating with the rank histogram becomes less gross when more quantiles are available. The cdf is then indeed estimated with more points.

Please cite this article in press as: Zamo, M., et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Sol. Energy (2014), http://dx.doi.org/10.1016/j.solener.2014.03.026

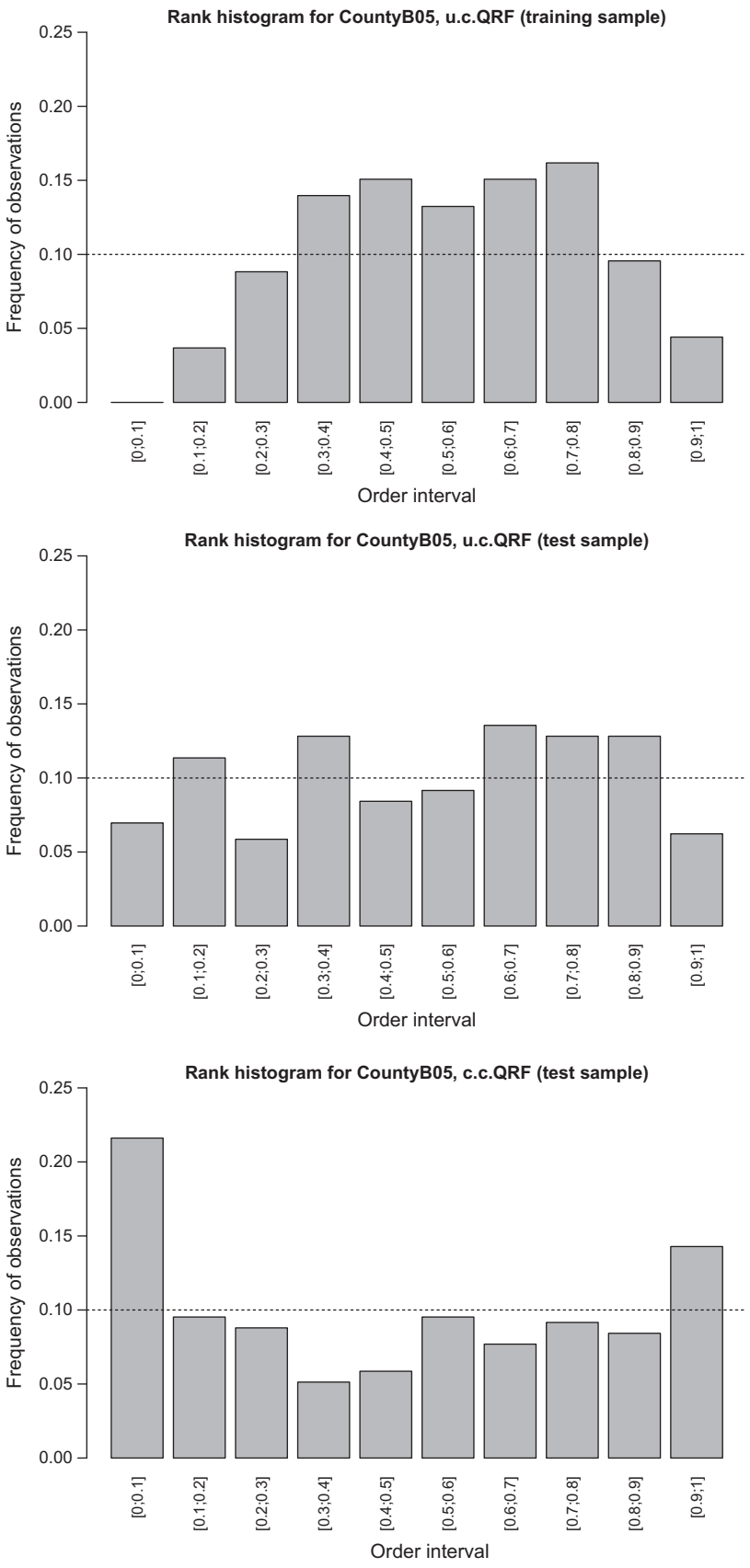

Fig. 3. Rank histograms of observations among forecast with QRF, for power plant CountyB05 and the coarse resolution order set. Upper rank histogram is for uncalibrated control forecast with QRF (u.c.QRF), for one training sample. Middle rank histogram is the same but for the associated test sample. Lower rank histogram is for the calibrated control forecast with QRF (c.c.QRF), for the same test sample. The horizontal dashed line is the expected frequency of observations for perfectly reliable forecasts.

Please cite this article in press as: Zamo, M., et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Sol. Energy (2014), http://dx.doi.org/10.1016/j.solener.2014.03.026

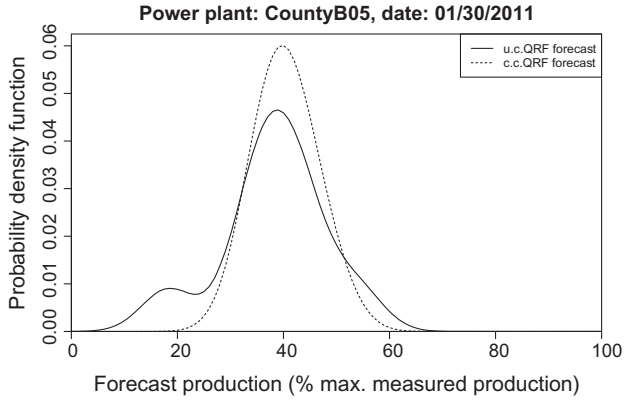

Fig. 4. Forecast probability density function (pdf) of daily electricity production at CountyB05, with two methods and for one day. Production are in percent of the maximum measured production. Pdfs have been smoothed.

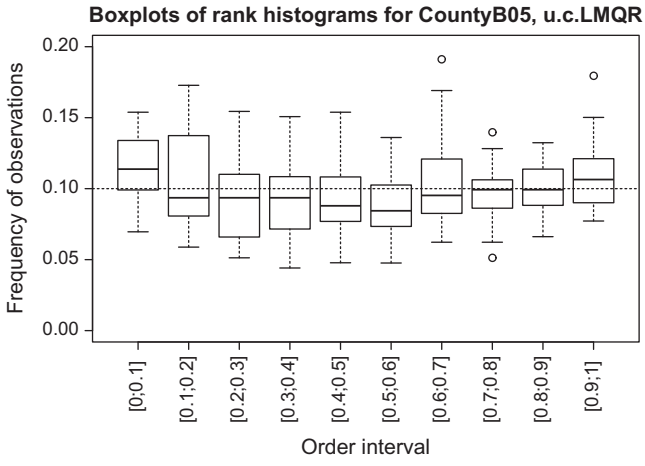

Fig. 5. Boxplots of rank histograms of observations among u.c.LMQR forecast, for power plant CountyB05 and the coarse resolution order set. Boxplots were computed with 2-fold cross-validation repeated ten times. The horizontal dashed line is the expected frequency of observations for perfectly reliable forecasts.

5. Summary, conclusion and perspectives

We built probabilistic forecast models for the daily PV production with an anticipation of 66 h at ten power plants in two counties of mainland France. Eight models were built using two quantile regression methods, by exploiting ensemble model PEARP in two ways and by post-processing or not processing each forecast with a calibration technique based on the rank histogram. A ninth forecast based on a monthly climatology of measured electricity productions was used as a reference forecast. The forecast took the form of a set of quantiles. Three sets of regularly spaced quantile orders were forecast in order to estimate the variability of performance with the resolution in quantile orders.

Table 4
Same legend as for Table 2 but for the medium resolution order set ($\tau = 0.05, 0.1, \ldots$).

	Power plant									
	CountyA					CountyB				
	02	06	10	11	16	03	05	06	07	08
clim	–	–	–	–	–	–	–	–	–	–
u.c.LMQR	5	10	**30**	**40**	20	**25**	15	10	10	–
c.c.LMQR	5	20	**30**	20	10	15	**30**	–	5	–
u.a.LMQR	–	–	15	20	–	10	10	–	10	–
c.a.LMQR	5	–	5	5	10	5	10	–	**30**	15
u.c.QRF	**80**	**40**	15	15	**35**	5	25	25	25	25
c.c.QRF	–	–	–	–	–	–	–	–	–	–
u.a.QRF	–	25	–	–	10	20	5	–	15	10
c.a.QRF	5	5	5	–	15	20	5	**65**	5	**50**

Table 5
Same legend as for Table 2 but for the fine resolution order set ($\tau = 0.01, 0.02, \ldots$).

	Power plant									
	CountyA					CountyB				
	02	06	10	11	16	03	05	06	07	08
clim	–	–	–	–	–	–	–	–	–	–
u.c.LMQR	15	10	**40**	**55**	20	10	**35**	5	10	–
c.c.LMQR	5	10	25	5	15	10	25	–	20	–
u.a.LMQR	–	–	–	5	–	5	–	–	15	–
c.a.LMQR	–	–	20	25	15	15	10	–	**25**	15
u.c.QRF	**55**	**50**	10	10	**35**	15	20	15	5	10
c.c.QRF	–	–	–	–	–	–	–	–	–	–
u.a.QRF	10	10	–	–	–	5	–	–	–	5
c.a.QRF	15	20	5	–	15	**40**	10	**80**	**25**	**70**

As expected, QR-based forecasts perform significantly better than the climatology, with a CRPS lowered by 25–50%. For most power plants, a QR-based forecast performs better than the others. But the most accurate forecast may vary from one power plant to another and with the number of forecast quantiles. For some power plants, no better forecast dominates clearly. It is thus not clear whether employing the full PEARP instead of the only control member improves significantly the predictive performance.

The calibration technique with the rank histogram performs in varied ways. First it requires enough orders of estimated quantiles in order to improve the reliability of one forecast. This is due to the linear interpolation used, approximation that may be too gross with a small number of available quantiles. But a high number of quantiles is not sufficient. Indeed, the corrected version of the control QRF forecast is never more reliable than the uncorrected forecast. This is due to large dissimilarities between the rank histograms computed on the training and test samples.

This study may be completed in several ways: first by including in our benchmark other quantile regression methods such as neural networks presented in Cannon

Please cite this article in press as: Zamo, M., et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Sol. Energy (2014), http://dx.doi.org/10.1016/j.solener.2014.03.026

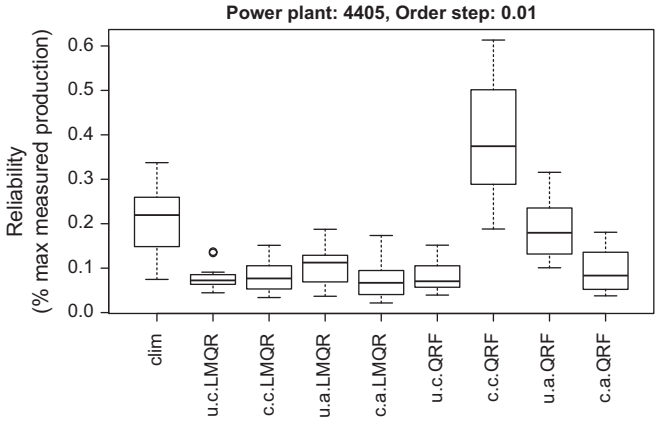

Fig. 6. Reliability term of all forecast methods, for power plant CountyB05 and the fine resolution order set. The score is given in percent of the maximum observed production. Boxplots are obtained after 10 repetitions of 2-fold cross-validation.

(2011) and White (1992) or binary trees as in Chaudhuri and Loh (2002). It could be possible to train a QR-based forecast for each of the ten physics used in PEARP and to apply the relevant model to PEARP's members with the associated physics. Nevertheless, computation times are long and this training would be time-consuming.

In order to improve the efficiency of the calibration technique, we could apply the rank histogram technique conditionally to some parameters. For example, a different rank histogram could be computed for each season and the forecasts of each season should be calibrated with the rank histogram of the same season. Other conditioning parameters may be tried, such as the weather regime (anticyclonic, perturbed,...) or intervals of statistics of the ensemble forecast itself (mean, spread and so on). However this would require much more data. Other calibration techniques may also be evaluated, such as Bayesian model averaging, presented in Raftery et al. (2005).

Finally, the usefulness of the PEARP may be evaluated for farther look-ahead times when ensemble prevision systems are usually expected to be more informative than a simple deterministic model.

Acknowledgements

We would like to thank two anonymous reviewers for their extensive reading of the first draft of this pair of articles. Their remarks helped to greatly improve the quality and clarity of these papers.

References

Anderson, J.L., 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. J. Climate 9 (7), 1518–1530.

Berre, L., Pannekoucke, O., Desroziers, G., Stefanescu, S., Chapnik, B., Raynaud, L., 2007. A variational assimilation ensemble and the spatial filtering of its error covariances: increase of sample size by local spatial averaging. In: Proc. ECMWF Workshop on Flow-Dependent Aspects of Data Assimilation. pp. 151–168.

Bracale, A., Caramia, P., Carpinelli, G., Di Fazio, A.R., Ferruzzi, G., 2013. A Bayesian method for short-term probabilistic forecasting of photovoltaic generation in smart grid operation and control. Energies 6 (2), 733–747.

Cade, B., Noon, B., 2003. A gentle introduction to quantile regression for ecologists. Front. Ecol. Environ. 1 (8), 412–420.

Cannon, A.J., 2011. Quantile regression neural networks: implementation in R and application to precipitation downscaling. Comput. Geosci. 37 (9), 1277–1284.

Chaudhuri, P., Loh, W., 2002. Nonparametric estimation of conditional quantiles using quantile regression trees. Bernoulli 8 (5), 561–576.

Chernozhukov, V., Fernandez-Val, I., Galichon, A., 2009. Improving point and interval estimators of monotone functions by rearrangement. Biometrika 96 (3), 559–575.

Chernozhukov, V., Fernández-Val, I., Galichon, A., 2010. Quantile and probability curves without crossing. Econometrica 78 (3), 1093–1125.

Descamps, L., Labadie, C., Bazile, E., 2011. Representing model uncertainty using the multiparametrization method. In: Proceedings of ECMWF Workshop on Representing Model Uncertainty and Error in Numerical Weather and Climate Prediction Models, 20–24 June 2011. pp. 175–182.

Hamill, T.M., 1997. Reliability diagrams for multicategory probabilistic forecasts. Weather Forecast. 12 (4), 736–741.

Hamill, T., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Mon. Weather Rev. 129 (3), 550–560.

Hamill, T.M., Colucci, S.J., 1996. Random and systematic error in NMC's short-range Eta ensembles. In: Preprints, 13th Conf. on Probability and Statistics in the Atmospheric Sciences, San Francisco, CA, Amer. Meteor. Soc. pp. 51–56.

Hamill, T., Colucci, S., 1998. Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. Mon. Weather Rev. 126 (3), 711–724.

Harrison, M.S.J., Richardson, D.S., Robertson, K., Woodcock, A., 1995. Medium-Range Ensembles using Both the ECMWF T63 and Unified Models – An Initial Report. UK Meteorological Office Tech. Rep. 153, 25.

Heimo, A., et al., August 2012. COST Action ES1002 Weather Intelligence for Renewable Energies (WIRE), Current State Report. COST Action ES1002, Current State Report.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast. 15 (5), 559–570.

Koenker, R., Bassett Jr., G., 1978. Regression quantiles. Econometrica: J. Econo. Soc., 33–50.

Koenker, R., d'Orey, V., 1987. Algorithm AS 229: computing regression quantiles. J. Roy. Stat. Soc. Ser. C (Appl. Stat.) 36 (3), 383–393.

Koenker, R., Hallock, K., 2001. Quantile regression. J. Econ. Perspect. 15 (4), 143–156.

Leutbecher, M., Palmer, T.N., 2008. Ensemble forecasting. J. Comput. Phys. 227 (7), 3515–3539.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Murphy, A., 1973. A new vector partition of the probability score. J. Appl. Meteorol. 12, 595–600.

Raftery, A., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133 (5), 1155–1174.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. <http://www.R-project.org/>.

Takeuchi, I., Le, Q., Sears, T., Smola, A., 2006. Nonparametric quantile estimation. J. Mach. Learn. Res. 7, 1231–1264.

Talagrand, O., Vautard, R., Strauss, B., 1997. Evaluation of probabilistic prediction systems. In: Proc. ECMWF Workshop on Predictability. pp. 1–25.

White, H., 1992. Nonparametric estimation of conditional quantiles using neural networks. Computing Science and Statistics. Springer, pp. 190–199.

Zadra, A., et al., May 2013. CAS/JSC Working Group on Numerical Experimentation – Research Activities in Atmospheric and Oceanic Modelling. WCRP Report 10/2013. <http://www.wcrp-climate.org/WGNE/blue_book.html>.

Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., in Press. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part I: Deterministic forecast of hourly production. Solar Energy, http://dx.doi.org/10.1016/j.solener.2013.12.006.

Please cite this article in press as: Zamo, M., et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Sol. Energy (2014), http://dx.doi.org/10.1016/j.solener.2014.03.026