

A Machine Learning Model for Long-Term Power Generation Forecasting at Bidding Zone Level

Michela Moschella, Mauro Tucci, Emanuele Crisostomi

Department of Energy, Systems, Territory
and Constructions Engineering, University of Pisa
Pisa, Italy
michela.moschella@ing.unipi.it

Alessandro Betti

i-EM s.r.l.
Livorno, Italy
alessandro.betti@i-em.eu

Abstract—The increasing penetration level of energy generation from renewable sources is demanding for more accurate and reliable forecasting tools to support classic power grid operations (e.g., unit commitment, electricity market clearing or maintenance planning). For this purpose, many physical models have been employed, and more recently many statistical or machine learning algorithms, and data-driven methods in general, are becoming subject of intense research. While generally the power research community focuses on power forecasting at the level of single plants, in a short future horizon of time, in this time we are interested in aggregated macro-area power generation (i.e., in a territory of size greater than 100000 km²) with a future horizon of interest up to 15 days ahead. Real data are used to validate the proposed forecasting methodology on a test set of several months.

I. INTRODUCTION

A. Motivations

As the penetration level of Renewable Energy (RE) sources is growing worldwide to meet ever tightening sustainability goals [1], the intermittent and uncertain nature of RE is posing increasing challenges to efficiently manage a power grid, eventually endangering its own stability. In this context, the availability of accurate forecasts of power generation from RE may mitigate the impact of the increasing penetration level and improve the operation of power systems [2]. In particular, in this paper we are interested in developing long-term RE power generation forecasting algorithms, up to 15 days ahead for aggregated areas. Such a long horizon of time ahead is convenient for maintenance scheduling, for planning tactic upgrades or for planning switching on/off of big conventional plants when future power generation from RE is expected to be particularly low or high, with respect to the load. In addition, we focus on forecasting algorithms operating at aggregated level, where a region here is a bidding zone in Italy (see Fig. 1). We aggregate data in terms of bidding zones since many measured and predicted data are available for this aggregation level (e.g., load consumption predicted and measured data are available at the Terna (Italian TSO) website¹).

B. State of the Art

There is a rich literature on power generation from RE, usually differing in terms of the future horizon of prediction that may range from 1 second to 6 hours, from 6 hours up to day ahead, and from 2 days ahead or longer, which correspond to the *intra-day*, the *day-ahead* and the *long-term* models, respectively. In this case a general overview of existing forecast methods can be found in [3] and [4] for Photovoltaic (PV) and Wind (WD) power generation, respectively.

We can also distinguish models with different data geographical resolution; the most popular spatial aggregation is the *pointwise* one, where data refer to a single power plant, but there is also a type of forecasting regarding regions and macro-areas; in the last case, the problem is completely different since the size and the location of all plants in a given region are not generally known.

Very few works make forecasts at large-scale regional areas; most focus on power generation from single PV or wind plants, that are successively aggregated following the so-called *up-scaling* method [5]. A similar paper is [6], where however only the PV case is considered, and for much smaller scales than in our case and only up to 2 days ahead. Another classic paper on this topic is [7], that however did not use information on energy production from Renewable Energy Sources (RES). This piece of information is on the other hand available in our case study with hourly resolution, thus allowing us to propose more accurate forecasting models.

C. Contributions

While plenty of papers have been written for power generation from RE, our problem is rather a peculiar one because we are trying to predict power generation in wide areas (in some cases greater than 100000 km²) for a long horizon of time. The problem is particularly challenging since we do not know where power generation plants are exactly located in such areas, nor their nominal size. Thus, pure data-driven methodologies, namely k-Nearest Neighbours (k-NN) and Quantile Regression Forest (QRF), are used to predict power generation on the basis of large historical data-sets.

While not too accurate results are obtained, especially for a horizon of forecast larger than 5 days, still the forecasting

¹<https://www.terna.it/en-gb/sistemaelettrico/transparencyreport/load.aspx>



Fig. 1. Bidding zones of Italy, identified by Terna.

results may be accurate enough to support classic power grid operations (e.g., maintenance planning).

II. DATA SET DESCRIPTION

In our problem we want to forecast power generated from solar and wind sources at an aggregated level (i.e. Italian bidding zones), using meteorological data as input variables (see Tab. I for a brief summary of used variables).

A. Power generation data

Power generation data are available from the aforementioned Terna website². In particular, we have at our disposal *hourly data* for each Italian bidding zone (indicated with acronyms NORD, CNOR, CSUD, SUD, SICI and SARD, as shown in Fig. 1), where the hourly value is the average power over the previous hour.

The previous data are known to contain some inaccurate information. In the pre-processing stage illustrated in Sec. III we also use the monthly power generation data (aggregated at national level) that is known to be more accurate.

B. Meteorological data

For the prediction of power generated from PV plants, we exploit the hourly satellite Global Horizontal Irradiance (GHI) and Global Horizontal Irradiance in Clear Sky conditions (GHI_{CS}) from 2015 on, provided by Fly s.r.l.³ (for more details on data validation see [8]), as well as the forecasting meteorological GHI provided by Aeronautica Militare (AM)⁴ from 2016 onwards. For WD farms we use the two components of the wind speed, i.e. the west-to-east component (UGRD), and the south-to-north one (VGRD).

Remark 1: Forecast meteorological data provided by AM come from two different models, depending on the forecast horizon they refer to. In particular,

- data referred to the forecast interval $[+0h, +72h]$, are outputs of the model *COSMO-ME*⁵, a local model on the south-central Europe and Mediterranean Sea;

- for data about the horizon $[+75h, +240h]$, the global IFS model⁶ (of ECMWF) is adopted.

Unfortunately, available forecast data do not cover the whole forecast horizon of 15 days, and consequently we have to fill such missing values; a simple *persistence* technique is used (i.e., the forecast of the last day is kept constant for the following days as well).

As the previous meteorological data (available in a *raster* format) cover wide areas, to compress the size of the input data, we aggregate the meteorological data to the level of Provinces (110 in Italy).

TABLE I
MODEL VARIABLES

Module	Variable	Time resolution	Spatial resolution
PV	GHI	1 hour	Province
	GHI_{CS}	1 hour	Province
	PV generated power	1 hour	bidding zone
	PV generated power	1 month	Italy
WD	UGRD	1 hour	Province
	VGRD	1 hour	Province
	WD generated power	1 hour	bidding zone
	WD generated power	1 month	Italy

III. METHODS

Following the notation of [9], we group variables into *predictor* and *response* variables classes. As in a *black box* identification procedure, we use machine learning algorithms to identify an unknown and most likely, non-linear, function that maps the predictors vector x into the response variable y , i.e.

$$y = f(x). \quad (1)$$

In our case, predictors correspond to the meteorological variables described in Sec. II-B, while the generated power is the response variable of the model. We use two models, one for PV and one for WD power generation forecasting, and in both cases data are divided in two parts: a *training set* with measured data, used to learn Eq. (1), and the *test set* with forecast data, exploited for evaluation and comparison.

A. Preprocessing and Missing data

A preprocessing phase is fundamental to remove outliers and prevent an identification methodology from learning wrong patterns in the training set [10].

First, the hourly values of the power time-series are proportionally scaled using the more accurate total national monthly generation. Additionally, for PV instances, outliers are identified as those irradiance-power pairs that fall out of a safety cone in the (\overline{GHI}, P) plane, where \overline{GHI} is the average satellite irradiance over each bidding zone, and P is the related hourly power, inspired by the procedure outlined in [11].

Very few data are missing in our data-set. Cubic spline interpolation is used to fill such gaps.

²<https://www.terna.it/SistemaElettrico/TransparencyReport/Generation/Ex-postdataontheactualgeneration.aspx>

³<http://www.flyby.it/>

⁴<http://www.meteoam.it/modelli-di-previsione-numerica>

⁵<http://www.cosmo-model.org/content/model/general/default.htm>

⁶<https://www.ecmwf.int/en/forecasts/datasets/set-i>

B. Machine Learning algorithms

A cooperative *ensemble* of *k-NN* and *QRF* methods is used to forecast power generation.

1) *k-NN*: *k-NN* is a parametric method based on the assumption that a given weather forecast is most likely to yield a power generation close to past instances with similar weather conditions. The k most similar samples in the past are selected and the corresponding historical powers are combined, with a weight depending on the similarity degree. Parameters of the *k-NN* are the number k of nearest neighbours, the distance metric, and the kernel used for weights modeling. In our case, Euclidean metric is used for the distance and hyperbolic kernel is exploited for the combination of k nearest neighbours [2]. Additionally, for the PV case we exploit the periodicity in the irradiance time series by selecting in the past the k nearest neighbours of a similar hour and month of the specific power sample to be predicted. In particular, if M and H are the month and the hour of the test sample to be predicted, respectively, only training instances with month and hour *close* to M and H are selected.

A summary of inputs used for the algorithm is shown in Table II.

2) *QRF*: The *QRF* algorithm is a variant of *Random Forest* (RF) developed by Breiman [12] which, unlike conventional RF, takes track of all the target samples and not just their average. RF is a collection of decision trees that are combined together to enhance the predictive capability of a single tree, by approximating the mean conditional distribution of the response variable. On the other hand, *QRF* [13] provides the full conditional distribution of the response variable. In particular, assuming that Y is the target variable and X the vector of predictors, then the final goal is finding the relationship between X and Y . A conventional RF estimates the conditional mean of the target Y , given the attribute $X = x$. Instead the *QRF*, given a certain level α ($0 < \alpha < 1$), estimates the conditional quantile

$$Q_\alpha(x) = \inf\{y : F(y|X = x) \geq \alpha\}, \quad (2)$$

where $F(y|X = x)$ is the conditional distribution function of Y given $X = x$. Consequently, *QRF* is a substantial improvement of conventional RF, because the α -quantile $Q_\alpha(x)$ gives a more complete information about the distribution of Y than the output of the conditional mean provided by RF [13]. The input variables used for this algorithm are shown in Table II.

TABLE II
INPUTS OF THE *k-NN* AND *QRF* ALGORITHMS

Algorithm	PV	WD
k-NN	GHI, GHI_{CS}	wind speed
QRF	GHI, GHI_{CS} , Month, Hour	UGRD, VGRD

C. PV model: post-processing phase

Only in the PV model, we post-process the *k-NN* and *QRF* before combining them in an ensemble.

The rationale for doing so is the dependence of power generation on year and season, so same irradiance values may correspond to different generated powers, at different time of the years.

The main steps of post-processing are the following:

- 1) get the mean ratio Q_{for} between the power generation forecasts referring to peak values of GHI and the related GHI values, in the interval of prediction (e.g. 360 h);
- 2) compute the mean ratio between measured power and irradiance data (corresponding to peak values of GHI) of the n weeks before the test set in training set (Q_{train});
- 3) finally compute the factor $K_{prod} = \frac{Q_{train}}{Q_{for}}$ and scale the prediction of the generated power by the same factor.

It is worth remarking that this rescaling is adequate only if the training set is immediately preceding the test period, otherwise the ratio Q would change and such post-processing would not be convenient.

D. Model tuning

Some hyperparameters need to be tuned for the procedure, such as:

- the threshold for removing outliers in the bi-variate pre-processing phase of the PV module;
- the level q of the quantile in the *QRF* algorithm;
- the number k of nearest neighbours in the *k-NN* algorithm;
- the number n of weeks in the post-processing of Sec. III-C.

In our case study we use a trial and error method on the training set.

IV. RESULTS AND DISCUSSION

We now validate our model performances on a test period of 6 months, from January to June 2017. We adopt a *semi-moving* window technique to select the training period: the start date is fixed (1st May 2015), whereas the end date is the day before the test period.

We evaluate the performances every month by considering 2 different error metrics: the Normalized Mean Bias Error (NMBE) and the Normalized Root Mean Squared Error (NRMSE). If we denote by \hat{y} the prediction of our model, and by y the actual generated power, if $e_i = \hat{y}_i - y_i$ is the forecasting error of the i -th hour within the horizon of forecast, then the error metrics are defined as:

$$(i) \text{ NMBE} = \frac{1}{N} \sum_{i=1}^N \frac{e_i}{M_m};$$

$$(ii) \text{ NRMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{e_i}{M_m} \right)^2},$$

where M_m was used for normalization purposes and was chosen as the historical maximum power value M in the month m . In addition, the number N is equal to the number of items considered for the computation of the scores; for instance, if we want to compute the error referred only to forecasts of 1

day ahead, we will select all the items in the test month m related to this forecast horizon (i.e., N will be equal to the number of days in the month multiplied by 24 (number of hourly values in 1 day)).

Since these two metrics provide normalized values, they allow a comparison between different bidding zones (with different installed nominal power). In addition, they also highlight specific characteristics of the model performances; actually, the NMBE metric provides information about the error polarization (i.e., if the model was overestimating or underestimating observed values), whereas the NRMSE takes account of the absolute error, avoiding balancing effects due to the pointwise errors signs.

A. NMBE and NRMSE: errors analysis

Fig. 2 and Fig. 3 show the values of (i)-(ii) errors as a function of the test month at national level, from one-day ahead up to 15 days ahead (for PV and WD power generation, respectively).

The two metrics NMBE and NRMSE emphasize two different aspects of the forecasting error. In particular, the NMBE allows one to appreciate the sign of the error, and to infer whether there is a constant bias in the provided forecasts (i.e., whether the provided forecasts systematically underestimate, or overestimate, power generation). At this regard, both Fig. 2 and Fig. 3 show that the error may take any of the two signs in different cases, with rare significant underestimates (the most evident occurs in the wind case in the month of February for the 15 days ahead forecast) or overestimates (the most evident occurs in the PV case in the month of April, for the 15 days ahead forecast again).

On the other hand, the NRMSE plots allow one to better appreciate the magnitude of the forecasting error. Several results can be appreciated in this case: first of all, as one would expect, the error systematically increases with the length of the future horizon of forecasting. In particular, the error ranges from 5% (1-day ahead) to 18% (15 day-ahead) in the PV case, and from 10% (1-day ahead) up to 35% (15-day ahead) in the WD case. In addition, one may note that the error is pretty much constant in the wind case (Fig. 3(b)) while seasonality effects can be easily identified in the PV case (Fig. 2(b)), as weather forecasts are uncertain in spring time, while it is simpler to make irradiance forecasts in Italy in June (most likely, it is going to be a sunny day). Finally, it is much simpler to predict irradiance than wind speed, as the NRMSE is much lower in the PV case than it is in the WD case.

B. Bidding zones and Italian country

Here NRMSE error is studied as a function of the test month separately for each bidding zone.

Fig. 4 shows the error computed on the whole forecast horizon of 15 days in order to gain a global performance indicator. It

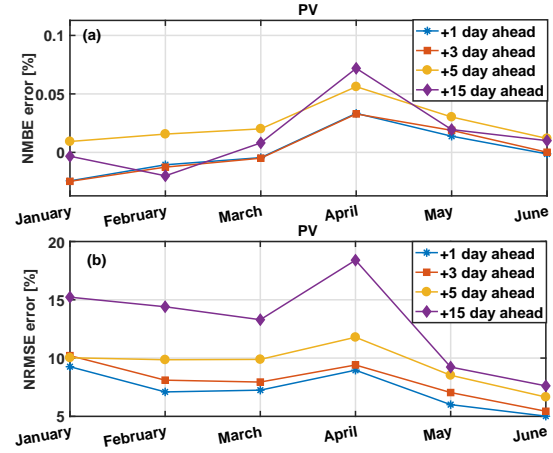


Fig. 2. NMBE (a) and NRMSE (b) averaged over bidding zones as a function of test month for PV plants.

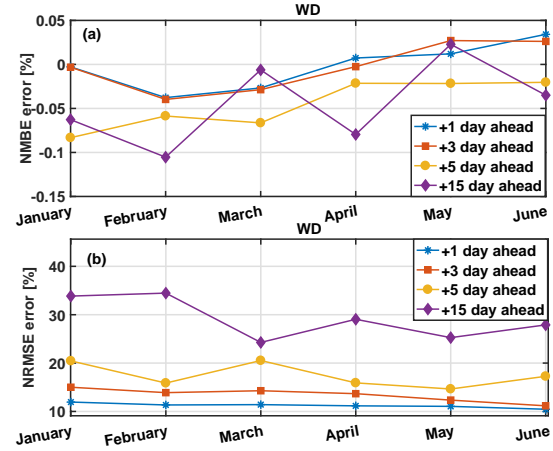


Fig. 3. NMBE (a) and NRMSE (b) averaged over bidding zones as a function of test month for WD farms.

is worth noticing that usually the trend of each single macro-area followed the overall trend at national level, identified by the blue curve, but some deviations may occur, as it can be seen for example in Fig. 4(a) for the month of April, where SUD exhibits a score 30% higher than the average global error (about 14.34 MW). Negative deviations may also occur, as it can be observed for example in Fig. 4(b) in April for CSUD (9% less with respect to the mean value). Such behaviour highlights the dependence of the model performance on the considered bidding area and hints that a benefit may be obtained by tuning the model specifically for different areas (we leave this for future work).

C. Forecasts versus Metering time series

In this section we show an example of comparison between our model forecasts and the actual measured power; in particular, Fig. 5 shows the results referring to the run of May 1, for both PV and WD power generation, considering the most productive bidding zone (NORD for PV, and SUD for WD [16], respectively). As we can see from Fig. 5, a good

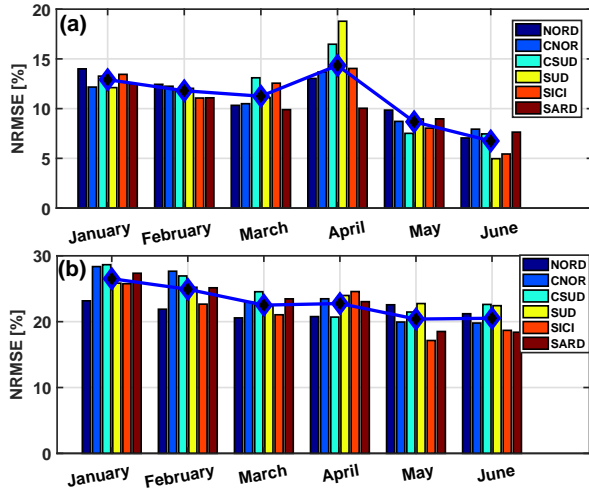


Fig. 4. NRMSE as a function of test month for PV (plot (a)) and WD (plot (b)) power generation of different bidding zones, computed on the whole forecast horizon of 15 days. The mean NRMSE at Italian level is also shown as a function of time (blue diamonds).

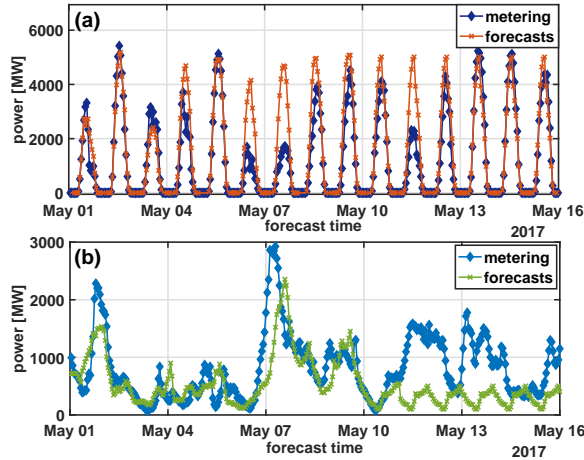


Fig. 5. (a) PV case: forecasts (orange curve) and metering (blue curve) as a function of time for PV power generation of NORD bidding zone; (b) WD case: forecasts (orange curve) and metering (blue curve) as a function of time for WD power generation of SUD bidding zone. The runtime of the 1st of May 2017 is shown.

agreement between the forecast and the actual value can be observed for the first 5 days ahead, whereas a degradation is observed from the 6-th day ahead on (after May 6). This effect is strongly related to the model dependence on the input accuracy: the larger the forecast horizons, the less reliable the meteorological data are. Finally, it should be observed that in the WD module, the last 5 days were affected by the persistence error, whereas in the PV case such effect was mitigated by the periodicity of the power curve.

V. CONCLUSION

In this paper we provide a power generation forecasting model at large-scale regional areas, where the macro-areas of interest are the Italian bidding zones.

Good results are achieved, thanks also to the use of a large historical data-set. More precisely, good performance can be observed for the first 5 days ahead, and results may be accurate enough to support classic power grid operation.

Further study can be however conducted; for instance, the implementation of specific tuning stages for each bidding zone, separately. Finally, another improvement may refer to the PV model; actually, the introduction of Global Tilted Irradiance (GTI) is expected to lead to a performance improvement, and for this reason will be eventually introduced in the model in a future work.

REFERENCES

- [1] Renewable Energy Policy Network for the 21st Century (REN21), "Renewables 2018 Global Status Report," Paris REN21 Secretariat, 2018.
- [2] L. Giloni, A. Betti, E. Crisostomi, A. Franco, M. Tucci, F. Bizzarri, and D. Mucci, "Day-Ahead Hourly Forecasting of Power Generation From Photovoltaic Plants," in *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 831-842, April 2018.
- [3] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. M. de Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78-111, 2016.
- [4] S. S. Soman, H. Zareipour, O. Malik and P. Mandal, "A review of wind power and wind speed forecasting methods with different time horizons," *North American Power Symposium 2010, Arlington, TX, 2010*, pp. 1-8.
- [5] S. Pelland, J. Remund, J. Kleissl, T. Oozeki, K. De Brabandere, "Photovoltaic and Solar Forecasting: State of the Art," Report IEA PVPS T14-01:2013, October 2013.
- [6] M. Pierro, M. De Felice, E. Maggioni, D. Moser, A. Perotto, F. Spada, C. Cornaro, "A New Approach For Regional Photovoltaic Power Estimation And Forecast," 33rd European Photovoltaic Solar Energy Conference and Exhibition, September 2017.
- [7] M. Marinelli, P. Maule, A. N. Hahmann, O. Gehrke, P. B. Nørgård, N. A. Cutululis, "Wind and photovoltaic large-scale regional models for hourly production evaluation," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 3, pp. 916-923, July 2015.
- [8] M. Morelli, A. Masini, F. Ruffini, M. A. C. Potenza, "Web tools concerning performance analysis and planning support for solar energy plants starting from remotely sensed optical images," *Environmental Impact Assessment Review*, vol. 52, pp. 18-23, 2015.
- [9] L. Breiman, "Statistical Modeling: The Two Cultures," *Statistical Science*, vol. 16, no. 3, pp. 199-215, 2001.
- [10] A. P. Engelbrecht, *Computational Intelligence: An Introduction*, 2nd ed., John Wiley & Sons Inc, 2007.
- [11] M. Fuentes, G. Nofuentes, J. Aguilera, D.L. Talavera, M. Castro, "Application and validation of algebraic methods to predict the behaviour of crystalline silicon PV modules in Mediterranean climates," *Solar Energy*, vol. 81, no. 11, pp. 1396-1408, Nov. 2007.
- [12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [13] N. Meinshausen, "Quantile regression forests," *The Journal of Machine Learning Research*, vol. 7, pp. 983-999, 2006.
- [14] S. S. Soman, H. Zareipour, O. Malik and P. Mandal, "A review of wind power and wind speed forecasting methods with different time horizons," *North American Power Symposium 2010, Arlington, TX, 2010*, pp. 1-8.
- [15] M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz, "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 65-76, 2013.
- [16] GSE S.p.A., "Rapporto Statistico, Energia da fonti rinnovabili in Italia, Anno 2016," Gennaio 2018.