Solar Energy 158 (2017) 1026–1038

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Data-driven upscaling methods for regional photovoltaic power estimation and forecast using satellite and numerical weather prediction data

Marco Pierro[a,e], Matteo De Felice[d], Enrico Maggioni[c], David Moser[e], Alessandro Perotto[c], Francesco Spada[c], Cristina Cornaro[a,b,*]

[a] Department of Enterprise Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy
[b] CHOSE, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy
[c] Ideam Srl, via Frova 34, Cinisello Balsamo, Italy
[d] ENEA Climate Modelling Laboratory, Bologna, Italy
[e] EURAC Research, Viale Druso, 1, 39100 Bolzano, Italy

ARTICLE INFO

Keywords:
Regional photovoltaic generation
Forecast
Neural networks
Spatial clustering

ABSTRACT

The growing photovoltaic generation results in a stochastic variability of the electric demand that could compromise the stability of the grid, increase the amount of energy reserve and the energy imbalance cost. On regional scale, the estimation of the solar power generation from the real time environmental conditions and the solar power forecast is essential for Distribution System Operators, Transmission System Operator, energy traders, and Aggregators.

In this context, a new upscaling method was developed and used for estimation and forecast of the photovoltaic distributed generation in a small area of Italy with high photovoltaic penetration. It was based on spatial clustering of the PV fleet and neural networks models that input satellite or numerical weather prediction data (centered on cluster centroids) to estimate or predict the regional solar generation. Two different approaches were investigated. The simplest and more accurate approach requires a low computational effort and very few input information should be provided by users. The power estimation model provided a RMSE of 3% of installed capacity. Intra-day forecast (from 1 to 4 h) obtained a RMSE of 5%–7% and a skill score with respect to the smart persistence from −8% to 33.6%. The one and two days ahead forecast achieved a RMSE of 7% and 7.5% and a skill score of 39.2% and 45.7%. The smoothing effect on cluster scale was also studied. It reduces the RMSE of power estimation of 33% and the RMSE of day-ahead forecast of 12% with respect to the mean single cluster value.

Furthermore, a method to estimate the forecast error was also developed. It was based on an ensemble neural network model coupled with a probabilistic correction. It can provide a highly reliable computation of the prediction intervals.

1. Introduction

Large share of photovoltaic (PV) power brings new challenges for the stability of the electrical grid, both at the local and national level, since it introduces into the electric load a stochastic variability dependent on meteorological conditions. Indeed, the electricity demand (residual load) that should be fitted by not intermittent generation results from the difference between the electric consumption and the distributed PV production.

Thus, in case of high PV generation higher secondary reserves and ready supply are needed to ensure electrical balancing and overcome the unpredictability and variability of the residual load. Moreover it implies an increase in costs related to transactions on the day-ahead and intra-day energy market and dispatching operations on the real-time energy market.

To sustain the growing PV distributed production, the use of modern power electronics, distributed control together with ancillary services like PV generation forecast is becoming essential for many European countries. Indeed, in Europe the PV penetration is now around 3% with Italy leading at 7.9% and International Energy Agency (IEA) scenarios predict for 2030 a PV generation of 10%–25% of the UE27 electric demand (IEA, 2014a,b).

* Corresponding author at: Department of Enterprise Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy.
E-mail addresses: marco.pierro@gmail.com (M. Pierro), matteo.defelice@enea.it (M. De Felice), enrico.maggioni@ideamweb.com (E. Maggioni), david.moser@eurac.edu (D. Moser), alessandro.perotto@ideamweb.com (A. Perotto), francesco.spada@ideamweb.com (F. Spada), cornaro@uniroma2.it (C. Cornaro).

http://dx.doi.org/10.1016/j.solener.2017.09.068
Received 22 May 2017; Received in revised form 26 July 2017; Accepted 28 September 2017

Nomenclature

Acronym	Meaning
DSO	Distribution System Operator
TSO	Transmission System Operator
NWP	Numerical Weather Prediction
WRF	Weather Research and Forecasting model
MOS	Model Output Statistic
ANNsE	Ensemble of Artificial Neural Networks
GNN and 6GNN	ANNsE models for power output estimation based on irradiance inputs (G) relative humidity inputs (RH)
RHNN and PCARHNN	ANNsE models for power output day-ahead forecast based on PM and KPM simple and smart persistence models

Variables	Meaning
GHI, GHIcs, RH and Tair	global horizontal irradiance, clear sky global horizontal irradiance, relative humidity and air temperature at ground level
PPK_{cs}	Pseudo clear sky performance index
$PO^{obs}, PO^{PM}, PO^{KPM}, PO^{for}$	PV power output observed and predicted by the simple and smart persistence, forecast models
$P_n(dd)$	daily plant capacity

On a regional scale, PV power estimation and forecast are relevant for Distribution System Operators (DSO), Transmission System Operator (TSO), energy traders, and Aggregators. In particular the estimation of regional PV power generation from the real time environmental conditions is needed since in Italy the actual energy meters used by DSO do not allow a real time power monitoring of the distributed photovoltaic production. Thus, power estimation can be used for PV power supervision, real time control of residual load and energy reserve activation in case of deviation. PV power forecast can be employed by users for transmission scheduling to reduce energy imbalance and related cost of penalties, residual load tracking, energy trading optimization, secondary energy reserve assessment.

An overview on benefit of PV power forecast in solving problems related to the grid integration of intermittent solar energy production can be found in Emmanuel and Rayudu (2017), Shivashankar et al. (2016), Alet (2015), Alet et al. (2016), Zhang et al. (2015).

For power estimation and intra-day forecast the use of ground measurements or satellite data is essential as for day-ahead forecasts Numerical Weather Prediction (NWP) data should be employed to obtain an acceptable accuracy level. The NWP data are generated by global or mesoscale simulation models able to provide the numerical integration of the coupled differential equations describing the dynamics of the atmosphere and radiation transport mechanisms (Lorenz et al., 2016).

Moreover, these data are usually corrected by post-processing algorithms (Model Output Statistics) that use past ground measurements to partially remove systematic errors of NWP (Pierro et al., 2015; Lorenz et al., 2009a,b).

Then PV power estimation or forecast can be achieved through deterministic i.e. (Pelland et al., 2011; Lorenz et al., 2011) or data-driven models based on machine learning or probabilistic approaches i.e. (Zamo et al., 2014a,b). For the deterministic models detailed information on the PV plant set up (geographic position, modules technologies, etc.) are needed. On the contrary for the data-driven models past power measurements are essential for training, validation and test while none or very few system information are required (Pierro et al., 2016a).

The starting point for Regional PV power estimation and forecast is the so-called bottom-up strategy. It consists in the estimation or forecast of all the distributed PV plants in the considered area. Nevertheless, it requires a large computational and data handling effort. Indeed, models should be implemented for each plant (even if the distance between two plants is lower than the spatial resolution of the irradiance or NWP data) and then the models should run for all the distributed systems. Moreover, when there are not enough historical data to train machine learning algorithms, a deterministic approach must be adopted. Nevertheless, it often happens that some system information needed for the model set up (such as orientation and tilt or module characteristic) are unknown. For these reasons, ongoing research is focused on upscaling methods that allow the estimation and forecast of distributed power of aggregates of PV plants through simplified approaches that reduce the computational effort and require less information on the PV fleet. For example, Fonseca et al. (2015) proposed four different upscaling method that can be used according to different plant information and data availability scenarios. Zamo et al. (2014a) developed a data-driven model for regional PV power forecast that only requires the whole installed capacity and the historical PV generation in the controlled area for model's training.

Upscaling methods are mainly based on the selection of a subsets of PV plants with a power output that can be considered representative of the regional photovoltaic production. Then the forecast of the subsets power output is rescaled taking into account the subsets capacity and total capacity to obtain the regional prediction.

Several strategies have been developed in order to select the representative subsets. In Lorenz et al. (2008) two different random selections were tested. In the first the spatial distribution of the selected subsets should reflect the regional distribution while in the second just a uniform distribution of selected systems was chosen. In Lorenz et al. (2011, 2012) a subsets selection was proposed so that their distribution with respect to the location, installed capacity and system characteristics (plane orientation and technology) reflects the distribution of the whole ensemble. In Fonseca et al. (2015) for the selection of representative subsets a stratified sampling method according to installed capacity and PV system location was developed.

Another upscaling method considered the PV generation in the controlled area as it was produced by a virtual PV plant. Then, the power output of this virtual plant is directly forecast by machine learning algorithms as reported in Zamo et al. (2014a).

Only recently, a hybrid upscaling strategy between the two above mentioned approaches has been tested. Instead of sampling strategy, clustering methods were used for spatial grouping of PV plants and then the power output of each cluster is considered produced by a virtual PV plant and directly predicted by deterministic or machine learning models (Wolff et al., 2016).

Moreover, the accuracy of regional forecast is greatly improved with respect to single site forecast due to the "ensemble smoothing effect". This effect is related to the forecasting errors correlation, the PV capacity distribution and the number of systems in the controlled area. The errors correlation between sites decreases with the distance (or with the size of the area) thus the regional forecast accuracy can be improved even by 50% with respect to the accuracy of single plant power prediction. For this reason, the performance of each site forecast only slightly affects the performance of regional prediction so that up-scaling methods can achieve similar accuracy of the bottom-up approach.

The smoothing effect in irradiance and PV power forecasting of ensemble of plants has been studied in Perez et al. (2011), Perez and Hoff (2013), Hoff and Perez (2012), Lorenz et al. (2008, 2009b) and Fonseca et al. (2014) while in Saint-Drenan et al. (2016) the smoothing effect was analyzed related to the spatial interpolation of the power yield produced by a random subsample of reference PV plants. The same smoothing effect can be observed in regional PV power estimation.

Another problem in the regional operative power estimation or

forecast is related to the data exchange between DSO or TSO and forecast providers. Indeed these electrical system operators are reluctant in sharing real time power generation or ground irradiance data (if available) thus machine learning algorithms that input past measurements (autoregressive models, recurrent neural network, etc.) as well as in general the use of time series (Kaur et al., 2016), cannot be employed by forecast providers.

An overview on PV power forecast techniques can be found in Paulescu et al. (2012), Kleissl (2013) and IEA (2013), while recent and complete reviews can be found in Raza et al. (2016), Antonanzas et al. (2016).

In this paper new upscaling methods for estimation and forecast are developed. The methods consist of two steps. First a PV spatial clustering was implemented and then models based on artificial neural networks ensemble (ANNsE) were developed for the estimation and day-ahead prediction of the regional power output with hours granularity. Spatial clustering allows the determination of the centroids i.e. the representative points in the controlled area on which the inputs or outputs of the ANNsE models should be provided or predicted.

For the second step two different approaches were investigated. The first estimates and forecasts the power output of each cluster and then models output are averaged to obtain the regional prediction (models output average). This can be considered as a bottom-up strategy on cluster level. The second approach provides directly the regional power prediction using inputs centered on each cluster centroid (model inputs average). This approach is based on the spatial smoothing of the inputs features. The performances of the two approaches were compared and the smoothing effect at cluster scale was investigated.

Satellite derived irradiance from METEOSAT-9 and numerical weather prediction from Weather Research and Forecasting model (centered on each cluster centroids) are used as inputs for the estimation and forecasting models. Furthermore, to provide the intra-day forecast, an ANNsE that makes use of past power estimation and day-ahead forecast was also set up.

The simplest upscaling strategy (model inputs average) requires very few input information that should be provided by users. For the training phase the total installed capacity, coordinates of each PV plants and one year of distributed power generation are needed while for operative forecast only the actual capacity is required. Moreover, it does not imply any real time data exchange between users and providers and it does not need to be periodically trained. Thus, the proposed method could be easily adopted by forecast providers to deliver the regional PV power estimation and forecast to DSO, TSO for grid management and balancing issues.

Finally a parametric probabilistic method to compute highly reliable prediction intervals of the day-ahead forecast was also developed. Thus, it can be effectively used by DSO, energy traders, and aggregator companies to estimate the probability of a specific PV generation bid on the energy market and by DSO and TSO to reduce the energy reserve and ready supply.

The data-driven upscaling methods were trained and tested on the real PV distributed generation of a small part of the South Tyrol region in the North of Italy, characterized by a PV penetration similar to the one achieved at national level.

The novelty of the upscaling methods was the use of a particular chain of machine learning algorithms. K-mean clustering and principal component analysis are adopted for features reduction while models based on ensemble of artificial neural networks are used for power estimation and day ahead power forecast. Then a new simple model for intra-day forecast that makes use of past power estimation and day ahead power prediction was presented. This study allows a complete assessment of the forecast accuracy at different horizons, from 0 h ahead (power estimation) to 72 h ahead.

Moreover, in literature, the ensemble smoothing effect was studied comparing the forecast accuracy of a single plant with the accuracy of the forecast of increasing ensembles of plants (Lorenz et al., 2008, 2009b; Fonseca et al., 2014). In this case, considering the PV systems of each cluster as a virtual power plant, the benefit of the smoothing is evaluated comparing the forecast performance of a single PV cluster with the accuracy obtained combining the forecast of different adjacent

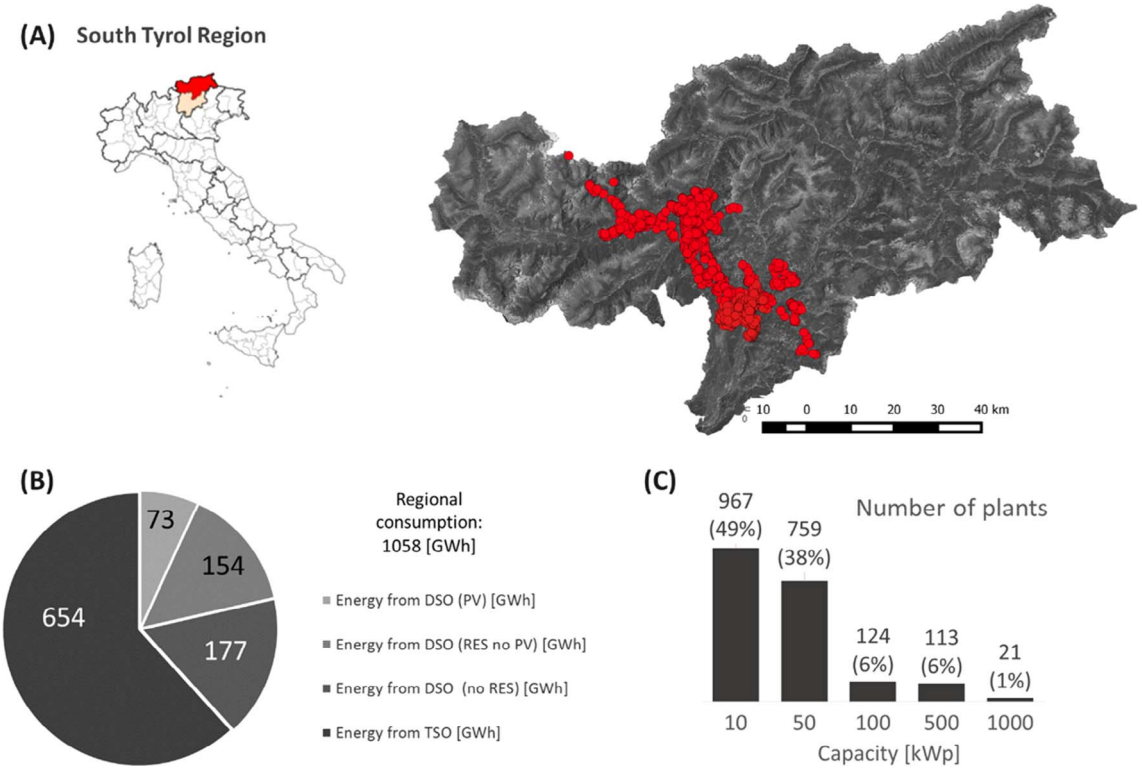

Fig. 1. (A) PV plants in the region of interest; (B) Regional electric consumption and energy supply from DSO using renewable energy sources (RES) and fossil energy sources and from TSO; (C) PV plants capacity distribution (reporting the maximum capacity above the bin).

PV clusters. The same study was carried out also for the accuracy of the power estimation. Finally, if a parametric probabilistic approach is used, the prediction intervals are usually estimated supposing a normal distribution of the forecast errors as in Lorenz et al. (2009b) or Marquez and Coimbra (2011). This hypothesis provides a correct estimation of the 95th quantile but is not reliable for other confidence levels. In this work, it was shown that the experimental error distribution is far to be Gaussian and a method to estimate the prediction intervals taking in to account the deviation from the normal distribution, was developed. It was proved that this method provides very reliable results at different confidence levels.

In Section 2 the data set used to training and test the machine learning models are described. In Section 3 the upscaling methods and the estimation and forecast models are explained in details. In Section 4 a short description of the used data-driven algorithms is provided. In Section 5 the metrics adopted to evaluate the estimation and forecast accuracy are summarized. In Section 6, the accuracy obtained applying the methods to the test data set is evaluated and discussed. Finally, in Section 7, summary and conclusions are given.

2. Data

2.1. PV power generation data

The upscaling method was used to estimate and forecast the distributed generation of 1985 PV plants in a small part of the South Tyrol Region in the North of Italy with an installed capacity of 68.2 MWp (at the end of 2015). This area of around 800 km^2 has a complex orography and variable weather conditions (see Fig. 1A).

In 2015 the electric demand in the controlled area was 1058 GWh while the PV generation provided 73 GWh (see Fig. 1B). Thus, the region has a photovoltaic penetration of 7% mainly due to small distributed PV plants with capacity lower than 50 kWp (see Fig. 1C).

The two years (2014–2015) of PV power generation data were provided by the local Distribution System Operator (Edyna) with a time resolution of 15 min.

2.2. Satellite derived irradiance data

The satellite derived irradiance data used for power estimation come from the Geostationary radiative fluxes products, under Météo-France responsibility. It was obtained by OSI SAF SSI algorithm (Ocean and Sea Ice - Satellite Application Facility - Surface Solar Irradiance) applied to the satellite images provided by METEOSAT-10 (MSG-3) at 0° longitude, covering 60S-60N and 60W-60E, with a 0.05° horizontal resolution (EUMETSAT, 2017), and hour granularity.

It should be remarked that in the satellite data the irradiance at sun elevation lower than 10° was always zero. Thus to reduce this error the data were post-processed with a cubic interpolation of the clear sky index at low sun elevation angles.

2.3. Numerical weather prediction data

The numerical weather predictions used for the day-ahead forecast were generated by the Weather Research and Forecasting (WRF–AWR 3.6.1) developed by NCAR (National Center of Atmospheric Research). The model is run operationally by the US National Weather Service and, being open source and easily portable, it is widely used around the world for research and weather forecasts (Skamarock et al., 2008). Daily hindcasts were performed for the year 2014 and 2015. The model was initialized at 12 UTC, analyzing the 24 h forecasts starting from the following 00 UTC, which is the typical procedure for the NWP solar day-ahead forecast. The model domain is centered over Italy with a horizontal resolution of 12 km, a higher resolution inner domain is nested centered on the region of interest, with a horizontal resolution of approximately 3 km. This horizontal resolution was necessary because of the complex orography of the region. The output was written every 20 min to achieve a better synchronization with the PV generation data.

Details on WRF physics configuration can be found in Pierro et al. (2016a).

3. Methodology

The developed upscaling method consists in the application of spatial clustering of PV plants and then in the use of satellite derived irradiance and numerical weather prediction (NWP) data (centered on each cluster centroid) as inputs for an Ensemble of Artificial Neural Networks (ANNsE) that estimates or predicts the regional PV power output (PO).

For the estimation and day-ahead forecast (24–48 h ahead), we tested the two following approaches:

1. Models output average: it provides the regional result by the average of the power estimation/forecast of each cluster. This can be considered a cluster bottom-up strategy.
2. Model inputs average: it uses the model inputs calculated on each cluster to directly provide the regional power estimation/forecast. This approach is based on the spatial smoothing of the inputs features.

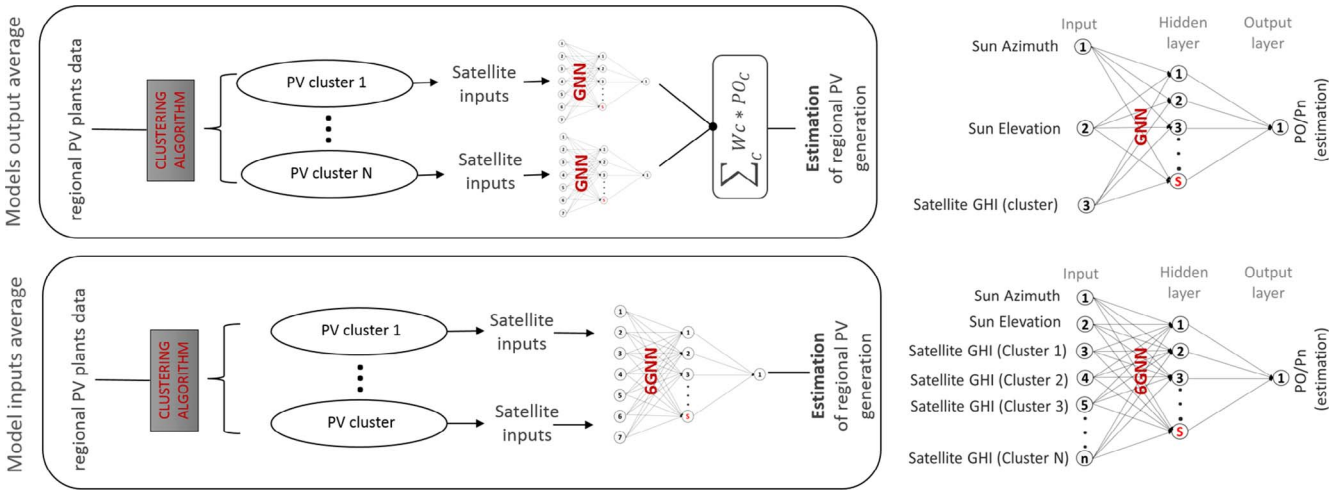

Fig. 2. The two approaches and the neural network models (GNN or 6GNN) used for estimation of PV normalized power output.

For intra-day forecast (1–4 h ahead) a further ANNsE model that makes use of past power estimation and one day ahead forecast (derived from the two previous models) was built. It provides the power output forecast from 1 to 4 h ahead.

The accuracy of the forecasts has been compared to the accuracy of two different persistence models (simple persistence and smart persistence) used as reference. Indeed both the persistence models are usually adopted in the literature as reference model for accuracy benchmarking i.e. (Lorenz et al., 2011; Wang et al., 2015).

Finally, a model to predict the error of the day-ahead forecast was developed. It was used to estimate the prediction intervals at different confidence levels.

All the ANNsE were trained on 2014 and tested on 2015.

All the approaches and the models described in details in the following subsections are originally developed for the present work. Nevertheless, the artificial neural network model used for the forecast of the PV generation of a single cluster (RHNN) is a modification of a previously developed model that was built for the prediction of the generation of a single plant (Pierro et al., 2016a).

3.1. Upscaling methods and models for estimation of regional PV power generation

The models output average consists in the estimation of the power yield of each cluster using an ANNsE based model (GNN). The estimation of the regional yield is then obtained by the average of the GNN output weighted with the hourly PV capacity: W_C = Pn(cluster)/Pn(regional).

The GNN model inputs the sun azimuth and elevation to account the sun position and the satellite derived irradiance centered on the cluster centroid.

The model inputs average provides directly the regional estimation of yield through a single ANNsE model (6GNN). The 6GNN uses the average clusters sun position (sun azimuth and elevation) and the satellite irradiance centered on each cluster centroid, as inputs.

Fig. 2 shows the two different approaches and models used to provide the estimation of the power yield (PO/P_n). Where P_n is the installed capacity.

3.2. Upscaling methods and models for day-ahead forecast of regional PV power generation

Similarly to power estimation, the models output average approach consists in forecasting the yield of each cluster using an ANNsE based model (RHNN). The regional forecast is obtained by weighted average of the cluster forecasts. The model RHNN is a slight modification of the neural network reported in Pierro et al. (2016a). It uses nine input features. The first four are the sun azimuth and elevation, the clear sky global horizontal irradiance (GHI_{cs}) and the ground air temperature (T_{air}) predicted by WRF. The other five inputs are the average values and the standard deviation of the relative humidity predicted by WRF for the vertical levels below 775 hPa and between 775 hPa and 400 hPa, and the prediction of the relative humidity at a higher level corresponding to 300 hPa. The last five features take into account the cloud formation at different atmosphere levels: low clouds approximately below 2500 m, mid clouds between 2500 and 7500 m and high clouds up to 9000 m.

The "model inputs average" provides directly the regional PV power forecast thought a single ANNsE model (PCARHNN). PCARHNN uses as inputs the sun azimuth and elevation, the clear sky irradiance (GHIcs) and the 2 m temperature (T_{air}) predicted by WRF on each cluster centroid and averaged over all the clusters. The other input features of the ANNsE result from the principal component analysis (PCA) pre-processing of the relative humidity of 20 vertical atmospheric levels (RH_{level}) predicted by WRF and calculated on the cluster centroids. It should be remarked that also additional WRF inputs (T_{air}, Geo-potential and wind speed of 20 vertical atmospheric levels) were tested for the day-ahead forecast model. Nevertheless the forecast accuracy did not improve so that the PCARHNN appeared the best compromise between input information and model complexity.

The clear sky irradiance (GHIcs) was used as input of the artificial neural network (ANN) models to take into account the geometric irradiance behavior so that the models need to forecast only the stochastic power variability.

Fig. 3 shows the two different approaches and models used to provide the day-ahead forecast of the power yield.

It should be noted that all the ANNsE models (RHNN and PCARHNN) do not input the WRF prediction of the GHI. Indeed it was proved in Pierro et al. (2015, 2016b) that the irradiance prediction provided by WRF radiation scheme should be post processed with a MOS to reach a satisfactory accuracy level.

3.3. Model for intra-day forecast of regional PV power generation

The intra-day forecast model essentially correct the day-ahead power prediction using previous power estimation. It is based on an ANNsE and for each hour it inputs the current and the past three yield estimation together with the day-ahead forecast of the next four hours. The model provides directly the forecast of regional power output from 1 to 4 h ahead.

Fig. 4 shows the model used for the intra-day forecast of the PV power yield.

To determine the number of past estimation data to use as ANN

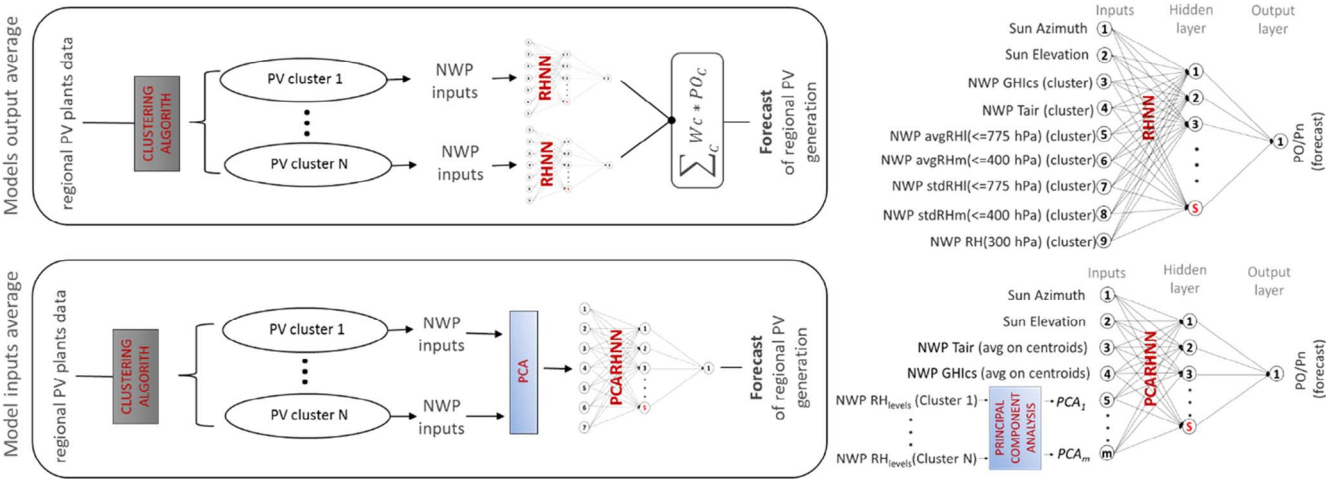

Fig. 3. The two approaches and the neural network models (RHNN or PCARNNN) used for day-ahead forecast of PV normalized power output.

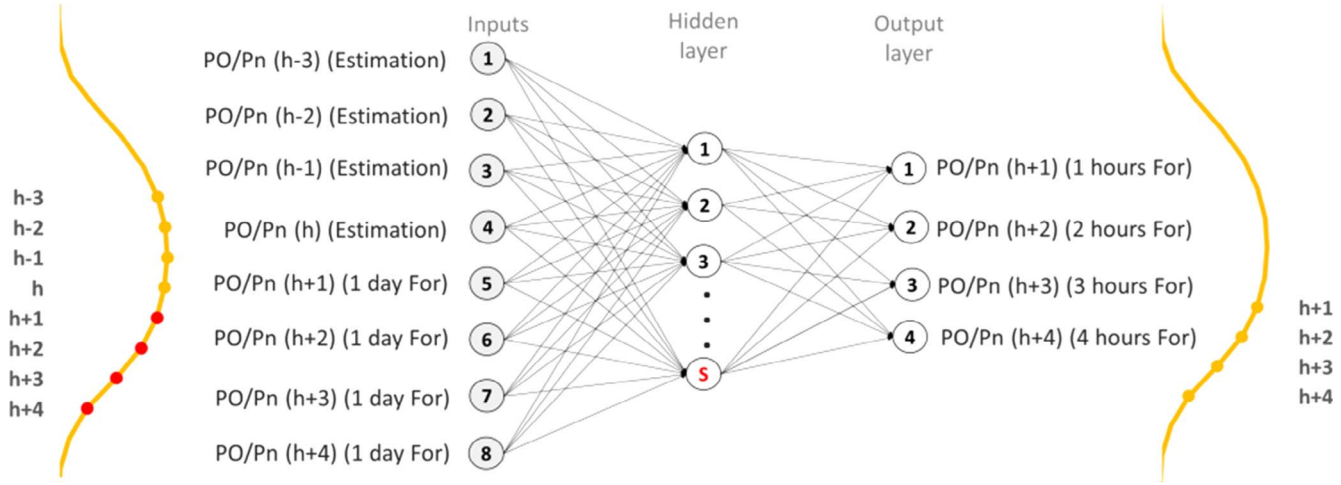

Fig. 4. ANN model for intra-day forecast of PV normalized power output.

inputs the autocorrelation function of the power output was analyzed. It appears that the maximum autocorrelation is reached for the first three past hours, thus only three past power estimation together with the current estimated production were used. This means that the day-ahead forecast of the first four hours of each day is only weakly corrected by the ANN model since at least one of the inputs is zero. Nevertheless models that input 8, 12, 18, 24 past power estimation data were also tested, but none of them obtained a sensible improvement of accuracy.

3.4. Persistence models

For day-ahead forecast of PV power two different persistence models are usually adopted as benchmarks to better assess the forecast performance: the simple persistence and the smart persistence. The smart persistence leads to a RMSE lower than the simple persistence.

The simple persistence (PM) is a trivial model that assumes power output of the day to forecast equals to the power output of the present day. On other hand, there are several smart persistences that can be adopted (Lorenz et al., 2011; Kaur et al., 2016) and (Pierro et al., 2016b). In analogy with the model proposed in Pierro et al. (2016b), in this paper a new smart persistence, namely clear sky persistence (KPM), was developed. It presumes the persistence of the daily pseudo clear sky performance index:

$$PPK_{cs}(dd) = \frac{\sum_{h=1}^{24} PO(h)/P_n}{\sum_{h=1}^{24} GHI_{cs}(h)/1000} \quad (1)$$

where dd is the present day.

Thus the PPK_{cs} is the ratio between normalized PV daily energy generation and the daily clear sky global horizontal radiation and it can be considered the equivalent of the daily clear sky index for regional PV power forecast.

Then the KPM can be calculated as:

$$PO^{KPM}(h+H) = P_n(dd) * PPK_{cs}(dd) * GHI_{cs}(h+H)/1000 \quad (2)$$

where h is the hour of the day, H is the forecast horizon (in this case 24 or 48 h), PO^{KPM} is the predicted power output, GHI_{cs} is the clear sky global horizontal irradiance and $PPK_{cs}(dd)$ is the daily pseudo clear sky performance of the present day (dd), defined in Eq. (1).

For intra-day forecast the simple persistence cannot be adopted since it brings to an hourly time shift that leads to an unrealistic prediction. On the contrary, the clear sky persistence can be calculated using the hourly pseudo clear sky performance index:

$$PPK_{cs}(h) = \frac{PO(h)/P_n}{GHI_{cs}(h)/1000} \quad (3)$$

So that the KPM for intra-day forecast can be obtained as:

$$PO^{KPM}(h+H) = P_n(dd) * PPK_{cs}(h) * GHI_{cs}(h+H)/1000 \quad (4)$$

where H is the intra-day forecast horizon that goes from 1 to 4 h.

In this case the hourly PPK_{cs} was calculated only for solar elevation greater than 10° to avoid singularity at low elevation angles. Then, to reconstruct the missing data, this index was interpolated with a cubic function for all the sun elevation lower than 10° considering a night value for the PPK_{cs} fixed at 0.5. This night value means that the power yield at low sun elevation angles is the half of the clear sky irradiance and leads to the lower persistence errors during sun rise and sun set.

3.5. Prediction of day-ahead forecast error

It is important not only to provide the forecast of power output but also the prediction intervals in which the real power output (PO) could be found with a fixed probability (confidence level). Under the hypothesis that the errors (residuals) of a given forecast model are normally distributed with zero mean value and standard deviation (σ) the actual yield should be found between: $(PO^{for}/Pn) \pm Z_{\alpha/2}\ \sigma$ with a

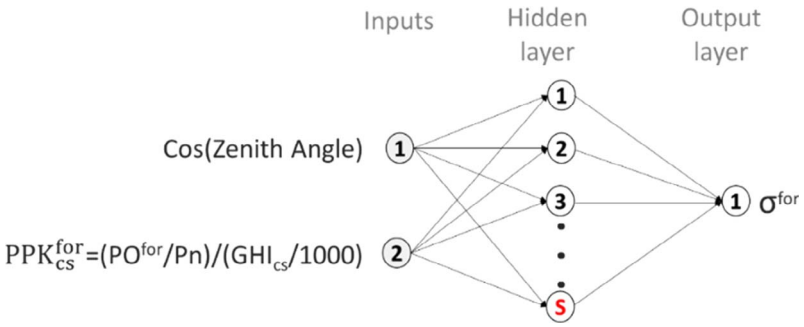

Fig. 5. ANN model for prediction interval forecast.

confidence level $1-\alpha$, where $Z_{\alpha/2}$ is the Z-score for the confidence level *1 - α* (i.e. 1- α it is equal to the integral of the standard normal distribution between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$).

Thus the prediction intervals could be calculated forecasting the standard deviation of the residuals (σ^{for}).

Fig. 5 shows the ANN model used for the prediction of the day-ahead forecast error.

In this case the ANNsE predicts the standard deviation of the errors using two input features: the cosine of the sun azimuth and the pseudo clear sky performance index (defined in Eq. (3)).

Nevertheless, the normal distribution is a poor assumption for solar forecast errors. Indeed, in the results section (Section 6.4), a method to improve the prediction intervals reliability taking in to account the deviations from normal distribution is reported.

4. Data driven algorithms

In this section, the main data-driven algorithms adopted for the upscaling methodology are briefly described.

4.1. Clustering algorithm

For spatial clustering the K-means method was applied to the geographical coordinate of each PV plant. K-means is one of the most popular unsupervised learning algorithm used to group a set $\mathbf{X} = (\underline{x}^1,\cdots\underline{x}^m)^T \in \mathbb{R}^{mxn}$ data points into K clusters centered on K representative points: centroids (where n is the dimension of each point vector $\underline{x}^i$ with i = 1:m and m > K is the number of points). It is an iterative procedure based on three steps:

1. Random selection among the m points ($\underline{x}^i$) of K initial centroids $\underline{C}^k$ with $k = 1{:}\ K$;
2. Identification of the index $k(i) = k$ corresponding to the nearest centroid to each $\underline{x}^i$ and then calculate the distortion cost function: $J(\underline{C}^1 \ldots \underline{C}^k)$;
3. Update $\underline{C}^k = <\underline{x}^i>$ with $i{:}\ k(i) = k$ and return to step 2 until the distortion J converge to its minimum J_0.

There are several distortion functions that can be used as reported in Munshi and Mohamed (2016), in the present work the mean error function was adopted:

$$J(\underline{C}^1 \ldots \underline{C}^k) = < ||\underline{x}^i - \underline{C}^{k(i)}||^2 > \text{ with } i = 1{:}\ m \tag{5}$$

The best number of cluster partitions (K_{best}) was selected, running the k-mean with K = 1:20 and calculating the $J_0(K)$ value so that K_{best} corresponds to the knee of the gradient of J_0: $dJ_0/dK = J_0(K+1) - J_0(K)$. Moreover to avoid singular initial conditions and to obtain a more reliable result, for each k-mean run, the random selection was repeated twenty times and the initial condition that obtains the lower distortion was selected.

It should be remarked that if the point coordinates have different ranges of variability they should be normalized with the respective mean value and variance to ensure equal weight in the distance calculation.

4.2. Artificial neural network

The ANN is a mathematical model that invokes the structure of biological neural connections as explained in Basheer and Hajmeer (2000). Several neural networks architectures have been developed and used in solar power and irradiance forecast application, see (Zhang et al., 1998; Mellit, 2008; Raza et al., 2016). In this work, an ensemble of multilayer perceptron (MLPNN) with one hidden layer was adopted as basic algorithm for all the estimation and forecast models. The MLPNN has the ability to imitate natural intelligence in its learning from existing sample data, so that the algorithm learns from sample data by constructing input–output connections. In an MLPNN with one hidden layer, the relation between input stimuli (X) and the neurons activities (Y) is modeled as follows:

$$Y = f^{(2)}(W^{(2)} f^{(1)}(W^{(1)} X + b^{(1)}) + b^{(2)}) \tag{6}$$

where (i = 1, 2) is the layer index, $f^{(i)}$ are transfer functions modeling the intensity of neurons activities, the weights matrices ($W^{(i)}$) mimics the strength of the synapse connections between neurons, and the bias vectors ($b^{(i)}$) stands for the neurons activation threshold. Thus a MLPNN is a nonlinear semi-empirical function dependent on a large number of parameters ($W^{(i)}$ and $b^{(i)}$). These parameters should be empirically derived by a training and validation procedure, minimizing the error between the input and the output of a known set of data (training and validation set). Moreover a stochastic component is introduced in the MLPNN function by a random choice of the initial condition of the minimization procedure and by a random partition of the training data into training and validation sets.

The Levenberg-Marquardt algorithm was used to minimize the mean square error (MSE) function using 60% of one year data for training and 40% for validation. The net structure was identified through an optimization process that provided the best number of neurons in the hidden layer through a further MSE minimization procedure. Once the best number of hidden neurons was identified, 500 ANNs were generated using the repeated random sample validation procedure. Subsequently, a qualified ensemble was selected (around 300 ANNs), choosing all the ANNs with the MSE lower than the average MSE of the 500 networks. Finally the forecast was obtained by averaging the ensemble outputs.

All the estimation and forecast models described in the previous section were trained and validated on the data of 2014 and tested on the data of 2015. Details of the method could be found in Cornaro et al. (2015).

4.3. Principal component analysis

It is good practice in the setup of a machine learning model to choose input features strongly correlated with the output but possibly uncorrelated with each other. Indeed a large number of redundant input information complicates the training phase increasing dramatically the number of parameters that should be estimated and the local minima of the cost function. This usually results in lower performance since the benefit of using more but correlated information are cancelled out by the increasing of model complexity. The principal component analysis (PCA) is a method developed by Pearson (1901) that could be employed to reduce the number of features retaining, at the same time, the relevant information.

The basic idea of PCA is to project the input features ($\boldsymbol{X} = (\underline{x}^1, \ldots \underline{x}^m)^T$) on the orthogonal base of the eigenvectors ($\boldsymbol{V} = (\underline{v}_1^T \ldots \underline{v}_n^T)$) of the covariance matrix, associated to the eigenvalues ($\lambda_1 \ldots \lambda_n$). The new features (obtained by the projection): $\boldsymbol{Z} = \boldsymbol{XV} = \{z_k^i\}$ are uncorrelated with each other and their variance is equal to the eigenvalues. Since the relevant input information is brought by the features that show the higher variances, the principal component is the subspace of the $\boldsymbol{K}$ component of $\boldsymbol{Z}$ with the higher eigenvalues:

$$\boldsymbol{PC} = \{z_k^i\}_{i=1:m;k=1:K} \text{ so that } Var(PC) < threshold * Var(Z) \tag{7}$$

where the *threshold* is the fraction of the total variance that has to be retained, $Var(PC) = \sum_{k=1:K}(\lambda_k)$ is the variance retained and $Var(Z) = \sum_{k=1:n}(\lambda_k)$ is the total variance.

In Fonseca et al. (2014) this technique was used to reduce the NWP inputs of their support vector regression model. These inputs were calculated over all the possible grid points of the region, so that the PCA was adopted to retain only the relevant difference between the numerical weather predictions of different points in the controlled area.

In the present case, the spatially reduction of the model inputs is mainly obtained by the clustering algorithm since the NWP were

calculated only on the centroid points. Instead PCA was used to retain only the relevant information about the relative humidity of 20 atmospheric levels predicted by WRF on the centroids. In the present work, the threshold was fixed to 0.95 so that the 95% of total variance was retained. This allows reducing the inputs of the day-ahead forecast models from 75 to 5.

5. Accuracy metrics

According to solar forecast literature, the main metrics used to evaluate the models' accuracy are reported in Table 1:

All the performance indexes in Table 1 are calculated excluding the night values (when the clear sky global horizontal irradiance provided by WRF is equal to zero).

6. Results

6.1. Clustering

The k-mean algorithm aggregates all the PV plants in six areas corresponding approximately to six municipalities: Naturno, Tirolo-Merano, Lana, Nalles, Bolzano, Collalbo-Soprabolzano.

Fig. 6 shows the spatial clustering of the PV systems in the controlled area.

In the present upscaling strategy, spatial clustering was used to determine the representative points in the region (centroids) on which the inputs or outputs of the ANNsE models should be provided or predicted. Thus, the power output of each cluster as well as the regional production is considered as the generation of a virtual power plant with growing PV capacity. Table 2 reports the regional and clusters capacity at the end of 2015.

It should be noted that in literature on PV power forecasting, clustering methods were mainly used to grouping typical time series of irradiance or PV production (Wang et al., 2015; Jiménez-Pérez and Mora-López, 2016; Azimi et al., 2016) and (Munshi and Mohamed, 2016). Only in few papers clustering was used in upscaling method for spatial grouping of PV plants or irradiance sensors (Fonseca et al., 2015) and (Lima et al., 2016).

6.2. Power estimation and forecast (overall results)

It should be specified that the local DSO provided the hourly PV generation data of five clusters: Naturno, Tirolo-Merano, Lana, Nalles and Bolzano-Collalbo-Soprabolzano. Therefore for estimation and day-ahead forecast the models output average approach was performed only on five clusters, grouping together the two clusters of Bolzano and Collalbo-Soprabolzano. In the following figures the fifth clusters (Bolzano-Collalbo-Soprabolzano) is simply called Bolzano cluster.

Fig. 7A shows the accuracy of the regional power estimation obtained by the two approaches: averaging the GNN estimation of each cluster (models output average) or estimating directly the regional generation through the model 6GNN (model inputs average). The first approach requires a greater computational effort with respect to the second. Moreover for operative use of the first method the actual capacity of each cluster should be provided while for the second method only the regional capacity is needed.

In this case, the model inputs average (6GNN) obtained an RMSE of 3% while the models output average (average of GNN output) achieved an RMSE of 3.2%. Thus, the simplest methodology was also the more accurate. From Fig. 7A it can be also observed that average RMSE of the clusters was 4.8% while the regional RMSE was 3.2%. The ensemble smoothing reduces the RMSE by 33% with respect to the mean clusters value.

Moreover Fig. 7B shows the error between the satellite GHI and the GHI ground measurements. The ground GHI was measured every 15 minutes by five reference cell placed in each cluster and was provided by local DSO. It can be noted that the estimation error is proportional to the satellite error with lower accuracy in Naturno and higher accuracy in Bolzano. Nevertheless, the site irradiance error (8–12% of G_0) is greatly reduced both on cluster (4–5.8% of P_n) and regional (3–3.2% of P_n) scales due to the smoothing effect. Since the regional estimation error is around 3%, the satellite derived irradiance could be used for real time power monitoring of the distributed photovoltaic production.

Fig. 8 shows an example of time series obtained by the two power estimation approaches. Greater estimation errors with respect to the observed power generation could be noted during variable days (15 and 16 of February 2015).

Fig. 9 reports the accuracy and the skill score (with respect to the RMSE of the PM) of the day-ahead regional power forecast obtained with the models output average approach (using the RHNN model) and the accuracy achieved by the model inputs average approach (using the PCARHNN model).

It can be observed that the two approaches lead to similar accuracy with an RMSE of 7.1%, thus the simplest model inputs average should be preferred. Moreover, also in this case, the RMSE on regional scale is higher than the mean cluster RMSE: 7.1% vs 8.1%. Thus the ensemble smoothing on cluster scale leads to a reduction of the regional RMSE of 12% with respect to the average cluster error. Furthermore, the regional skill score of 42.8% is higher than the skill score obtained by each cluster forecast (Fig. 8B).

Fig. 10 shows an example of time series obtained by the two day-ahead approaches. In this case, greater estimation errors with respect the observed power generation could be found during overcast and variable days (15, 16 and 17 of February 2015).

Fig. 11 shows the RMSE and MAE of regional PV power estimation (6GNN model) and intra-day and day-ahead forecast (PCARHNN model) vs forecast horizon, while in brackets the skill score (SS) with respect to the RMSE of the clear sky persistence (smart persistence) is reported.

First of all, comparing the 1 day-ahead simple and clear sky persistence from Figs. 7 and 8, it should be noted that the smart model improves the accuracy of almost 6% with respect the simple model. For a single site in Bolzano this improvement reaches 10% as reported in Pierro et al. (2016a).

Furthermore, power estimation model using satellite derived irradiance achieves 3% of RMSE and 2% of MAE. Intra-day forecast obtains a RMSE of 5%–7% (SS from −8% to 34%) and a MAE of 3%–4% (SS from 0.4% to 40%). Day-ahead forecast achieves a RMSE of 7% and 7.6% (SS from 39% to 45%) and MAE of 4%–5% (SS from 43% to 44%).

For horizon longer than 4 hours the intra-day forecast reduction of RMSE results in a lower accuracy with respect to the day-ahead forecast. Thus satellite derived data can be used to correct NWP up to 4 h

Table 1
accuracy metrics.

Acronym and formulae
$e_h = (PO^{for}(h) - PO^{obs}(h))/P_n(dd)$
$Root\ Mean\ Square\ Error = RMSE = \sqrt{\frac{\sum_{h=1}^{n} e_h^2}{n}}$
$Mean\ Absolute\ Error = MAE = \frac{\sum_{h=1}^{n} \lvert e_h \rvert}{n}$
$Mean\ Bias\ Error = MBE = \frac{\sum_{h=1}^{n} (e_h)}{n}$
$Skill\ Score = SS(RMSE) = 100\left(\frac{RMSE(PM) - RMSE(for)}{RMSE(PM)}\right)$%with respect to RMSE
$Skill\ Score = SS(MAE) = 100\left(\frac{MAE(PM) - MAE(for)}{MAE(PM)}\right)$%with respect to MAE

Where
$PO^{obs}(h)$ = hourly observed PV power output [kW]
$PO^{for}(h)$ = hourly forecast of PV power output [kW]
$P_n(dd)$ = Daily plant capacity [kWp]
n = number of sun hours .

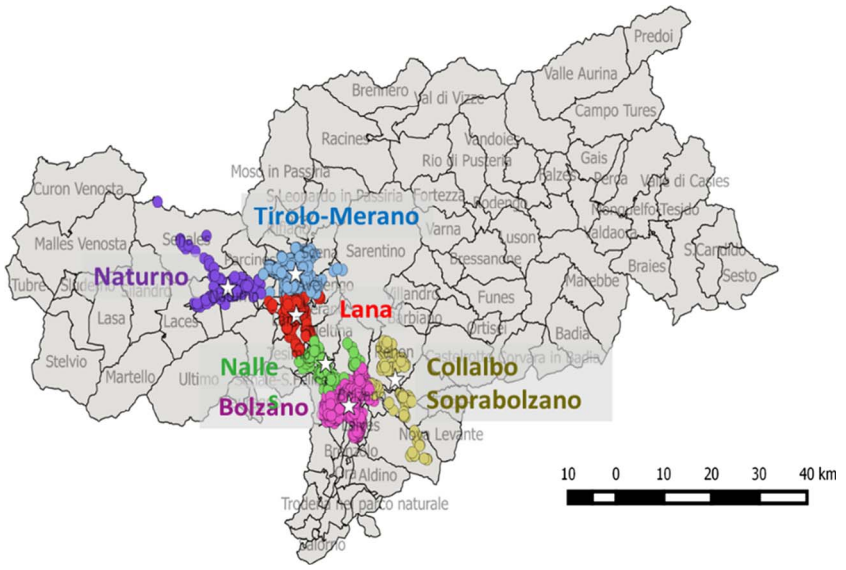

Fig. 6. PV plants spatial clustering.

Table 2
Regional and clusters PV capacity at the end of 2015.

Cluster name	PV capacity
Naturno	8.00 MWp
Tirolo-Merano	14.03 MWp
Lana	12.78 MWp
Nalles	5.82 MWp
Bolzano	24.42 MWp
Collalbo-Soprabolzano	3.19 MWp
Region	68.16 MWp

ahead. Similar result was found in Wolff et al. (2016) as well as in other papers. On the contrary, the lower the horizon the higher the persistence accuracy. Thus in this case for 1 h ahead the smart persistence provides a slightly better accuracy.

For intra-day PV forecast of two different region in Germany, a RMSE of 3.9%–4.3% of P_n with a skill score of 40%–42.3% were provided. In Wang et al. (2015) a RMSE of the best intra-day forecast on regional scale between 1.8% and 3.8% and a SS between 0% and 11.6% (to 1 up 4 h ahead) were found. The reachable RMSE on regional scale depends not only on the persistence but also on the size of the controlled areas. In this case the controlled area is quite small thus higher RMSE are reached: 5%–7%. Nevertheless the skill score from −8% to 34% are inside the state of art range.

For 1 day-ahead forecast in Germany, in Lorenz et al. (2011) the authors obtained an RMSE of 4.1%–4.3% of P_n with a skill score of 48%–52.8% (with respect to the simple persistence). Fonseca et al. (2014) found for a region in Japan a best RMSE of 10.24%. The same authors in Fonseca et al. (2014) found for four regions in Japan a RMSE between 6% and 7% with a SS between 50% and 60% depending from the region size. Once again in Fonseca et al. (2015) for three different areas of size from 32500 to 104996 km^2 a skill score between 56% and 54% (with respect to the simple persistence) was found. Finally in Zamo et al. (2014a) for two counties in France a RMSE of 6% and 5.8% with an upscaling and bottom-up approach was obtained. Thus, considering the controlled area of 800 km^2, the results presented here: RMSE of 7.1% and SS of 42.8% (with respect to the simple persistence) can be considered inside the state of art range of accuracy.

Moreover, because of smoothing effect, the accuracy of regional forecast leads to a reduction of RMSE between 30% and 50% with respect to the performance obtained with the forecast of a single PV plant generation (Lorenz et al., 2008; Fonseca et al., 2014). The RMSE of 7.1% obtained for the regional day-ahead forecast can be compared with the RMSE of 11.8% achieved in the forecast of the power output of an optimal tilted PV plant located in Bolzano (Pierro et al., 2016a). Thus the regional forecast provides a reduction of 40% with respect to the single power output site prediction, coherently with the literature results.

6.3. Ensemble smoothing effect

The ensemble smoothing effect is related to the number of systems in the controlled area (N_{plants}), the PV capacity distribution ($P_n^i with i = 1: N_{plants}$) and the pair correlation of the error (ρ^{ij}).

Considering that the forecast errors are usually almost unbiased, the RMSE of a PV fleet can be calculated as follows:

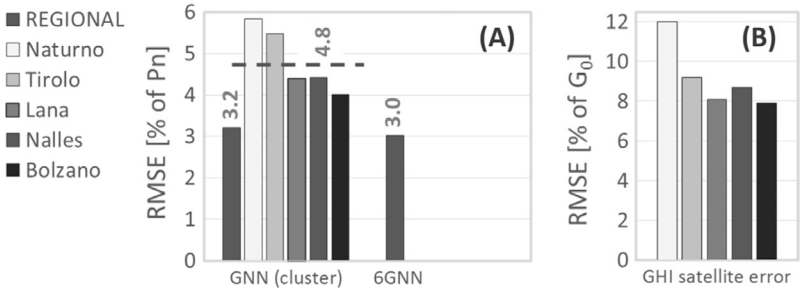

Fig. 7. (A) accuracy of the GNN model applied to each cluster (models output average approach) and accuracy of 6GNN model for regional PO estimation (model inputs average approach) (B) error between the hourly satellite GHI and the hourly GHI ground measurements (where Go = 1000 W/m^2).

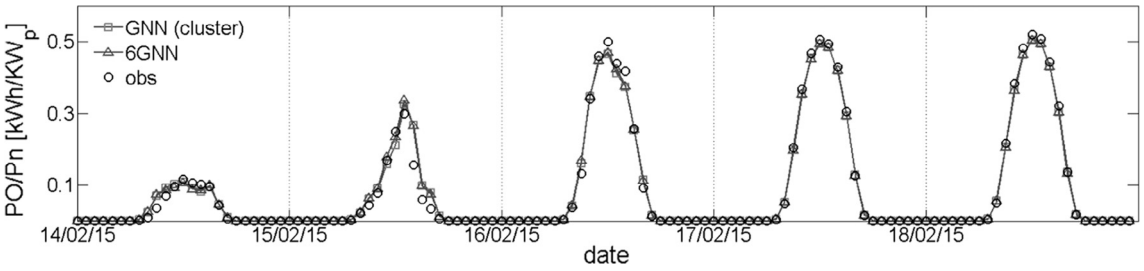

Fig. 8. Example of power estimation trend for five days of February 2015.

$$RMSE_e = \frac{1}{P_n^e}\sqrt{\frac{1}{N_h}\sum_{h=1}^{N_h}(PO_e^{for}-PO_e^{obs})^2} = \frac{1}{P_n^e}\sqrt{\frac{1}{N_h}\sum_{h=1}^{N_h}\left(\sum_{i=1}^{N_{plants}} e_i\right)^2}$$
$$= \frac{1}{P_n^e}\sqrt{\frac{1}{N_h}\sum_{h=1}^{N_h}\left(\sum_{i=1}^{N_{plants}} e_i\right)\left(\sum_{j=1}^{N_{plants}} e_j\right)}$$
$$= \frac{1}{P_n^e}\sqrt{\sum_{i=1}^{N_{plants}}\sum_{j=1}^{N_{plants}}\left(\frac{1}{N_h}\sum_{h=1}^{N_h} e_i e_j\right)} \cong \frac{1}{P_n^e}\sqrt{\sum_{i=1}^{N_{plants}}\sum_{j=1}^{N_{plants}}(\sigma_i\sigma_j\rho^{ij})}$$
$$\cong \sqrt{\sum_{i=1}^{N_{plants}}\sum_{j=1}^{N_{plants}}\left(\frac{P_n^i P_n^j}{(P_n^e)^2}RMSE_i RMSE_j\rho^{ij}\right)} \quad (8)$$

where N_h is the number of sun hours, $RMSE_e$ is the root mean square error of the forecast of PV plant ensemble, $P_n^e = \sum_{i=1}^{N_{plants}} P_n^i$ is the total PV capacity, e_i and σ_i are the forecast error and the error standard deviation of each plant of the fleet.

If the forecasting errors of the plants are completely correlated ($\rho^{ij} = 1$ *thus* $RMSE_i = RMSE$), the error of the ensemble is: $RMSE_e = RMSE\sqrt{\sum_{i=1}^{N_{plants}}\sum_{j=1}^{N_{plants}}\left(\frac{P_n^i P_n^j}{(P_n^e)^2}\right)}$ so that it is equal to the RMSE of each plant if all the systems have the same capacity ($P_n^e = N_{plants}P_n$). In this case no smoothing effect can be observed. On the contrary if the prediction error are perfectly uncorrelated ($\rho^{ij} = \delta^{ij}$) the error of the ensemble is:$RMSE_e = \sqrt{\sum_{i=1}^{N_{plants}}\left(\frac{P_n^i}{P_n^e}\right)^2 MSE_i}$ so that it is equal to $\frac{\sqrt{\langle MSE_i\rangle_{ensembe}}}{\sqrt{N_{plant}}}$ if all the systems have the same capacity. Thus in this case, the RMSE of the ensemble will decrease as $1/\sqrt{N_{plant}}$ and the large number value of smoothing effect is reached.

In real case, error correlation between sites depends on the distance between sites (or controlled area size), the forecast horizon and, the speed of clouds in the region. In particular, in Lorenz et al. (2009b). It was showed that the mean correlation of the forecasting error between two different site decreased exponentially with the distance as:

$\rho^{ij} = e^{a(d_{ij})^b}$ so that the RMSE of the ensemble decreases with the size of the controlled area following the same law.

To study the smoothing effect on a cluster scale, the RMSE of different ensembles of adjacent clusters was calculated. Fig. 12A and B shows the estimation and forecast error as a function of the maximum distance between the centroids of each ensemble. Fig. 12C reports the mean pair-correlation between forecast errors of all the adjacent clusters belonging to the same ensemble. Also on a cluster scale the correlation as well as the RMSE of the ensembles decrease exponentially with the size of the area (maximum distance between centroids). In this case, the pair-correlation between cluster errors shows a slower decay with respect to the correlation between two single plants errors (Lorenz et al., 2009b). Indeed, the meteorological conditions at clusters scale remains strongly correlated for larger distances with respect to the single points.

Fig. 12C also reports the error reduction factor of the forecast and persistence models and Fig. 12D shows the skill score. It can be observed that the persistence factor decreases with the distance with lower rate than the forecast factor while the skill score increases linearly of 0.1% per km. The skill score is usually calculated to compare the forecast accuracy achieved in different sites or in different years since it doesn't depend on the irradiance variability. In case of plants with similar characteristics it remains almost constant between different sites or years) (Pierro et al., 2016b). On the contrary, for the regional

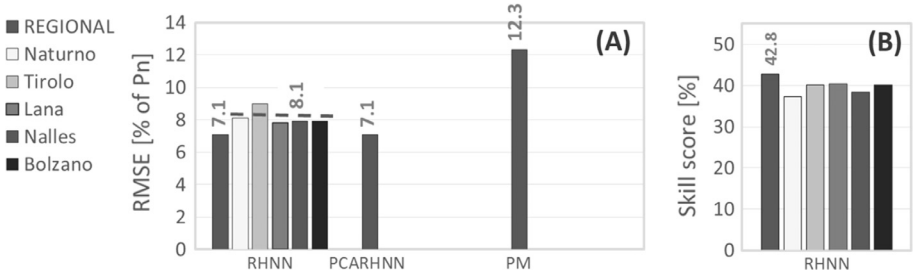

Fig. 9. Accuracy (A) and skill score (B) of the RHNN model apply to each cluster (models output average approach), accuracy of PCARHNN model for 1 day-ahead regional power forecast (model inputs average approach) and accuracy of the simple persistence model (A).

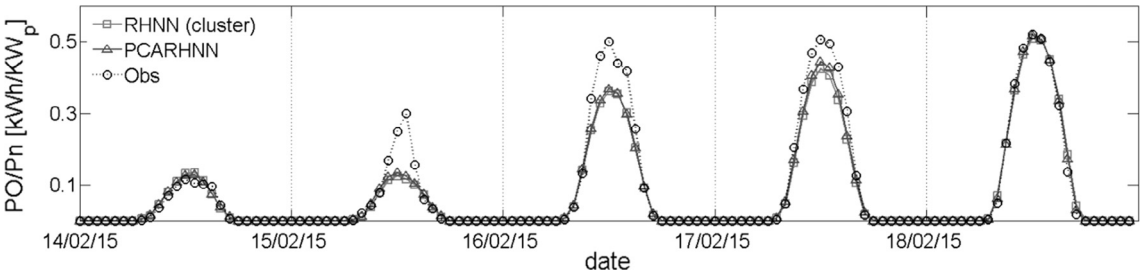

Fig. 10. Example of day ahead forecast trend for five days of February 2015.

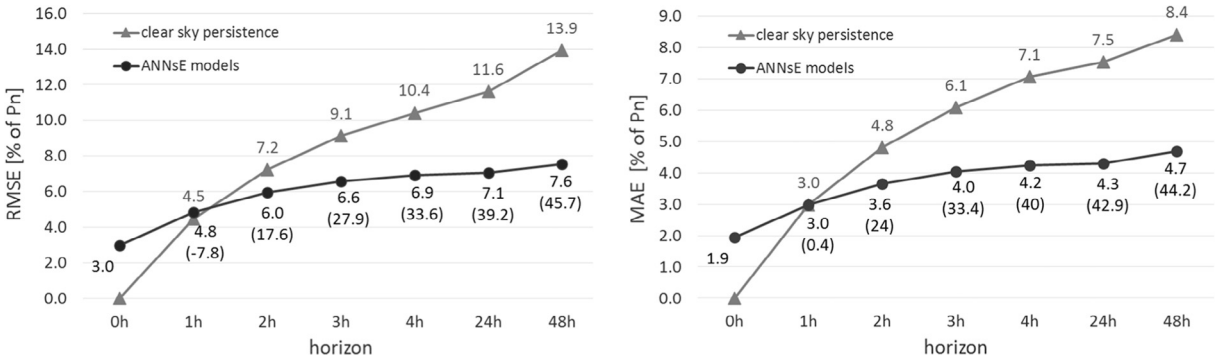

Fig. 11. RMSE and MAE of regional PV power estimation (6GNN model) and intra-day and day-ahead forecast (PCARHNN model) vs forecast horizon (in brackets the skill score with respect to the RMSE and MAE of the clear sky persistence).

forecast the skill score increases linearly with the controlled area, thus the reduction of RMSE is achieved not only with the decrease of solar variability (RMSE of persistence model) but also with the improvement of forecast capability of the models due to the smoothing effect.

Finally, the fits of the mean RMSE (all with coefficients of determination R^2 greater than 0.985) allow the extrapolation of the $RMSE_e$ of larger ensemble areas. For power estimation, the $RMSE_e$ could be reduced to 1% of P_n for ensemble areas with maximum distance between cluster centroids around 150 km. For day-ahead power forecast a $RMSE_e$ of 4% of P_n and a skill score of 60% could be achieved with distance around 200 km. This confirms that the de-correlation distance, for hourly values, is around 100–200 km (Perez et al., 2011; Hoff and Perez, 2012). Moreover the extrapolation proves that the obtained accuracy is in the "state of the art" range. Indeed, as reported in the previous subsection, in Lorenz et al. (2011) an $RMSE_e$ of 4.1%–4.3% of Pn with a skill score of 48%–52.8% was obtained for distances around 500 and 750 km while in Fonseca et al. (2014) a $RMSE_e$ between 6% and 7% with a skill score between 50% and 60% was achieved for distances between 200 and 400 km.

6.4. Prediction intervals

Fig. 13A shows the standard deviation of the forecast error predicted by the model described in Section 3.5. The maximum standard deviation values are reached near noon with variable irradiance corresponding to PPK_{cs} between 0.3 and 0.6.

Fig. 13B shows the reliability plot i.e. the frequency of observation that lays inside of each prediction interval versus the respective confidence levels (expected probability). The frequency of observation was evaluated during the years 2015. A prediction of the forecast errors is completely reliable if the observed frequency is equal to the corresponding confidence level (gray dash line in Fig. 13B). Fig. 13C reports the probability distribution function (PDF) of the error normalized by the predicted standard deviation ($Ne = e/\sigma^{for}$) of the year 2015 (black line) and the standard normal distribution (red line).

In the present case, the model provides an overestimation of the prediction interval since the observed frequency is almost always greater than the expected one (Fig. 13B). Indeed, the PDF of the normalized error is considerably different from the standard normal distribution and the probability of $|Ne| \leqslant Z_{\mp\alpha/2}$ is greater than $(1-\alpha)$. As mentioned in Section 3.5, the normal distribution of forecast error is a

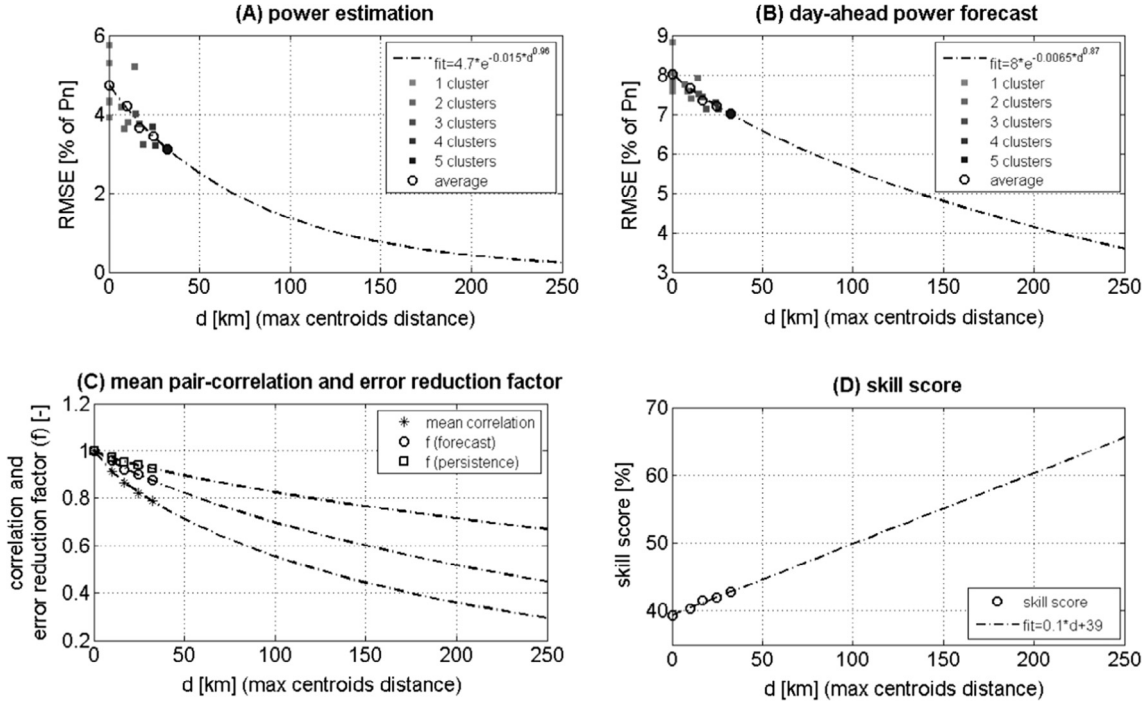

Fig. 12. (A) and (B) RMSE of estimation and day-ahead forecast of different ensembles of adjacent clusters; (C) mean pair-correlation and error reduction factors of forecast and persistence models; (D) skill score with respect to the RMSE of the persistence model.

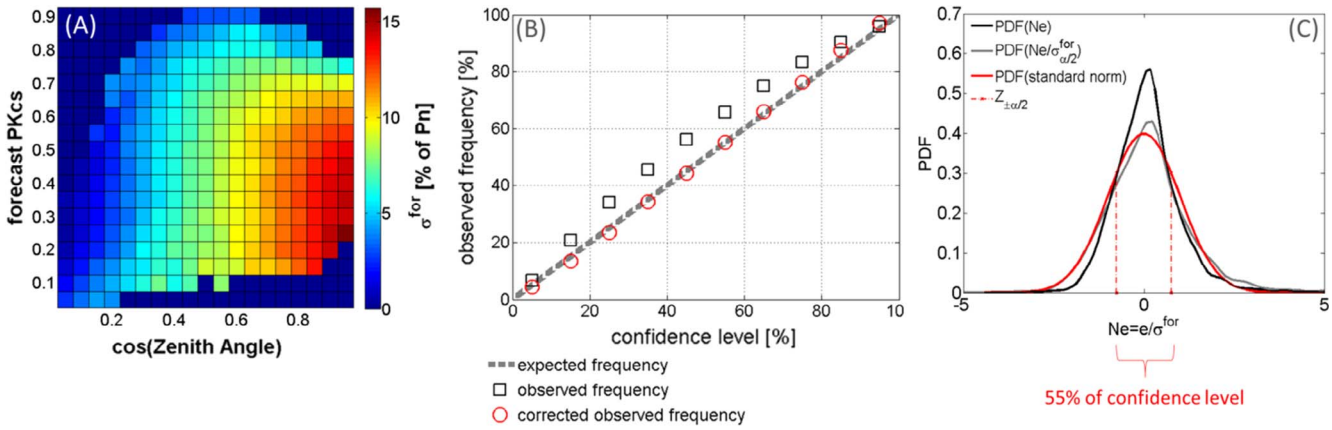

Fig. 13. (A) predicted standard deviation of the forecast error; (B) reliability plot; (C) PDF of the error normalized by the predicted standard deviation.

too strong hypothesis. It can be adopted, as in Lorenz et al. (2009b) or Marquez and Coimbra (2011), only to provide the 95th percentile, since the observed PDF of the error is contained inside the standard normal one. Therefore, the observed frequency is equal to the expected 95% of confidence (see the black squares in Fig. 13B). On the contrary, this assumption does not work as well for the prediction intervals corresponding to different confidence levels. Nevertheless, for each confidence level it was possible to predict a correction coefficient ($\sigma^{for}_{\alpha/2}$) so that:

$$P\left(\left|\frac{Ne}{\sigma^{for}_{\alpha/2}}\right| \leqslant Z_{\alpha/2}\right) = P\left(\left|\frac{e}{\sigma^{for}\sigma^{for}_{\alpha/2}}\right| \leqslant Z_{\alpha/2}\right) \cong 1-\alpha. \tag{9}$$

In this way, the forecast of the standard deviation of the error depends not only on the zenith angle and the pseudo clear sky performance index but also on the considered confidence level: $\sigma^{for}(\vartheta_{zenith},PPK_{cs},\alpha) = \sigma^{for}_{\vartheta_{zenith},PPK_{cs}}\sigma^{for}_{\alpha/2}$. The correction was calculated on the PDF of *Ne* of the year 2014 and tested on the year 2015. Fig. 11C also shows the PDF of ($Ne/\sigma^{for}_{\alpha/2}$) of the test year (gray line). It can be observed that integral between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ of the modified PDF is now almost equal to integrals of the standard normal distribution. Fig. 13B also reports the reliability plot of the corrected prediction intervals. This good result was possible since the PDF of *Ne* of different years (2014 and 2015) are very similar, so that the corrections factors calculate on the PDF of the year 2014 are effective also for the year 2015.

Fig. 14 reports the trend of the prediction interval for five days of February 2015. It can be observed that the extent of the interval is reduced when passing from overcast to clear sky days.

7. Summary and conclusion

A new upscaling method was developed and used for estimation and forecast of the PV distributed generation of 1985 PV plants in a small area of the South Tyrol region (800 km^2) in Italy. It was based on spatial clustering of the PV fleet and neural networks models making use of satellite or NWP data. Two different approaches were investigated. In the first approach, estimation and forecast of the power generation of each cluster were averaged to obtain the regional prediction (models output average). The second approach provided directly the regional power prediction using inputs centered on each cluster centroid (model inputs average). In this case, the model inputs average gave slightly better results.

This study allowed a complete assessment of the forecast accuracy at different horizons, from 0 h ahead (power estimation) to 48 h ahead. The power estimation model achieved 3% of P_n in RMSE and 2% in MAE. Intra-day forecast (from 1 to 4 h) obtained a RMSE of 5%–7% and a MAE of 3%–4%. The skill score with respect to the accuracy of the smart persistence model was between −8% and 34% in RMSE and from 0.4% to 40% in MAE. One and two days-ahead forecast provided a RMSE of 7% and 7.5% (with a skill score of 39% and 45%) and a MAE of 4% and 5% (with a skill score of 43% and 44%).

The model inputs average is the simplest upscaling strategy that requires the lower computational effort and needs very few input information that should be provided by users. It could be easily adopted by forecast providers to deliver the regional PV power estimation and mid-term forecast to DSO, TSO, energy traders, and aggregators.

The models output average was also used to study the smoothing effect on cluster scale. This effect improved the accuracy of regional power estimation and forecast with respect to the mean clusters value. It reduced the RMSE of power estimation of 33% and the RMSE of day-ahead forecast of 12%. It was found that each cluster behaved like a single plant since the accuracy improvement due to the ensemble smoothing followed the same exponential law found for the ensemble of plants. Extrapolating the accuracy improvement on wider areas, the developed upscaling methods brought to an accuracy comparable with the one found in literature for areas greater than 200 × 200 km^2.

Also persistence followed the same exponential law but it showed a slower decrease of error with respect to the forecast. It appeared that the skill score increased linearly with the size of the region with a rate

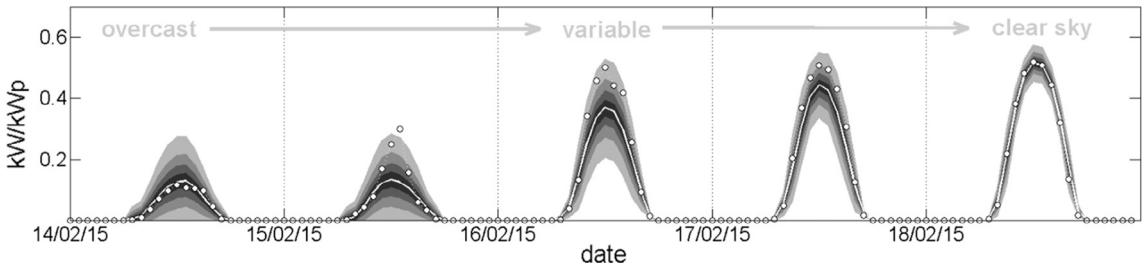

Fig. 14. Example of prediction interval trend for five days of February 2015. The gray colours correspond to different confidence levels: 95%–75%–50%–25%; from the clearest gray showing the interval with 95% of confidence to the darker gray showing the interval with 25% of confidence. The dots represent the observed values while the white line is the forecast.

of 0.1% per km. Thus, the wider is the considered area the greater is the benefit of using a more complex forecast model instead of persistence.

The ensemble smoothing of power estimation is much more effective compared to the PV forecast, since the accuracy increases with the increasing of the area faster than the accuracy of the forecast.

A model to estimate the forecast error was also developed. It is based on an ensemble neural network model that forecasts the standard deviation of the error coupled with a probabilistic correction that takes in to account deviations of the PDF of the error from the normal distribution. Indeed using a parametric probabilistic approach, the prediction intervals are usually estimated supposing a normal distribution of the forecast errors but this is a poor assumption. This hypothesis provides a correct estimation of the 95th quantile but is not reliable for other confidence levels. The proposed model allows a very reliable computation of the prediction intervals so that the frequency of the observations that falls inside each interval is almost equal to the associate confidence level. Thus prediction intervals can be employed not only to reduce the energy reserve (by the use of 95th quantile of the forecast errors) but also to estimate the probability of a specific solar energy bid on the energy markets.

Acknowledgement

The authors acknowledge the Distributor System Operator Edyna Srl, for providing the data and for co-funding the work. A special thanks to the technical director Andreas Bordonetti for the trust granted to Eurac and to its staff Tschager David, Bruno Fasoli and Danilo Pederiva for the useful discussions.

References

Alet, P.-J., 2015. Photovoltaics merging with the active integrated grid: a white paper. European Photovoltaic Technology Platform.

Alet, P.-J., Efthymiou, V., Graditi, G., Henze, N., Juel, M., Moser, D., et al., 2016. Forecasting and observability: critical technologies for system operations with high PV penetration. In: Proceedings of the 32nd EuropeanPhotovoltaic Solar Energy Conference and Exhibition, Paris, pp. 1444–1448.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F., Antonanzas-Torres, F., 2016. Review of photovoltaic power forecasting. Sol. Energy 136, 78–111.

Azimi, R., Ghayekhloo, M., Ghofrani, M., 2016. A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting. Energy Convers. Manage. 118, 331–344.

Basheer, I., Hajmeer, M., 2000. Artificial neural networks: fundamentals, computing, design, and application. J. Microbiol. Methods 43 (1), 3–31.

Cornaro, C., Pierro, M., Bucci, F., 2015. Master optimization process based on neural networks ensemble for 24-h solar irradiance forecast. Sol. Energy 111, 297–312.

Emmanuel, M., Rayudu, R., 2017. Evolution of dispatchable photovoltaic system integration with the electric power network for smart grid applications: A review. Renewable Sustainable Energy Rev. 67, 207–224.

EUMETSAT, 2017. Operational Services Specitication. EUMETSAT.

Fonseca, J.G., Oozeki, T., Ohtake, H., Ichi Shimose, K., Takashima, T., Ogimoto, K., 2014. Regional forecasts and smoothing effect of photovoltaic power generation in Japan: An approach with principal component analysis. Renewable Energy 68 (403–413).

Fonseca, J., Oozeki, T., Ohtake, H., Takashima, T., Ogimoto, K., 2015. Regional forecasts of photovoltaic power generation according to different data availability scenarios: a study of four methods. Prog. Photovoltaics: Res. Appl. 23 (10), 1203–1218.

Hoff, T., Perez, R., 2012. Modeling PV fleet output variability. Sol. Energy 86, 2177–2189.

IEA, 2013. Photovoltaic and Solar Forecasting: State of the Art. IEA PVPS Task 14.

IEA, 2014a. Snapshot of Global PV Markets. IEA PVPS Task1.

IEA, 2014b. Technology Roadmap Solar photovoltaic energy. IEA Renewable Energy Division.

Jiménez-Pérez, P., Mora-López, L., 2016. Modeling and forecasting hourly global solar radiation using clustering and classification techniques. Sol. Energy 135, 682–691.

Kaur, A., Nonnenmacher, L., Coimbra, C., 2016. Net load forecasting for high renewable energy penetration grids. Energy 114, 1073–1084.

Kleissl, J., 2013. Solar Energy Forecasting and Resource Assessment. Academic Press.

Lima, F., Martins, F., Pereira, E., Lorenz, E., Heinemann, D., 2016. Forecast for surface solar irradiance at the Brazilian Northeastern region using NWP model and artificial neural networks. Renewable Energy 87, 807–818.

Lorenz, E.H., 2012. Local and regional photovoltaic power prediction for large scale grid integration: assessment of a new algorithm for snow detection. Prog. Photovolt 20, 760–769.

Lorenz, E., Hurka, J., Karampela, G., Heinemann, D., Beyer, H.S., 2008. Qualified forecast of enesemble power production by spatially dispersed gri-connected PV systems. In: 23rd EU PVSEC section 5AO.8.6.

Lorenz, E., Remund, J., Muller, S.C., Traunmull, W., Steinmaurer, G., Pozo, D., et al., 2009a. Benchmarking of different approaches to forecast solar irradiance. In: European Photovoltaic Solar Energy Conference. Hamburg, Germany, pp. 4199–4208.

Lorenz, E., Hurka, J., Heinemann, D., Beyer, H.G., 2009b. Irradiance forecasting for the power prediction of grid connected photovoltaic systems. IEEE J. Selected Topics Appl. Earth Observations Remote Sensing 1 (2), 2–10.

Lorenz, E., Scheidsteger, T., Hurk, J., Heinemann, D., Kurz, C., 2011. Regional PV power prediction for improved grid integration. Prog. Photovoltaics: Res. Appl. 19, 757–771.

Lorenz, E., Heinemann, D., Kurz, C., 2012. Local and regional photovoltaic power prediction for large scale grid integration: assessment of a new algorithm for snow detection. Prog. Photovolt 20 (6), 760–769.

Lorenz, E., Kühnert, J., Heinemann, D., Nielsen, K., Remund, J., Stefan, C., 2016. Comparison of global horizontal irradiance forecasts based on numerical weather prediction models with different spatio-temporal resolutions. Prog. Photovoltaics: Res. Appl. 24 (12), 1626–1640.

Marquez, R., Coimbra, C., 2011. Forecasting of global and direct solar irradiance using stochastic learning methods , ground experiments and the NWS database. Sol. Energy 85 (5), 746–756.

Mellit, A., 2008. Artificial intelligence technique for modelling and forecasting of solar radiation data: a review. Int. J. Artif. Intell. Soft Comput. 1 (1), 52–76.

Munshi, A., Mohamed, Y., 2016. Photovoltaic power pattern clustering based on conventional and swarm clustering methods. Sol. Energy 124, 39–56.

Paulescu, M., Paulescu, E., Gravila, P., Badescu, V., 2012. Weather modeling and forecasting of PV systems operation. Springer Science & Business Media.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philos. Mag. 2 (11), 559–572.

Pelland, S., Galanis, G., Kallos, G., 2011. Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model. Prog. Photovolt: Res. Appl. 21 (3), 284–296.

Perez, R., Hoff, T., Kivalow, S., 2011. Spatial and Temporal Characteristics of Solar Radiation Variability. International Solar Energy (ISES) World Congress, Kassel, Germany.

Perez, R., Hoff, T., 2013. Solar resource variability. In: Kleissl, J., (Ed.), Solar Energy Forecasting and Resource Assessment. first ed. Academic Press, Waltham, pp. 133–148.

Pierro, M., Bucci, F., Cornaro, C., Maggioni, E., Perotto, A., Pravettoni, M., et al., 2015. Model output statistics cascade to improve day ahead solar irradiance forecast. Sol. Energy. 117, 99–113.

Pierro, M., Bucci, F., De Felice, M., Maggioni, E., Perotto, A., Spada, F., et al., 2016a. Deterministic and stochastic approaches for day-ahead solar power forecasting. J. Sol. Energy Eng. 139 (2), 021010.

Pierro, M., Bucci, F., De Felice, M., Maggioni, E., Perotto, A., Spada, F., et al., 2016b. Multi-model ensemble for day ahead prediction of photovoltaic power generation. Sol. Energy 134, 132–146.

Raza, M., Nadarajah, M., Ekanayake, C., 2016. On recent advances in PV output power forecast. Sol. Energy 136, 125–144.

Saint-Drenan, Y.M., Gooda, G.H., Brauna, M., Freisingerc, T., 2016. Analysis of the uncertainty in the estimates of regional PV power generation evaluated with the upscaling method. Sol. Energy 135, 536–550.

Shivashankar, S., Mekhilef, S., Mokhlis, H., Karimi, M., 2016. Mitigating methods of power fluctuation of photovoltaic (PV) sources – A review. Renewable Sustainable Energy Rev. 59, 1170–1184.

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., 2008. A description of the Advanced Research WRF version 3. NCAR Tech. Tech. rep., Note NCAR/TN-4751STR.

Wang, F., Zhen, Z., Mi, Z., Sun, H., Su, S., Yang, G., 2015. Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting. Energy Build. 86, 427–438.

Wolff, B., Kühnert, J., Lorenz, E., Kramer, O., Heinemann, D., 2016. Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. Sol. Energy 135, 197–208.

Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., 2014a. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production part I: Deterministic forecast of hourly production. Sol. Energy 105, 792–803.

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

Zhang, G., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks: the state of the art. Int. J. Forecasting 14 (1), 35–62.

Zhang, J., Hodge, B., Lu, S., Hamann, H., Lehman, B., 2015. Baseline and target values for regional and point PV power forecasts: Toward improved solar forecasting. Sol. Energy 122, 804–819.