Forecast of Distributed Energy Generation and Consumption in a Partially Observable Electrical Grid: A Machine Learning Approach

Elvio G. Amparore*†, Federico Cinus§*, Cristiano Maestri‡, Leonardo Petrocchi‡,
Dario Polinelli‡, Fabio Scarpa‡, Alan Perotti*, André Panisson*, Paolo Bajardi*
*ISI Foundation, Turin, Italy
†Università degli Studi di Torino, Turin, Italy
§Sapienza University of Rome, Rome, Italy
‡Terna S.p.A., Rome, Italy
*{*elvio.amparore, federico.cinus, alan.perotti, andre.panisson, paolo.bajardi*}*@isi.it*,
‡{*cristiano.maestri, leonardo.petrocchi, dario.polinelli, fabio.scarpa*}*@terna.it*

2021 IEEE Madrid PowerTech | 978-1-6654-3597-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/PowerTech46648.2021.9494887

***Abstract*—With a radical energy transition fostered by the increased deployment of renewable non-programmable energy sources over conventional ones, the forecasting of distributed energy production and consumption is becoming a cornerstone to ensure grid security and efficient operational planning. Due to the distributed and fragmented design of such systems, real-time observability of Distributed Generation operations beyond the Transmission System Operator domain is not always granted. In this context, we propose a Machine Learning pipeline for forecasting distributed energy production and consumption in an electrical grid at the HV distribution substation level, where data from distributed generation is partially observable. The proposed methodology is validated on real data for a large Italian region. Results show that the proposed model is able to predict up to 7 days ahead the amount of load and distributed generation (and the net power flux by difference) at each HV distribution substation with a 24%-44% mean gain in out-of-sample accuracy against a non-naive baseline model, paving the way to advanced and more efficient power system management.**

***Index Terms*—Distributed Power Generation, Load modeling, Machine Learning, Time series analysis.**

I. Introduction

Power Systems worldwide are undergoing a radical energy transition: on one hand, the number of renewable non-programmable energy sources and small power plants connected to the distribution grids – known as Distributed Generation (DG) - has increased significantly. On the other hand, the number of conventional power plants, capable of providing grid regulating services for the System security, has been gradually decreasing in many countries. This evolution is already causing severe security issues to Power Systems. National Transmission System Operators (TSOs) are responsible for the management of issues like: (i) the increasing periods of over-generation from renewable power plants during the central hours of the day, which may result in their curtailment; and (ii) the increasing frequency of power flow inversions at the HV distribution substations, which connect the transmission grid to the distribution grids. This occurs when the DG becomes greater than the local load absorbed by passive users. These phenomena implies new approaches to the Power System management since the Power System has been designed in the past considering distribution systems as passive grids.

In order to guarantee the Power System security in the most cost-effective way, TSOs need to observe in real-time the DG for operational purposes (static and dynamic security analyses, inertia assessments, etc.) and forecast the production of DG for planning purposes (maintenance scheduling on the sub-transmission grid, market efficiency, etc.).

This paper presents an analytic pipeline to forecast the DG and the users' load at the HV distribution substations in Italy, from the TSO point-of-view. Several observability constraints need to be considered to perform load/DG forecasts:

- The TSO monitors the net power exchange at each HV distribution substation, but has no real-time DG visibility;
- Distribution System Operators (DSOs) send energy measurements for a subset of the generation portfolio connected to the distribution grid, with a 1-month delay;
- Typically, these historical series assume continuous values for production units with an installed capacity greater or equal to 55 kW. Smaller production units are usually aggregated by DSO area, with some exceptions;
- The connection of MV/LV production units to their substation is generally not known by the TSO;
- Only a few statistical data are known for load consumption in the distribution grid and are generally aggregated at a much higher level (regional, zonal, national).

Therefore, the TSO has observability of the net power flux at the transformers of each HV distribution substation but does not have all the information needed to split the load from DG[1].

[1]All the above considerations were true at the time of experimentation. In this regard, the Italian Regulator ARERA started to address some of these issues with Deliberation 36/2020/R/EEL.

978-1-6654-3597-0/21/$31.00 ©2021 IEEE

Contribution:

The contribution of this paper is threefold: ① a system to estimate the load/DG on past data (older than 1 month), referred to periods for which power generation time series are available to the TSO at the time of the analysis; ② a heuristic algorithm to approximate the association between DG units and HV distribution substations to obtain the aggregated production/consumption time series at the HV substation level; ③ a Machine Learning model (named "ARMBTEX") that forecasts in real-time the production and consumption time series at HV distribution substation level in a very short (1 hour ahead) and medium (1 week ahead) time ranges.

Novelty: The combination of ① and ② allows the Italian TSO to estimate the unknown load/DG from the measured net flux on historical data. The reconstructed ground truth is then used and evaluated by ③ to forecast the real-time load/DG in a partially-observable large-scale grid, where only flux/weather information is available. This new system allows a TSO to obtain valuable forecasts for operational planning, within the regulatory framework and the technological limitations of the current system.

II. Background

In recent years the number of papers that addressed the problem of predicting energy production and consumption has drastically increased. It is widely acknowledged that machine learning (ML) can be used to model, design, and predict the behavior of these systems. The development of new ML techniques has considerably raised the accuracy, robustness, precision, and generalization ability of such models [1].

Renewable energy forecasting at a local scale is a complex and challenging task. Recently, the use of ML to forecast wind energy production [2]–[6], solar energy [7]–[15], and marine currents [16] has been extensively tested, showing significant improvements over non-ML methods. The relevant aspect of these approaches is that renewable energy generation can be successfully predicted, as long as the corresponding environmental variables (wind speed, solar radiance, cloud cover, etc.) are properly forecast. The forecast of energy consumption at a specific site is also considered by many papers [17]–[21]. These approaches can only be carried out by having many technical details of the considered plants/sites, which could not always be the case for the TSO. A more general approach is therefore needed to make forecasts on a large regional scale. In this regard, a ML model can be designed to provide the global energy output of many sources from several environmental variables, learning the energy/variables relation from multi-year data, as in [22]. Similar approaches have also been used to predict transmission line congestion with a high share of renewables [23]. Cyclical patterns also play a significant role in accurate forecasts of both energy generation and consumption [18], [24], at various time scales (hourly, daily, weekly, yearly, etc.) and considering holidays.

Several ML approaches have been investigated and compared in [5], [6], [25], ranging from SVR models, PSO, fuzzy Neural Networks, Gradient Boosting, and Random Forest.

F	Time series of the net flux.	$\mathbf{X}^{[t]}$	Input features at time t.
L	Time series of the net load.	$\mathbf{Y}$	Forecast targets.
$G^{[s]}$	Time series of the DG of type s.	$\mathbf{W}$	Weather time series.
G	Sum of all DG time series.	ϕ	A ML estimator function.
$\mathbf{G}$	Vector of all DG time series.	$\mathbf{\Phi}$	Forecast model.
$C^{[s]}$	Capacities of all units of type s.	$\mathcal{F}$	A set of forecast models.
$\mathbf{C}$	Vector of installed capacities.	$\mathcal{D}_m^{\text{train}}$	Monthly training dataset.
$\Gamma^{[u]}$	Time series of a prod. unit u.	D+J	Prediction at day J ($1 \leq J \leq 7$).
$\mathbf{\Gamma}^{[s]}$	Vector of all $\Gamma^{[u]}$ of type s.	$\bar{\bullet}$	Forecast value of $\bullet$
$\mathbf{a}^{[s]}$	Substation association vector for units of type s.		
$\hat{\mathbf{a}}^{[s]}$	Approximated $\mathbf{a}^{[s]}$.	$\mathcal{U}^{[s]}$	Selected units of type s.
$\mathcal{L}_J$	Set of lags of predictors at day J.		

Table I. Notation reference.

From the point-of-view of the TSO, the local renewable energy generation acts as a variable component that hides a part of the actual customer net load, making it harder for the TSO to estimate the real load and meet its demand if the renewable generation drops suddenly [26], [27].

III. Algorithmic pipeline

This section goes through the problem definition (III-A), the description of the available dataset (III-B), and the two consequential tasks, involving respectively: the inference of the target data (III-C), and the definition of the predictive model and its validation (III-D, III-E).

A. Problem Statement

Let us consider the time series F up to time t: $F_0, F_1, ..., F_t$ summarizing the active power flux (in MW) flowing from the TSO to the DSO measured at each individual HV distribution substation transformers. A value F_t can be seen as the difference between the local load L_t and the local generation G_t at the substation: $F_t = L_t - G_t$, where the series L and G are unknown. The primary goal is to forecast the future values $\overline{L}_{t'}$, $\overline{G}_{t'}$ and $\overline{F}_{t'}$ for a set of time points $t' > t$. A time series F at 15-min granularity is used and 1 to 7 days ahead predictions are considered for the experiments. Since both L and G are necessary to build and validate a forecasting model, a preliminary problem consists in estimating the past time series of L and G. In the following, we refer to the primary goal as the *forecast problem*, and to the estimation of past L and G data as the *target estimation problem*.

The information available for solving the *target estimation problem* sets the context of a partially observable electrical grid. Information about the production units that could be used to reconstruct G is partial. Only a subset of all DG units' time series is known, and there is no direct information about the connections between DG units and substations.

The DG contributions from different energy sources are collected in separate time series $G^{[s]}$ (s.t. $G = \sum_s G^{[s]}$), with $s \in \{solar, thermal, hydro, wind\}$. The series $\mathbf{G} = [G^{[s]}]$ are estimated using additional data sources with the methodology described in Section III-C. Finally, we associate to each substation a weather time series $\mathbf{W}$, and the aggregated installed capacity at the substation $\mathbf{C} = [C^{[s]}]$, with one value $C^{[s]}$ (in MW) for each energy source s.

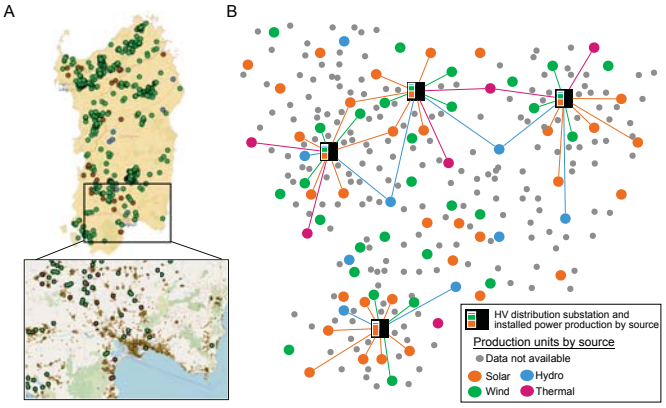

Fig. 1. (A) Spatial distribution of production units of Sardinia. The map shows all sources except solar. The zoom shows the vast diffusion of solar installations. (B) Sketch of the sampling procedure to infer energy generation at HV distribution substations.

B. Dataset

The considered dataset comes from a set of heterogeneous sources related to the TSO, plus weather data. Data were collected for the 2017–2019 period for Sardinia's territory, Italy. Sardinia is an (electric) island, connected to the mainland only through HVDC submarine cables. Therefore, all DG production units connected to Sardinian substations are located on the island. In the considered area there are ∼70 HV distribution substations [28]. For each of these substations, the collected dataset consists of (1) the net power flux time series F with 15-min granularity and corresponding quality codes; (2) the estimation of the installed capacities $C^{[s]}$, updated weekly by the TSO; (3) weather data $\mathbf{W}$ containing temperature, rain probability, wind speed and direction, and cloud cover from an observation point close to the substation.

A national database stores all the production unit details (nominal power, type, address). In the considered area, there are about 37,000 DG units [29], and less than 3% of them are required to send energy production data continuously. For some of these units, the TSO has access to energy time series that are provided by the DSO, but only at the beginning of every month and with a one-month delay. Moreover, the TSO does not know the HV distribution substation to which each production unit u is (indirectly) connected (see Figure 1).

C. Targets estimation of each HV distribution substation

In a substation, the aggregated generation $G^{[s]}$ for the energy source s is approximated as follows. Let $\Gamma^{[u]}$ be the time series of a production unit u, and let $\mathbf{\Gamma}^{[s]} = [\Gamma^{[u]}]$ be the matrix of all the generation time series of units of type s. The generated energy of type s is given by $G^{[s]} = \mathbf{a}^{[s]} \cdot \mathbf{\Gamma}^{[s]}$, where $\mathbf{a}^{[s]}$ is the *real substation association vector* defined s.t. the value of $\mathbf{a}_u^{[s]}$ is 1 if unit u of type s is assigned to that substation, and 0 otherwise. However, the partially observable system considered in this paper requires $\mathbf{a}^{[s]}$ to be approximated because only a subset of $\mathbf{\Gamma}^{[s]}$ is available, and the exact unit-substation association is unknown.

The adopted solution computes an *approximated substation association vector* $\hat{\mathbf{a}}^{[s]} \approx \mathbf{a}^{[s]}$ by following these steps: (I) defining a set $\mathcal{U}^{[s]}$ of units, selected among the 3% that actually send data (see Sec. III-B), using a sampling strategy with replacement; and (II) assigning to each selected unit u a weight proportional to its nominal power P_u.

The goal is to extract a set of *probe units* $\mathcal{U}^{[s]}$ which is representative of all the units connected to a specified HV distrib. substation (see Figure 1/B). Since the association between each unit u and its substation is guessed, rule (I) does not impose u to be associated with a single substation.

The set $\mathcal{U}^{[s]}$ of selected units of type s is chosen as the smallest set that satisfies the following criteria: (i) Only units that send data to the DSO can be picked; (ii) Units are taken in distance order; (iii) The total capacity of the selected units must be at least a fraction (chosen to be 50%) of the installed capacity $C^{[s]}$; (iv) A minimum number of units (chosen based on the average unit availability) is always associated, to average the approximation error.

The approximated substation assoc. vector $\hat{\mathbf{a}}^{[s]}$ is then

$$\hat{\mathbf{a}}_u^{[s]} = C^{[s]} \cdot \left(P_u / \textstyle\sum_{j \in \mathcal{U}^{[s]}} P_j \right) \text{ if } u \in \mathcal{U}^{[s]}, \text{ 0 otherwise}$$

The data acquisition defines the algorithm's pipeline described in the next sections. The one-month delay deriving from the $\Gamma^{[u]}$ time series availability to the TSO implies that G can be fully constructed as $G^{[s]} \approx \hat{\mathbf{a}}^{[s]} \cdot \mathbf{\Gamma}^{[s]}$, up to the previous month. Similarly, L can be derived wherever G is estimated as $F + G$.

D. Forecast Problem Formulation

This section introduces the definition of the training and the test sets in the load/DG forecasting problem for a given HV distribution substation. This formulation deeply depends on the temporal flow of the data, which defines the constraints for the real-time forecasting pipeline.

The input features $\mathbf{X}$ are $[\mathbf{C}, \mathbf{W}, F]$, where $\mathbf{C} = [C^{[s]}]$ is the vector of installed capacity (of all source types), $\mathbf{W}$ is the observed weather vector at the nearest observation point to the substation, and F is the net power flux time series measured at the transformers of the substation. The target output $\mathbf{Y}$ is a multivariate series defined as $[\mathbf{G}, L]$, where $\mathbf{G} = [G^{[s]}]$.

Each time point represents a 15-minute slot. The beginning of the current day is indicated with t_0, the current quarter-of-hour with t_{now}. The point $t_0 - k$ denotes the last instance where the $\Gamma^{[u]}$ series are given. Notation D+J indicates the forecast target is J-days ahead.

The data flow follows these dynamic constraints:

1) The training target $\mathbf{Y}$ can be established up to $t_0 - k$, where all data are known or estimated;
2) Starting from $t_0 - k + 1$, the target $\mathbf{Y}$ cannot be reconstructed since the series $\Gamma^{[u]}$ are not yet available;
3) The net flux is not available after the current time t_{now}.

The goal is to compute a set of forecasts for all time points t' of the day D+J. To compute a single forecast at time t' it is possible to define the set $\mathbf{X}_{t'}^{[t_{\text{now}}]}$ of all the data available at the

current time t_{now}, and the input forecast values for the time points up to t' (i.e. $\overline{\mathbf{C}}$ and $\overline{\mathbf{W}}$). Let $\overline{\mathbf{W}}_{t'}$ be the weather forecast at time t', and let $\overline{\mathbf{C}}_{t'}$ be the vector of estimated installed capacity at time t'. The *forecast model* $\mathbf{\Phi}_J : \mathbf{X}_{t'}^{[t_{\text{now}}]} \rightarrow \overline{\mathbf{Y}}_{t'}$ is a machine learning predictor that, given the input vector $\mathbf{X}_{t'}^{[t_{\text{now}}]}$ at time t', produces the prediction pair $\overline{\mathbf{Y}}_{t'} = [\overline{\mathbf{G}}_{t'}, \overline{L}_{t'}]$ for the same 15-minute interval, J days ahead from t_0.

The model $\mathbf{\Phi}_J$ is a predictor for a single time point t'. All the 96 different forecasts (one for each 15 minutes interval in a day) for a day D+J can be generated by evaluating $\mathbf{\Phi}_J(\mathbf{X}_{t'}^{[t_{\text{now}}]})$ for the 96 time points t'.

A *weekly model set* $\mathcal{F}$ is a set of models $\mathbf{\Phi}_J$ that produce forecasts with jump J ranging from +1 to +7 days. Each HV distribution substation will have its specific model set $\mathcal{F}$.

A training dataset can be defined for the month m. Let t_m be the last time point of m, with $t_m \leq t_0 - k$. The *monthly training dataset* $\mathcal{D}_m^{\text{train}}$ is defined as

$$\mathcal{D}_m^{\text{train}} = \left\{ (\mathbf{X}_{t'}^{[t_m]}, \mathbf{Y}_{t'}) \;\middle|\; \forall t' \leq t_m \right\}$$

E. Forecasting approach

Figure 2 shows the temporal structure of the involved variables, which are: *(a)* the installed DG power at the HV distribution substation; *(b)* the estimated distributed generation time series (available with a one-month delay); *(c)* the weather time series (exogenous features), where data points $> t_{\text{now}}$ represent weather forecasts; *(d)* the inferred DG (green), the reconstructed consumers' load (blue), and the real-time net flux (red). The solid series in *(d)* are the values for G, L and F, while the dotted ones are forecasts of a ML model.

This section describes the ML models used to make the forecasts. The presented models are derived from a comparative selection between several techniques, including gradient boosting, neural networks, SARIMA models, and univariate/multivariate approaches. In the following, the benchmark is built upon a univariate baseline model, which exploits seasonal patterns, and two supervised ML models based on gradient boosting, which differ for an input feature (the net flux F).

PMA *model*: While standard benchmarks for time-series forecasting tasks are usually dummy models based on persistence

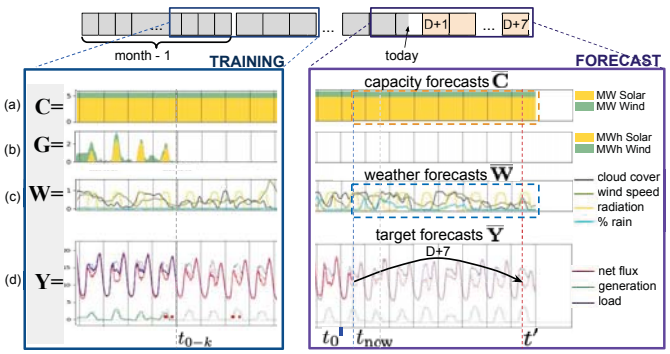

Fig. 2. Schematic timeline representation of the input $\mathbf{X}_{t'}^{[t_{\text{now}}]}$ and output $\mathbf{Y}_{t'}$ variables, given a current time t_{now}.

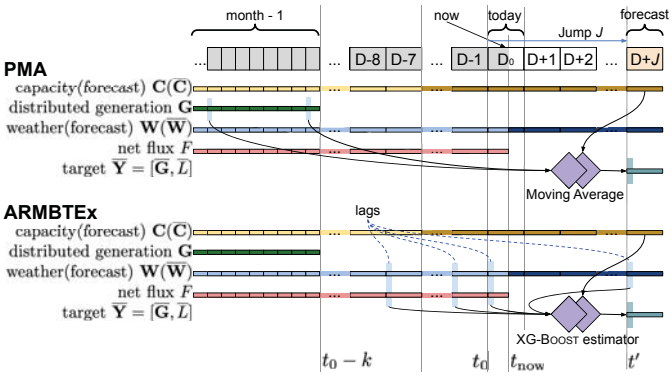

Fig. 3. High-level schema of the PMA and the ARMBTEx models.

(i.e. predicting the next value as the last observed one), the proposed pipeline is compared against a more realistic baseline, representing the best-effort forecast of a domain expert. The *Periodic Moving Average* (PMA) model exploits the fact that generation and load are recurrent processes [30], [31] with a strong daily and weekly seasonality. A graphical representation of the model logic is provided in Figure 3(top). The predicted values for generation and load are defined as the moving average of the latest 4 available samples *at the same time the same day of the week* in the past. However, because of a one-month delay in the communication of the $\Gamma^{[u]}$ to the TSO, these last values could be several weeks old.

ARMBTEx *model*: The proposed ML model $\mathbf{\Phi}^{\text{ML}}$ is an *Auto-Regressive Multivariate model based on Boosted Trees with Exogenous variables* (ARMBTEx) built on eXtreme Gradient Boosting algorithm (XG-BOOST). A graphical representation is given in Figure 3(bottom). The model input consists of multiple time points, arranged according to a set of *lags*. A lag z is a time shift from the target time point t'. Let $\mathcal{L}_J$ be a set of lags defined for the D+J forecast model.

The XG-BOOST package[2] is a scalable implementation of the gradient boosting algorithm [32], [33] for categorical (continuous) supervised learning. These models $\phi : \mathbf{X} \rightarrow \mathbb{R}$ are used as multivariate *regressor* functions. Following the tree ensemble paradigm, the generic prediction entry of the model ϕ is the sum of independent trees, learned iteratively by minimizing a regularised additive objective function.

ARMBTEx is built on one ML estimator ϕ_L to forecast the load $\overline{L}_{t'}$, and one estimator $\phi_{G^{[s]}}$ (for each source s) to estimate the value $\frac{\overline{G}_{t'}^{[s]}}{C_{t'}^{[s]}}$, with $C_{t'}^{[s]}$ being the installed capacity at time t'. Estimators take in input a masked vector

$$\mu(\mathbf{X}_{t'}^{[t_{\text{now}}]}) = \left(x_{t'-z} \;\middle|\; \forall z \in \mathcal{L}_J\right), \quad x_t = \begin{cases} (\mathbf{W}_t, F_t) & \text{if } t < t_{\text{now}} \\ \overline{\mathbf{W}}_t & \text{otherwise} \end{cases} \quad (1)$$

i.e. μ selects the features relative to the lags in $\mathcal{L}_J$.

The forecast $\overline{\mathbf{Y}}_{t'} = [\overline{\mathbf{G}}_{t'},\ \overline{L}_{t'}]$ is then computed as

$$\overline{\mathbf{G}}_{t'} = \left[C_{t'}^{[s]} \cdot \phi_{G^{[s]}}\left(\mu(\mathbf{X}_{t'}^{[t_{\text{now}}]})\right) \middle| \forall s \right], \quad \overline{L}_{t'} = \phi_L\left(\mu(\mathbf{X}_{t'}^{[t_{\text{now}}]})\right)$$

[2]https://xgboost.readthedocs.io

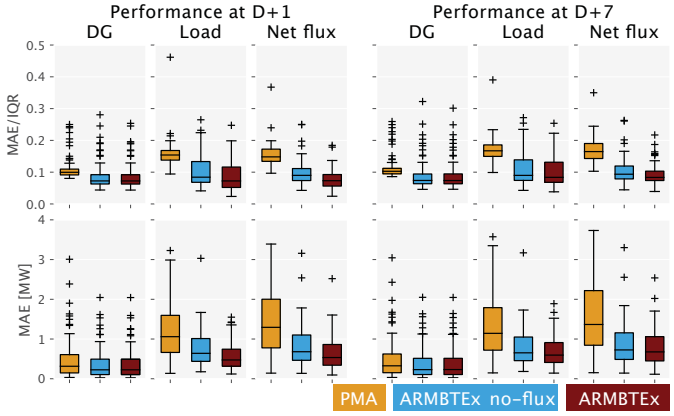

Fig. 4. Model comparison for the D+1 (left) and D+7 (right) estimators, for the three targets (columns), in terms of MAE/IQR (top) and MAE (bottom). Lower scores are better.

ML estimators ϕ are trained on all tuples $\{(\mu(\mathbf{X}), \mathbf{Y}) \mid \forall(\mathbf{X}, \mathbf{Y}) \in \mathcal{D}^{\text{train}}_{m-1}\}$ of the *monthly training dataset* $\mathcal{D}^{\text{train}}_{m-1}$ of the previous month $m-1$. Let $\mathbf{\Phi}^{\text{PMA}}_{m}$ and $\mathbf{\Phi}^{\text{ML}}_{J,m}$ denote the PMA and ARMBTEx models trained on the $\mathcal{D}^{\text{train}}_{m}$ dataset.

Parameter selection: The XG-Boost ϕ estimators use several hyper-parameters to control the amount of over-/under-fitting of the decision trees ensemble. The hyper-parameter selection has been carried out following a grid search approach, i.e. evaluating the training and test performances of several possible parameter combinations. The grid search resulted in estimators with 100 decision trees of 10 maximum levels depth each.

The *lags* hyper-parameters control the amount of periodicity each estimation model ϕ may learn from. The selected combination of lags is the following: Lag z_0 is always the data of the forecast day. The other lags are chosen to capture the correlation with the daily and weekly seasonality of the data. z_1 is the most recent data available (today), z_2 is the same hour but in the previous day, and z_3 is the same hour of z_0, 7– or 14–days back, to capture the weekly seasonality.

IV. Computational Results

Model validation is performed on each substation, comparing the performance of PMA against ARMBTEx. The errors evaluation considers the two reconstructed targets ($\overline{G} - G$, $\overline{L} - L$) and the error on their difference $\overline{F} - F$ (the net flux), producing 3 different residuals to estimate the errors and validate the models.

The validation can be summarized as follows:

1) Every month m a new training dataset $\mathcal{D}^{\text{train}}_{m-1}$ is built using all data available up to a month $m-1$;
2) The weekly model sets $\mathcal{F}^{\text{baseline}}_{m}$ and $\mathcal{F}^{\text{ML}}_{m}$ are trained on the $\mathcal{D}^{\text{train}}_{m-1}$ dataset;
3) The forecast models generate all the short-time (D+1) to the medium-time (D+7) out-of-sample estimates for month m (*walk-forward* validation);
4) Estimates $\overline{\mathbf{Y}}$ can be compared against the reconstructed data $\mathbf{Y}$ in the 2017–2019 period, to verify the achieved target accuracy of each D+J model.

The training window used by ARMBTEx is *800* days in the past (if available), with a warm-up period of three months.

D+1 and D+7 are the two extreme cases of prediction in terms of error, respectively minimum and maximum. The evaluation is performed using the *mean absolute error* (MAE), as well as a *dimensionless error* obtained as the MAE divided by a substation-specific robust scale factor (interquartile range).

Figure 4 shows the average out-of-sample performances of the D+1 and D+7 estimators (left and right, respectively) for the three targets G, L and F. Each boxplot sample is a substation. The models in the figure are, in order: the PMA model; ARMBTEx without the current net flux in the input variables, i.e. Eq. (1) never includes F_t in the input; ARMBTEx with the most recently available information about net flux. Lower scores are better, meaning lower average errors. Training a monthly $\mathcal{F}^{\text{ML}}_{m}$ model for a target substation requires about 60 seconds on a single CPU on the test hardware. Each monthly training dataset has about 80,000 samples. All monthly predictions of $\mathcal{F}^{\text{ML}}_{m}$ are then computed in less than 1 second, once the model is trained. All tests were performed on modern commodity hardware.

ARMBTEx provides significant improvements over the baseline model (PMA) for all three targets, in particular for the load estimate. In this case, the availability of the net flux of the current day among the input variables further increases the average estimator accuracy. The results are consistent among different metrics, as expressed in Figure 4.

The availability of critical information with a 1-month delay makes it difficult for a univariate estimator to provide accurate results. Therefore it is not surprising that well-trained multivariate models outperform univariate ones in this context. Figure 5 shows the average error distribution of the target estimates over the course of the months. The PMA estimator shows the largest fluctuations since it takes at least one month to adapt. ARMBTEx models provide more stable estimates over the course of the years.

Finally, Table II contains the estimation of the accuracy gain of ARMBTEx with respect to the baseline model (PMA). The percentage error gain is calculated for each HV distribution substation and each target, and the mean values for the

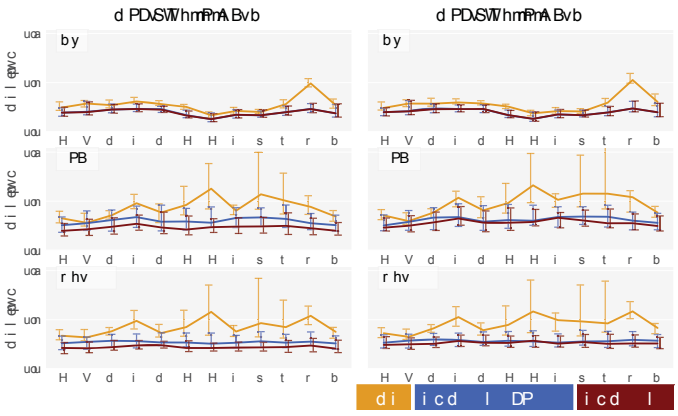

Fig. 5. Error distribution over the prediction months.

-	DG	Load	Net flux
MAE reduction	24.065%	44.335%	48.350%
RMSE reduction	21.949%	40.512%	44.595%

Table II. Mean percentage error reduction going from the baseline (PMA) to the ARMBTEx model, for the three targets at D+1.

correspondent distributions are displayed. The results for both the net flux and the load are remarkable (48%–44% of gain, respectively) while the DG shows an improvement greater than 24%. It is worth noting that the out-of-sample performances are evaluated on the *estimated* load and DG, as they are the targets used to train the models, even if the load/DG do not constitute a validated ground truth (only the net flux is). As an additional validation, the accuracy for the reconstructed net flux has been further evaluated via the predicted targets. Such external validation confirms the fairly good performances of the whole algorithmic pipeline. The validation with real-time measures of the net flux produced low mean errors and almost 50% of accuracy gain.

V. Conclusions and Future Work

This work presents a data-driven pipeline from the TSO perspective to infer and forecast useful information to enhance grid security and management. Starting from aggregated and partially available measures, the system develops substation-specific predictive models up to 7 days ahead that are able to separate the contribution of DG and load from the net flux. ARMBTEx consistently outperformed the benchmark on several metrics, increasing the accuracy of the DG forecast by 24%, and increasing the accuracy of the load forecast by 44%. We acknowledge some limitations in the present research. Unfortunately, an external ground truth to validate the entire pipeline was only available for the net flux. Future work could be devoted to performing extensive sensitivity analysis on the association vector $\hat{\mathbf{a}}^{[s]}$, to investigate how different rules might affect the ground truth and validation accuracies. Moreover, future research may quantify the impact of weather forecasts on DG forecasts, as current computational experiments were limited to assess the model accuracy considering actual weather conditions.

References

[1] A. Mosavi *et al.*, "State of the art of machine learning models in energy systems, a systematic review," *Energies*, vol. 12, no. 7, p. 1301, 2019.

[2] P. Chatziagorakis *et al.*, "Enhancement of hybrid renewable energy systems control with neural networks applied to weather forecasting: the case of Olvio," *Neural Computing and Applications*, vol. 27, no. 5, pp. 1093–1118, 2016.

[3] Q. He, J. Wang, and H. Lu, "A hybrid system for short-term wind speed forecasting," *Applied energy*, vol. 226, pp. 756–771, 2018.

[4] Z. Qu *et al.*, "A hybrid model based on ensemble empirical mode decomposition and fruit fly optimization algorithm for wind speed forecasting," *Advances in Meteorology*, 2016.

[5] A. Khosravi *et al.*, "Time-series prediction of wind speed using machine learning algorithms: A case study Osorio wind farm, Brazil," *Applied Energy*, vol. 224, pp. 550–566, 2018.

[6] Sharifian *et al.*, "A new method based on type-2 fuzzy neural network for accurate wind power forecasting under uncertain data," *Renewable energy*, vol. 120, pp. 220–230, 2018.

[7] A. Ahmad *et al.*, "Hourly global solar irradiation forecasting for New Zealand," *Solar Energy*, vol. 122, p. 1398, 2015.

[8] H. Loutfi *et al.*, "Generation of horizontal hourly global solar radiation from exogenous variables using an artificial neural network in Fes (Morocco)," *IJRER*, vol. 7, no. 3, pp. 1097–1107, 2017.

[9] N. Premalatha and A. Valan Arasu, "Prediction of solar radiation for solar systems by using ANN models with different back propagations," *Applied research and tech.*, vol. 14, no. 3, pp. 206–214, 2016.

[10] M. David *et al.*, "Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models," *Solar Energy*, vol. 133, pp. 55–72, 2016.

[11] C. Voyant *et al.*, "Uncertainties in global radiation time series forecasting using machine learning," *Energy*, vol. 125, pp. 248–257, 2017.

[12] Salcedo-Sanz *et al.*, "An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the Sunshine State of Australia," *Applied Energy*, vol. 209, pp. 79–94, 2018.

[13] E. Ogliari *et al.*, "Physical and hybrid methods comparison for the day ahead PV output power forecast," *Renewable Energy*, vol. 113, pp. 11–21, 2017.

[14] Wang *et al.*, "Deterministic and probabilistic forecasting of photovoltaic power based on deep convolutional neural network," *Energy conversion and management*, vol. 153, pp. 409–422, 2017.

[15] B. P. Mukhoty *et al.*, "Sequence to sequence deep learning models for solar irradiation forecasting," in *PowerTech*. IEEE, 2019, pp. 1–6.

[16] M. B. Anwar *et al.*, "Novel power smoothing and generation scheduling strategies for a hybrid wind and marine current turbine system," *IEEE Trans. on Power Systems*, vol. 32, no. 2, pp. 1315–1326, 2016.

[17] F. Chahkoutahi and M. Khashei, "A seasonal direct optimal hybrid model of computational intelligence and soft computing techniques for electricity load forecasting," *Energy*, vol. 140, pp. 988–1004, 2017.

[18] A. Bagnasco *et al.*, "Electrical consumption forecasting in hospital facilities: An application case," *Energy and Buildings*, vol. 103, pp. 261–270, 2015.

[19] I. M. Coelho *et al.*, "A GPU deep learning metaheuristic based model for time series forecasting," *Applied Energy*, vol. 201, pp. 412–418, 2017.

[20] E. Mocanu *et al.*, "Deep learning for estimating building energy consumption," *Sustainable Energy, Grids and Networks*, vol. 6, pp. 91–99, 2016.

[21] M. H. Alobaidi *et al.*, "Robust ensemble learning framework for day-ahead forecasting of household based energy consumption," *Applied energy*, vol. 212, pp. 997–1012, 2018.

[22] F. Touati *et al.*, "Long-term performance analysis and power prediction of PV technology in the state of Qatar," *Renewable Energy*, vol. 113, pp. 952–965, 2017.

[23] P. Staudt *et al.*, "Predicting transmission line congestion in energy systems with a high share of renewables," in *PowerTech*. IEEE, 2019, pp. 1–6.

[24] R. K. Jain *et al.*, "Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Applied Energy*, vol. 123, pp. 168–178, 2014.

[25] S. Aggarwal and L. Saini, "Solar energy prediction using linear and non-linear regularization models: A study on AMS 2013–14 Solar Energy Prediction Contest," *Energy*, vol. 78, pp. 247–256, 2014.

[26] A. Kaur *et al.*, "Impact of onsite solar generation on system load demand forecast," *Energy Conversion and Management*, vol. 75, pp. 701–709, 2013.

[27] ——, "Net load forecasting for high renewable energy penetration grids," *Energy*, vol. 114, pp. 1073–1084, 2016.

[28] Terna, "Electrical grid statistics of Italy 2018," link to document.

[29] ——, "Renewable sources statistics of Italy 2019," link to document.

[30] L. F. Ochoa *et al.*, "Evaluating distributed time-varying generation through a multiobjective index," *IEEE Transactions on Power Delivery*, vol. 23, no. 2, pp. 1132–1138, 2008.

[31] W. L. Theo *et al.*, "Review of distributed generation (DG) system planning and optimisation techniques: Comparison of numerical and mathematical modelling methods," *Renewable and Sustainable Energy Reviews*, vol. 67, pp. 531–573, 2017.

[32] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *22nd ACM SIGKDD*, 2016, pp. 785–794.

[33] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.