1 An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting

2
3 Simone Sperati^{a*}, Stefano Alessandrini^b, Luca Delle Monache^b

4 ^a Ricerca Sistema Energetico (RSE) SpA, Milano, Italy

5 ^b National Center for Atmospheric Research (NCAR), Boulder, CO, USA

6
7 * Correspondence to: simone.sperati@rse-web.it

8 9 **Abstract**

10
11 Solar energy production is steadily growing in several countries. Depending on meteorological
12 variables such as solar irradiance, cloud cover and temperature, solar power production
13 predictability is limited by the chaotic nature of these parameters. This work presents an application
14 of the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction
15 System (EPS) to produce short-term probabilistic solar power forecasts (SPF). The EPS is based on
16 running a meteorological model multiple times, starting from slightly perturbed initial conditions.
17 The distribution of these different runs allows estimating the prediction uncertainty. We use about
18 two years of power data from three solar farms located in different parts of Italy, characterized by
19 different climatic conditions. We retrieve ensemble members of meteorological variables to build a
20 power probability density function (PDF) for 0-72 h forecast horizons. A neural network (NN) is
21 applied to reduce model bias and to generate a PDF of power starting from the ensemble of
22 meteorological variables. Then two statistical methods are applied to calibrate further the ensemble
23 of power predictions, which results in the final probabilistic power forecast. The first technique has
24 been already tested for wind power forecasting (WPF), and it is based on estimating the variance
25 deficit (VD) or “lack of spread” of the ensemble distribution. The second is a well-established
26 method to calibrate probabilistic predictions based on Ensemble Model Output Statistics (EMOS).
27 Also, a persistence ensemble (PE) is used as a baseline technique. PE builds the ensemble members
28 from the most recent power measurements available at the same lead time. It is shown that EPS
29 coupled with VD and EMOS is a valid method to produce skillful probabilistic SPF in different
30 climatic conditions. VD and EMOS show a similar level of statistical consistency and forecast skill
31 for each test case. Overall, the performances of VD and EMOS appear significantly better than
32 those of PE.

33
34 Keywords: solar energy, ensemble forecasting, probabilistic forecasting

35 36 **1. Introduction**

37
38 The use of solar photovoltaic (PV) energy as a renewable source is constantly increasing in several
39 countries. PV is highly dependent on the availability of its primary resource (i.e., solar irradiance)
40 and other meteorological factors such as cloud cover and the air temperature near the surface. The
41 total amount of solar radiation reaching the earth surface (Global Horizontal Irradiance, GHI)
42 depends on several factors. The amount of liquid water and the presence of aerosols are among the
43 most important because they can reflect or absorb solar radiation. Aerosols have both natural and
44 anthropogenic origins. GHI determines the PV production at the surface and has a certain grade of
45 variability, which cannot be exactly predicted. Furthermore, the atmosphere is a chaotic system
46 where small errors in estimating its initial state can lead to large predictions errors at increasing
47 forecast lead times [1]. Therefore, it is challenging to generate an accurate forecast due to our
48 limited ability in describing the initial state of the atmosphere.

49 In a deterministic framework, a single value is provided for each location and time horizon in the
50 future. No information about the prediction uncertainty is available, and the end-user relies only
51 upon the accuracy of the forecast system in use. On the other hand, a probabilistic approach allows

estimating the probability of occurrence of a given event, along with information about the expected uncertainty affecting the forecast. Probabilistic methods consist in applying statistical techniques to the outputs of a deterministic meteorological model [2] or on the use of ensemble meteorological systems like the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) [3]. The EPS is based on running a meteorological model multiple times, adding stochastic model perturbations to the initial conditions. By integrating in time and space each of these runs, it is possible to estimate the prediction uncertainty by looking at the distribution in space and time of the ensemble members.

Probabilistic forecasting of solar power is a relatively recent topic, with only a few publications in the literature. For instance, in [2] the authors use an analog ensemble to produce probabilistic solar power forecasts (SPF). In [4] the use of an ensemble approach based on different machine learning algorithms is proposed, while in [5] SPF are produced for 1 h ahead using a Bayesian autoregressive time series model without using Numerical Weather Prediction (NWP) data. [6] proposes an algorithm to produce 6 h ahead SPF using a vector autoregression framework. [7] analyzes the performance of eight forecasting methods for probabilistic PV forecasting using predictors from an NWP-based ensemble system.

In this work, we focus on producing probabilistic SPF for a short-term time horizon (0-72 h ahead) using the ECMWF EPS. The EPS model is constituted by a set of forecasts (i.e., the ensemble members), each one initialized adding small perturbations to the analysis [8]. Each member is designed to be statistically indistinguishable. At short forecast lead times (i.e., 0-1 h) the differences between the members are likely to be small and consistent with the observations uncertainty. The differences among the members increase with longer forecast horizons, and it is referred to as dispersion or spread. The ECMWF EPS, like many other NWP-based ensembles, is known to be under-dispersive [8], meaning that the ensemble spread tends to underestimate the prediction uncertainty by lacking (on average) enough spread, especially for the short-range horizons (i.e., up to three days ahead). For this reason, it is necessary to apply statistical calibration procedures to increase the EPS reliability and statistical consistency [9].

First, we first use a neural network (NN) to reduce model bias and to create a probability density function (PDF) of solar power. Then, we apply and compare two statistical techniques to obtain calibrated ensemble power data and to produce the final SPF. The first technique was initially introduced by [10]. The method was then applied to calibrate wind power forecasts (WPF) [11]. It is based on the estimation of the variance deficit (VD) of the ensemble system. The “lack of spread” of the power prediction produced by the NN is adjusted to enhance the ensemble distribution statistical consistency and reliability. The second method is based on Ensemble Model Output Statistics [12, 13], a well-established calibration technique based on minimizing the Continuous Ranked Probability Score (CRPS) of the ensemble data set. A definition of the CRPS is given in section 4. Also, we define a reference method by building a persistence ensemble (PE), which includes the 51 most recent power measurements available at the same lead time and location. We assess relevant statistical attributes of the ensemble forecasts analyzed in this study, including statistical consistency, skill and forecast value.

The paper is organized as follows. Section 2 describes the observational and meteorological data sets. Section 3 explains the models. Section 4 shows the evaluation framework, while results are reported in section 5. Conclusions are then drawn in section 6.

2. Data description

We use data from three PV power plants with Nominal Power (NP) of 5.21 kW, 5.21 kW and 4.99 MW, respectively. The facilities are located in Milano (Northern Italy) Catania, and Calabria (Southern Italy). These locations are characterized by distinct climatic conditions. The Milano plant is installed inside the RSE SpA [14] headquarter, in the suburban industrial area of the city. The Catania plant has the same characteristics as Milano, regarding NP and electronic components, and

is located in the suburban area of the town of Catania, close to the Etna volcano. While Milano and Catania are PV research facilities, the Calabria solar PV installation is an industrial plant with a higher NP owned by Enel Green Power Spa. The map in Figure 1 shows the location of each power plant.

Figure 1 Location of the power plants Milano (A), Catania (B) and Calabria (C).

Both Milano and Catania are characterized by conditions that are more challenging to predict than Calabria. At Milano this is due to higher levels of pollution in the area surrounding the power plant, where the presence of aerosol particles can influence solar radiation at the ground. At Catania the Etna volcano releases at times volcanic ashes, which may influence power production. About 20 eruptions happened in the available data set. Catania experiences more favorable climatic conditions for PV production than the other two locations, with an average number of sunny days per year higher than in Milano where, on the contrary, the weather tends to be cloudier especially during the winter. This is consistent with the ratio between mean produced power (MP) and NP, which is 22% for Milano, 24% for Calabria and 30% for Catania over the available data set. For each plant, hourly averaged power data are available for the period January 2010 – December 2011 (Catania), July 2010 – December 2011 (Milano) and April 2011 – June 2013 (Calabria). For Milano and Catania, hourly averaged measurements of temperature at 2-m above ground (T2M), Global Horizontal Irradiation (GHI), and Direct Normal Irradiation (DNI) are also available. We retrieved EPS forecast data with 0.25° horizontal resolution and 3-hourly temporal frequency at the 4 nearest grid points of each site. In each case, we consider a model run starting at 00 UTC for the 72 hours ahead. The selected meteorological variables are integrated over time GHI (J/m^2), T2M and Total Cloud Cover (TCC). Each variable is then interpolated at each site location using a bilinear function. The azimuth angle (AZ) and solar elevation angle (EL) for each location are computed and averaged on each 3-hourly time step using a library available in R [15]. AZ and EL are added to the set of forecast variables. To work with instantaneous ensemble irradiance instead of cumulative data, GHI is converted to W/m^2 for every 3-hour time step dividing the values in J/m^2

by 10800 s (i.e., the 3-hour time step in seconds). In fact, instantaneous GHI values were not available on the ECMWF archive. The TCC forecast data are consistent with the observed ratios between MP and NP and reflect the different climatic conditions of the sites. In fact, the percentage of hours with forecasted TCC lower than 0.5 is about 53% for Milano, 57% for Calabria and 63% for Catania.

The data sets are split into two periods, training and verification. For Catania and Calabria, the training and the test periods are 1-year long. For Milano, the training period is reduced to 6 months due to the shorter available data set, using the whole 2011 as test period (see Table 1).

Table 1 Training and test periods for the different data sets.

Power plant	Training	Test
Milano	2010/07/01 – 2010/12/31	2011/01/01 – 2011/12/31
Catania	2010/01/01 – 2010/12/31	2011/01/01 – 2011/12/31
Calabria	2011/04/01 – 2012/06/30	2012/07/01 – 2013/06/30

Only lead times corresponding to a positive solar elevation are included in the data set analyzed in this study.

3. Models

3.1 Ensemble Prediction System (EPS)

The EPS is based on perturbations of the initial conditions of the atmospheric system to produce probabilistic forecasts. In this work we use the T639/T319 version of EPS, upgraded in January 2010 [16]. The main update in this release consists of an increased horizontal resolution of EPS, from 0.5° to 0.25° (i.e. about 32 km at mid-latitudes).

EPS simulates 50 different evolutions of each atmospheric variable of interest plus an unperturbed forecast (i.e., the control run). This forecast is driven directly from the ECMWF analysis and differs from the ECMWF deterministic forecast only for its coarser resolution. Initial conditions for the other 50 forecasts are created adding small perturbations to the analysis, using a mathematical method based on singular vector decomposition and stochastic parameterization to represent model uncertainty [8]. This kind of approach looks for perturbations that maximize the impact, which can be either of reinforcement or weakening, on individual weather systems for a 48 h ahead forecast. The impact is measured by the total energy above the reference hemisphere, at 30° latitude.

3.2 Neural network (NN)

An artificial NN first processes ensemble meteorological data to remove model bias. Also, the NN is applied to create a PDF of solar power given as input the meteorological variables. We use a feed-forward NN with a single hidden layer [17] developed in R environment. The NN has a simple structure that connects the input variables with a defined response variable using different weights. The information is processed by one hidden layer of 8 neurons. The fitting criterion of the NN is least squares plus a weight decay term, which uses as a penalty the sum of squares of the weights. To allow this, the inputs are rescaled in the [0,1] interval to be comparable with the outputs of the internal unit [17]. This type of NN was already used for WPF in [11] and for SPF in [2]. Similar approaches are shown in [18, 19].

The NN is configured using GHI, TCC and T2M as input variables, along with forecast lead time, solar AZ and EL. We set up a single NN using forecast variables given by the EPS control run, instead of setting up a different NN for each of the EPS members and for each forecast lead time.

As a matter of fact, the ECMWF EPS perturbation strategy generates indistinguishable members and we verified (not shown) that training the NN on different members leads to very similar results. Solar power is defined as the response variable. The initial weights to configure the NN are randomly initialized performing multiple trials. In each trial, the NN is trained starting from a different set of random weights. The number of iterations of the NN to reach a local convergence is limited to 2000 to avoid over-fitting. The NN is trained using the whole training period except 75 random days (Training_1). These 75 random days (Training_2) are used to select the best NN from the trials, by the root mean square error (RMSE, see Appendix A) calculated from the NN-corrected EPS control run and the measurements. The NN with the lowest RMSE is then applied to each ensemble member over the whole test period, obtaining 51 members of solar power. Also, the best NN is applied over Training_2 data. This period is then used to train the calibration (VD and EMOS) algorithms.

After the NN, we apply a logit function [20] to each member to make them reproducing a Gaussian distribution. The generalized logit function takes x values within a defined range $[x_{min}, x_{max}]$ and transforms them to span $[-\infty, \infty]$. It is defined as follows:

$$y = \log\left(\frac{p}{1-p}\right) \quad (1)$$

where $p = \frac{(x-x_{min})}{(x_{max}-x_{min})}$. This simplifies the calibration process since Gaussian distributions are characterized by their first two moments only. The PDFs are then transformed back with the inverse logit function at the end of the calibration process. Figure 2 shows the effect of the logit transformation, plotting the ensemble distribution around the mean before and after the logit transform.

Figure 2 Ensemble distribution before (left) and after (right) logit transform. The case shown is January 1, 2011, lead time 39 h (i.e., 15 UTC), Catania.

Finally, Figure 3 summarizes the whole prediction process.

Figure 3 Scheme of the whole prediction process. In the training (left), 30 NN are trained on Training_1 period with EPS control run data, along with forecast lead time, EL, AZ and solar power. The NN with the best performance is selected on Training_2 period and then applied over the test period (right) on the 51 EPS members of GHI, TCC and T2M to obtain 51 EPS members of solar power. VD and EMOS techniques are applied after the logit transform. Finally, 51 calibrated EPS members of solar power are obtained after inverse logit transform.

3.3 Variance deficit (VD)

The VD method [11] adjusts the ensemble spread at each forecast lead time and location, trying to make it similar to the standard deviation of the observations at the corresponding past lead times. We apply the logit function to transform the post-processed ensemble members of Training_2 and the corresponding power measurements. After that, the vd coefficient is calculated from the ratio between the RMSE of the ensemble mean versus the power measurements and the average standard deviation of the ensemble, as follows:

$$vd = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2}}{\frac{1}{N} \sum_{i=1}^N \sigma_i} \quad (2)$$

where p_i is the ensemble mean, o_i are the power measurements and σ_i is the standard deviation of the ensemble members; vd is then applied to each ensemble member during the test period using the following formula:

$$ec_j = \bar{e} + vd(e_j - \bar{e}) \quad (3)$$

where ec_j is the calibrated ensemble set, $\bar{e}$ is the ensemble mean, and e_j represents each of the ensemble members.

3.4 Ensemble Model Output Statistics (EMOS)

The EMOS technique was introduced in [12]. An example of EMOS applied to EPS data can be found in [13], where the authors use it to calibrate 100-m a.g.l. winds. Let $X_1, \dots, X_m$ be an ensemble of forecasts for a univariate quantity Y and that S_2 their ensemble variance, then the EMOS PDF is given by the following normal distribution:

$$N(a + b_1 X_1 + \dots + b_m X_m, c + d S_2^2) \quad (4)$$

where $a + b_1 X_1 + \dots + b_m X_m$ (i.e., the EMOS predictive mean) is a bias-corrected average of the ensemble members and $c + d S_2^2$ (i.e., the EMOS predictive variance) is a linear function of the ensemble variance. Various techniques to estimate the coefficients, $a, b_1, \dots, b_m, c$ and d are shown in [12]. In this work we focus on the minimum CRPS estimation, i.e., we look for the coefficients that minimize the CRPS value over Training_2. The CRPS can be expressed as an analytic function of the coefficients as follows:

$$\Gamma(a; b_1, \dots, b_m; c; d) = \frac{1}{k} \sum_{i=1}^k (c + d S_i^2)^{1/2} \left\{ Z_i [2\Phi(Z_i) - 1] + 2\varphi(Z_i) - \frac{1}{\sqrt{\pi}} \right\} \quad (5)$$

where φ and Φ denote the PDF and CDF, respectively, of a normal distribution with 0 mean and variance equal to 1, and Z_i is a standardized forecast error defined as follows:

$$Z_i = \frac{Y_i - (a + b_1 X_{i1} + \dots + b_m X_{im})}{(c + d S_i^2)^{1/2}} \quad (6)$$

Equation (8) is minimized numerically using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [21] using a library available in R. The variance coefficients c and d are also constrained to be non-negative. After applying EMOS, we obtain a 51-member ensemble taking the CDF quantiles at $m/(M+1)$ for $m=1, \dots, M$.

3.5 Persistence ensemble (PE)

We use PE as a reference for the forecasts obtained using VD and EMOS. We build PE for the time horizons from 0 to 72 h ahead as shown in [2], using the most recent 51 power values measured at the same lead time. The 51 past measurements are sorted to define the power quantile intervals. The number of observations is set to 51 to create a probabilistic distribution with the same number of members as the EPS model. For a particular period and in the case of continuous clear sky conditions, at each hour of the day, it is reasonable to expect a similar level of production since the sun position is similar. Hence, PE can be an accurate forecasting method depending on these conditions.

4. Evaluation methods

The statistical consistency of the ensemble forecasts is assessed using rank histograms, which are standard probabilistic verification tools that indicate whether the ensemble members are statistically indistinguishable from the observations [22]. In a statistically consistent system, a measured value ranked among the corresponding ordered ensemble members should equally likely take any position in the range of the whole PDF. As a result, the rank histogram would be flat, with a uniform rank probability of $1/(n+1)$.

The ability of a probabilistic prediction to quantify its uncertainty is checked using binned spread/skill diagrams, in which the ensemble spread (i.e. the standard deviation of the ensemble members) is plotted against the RMSE of the ensemble mean over small bins of spread [23], each

one having the same number of data points. A good correlation between spread and RMSE indicates that the ensemble system is capable of forecasting its own error [23] and is a necessary condition for statistical consistency.

In [24] the author uses signal detection theory to explore the usefulness of probabilistic weather forecasts when translated into binary decisions. The importance of verifying the quality of a particular binary event is related to stakeholders' needs. For example, Transmission System Operators (TSOs) may be interested in forecasting production rates above/below a given threshold that might be critical to ensure the grid's stability. They might take preventive measures, such as buying energy generated by different sources than solar, to guarantee the balance between load and demand. The overall value of a probabilistic forecast is described in [25] by its lower and upper bounds, which are defined by the Brier skill score (BSS) and the relative operating characteristic (ROC) skill score (ROCSS), respectively. The BSS is a positive-oriented score calculated from the Brier Score (BS), which is given by the following formula:

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2 \quad (7)$$

where p_i is the predicted probability of the i -th forecast and o_i states if the event occurred or not. The BS is negatively oriented, with perfect forecasts exhibiting $BS = 0$. The BSS is then given by:

$$BSS = 1 - \frac{BS}{BS_{ref}} \quad (8)$$

where BS_{ref} is the Brier score of a reference forecast. The BS can be further split into reliability, resolution and uncertainty components [26, 27]. Reliability checks for the occurrence rate of an event (i.e., its observed relative frequency) on the forecasted probability of the same event, for a given level of probability. It corresponds to the bias of a continuous variable. Hence, values close to 0 are better. Resolution measures the ability of a forecast to a priori sort out when an event occurs or not. It is perfect when a probabilistic system forecasts 100% on occasions when the event occurs and 0% when the event does not occur. Finally, uncertainty is the variability of the observations, which depends only on the climatology of the sample and corresponds to the highest possible (i.e., perfect) value of resolution. The ROCSS is based on using the ROC curve [28] to plot the false alarm rate (i.e., the false alarms divided by total non-occurrences of a determined event) against the hit rate (correct forecasts divided by total occurrences of the same event). In this way, it is possible showing the ability of a forecast to discriminate. The ROC curve and the ROCSS thus depend upon resolution and not reliability. The ROCSS is used to translate the ROC score into a skill score where 1 equals to perfect forecasts and values lower than 0 indicate lower performance than climatological forecasts. Both the BSS and the ROCSS are applied here using PE as a reference method.

Reliability is assessed in more detail using reliability diagrams, where the sharpness of the forecasting system is also checked. A sharper (i.e., narrower) forecast can produce probability values more towards the extremes (i.e., close to 0 or 100 %) for any given event threshold.

Sharpness is diagnosed plotting the relative frequency of each class interval of probability. A sharper forecast leads then to a better resolution, if the forecasts are reliable [29].

Further assessment is performed by evaluating the continuous ranked probability score (CRPS) of the forecasting methods. The CRPS is a common verification index that compares a full probabilistic distribution with the measurements in case both are represented as cumulative distribution functions (CDF) [30]. Therefore, the index is not computed for only a single threshold value, but it provides a score for the overall PDF. It is a negative-oriented score. In fact, the index is lower when the PDF originating the CDF is sharp and centered on each observation. The CRPS is given by the following formula:

$$CRPS = \frac{1}{N} \sum_{i=1}^N \int_{-\infty}^{\infty} (F_i^f(x) - F_i^0(x))^2 dx \quad (9)$$

where $F_i^f(x)$ is the CDF of the probabilistic forecasts for the i -th value at each time step and $F_i^0(x)$ is the CDF of the measurements. CRPS corresponds to the mean squared error of the cumulative distribution, and can be reduced to the mean absolute error (MAE, see Appendix A) for a deterministic forecast [31]. The index is computed by numerical integration techniques, based on a discretization of the forecast CDFs, as well as the corresponding power measurements.

5. Results

5.1 Statistical consistency

Figure 4 shows the rank histograms calculated for each test case and each forecasting method over the lead times corresponding to positive solar elevation.

Figure 4 Rank histogram for PE (left), VD (center) and EMOS (right) computed considering all lead times with positive solar elevation together for Milano (top), Catania (middle) and Calabria (bottom).

The horizontal line in the diagrams indicates the frequency of perfectly consistent ensemble. The vertical bars in the charts are the confidence intervals calculated with a quantile function of a binomial distribution [32, 33]. The intervals delimit the 5%-95% quantile interval of the binomial distribution. They are used to define a range in which deviations from the right frequency are consistent with the number of samples in each bin. In fact, a limited number of samples allows considering deviations that could occur by chance whereas an infinite number of samples would reduce the confidence intervals to zero. The missing rate error (MRE) [34] reported in the diagrams indicates the percentage of measurements falling outside the lowest-highest members. It is expressed as follows:

$$MRE = 100 \left(\frac{N_1 + N_{n+1}}{M} - \frac{2}{n+1} \right) \quad (10)$$

where n are the ensemble members, N_1 and N_{n+1} are the number of observations respectively below the lower and above the higher member and M is the total number of verifying values (forecast/observation pairs). Larger positive (negative) values of the MRE reveal a more under-dispersive (over-dispersive) ensemble. Ensemble models such as the ECMWF-EPS usually show better statistical consistency in the medium forecast range, i.e. after 72 h ahead [11]. The under-dispersive conditions observed here confirm this issue. In fact, even after calibration, the rank histograms show in each case slightly over-confident behaviors of the model, with the first and the last bins more populated than the others.

5.2 Spread/skill consistency

Figure 5 depicts spread/skill diagrams. The charts are created using 10 bins, each one having the same number of data points. For Milano, all models are under-dispersive for smaller spread values. VD shows overspread condition at higher spread values while EMOS remains slightly under-dispersive even at higher values. For Catania and Calabria, both VD and EMOS show smoother trends. VD is still over-dispersive for higher values while EMOS remains slightly under-dispersive for almost all the bins. PE shows a good spread/skill consistency for Calabria. Overall, the diagrams indicate that all methods have some ability to quantify the uncertainty of power predictions.

Figure 5 Binned spread/skill diagrams for the three solar farms and three different forecast methods. Both root mean square error (RMSE) and spread values are normalized by nominal power (NP). The 5-95% bootstrap confidence intervals are plotted only for VD to reduce clutter.

5.3 Forecast value

Figure 6 shows the BSS of VD and EMOS using PE as reference forecast. The event considered to calculate the BSS is SPF lower than MP, where MP is computed independently for every lead time. We set these thresholds to define events potentially different from climatology that may be critical for TSOs. In fact, to guarantee the grid stability in those events, it would be necessary to replace the solar energy production with other sources. We plot the BSS values as a function of the forecast lead time for all the power plants. Also, 5-95% bootstrap confidence intervals are computed considering all the lead times with positive solar elevation together and reported on the proximity of the left axis. Given the diurnal variation of MP, different thresholds are used for each lead time.

Figure 6 BSS of VD and EMOS, using PE as reference, as a function of forecast lead time for the three different solar power plants. Next to the left axis is shown the BSS 5-95% bootstrap

confidence interval computed considering all the lead times with positive solar elevation together. The event considered is solar power forecasts lower than MP at each lead time.

Both models exhibit very similar BSS values for Milano and Catania while for Calabria EMOS shows slightly better results. The positive BSS indicates that EMOS and VD have a better BS than PE. At lead times corresponding to 18 UTC both models show negative BSS values for Calabria, where they perform slightly worse than PE. Given the low solar elevation in correspondence of 18 UTC, this could be addressed to lower correlation between the forecasts and the observations caused by some obstacles shading. In fact, the Pearson correlation [35] calculated from observed power and GHI of the EPS control run for Calabria, at the lead times corresponding to 18 UTC, is equal to 0.59. This issue is not observed for Milano and Catania, where BSS values at 18 UTC still indicate better performance than PE and the correlation index is equals to 0.75 and 0.78 respectively.

For each approach and power plant, the BS is then decomposed in its reliability, resolution and uncertainty components as shown in Table 2. The BS components are computed considering all the lead times with positive solar elevation together. The 5-95% bootstrap confidence intervals are also reported.

Table 2 Decomposition of the BS in its reliability, resolution and uncertainty parts for the three different solar power plants considering all the lead times with positive solar elevation together. The 5-95% bootstrap confidence intervals are reported as [lower value, mean value, higher value]. The event considered is solar power forecasts lower than MP at each lead time. Best results are in bold.

		Milano	Catania	Calabria
Reliability	VD	[3.25, 5.43 , 6.50]*10 ⁻³	[0.88, 2.44 , 3.05]*10 ⁻³	[0.35, 2.01 , 2.53]*10 ⁻³
	EMOS	[3.66, 6.47, 7.84]*10 ⁻³	[1.67, 3.81, 4.78]*10 ⁻³	[0.69, 2.37, 3.02]*10 ⁻³
	PE	[2.09, 1.57, 2.44]*10 ⁻²	[0.69, 1.07, 1.29]*10 ⁻²	[3.64, 6.29, 7.69]*10 ⁻³
Resolution	VD	[1.09, 1.16, 1.22]*10 ⁻¹	[1.01, 1.07, 1.14]*10 ⁻¹	[1.09, 1.16, 1.21]*10 ⁻¹
	EMOS	[1.12, 1.19 , 1.25]*10 ⁻¹	[1.03, 1.10 , 1.16]*10 ⁻¹	[1.19, 1.27 , 1.32]*10 ⁻¹
	PE	[3.31, 3.99, 4.44]*10 ⁻²	[3.49, 4.03, 4.43]*10 ⁻²	[6.44, 7.04, 7.58]*10 ⁻²
Uncertainty		2.49*10 ⁻¹	2.34*10 ⁻¹	2.47*10 ⁻¹

For each case, VD shows slightly better levels of reliability (note that a lower value is better) than other models while the EMOS forecasts have slightly more resolution (note that higher is better) than VD. Also, PE exhibits in each case worse indices than VD and EMOS. Looking back at the BSS diagrams in Figure 6, it can be said that the main contribution to the better BSS of EMOS over VD for Calabria is due to its higher resolution. However, the resolution of VD and EMOS for Milano and Catania is somewhat similar.

Figure 7 shows the ROCSS values for VD and EMOS using PE as a reference forecast. The event considered is again SPF lower than MP at each lead time. ROCSS is shown as a function of forecast lead time for all test cases; 5-95% bootstrap confidence intervals close to the left axis are computed considering all the lead times with positive solar elevation together. For Milano and Catania, both VD and EMOS show similar ROCSS values, with almost no significant differences according to the confidence intervals. As for BSS, Calabria shows slightly better ROCSS for EMOS than VD. The lack of correlation between forecasted GHI and observed power is still evident at lead times corresponding to 18 UTC. In each case, both VD and EMOS report positive ROCSS values, indicating a better ability to resolve the event than PE.

Figure 7 ROCSS of VD and EMOS, using PE as reference, as a function of forecast lead time for the three different solar power plants. Next to the left axis is shown the ROCSS 5-95% bootstrap confidence interval computed considering all the lead times with positive solar elevation together. The event considered is solar power forecasts lower than MP at each lead time.

5.4 Reliability and sharpness

Figure 8 reports the reliability diagrams for the three test cases, where the event considered is still SPF lower than MP, where MP is computed independently at each lead time. In the charts, confidence intervals are the 5% - 95% quantiles range derived from a binomial distribution with the same probability as the corresponding class. They show a range in which deviations due to sampling effects are still consistent with the right reliability.

Figure 8 Reliability diagrams for the three solar farms and three different forecast methods, computed for all the lead times with positive solar elevation. The event considered is solar power forecasts lower than MP where MP is computed independently at each lead time. Confidence intervals are plotted only for VD to reduce clutter. The low lines depict the sharpness of each forecast method.

The forecasts of both VD and EMOS exhibit similar levels of sharpness. They can generate more predictions with a probability of zero or one than PE. In particular, PE produces fewer forecasts with a zero probability than VD and EMOS, while it performs somewhat similarly when predicting with a probability equal to one. VD appears slightly more reliable than the other methods in the Milano test case while at Catania and Calabria the performance of VD and EMOS are comparable, still outperforming PE.

5.5 CRPS

Figure 9 depicts the CRPS as a function on forecast lead time for the three test cases, comparing the performance of PE, VD and EMOS. The index is normalized by both NP (left axis) and MP (right axis).

Figure 9 CRPS normalized by NP (left axis) and MP (right axis) as a function of forecast lead time; 5-95% bootstrap confidence intervals are plotted only for VD to reduce clutter.

At Milano and Catania VD and EMOS have a similar performance and are better than PE, with statistically significant differences, particularly during early afternoon hours. The Calabria case doesn't show the same performance difference between the three forecasting systems. Also, EMOS seems to perform slightly better than VD during the hours with higher solar elevation. For all test cases, the CRPS peaks are around 8-9% on NP for both VD and EMOS. However, looking at the results on MP, the higher predictability of Catania is noticed, where the CRPS peaks reach around 22%, compared to 30% for Calabria and 31% for Milano.

6. Conclusions

In this work we evaluate the performance of the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) to produce 3-hourly, 0-72 h ahead probabilistic solar power forecasts (SPF) for three power plants (named Milano, Catania, and Calabria) located in different parts of Italy. The Milano plant is located in Northern Italy, where cloudier conditions tend to be more frequent. Catania and Calabria facilities are located in the South of Italy where the average solar irradiance is higher.

After applying a Neural Network (NN) to the ensemble forecasts to reduce the model bias, we test two statistical calibration methods to obtain the final SPF. One method is based on the estimation of the ensemble variance deficit (VD), and it was originally applied to wind power forecast (WPF). The other method is the Ensemble Model Output Statistics (EMOS) which is a state-of-the-art technique for calibrating ensemble forecasts. Both approaches are compared with a persistence ensemble (PE) forecast, which is made by the 51 most recent observations available at the same forecast lead time.

VD and EMOS show similar levels of statistical consistency. They tend to be slightly under-dispersive and show an acceptable level of consistency between spread and error. Hence, they can predict the uncertainty of power forecasts, especially at Catania and Calabria. For Milano, the lower predictability conditions of the site affect the models skills. Also, shorter availability of training data for Milano should be taken into account when analyzing the forecast performance.

VD and EMOS report similar forecast value, which is expressed by the Brier Skill Score (BSS) and the Relative Operating Characteristics Skill Score (ROCSS). These indices are computed for events described by SPF lower than the mean produced power (MP) computed independently at each lead time. We considered this particular threshold to select events different from climatology. Reliability and sharpness of VD and EMOS are also better than those of PE. Also, the Continuous Ranked Probability Score (CRPS) shows similar performances of VD and EMOS at all the facilities.

The overall performances of VD and EMOS are significantly better than those of PE. At Calabria, during forecast lead times corresponding to a low solar elevation, VD and EMOS don't outperform PE. A lack of correlation between the forecasted GHI from EPS and the observed power is responsible for this, preventing VD and EMOS to work properly.

In conclusion, ECMWF EPS combined with postprocessing methods can provide skillful SPF in different climatic conditions. Future work should expand the presented analysis to other test cases to consider a wider range of conditions. Also, the potential improvement given by using ensemble models in the medium range (i.e., after 72 h ahead) should be analyzed.

Acknowledgments

This work has been partly funded by the U.S. Department of Energy's SunShot Initiative, under award DE-EE0006016, and by the Research Fund for the Italian Electrical System under the Contract Agreement between RSE S.p.A. and the Ministry of Economic Development-General Directorate for Nuclear Energy, Renewable Energy and Energy Efficiency in compliance with the Decree of 8 March 2006. The COST Action WIRE ES1002 is acknowledged for sponsoring a Short-term Scientific Mission of Simone Sperati to NCAR, in which part of this work was performed. Dr. Sue Ellen Haupt is acknowledged for hosting Simone Sperati at NCAR. Dr. Pierre Pinson of DTU, Denmark, is acknowledged for his contribution in developing the variance deficit method.

Appendix A

RMSE and MAE are defined as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2} \quad (11)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |p_i - o_i| \quad (12)$$

where p_i is the i -th predicted value and o_i is the i -th forecasted value.

References

1. Lorenz, E.N., 1963. Deterministic nonperiodic flow. J. Atmos. Sci., 20, 130-141.
2. 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.
3. Molteni, F., Buizza, R., Palmer, T.N., Petroliagis, T., 2006. The ECMWF Ensemble Prediction System: Methodology and validation. Quarterly Journal of the Royal Meteorological Society 122, 529, 73-119.
4. Mohammed, A.A., Yaqub, W., Aung, Z., 2015. Probabilistic Forecasting of Solar Power: An Ensemble Learning Approach. Intelligent Decision Technologies - Smart Innovation, Systems and Technologies 39, 449-458.
5. 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, 733-747.
6. Bessa, R.J., Trindade, A., Silva, C.S.P., Miranda, V., 2015. Probabilistic solar power forecasting in smart grids using distributed information. International Journal of Electrical Power & Energy Systems 72, 16-23.
7. 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.
8. Leutbecher, M., Palmer, T.N., 2008. Ensemble forecasting. Journal of Computational Physics 227, 3515-3539.
9. McSharry, P., Pinson P., Girard R., 2009. Methodology for the evaluation of probabilistic forecasts. SafeWind Deliverable Dp-6.2, 2009.
10. Buizza, R., Richardson, D.S., Palmer, T.N., 2003. Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's ensembles. Quarterly Journal of the Royal Meteorological Society 589, 1269-88.
11. Alessandrini, S., Sperati, S., Pinson, P., 2013. A comparison between the ECMWF and COSMO Ensemble Prediction Systems applied to short-term wind power forecasting on real data, Applied Energy 107, 271-280.
12. 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.
13. Junk, C., Alessandrini, S., Delle Monache, L., 2015. Analog-Based Ensemble Model Output Statistics. Monthly Weather Review 143, 2909-2917.
14. Ricerca Sistema Energetico (RSE) SpA, URL <http://www.rse-web.it/home.page>
15. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>
16. Miller M, Buizza R, Haseler J, Hortal M, Janssen P, Untch A. Increased resolution in the ECMWF deterministic and ensemble prediction systems. ECMWF Newsletter 2010; 124:10-16.
17. Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Fourth edition. Springer.

18. Lorenz, E., Hurka, J., Heinemann, D., Beyer, H.G., 2009. Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 2, 2-10.
19. Lauret, P., Diagne, M., David, M., 2014. A Neural Network Post-processing Approach to Improving NWP Solar Radiation Forecasts. *Energy Procedia* 57, 1044-1052.
20. Cramer, J.S., 2003. The origins and developments of the logit model. In J.S. Cramer (Ed.), *Logit models from economics and other fields*, 149-158, Cambridge University Press.
21. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. 2d ed. Cambridge University Press, 963 pp.
22. Hamill, T., 2001. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review* 129, 550-560.
23. Hopson, T.M., 2014. Assessing the Ensemble Spread–Error Relationship. *Monthly Weather Review*, 142, 1125-1142.
24. Mason, I. B., 1980. Decision-theoretic evaluation of probabilistic predictions. *WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*, Nice, 8-12 September, 219-228.
25. Jolliffe, I., Stephenson, D., 2003. *Forecast Verification: A Practitioner’s Guide in Atmospheric Science*. Wiley and Sons, Chichester, 240 pp.
26. Wilks, D.S., 2006. *Statistical Methods in the Atmospheric Sciences*. 2nd Ed. Academic Press, 627 pp.
27. Murphy, A. H., 1973. A new vector partition of the probability score. *J. Appl. Met. Meteor.*, 12, 534–537.
28. Mason, I. B., 1982. A model for the assessment of weather forecasts. *Aust. Meteorol. Mag.*, 30, 291–303.
29. Gneiting, T., Raftery, A.E., Balabdaoui, F., Westveld, A., 2003. Verifying probabilistic forecasts: Calibration and sharpness. *Proc. Workshop on Ensemble Weather Forecasting in the Short to Medium Range*, Val-Morin, QC, Canada.
30. Carney, M., Cunningham, P., 2006. *Evaluating Density Forecasting Models*. Dublin, Trinity College Dublin, Department of Computer Science, TCD-CS-2006-21, 12.
31. Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559-570.
32. Bröcker, J., Smith, L. A., 2007. Increasing the reliability of reliability diagrams. *Weather and forecasting* 22(3), 651-661.
33. Pinson, P., McSharry, P., & Madsen, H., 2010. Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. *Quarterly Journal of the Royal Meteorological Society* 136(646), 77-90.
34. Eckel, F.A., Mass, C.F., 2005. Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, 20, 328–350.
35. Becker, R.A., Chambers, J.M., Wilks, A.R., 1988. *The New S Language*. Wadsworth & Brooks/Cole.