energies

MDPI

Article

High-Resolution PV Forecasting from Imperfect Data: A Graph-Based Solution

Rafael E. Carrillo [1], Martin Leblanc [1], Baptiste Schubnel [1], Renaud Langou [1], Cyril Topfel [2] and Pierre-Jean Alet [1,*]

[1] CSEM PV-Center, Rue Jaquet-Droz 1, 2000 Neuchâtel, Switzerland; rafael.carrillo@csem.ch (R.E.C.); martin.leblanc@alumni.epfl.ch (M.L.); baptiste.schubnel@csem.ch (B.S.); renaud.langou@csem.ch (R.L.)

[2] BKW AG, Viktoriaplatz 2, 3013 Bern, Switzerland; cyril.topfel@bkw.ch

* Correspondence: pierre-jean.alet@csem.ch

Received: 5 October 2020; Accepted: 27 October 2020; Published: 3 November 2020

Abstract: Operating power systems with large amounts of renewables requires predicting future photovoltaic (PV) production with fine temporal and spatial resolution. State-of-the-art techniques combine numerical weather predictions with statistical post-processing, but their resolution is too coarse for applications such as local congestion management. In this paper we introduce computing methods for multi-site PV forecasting, which exploit the intuition that PV systems provide a dense network of simple weather stations. These methods rely entirely on production data and address the real-life challenges that come with them, such as noise and gaps. Our approach builds on graph signal processing for signal reconstruction and for forecasting with a linear, spatio-temporal autoregressive (ST-AR) model. It also introduces a data-driven clear-sky production estimation for normalization. The proposed framework was evaluated over one year on both 303 real PV systems under commercial monitoring across Switzerland, and 1000 simulated ones based on high-resolution weather data. The results demonstrate the performance and robustness of the approach: with gaps of four hours on average in the input data, the average daytime NRMSE over a six-hour forecasting horizon (in 15 min steps) and over all systems is 13.8% and 9% for the real and synthetic data sets, respectively.

Keywords: multi-site photovoltaic forecasting; spatio-temporal correlation; graph signal processing; signal reconstruction

1. Introduction

Production forecasting is a critical technology for enabling large-scale penetration of PV generation into the power grid [1]. In particular, improved forecasting leads to lower net generation costs in the power system and to reduced curtailment of PV production [2]. Forecasting accuracy is limited both by the chaotic nature of the weather and by the uncertainty on the physical response of PV systems to given weather conditions due to variable temperature coefficients, nonlinear behavior at low irradiance, or complex local shading.

To address this dual challenge, the state of the art for the forecasting of single PV systems is the combination of multiple machine learning approaches with numerical weather predictions (NWP) as inputs. In a recent benchmarking study on 152 PV systems in the Netherlands, such techniques yielded a daytime normalized root-mean-square error (NRMSE) between 9% and 17% of peak power, depending on the method and system [3]. Forecasts for grid operation purposes have focused on a regional level,

at which they benefit from a strong smoothing effect to yield an NRMSE from 5% (1 h ahead) to 7% (24 h ahead) of the daytime peak power in the North of Italy [4]. They estimate the current and near-future levels of regional PV power production using data from sampled systems, static information on all installed PV systems, NWP, and machine learning algorithms.

Many approaches have been proposed to improve PV forecasting accuracy. Since the high variability of solar irradiance mainly comes from cloud movement and its stochastic blocking of sunlight, the works in [5,6] use cloud motion vector information extracted from sky imagers for intra-hour irradiance forecasting. However, sky imagers are too expensive to be deployed at all PV sites, and moreover, they are only useful for intra-hour prediction horizons. The works in [7,8] use satellite images for hourly PV production forecasting. However, wide-area satellite images are not capable of capturing site-specific information, and thus not good for site-specific PV forecasting. The works in [9,10] use forecasted cloud information to improve forecasting accuracy. However, such data may be expensive and require heavy processing.

Most recently, several studies have proposed data-driven forecast methods that use multi-site spatio-temporal historical data for PV forecasting without requiring NWP or cloud movement information [11–19]. Simple and fast multi-site forecasting techniques using linear autoregressive (AR) models are proposed in [11–13]. The work in [12] proposes a lasso regularization in the AR model to automatically select the input variables, while the work in [13] further develops the AR model into a local vector AR model with ridge regularization that considers local weather changes. However, nonlinear methods outperform the linear techniques for forecasting horizons longer than an hour ahead [20], and recent works have used advanced deep learning techniques to forecast multi-site PV power production for hourly horizons. Feed forward neural networks (FNN) [14] and long short-term memory (LSTM) networks [15,21] are proposed for multi-site spatio-temporal PV forecasting. Particularly, LSTMs have improved the forecasting accuracy since they are well suited for time-series forecasting by processing sequence information using internal gating mechanisms [22,23]. In addition, recent works use convolutional neural networks (CNN) [17–19] to learn the complex spatial dependencies of the production data and indirectly track cloud motion. By considering spatio-temporal relations, they can capture cloud cover and cloud movement, since passing clouds influence neighboring PV sites sequentially, thus yielding an improved forecasting accuracy. A limitation of all these techniques is that they rely on the availability of clean data for both training and forecasting.

Yet in real life, validation and cleaning of data from grid sensors is a prerequisite for any data-driven solution, which may be affected by measurement or transmission errors. Published solutions for PV rely on the knowledge of characteristics of the systems [24], which themselves may be faulty or missing. The work of [25] proposed a processing pipeline that takes into account data cleaning and treatment of missing data. They propose to fill the missing gaps by finding a representative PV signal from the same geographical area and replacing the missing segment by the representative signal. The work in [26] proposed a low-rank tensor learning scheme to reconstruct the missing gaps. The low-rank tensor model implicitly considers the spatio-temporal correlations of multi-site PV production signals. However, none of the aforementioned works consider a spatial model for the non-regular spatial distribution of PV plants, thus they do not fully exploit the spatio-temporal correlations of PV production signals to reconstruct the missing segments.

In this paper, we propose a robust framework for multi-site PV forecasting, which can cope with incomplete and noisy data. The proposed forecasting approach is based on a linear spatio-temporal autoregressive (ST-AR) model that only uses past production data from neighboring systems to predict the production of a specific plant. The underlying idea is that temporal correlation between the outputs of PV systems is linked with their spatial relationship, and that PV systems can effectively be used as ground-level weather stations to forecast the production of their neighbors. In order to learn the ST-AR model, we use a group Lasso estimator to automatically select the most informative nodes for prediction

as well as to estimate their coefficients. To address the problem of incomplete and noisy data, we propose a graph-based method to preprocess the input data to the ST-AR forecast model. Our method is based on building a graph model to capture the spatial dependencies among the PV systems and exploit the spatio-temporal relations to reconstruct the missing parts of the data. The ST-AR model is enhanced by a novel, data-driven method to calculate the clear-sky performance of PV systems, which takes into account the effects of their local environment and physics. The proposed framework was evaluated in two data sets for an entire year: (1) production data from 303 real PV systems, and (2) simulated production of 1000 PV systems, in both cases distributed over Switzerland.

The organization of the remainder of the paper is as follows. Section 2 presents the proposed robust framework for PV production forecasting and details both the graph-based reconstruction and forecasting methods. Results for PV signal reconstruction and forecasting are presented in Section 3. Finally, we conclude in Section 4 with a discussion of the results and future directions.

2. Methods

In this section, we present the proposed robust framework for PV production forecasting. Figure 1 depicts a block diagram of the framework. The gap reconstruction and preprocessing module filters the data (i.e., fills the gaps from missing data as well as removes noise) and preprocesses the data (normalization and de-trending) such that its output is in the correct form for the learning or the forecast modules. The learning module learns the graph models used for forecasting from historical data. Finally, the forecast module provides a forecast of the PV production over each point in the area based on the learned models and the cleaned data.

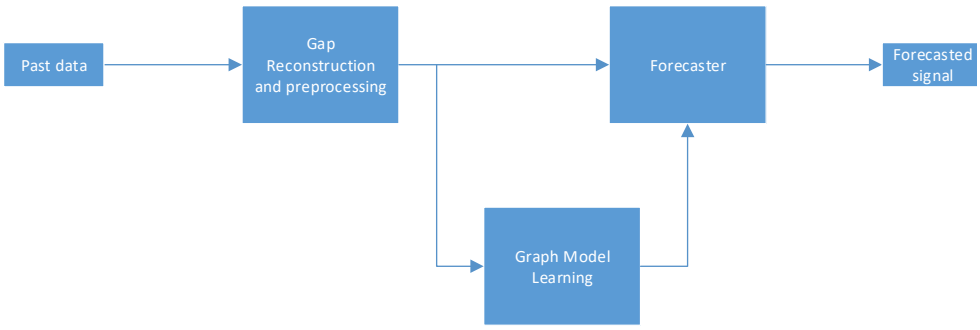


Figure 1. Block diagram of the proposed approach for robust photovoltaic (PV) production forecasting.

2.1. Graph-Based Data Reconstruction: Filling the Gaps

In the following, we detail our proposed approach to fill the gaps in PV power production time series, which is inspired by the work in [27]. Our method is based on building a graph model to capture the spatial dependencies among the PV systems and exploit the spatio-temporal relations to reconstruct the missing parts of the data. The main assumption of our method is that the normalized production data (normalized by peak production or size) are smooth in the temporal as well as in the spatial domain.

Let us begin with some definitions. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ denotes an undirected weighted graph where $\mathcal{V}$ denotes the set of vertices or nodes, $|\mathcal{V}| = \mathcal{N}$, $\mathcal{E}$ denotes the set of edges or links, $\mathbf{A}$ denotes the weighted adjacency matrix (symmetric $N \times N$), where $A_{ij} > 0$ if nodes i and j are connected ($i \neq j$). Thus, in our case each plant corresponds to a node in the graph $\mathcal{G}$ and the vertices models the spatial relations between N PV plants. We construct the adjacency matrix $\mathbf{A}$ by placing edges for the 10 nearest geographical neighbors of each plant and the weights are computed as a function of the distance between plants using a Gaussian kernel. Figure 2 shows an example of such a graph.

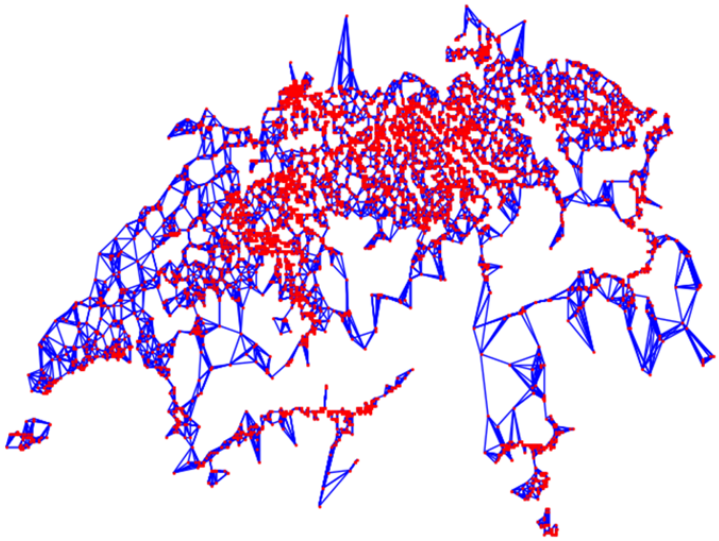


Figure 2. Example of a graph constructed using the locations (coordinates) of the monitored PV plants over Switzerland. The edges are selected using the 10 nearest neighboring plants.

Each node has an associated time-series signal representing the power production over some time interval. Assuming the data is sampled at a regular time, the discrete time series is represented as an M-dimensional vector, where M is the number of time samples. Thus, we can model the graph time-series signal as the matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$, where each row represents the time-series data associated to the nodes. All time series are normalized by the peak production such that the maximum in each row (node) is one. The core assumption in our model is that the temporal derivative of the signal $\mathbf{X}$ (in our case, power production) is smooth in the graph (vertex) domain. This corresponds to the assumption that the effect of clouds on one system propagates mostly unchanged onto neighboring systems. The smoothness of graph signals is a qualitative characteristic that expresses how much the node samples vary with respect to the underlying graph [28]. It can be formalized as follows.

Let $\mathbf{L} \in \mathbb{R}^{N \times N}$ be the graph Laplacian matrix associated to the graph $\mathcal{G}$, defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the (diagonal) degree matrix of the graph in which the main diagonal is defined as $D_{ii} = \sum_{j=1}^{N} A_{ij}$. The graph Laplacian can be interpreted as a difference operator for signals defined on the graph. Let $\mathbf{x} \in \mathbb{R}^N$ be a column vector representing a column of $\mathbf{X}$. A typical measure of graph smoothness for $\mathbf{x}$ is the Laplacian quadratic form defined as:

$$S_2(\mathbf{x}) := \mathbf{x}^T \mathbf{L} \mathbf{x} = \sum_{(i,j) \in \mathcal{E}} A_{ij} (x_i - x_j)^2 \geq 0, \tag{1}$$

which is a weighted sum of neighborhood variation over all nodes. Let $\mathbf{G} \in \mathbb{R}^{M \times (M-1)}$ denote the temporal difference operator such that every row of $\mathbf{Z} = \mathbf{X}\mathbf{G}$ contains the time difference signal for each node. Let $\mathbf{z}_i$ denote the i-th column of $\mathbf{Z}$, thus we can define the graph time-series smoothness function as:

$$F(\mathbf{Z}) = \sum_{i=1}^{M-1} \mathbf{z}_i^T \mathbf{L} \mathbf{z}_i = tr[\mathbf{Z}^T \mathbf{L} \mathbf{Z}]. \tag{2}$$

The function $F(\mathbf{Z}) = F(\mathbf{X}\mathbf{G})$ sums over time the graph smoothness (Laplacian quadratic form) of the time difference signals. The smaller $F(\mathbf{X}\mathbf{G})$ is, the smoother the graph time-series signal $\mathbf{X}$ is.

In addition, the corrupted signal with gaps can be modeled as:

$$\mathbf{Y} = \mathbf{S} \circ (\mathbf{X}_o + \mathbf{W}), \tag{3}$$

where $\mathbf{S} \in \{0,1\}^{N \times M}$ is a binary sampling operator that models the gaps in the signal, $\mathbf{X}_o$ denotes the original complete clean signal, $\mathbf{W}$ is a realization of white additive Gaussian noise and $\circ$ denotes the Hadamard product operator.

In order to reconstruct the signal $\mathbf{X}_o$ from the observed (incomplete) signal $\mathbf{Y}$, we solve the following convex problem:

$$\min_{\mathbf{X}} tr[(\mathbf{XG})^T \mathbf{LXG}] \text{ subject to } \|\mathbf{S} \circ \mathbf{X} - \mathbf{Y}\|_F \leq \epsilon, \tag{4}$$

where $\|\cdot\|_F$ denotes the Frobenius (or L2) norm for matrices. The problem finds the smoothest graph time-series signal that is consistent with the available data. The constant ϵ defines the radius of the data fidelity constraint that can be related to the measurement noise level, i.e., the closer ϵ is to zero the more accurate we assume our measurements to be. In case $\epsilon = 0$, we assume our measurements are noise free. We developed an efficient algorithm to solve the convex problem based on the projected gradient descent method [29]. The algorithm is implemented in Python and is capable of scaling to large networks of tens of thousands nodes.

2.2. Spatio-Temporal Forecast Method

The developed forecast method is based on the assumption that the current power production in a system can be modeled as a linear combination of the past production data of a subset of nodes over a predefined time interval. The rationale behind this model is that events measured in the past production of some nodes, for example, clouds or storms passing by, are informative to predict the production in other nodes. In the following, we describe the main steps in the proposed forecast method.

2.2.1. Normalization and De-Trending

Normalization and de-trending of time-series data become a key step because the main assumption of statistical linear methods is that the data is stationary. However, the power production data is not stationary and has a strong dependency on time (both daily and seasonally). The normalization (and de-trending) method we chose is based on a data-driven computation of the clear sky production for each individual node. We construct for each node and each day of the year a normalization profile based on historical data from the previous year as follows. We first compute over a year of historical data a maximal 24h-production profile at each node x, defined as:

$$P^x_{\max}(t) = \max_{d \in \text{year}} \bar{P}^x_d(t), \tag{5}$$

where $\bar{P}^x_d(t)$ is the PV power produced at node x at day d, and t is the day timestep. The max-profile $P^x_{\max}$ undergoes three further transformations. It is first smoothed out using the Savitzky–Golay filter. Then, the profile is stretched in a day-dependent fashion to reflect seasonal variations. Our stretching/squeezing technique takes into account historical sunrise and sunset times, i.e., defined as the first and last non-zero historical production times of the day (these values are inferred over an interval centered on the specific day, to avoid issues on very cloudy days). It aligns the normalization signal so as to ensure that the tails of the normalization profile are very close to the expected daily sunrise and sunset, see Figure 3. Finally, it is multiplied by a daily scaling factor

$$\alpha^x_d = \frac{\max_{t \in d} \text{GHI}^x_t}{\max_t \text{GHI}^x_t}, \tag{6}$$

where GHI_t^x is the clear sky global irradiance at node x and time t, computed with Ineichen–Perez turbidity model and PV-lib [30]. This last step ensures that the normalized signals are in the same range, independently of the period of the year. By doing so, we end up with a normalization profile $\check{P}_t^x$ for each time step t of the original (yearly) time series.

In the forecasting method, we consider the normalized data $P_t^x / \check{P}_t^x$ that captures the cloud motion without having a strong dependency on seasonal and daily patterns, hence is expected to be relatively well described with linear autoregressive models, see Section 2.2.2. In the night time both values are zero so the ratio is undefined yet it must have a value for forecasting to be possible in the first hours of the day. We decided to define the night-time normalized value as the average value of the ratio in the previous day since this is an unbiased estimator, which assumes a form of persistence. Unless confusions may arise, we denote for the rest of the paper the normalized signals with P_t^x. Figure 3 shows an example of a normalized signal for a particular node.

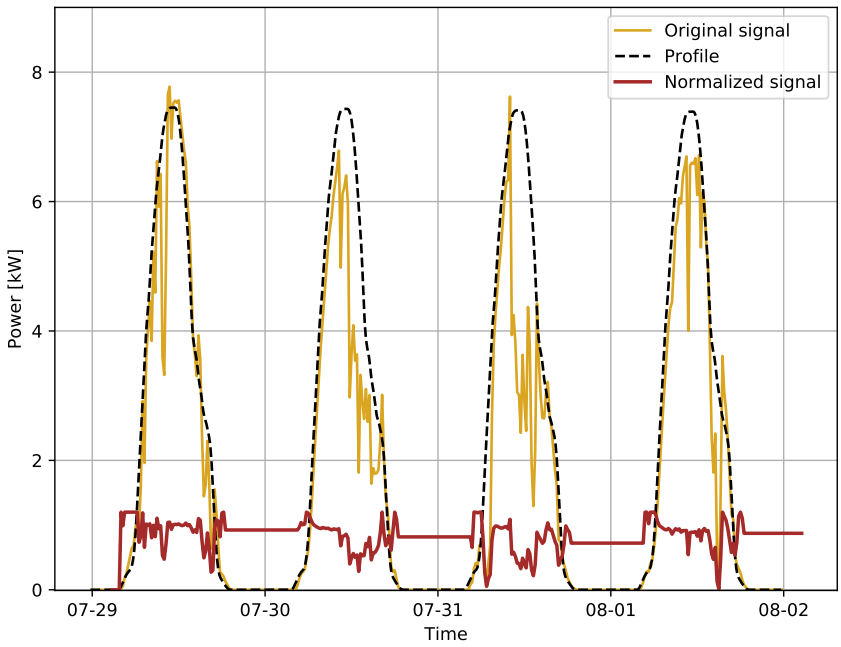


Figure 3. Example of normalization of the PV production signal at one site over four days: original signal (gold), clear sky profile (black, dashed) and normalized signal (brown).

2.2.2. Spatio-Temporal Auto-Regressive Forecast Model

The linear spatio-temporal auto-regressive (ST-AR) model used for predicting the production at site $x \in \mathcal{N}$, where $\mathcal{N} = \{x_1, x_2, \ldots, x_N\}$ denotes the set of N nodes (sites), at time t is the following:

$$P_t^x = \beta_0^x + \sum_{y \in \mathcal{N}} \sum_{i=1}^{Q} \beta_i^{x,y} P_{t-i}^y + \eta_t^x, \tag{7}$$

where P_t^x denotes the normalized PV power production in site x at time t, Q is the past history horizon (in discrete samples), i.e., the order of the AR model or the number of past samples we keep in the model, η_t^x is the error term, and $\beta_i^{x,y}$ is the model coefficient for lag i and site y. We assume that all time series are

sampled at the same rate and at the same time. To learn the model coefficients $\beta_i^{x,y}$ from R measurements, we would need to solve the problem

$$\begin{aligned}\hat{\boldsymbol{\beta}}^x &= \arg\min_{\boldsymbol{\beta}} \sum_{k=1}^{R} \left(P_{t+k}^x - \left(\beta_0^x + \sum_{y \in \mathcal{N}} \sum_{i=1}^{Q} \beta_i^{x,y} P_{t+k-i}^y \right) \right)^2 \\ &= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{P}^x - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 .\end{aligned} \tag{8}$$

In this vectorial form, the vector $\beta^x \in \mathbb{R}^{QN+1}$ is formed by grouping all $\beta_i^{x,y}$ that belong to the same node y, i.e.,

$$\beta^x = [\beta_0^x, \beta_1^{x,x_1}, \ldots, \beta_Q^{x,x_1}, \beta_1^{x,x_2}, \ldots, \beta_Q^{x,x_2}, \ldots, \beta_1^{x,x_N}, \ldots, \beta_Q^{x,x_N}]^T.$$

The measurement vector is defined as $P^x = [P_{t+1}^x, P_{t+2}^x, \ldots, P_{t+T}^x]^T$ and the regressor or design matrix is defined as:

$$\boldsymbol{X} = \begin{bmatrix} 1 & P_t^{x_1} & \cdots & P_{t+1-Q}^{x_1} & \cdots & P_t^{x_N} & \cdots & P_{t+1-Q}^{x_N} \\ 1 & P_{t+1}^{x_1} & \cdots & P_{t+2-Q}^{x_1} & \cdots & P_{t+1}^{x_N} & \cdots & P_{t+2-Q}^{x_N} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & P_{t+R}^{x_1} & \cdots & P_{t+R-Q}^{x_1} & \cdots & P_{t+R}^{x_N} & \cdots & P_{t+R-Q}^{x_N} \end{bmatrix}.$$

The production is modeled as a linear combination of the Q past production values for all nodes in $\mathcal{N}$, i.e., all available nodes. However, a large portion of the nodes might be redundant thus not adding valuable information for the prediction. Therefore, we need a tractable method to learn or infer which subset of nodes is the most informative for prediction. In order to select the most informative coefficients, we proposed to use the group LASSO (Least Absolute Shrinkage and Selection Operator) estimator [31].

The group LASSO adds a regularization term to the classical least squares problem to promote sparse group solutions, i.e., it will drive entire groups of coefficients belonging to the same node to zero, thus having the desired selection effect. The group LASSO estimator solves the convex problem:

$$\hat{\boldsymbol{\beta}}^x = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{P}^x - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda^x \sum_{y \in \mathcal{N}} \|\boldsymbol{\beta}^y\|_2 , \tag{9}$$

where λ^x is a regularization parameter that controls the sparsity of the solution, i.e., the number of nodes selected in the model, and $\boldsymbol{\beta}^y \in \mathbb{R}^Q$ denotes the subvector of coefficients associated to node y. In order to learn the model coefficients for all nodes, we need to solve the group LASSO problem for all nodes in $\mathcal{N}$. By doing so, the problem can be interpreted as a graph learning problem [32] where we learn the topology of a directed graph that minimizes the prediction error. To solve it, we implemented an efficient and scalable algorithm based on the proximal gradient descent method [29].

As is visible from the developed expression in (8), the vectors $\hat{\boldsymbol{\beta}}^x$ depend on the past history horizon. Intuitively, the radius around site x within which nodes can provide valuable information depends on the distance over which clouds can travel. For a very short history horizon (e.g., $Q = 1$), the most informative nodes will therefore be close neighbors to site x. For longer horizons (e.g., $Q = 12$, which corresponds to the longest history horizon in our study) weather changes at some remote nodes will have enough time to propagate to x. The set of nodes selected by the group LASSO will therefore extend over a large area. As an example, Figure 4 shows a set of edges obtained for a central node in Switzerland for $Q = 12$ (three hours of history).

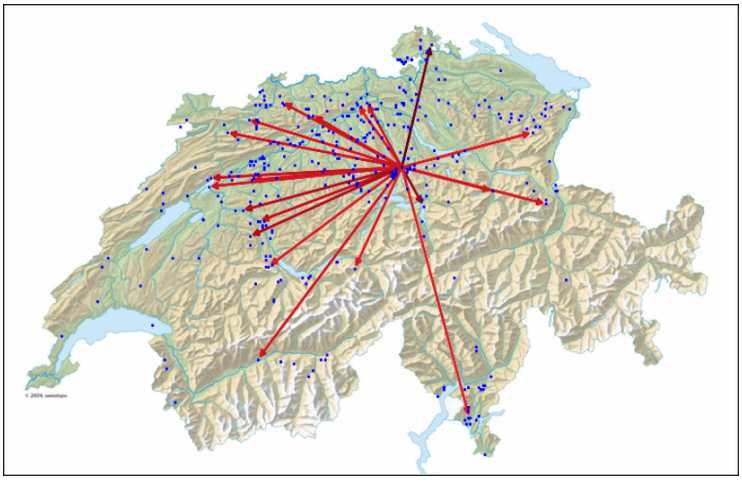


Figure 4. Example of a set of edges obtained by the group LASSO approach for a node in central Switzerland with a past history horizon of three hours.

2.2.3. Prediction for Short-Term Horizon

Once the graph model is learned, we use the model coefficients to forecast the production for all nodes in $\mathcal{N}$ for a horizon of H time steps ahead. In order to do so, we first predict the production for one time step ahead as:

$$\hat{P}^x_{t+1} = \beta^x_0 + \sum_{y \in \mathcal{N}} \sum_{i=0}^{Q-1} \beta^{x,y}_i P^y_{t-i}, \forall x \in \mathcal{N}. \tag{10}$$

Predictions for h time steps ahead, $h \in [2, \ldots, H]$, use the past measured data, P^y_{t+h-i}, $i \in \{h, \ldots, Q-1\}$ as well as the predictions at previous times, i.e., $\hat{P}^y_{t+h-i}$, $i \in \{1, \ldots, h-1\}$. The predictions are computed with the following relation:

$$\hat{P}^x_{t+h} = \beta^x_0 + \sum_{y \in \mathcal{N}} \left(\sum_{i=1}^{h-1} \beta^{x,y}_i \hat{P}^y_{t+h-i} + \sum_{i=h}^{Q-1} \beta^{x,y}_i P^y_{t+h-i} \right), \forall x \in \mathcal{N}. \tag{11}$$

Since the model coefficients were learned using normalized data, the same normalization is applied to the input data to produce normalized predictions. After the predictions are computed a de-normalization step is needed, i.e., multiplication by the daily profiles for each node.

3. Results

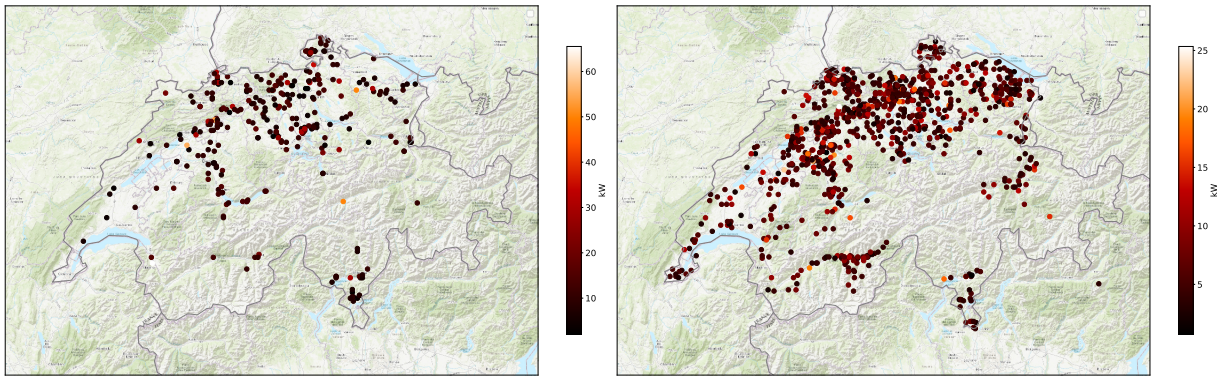
3.1. Evaluation Data and Metrics

We trained and evaluated our methods over two data sets: the "real database" and the "synthetic database".

The "real database" consists in production data from 303 real PV systems distributed over Switzerland and monitored with Solar-Log devices from BKW's subsidiary Solare Datensysteme GmbH (SDS). All data were anonymized before the authors accessed them: any field containing personal information were excluded and spatial coordinates were rounded to 0.01 degrees. All sites have uninterrupted data for 2016 and 2017 with a temporal resolution of 15 min.

The "synthetic dataset" consists in the simulated production of 1000 PV systems distributed in Switzerland, dubbed "synthetic database" for the rest of the article. The synthetic dataset is included to evaluate the scalability of the proposed methods to a larger number of nodes. We used the ModelChain class in the pvlib python library [30] to generate the power production time series for each system

in the database. Using this class requires describing the location of the PV system, its configuration, and providing prepared weather data. To define the locations we fitted a Gaussian mixture model (GMM) with 80 components on the locations of the PV plants in the real database. From this GMM we sampled the locations of the synthetic database. Then we inferred the distributions of the size, orientation and pitch angles of the systems in the real dataset and sampled from these distributions to define the configurations of the PV systems in the synthetic datasets. Finally we used as inputs historical weather data in 15 min resolution from the HelioClim 3 database (http://www.soda-pro.com/help/helioclim/helioclim-3-overview) for the years 2016 to 2018. Figure 5a,b show the spatial distribution of the synthetic and real datasets, respectively. Colors indicate the peak production of each plant.



(**a**) (**b**)

Figure 5. Spatial distribution of the different datasets. (**a**) Real dataset. (**b**) Synthetic dataset. Colors indicate the peak production of each plant.

In order to evaluate over a full year, we divided the test year (2017) into 24 test periods (batches) of two weeks each. The prediction models were trained with data from the two months prior to each test period. The performance metric for both reconstruction accuracy and forecasting accuracy is the daytime normalized root mean square error (NRMSE) defined at site x and forecasting step (horizon) i as:

$$NRMSE(x,i) = \sqrt{\frac{1}{T}\sum_{t\in\mathcal{S}}\left(\frac{\hat{P}^x_{t+i} - P^x_{t+i}}{P^x_{max}}\right)^2},$$

where P^x_t and $\hat{P}^x_t$ are the ground truth power and predicted power (de-normalized), respectively, of site x at time t, P^x_{max} is the maximum power of site x over the evaluation period $\mathcal{S}$, i.e., the 2017 year, and T is the number of time steps in the evaluation interval $\mathcal{S}$. We exclude night periods from the evaluation period $\mathcal{S}$.

3.2. Results of the Graph-Based Algorithm for Data Reconstruction

First we present the performance of the graph reconstruction method. We evaluated it by simulating gaps in the time-series. The simulated gaps are drawn from a statistical model such that the expected length of the gaps per day can be varied. The spatial graphs for each data set (real and synthetic) were constructed using the 20 nearest neighbors and using a Gaussian kernel to compute the weights in the adjacency matrix. We varied the expected length of gaps from 2 to 16 h per day to evaluate the performance of the algorithm over a period of one year in batches of two weeks. The upper bound for the residual $L2$ norm constraint, ϵ, in Equation (4) was set as $\epsilon = 10^{-2}\|\mathbf{Y}\|_F$, i.e., 1% of the total Frobenius norm of the measurement matrix $\mathbf{Y}$.

Figure 6 shows the results of the tests. We compare our reconstruction algorithm to a simple linear interpolation to illustrate the effectiveness of the proposed method when large gaps are present in the data. The results show that the reconstruction error is lower for the synthetic dataset than for the real dataset. In the synthetic dataset the proposed method achieves errors below 20% for gaps with expected lengths up to 8 h. In the real dataset, the proposed method yields errors below 20% for gaps with expected lengths up to 4 h. Figure 7 shows a visualization of a reconstructed signal and the original signal for three nodes in a window of six days. The signals come from the real dataset with simulated gaps with an expected length of 8 h per day.

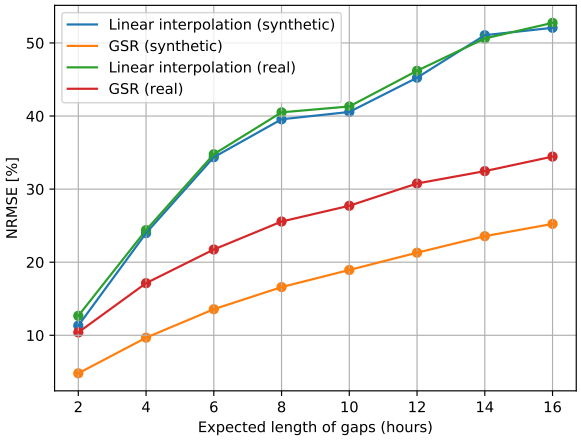


Figure 6. Daytime normalized root-mean-square error (NRMSE) for signal reconstruction against expected length of gaps (in hours) per day in the data for both the real and synthetic datasets. "GSR" stands for the graph signal reconstruction technique introduced in this paper.

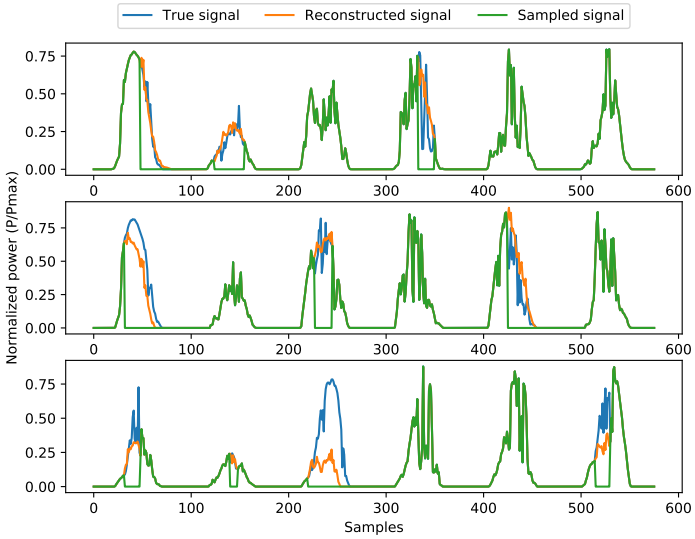
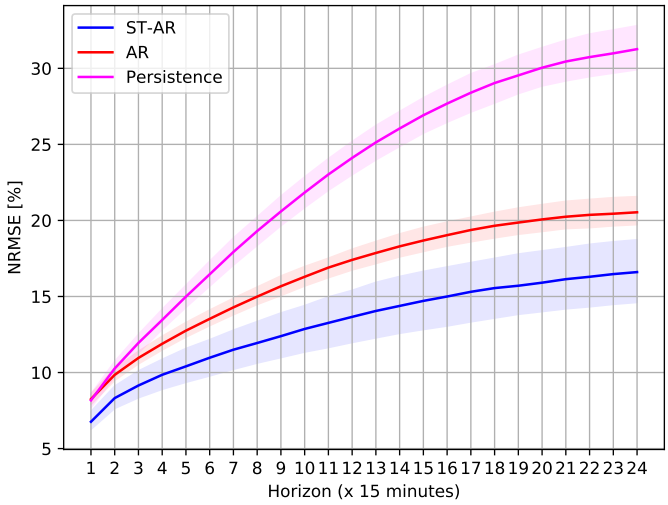


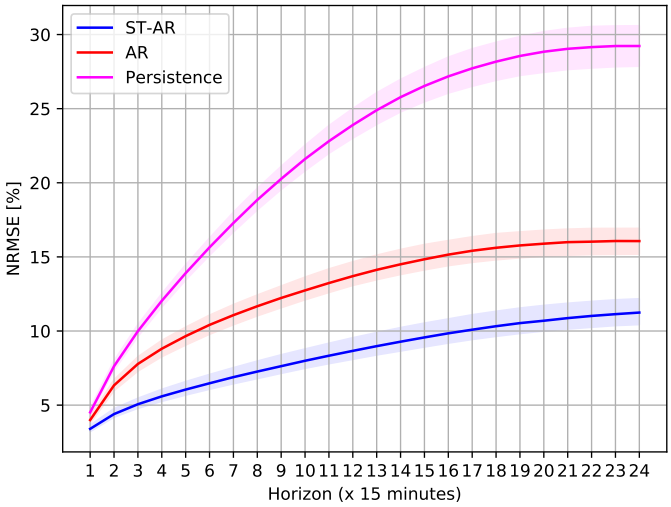
Figure 7. Reconstruction example of a corrupted signal with gaps of expected length of 8 h per day. Visualization for three nodes in a window of six days. The signals are normalized by the peak production.

3.3. Forecasting Results Using Uninterrupted Data

We first evaluated the performance of the proposed ST-AR model using clean and uninterrupted data. The forecast horizon was set to 6 h ahead, i.e., $H = 24$ samples ahead with a sampling time of 15 min. The order of the model was set to $Q = 12$, i.e., we used the past three hours to forecast. We first compared the ST-AR model against the persistence model and the single-site linear auto-regressive (AR) model, i.e., similar to the model defined in Equation (7) but using only information from the same node. The data for the AR model was normalized by the same data-driven clear-sky profiles described in Section 2.2.1. Figure 8a,b shows the forecast error evolution over the prediction horizon (in discrete time steps) for the real dataset and the synthetic dataset, respectively. For each prediction step the median value (solid line) and the interquartile distance (shadow bounds) over all sites are shown.



(**a**)



(**b**)

Figure 8. Comparison between spatio-temporal autoregressive (ST-AR), single autoregressive (AR) and persistence models. (**a**) Forecast NRMSE for the real dataset. (**b**) Forecast NRMSE for the synthetic dataset. The forecast horizon is six hours in steps of 15 min. Solid lines show the median error while the shaded areas show the inter-quantile distance of the errors.

For both datasets, the ST-AR model outperforms the persistence and single-site AR models for the entire forecasting horizon. The maximum errors were obtained for the last time step of the prediction horizon, i.e., 6 h ahead prediction, as expected. In the case of the synthetic dataset, all median errors are below 12% whereas for the real dataset all median errors are below 17%, which is better than state-of-the-art methods for the same horizon [1]. Moreover, for forecast horizons up to 16 steps ahead, i.e., 4 h, the median errors are below 10% for the synthetic dataset and 15% for the real data set.

For the real dataset, we have also included the forecasts for single sites and benchmarked the performance of the proposed method against a method that uses NWP. The forecasts were computed for two locations at which measured PV data and NWP were available, Bern and Bätterkinden. The two locations are about 25 km from each other. As comparison we use forecasts based on support vector regressors (SVR) with NWP as inputs (global irradiance and temperature). As reported in [33], this technique outperforms several state-of-the-art methods such as SARIMA and two neural network models with different weather forecasts. It is therefore a representative benchmark. The forecasts for Bern were computed using historical NWP from the global forecast system (GFS) that have a temporal resolution of 3 h, thus only two points are available for the 6 h ahead horizon. The forecasts for Bätterkinden were computed using historical NWP from Meteotest (https://meteotest.ch/en/) with a temporal resolution of 1 h. Figure 9 presents the time series of actual PV production and six-hour-ahead forecasts using both the ST-AR approach and SVR with NWP. It illustrates the variability in production profiles and forecast errors, which explains the spread seen on Figure 8. Figure 10 shows the dependence of the corresponding NRMSE values on the forecast horizon. The results highlight the dependency of forecasts models that use NWP on the temporal resolution of available weather predictions, whereas the ST-AR overcomes this issue by using past PV production data from other PV stations. The ST-AR approach outperforms the combination of NWP and SVR up to forecast horizons of four to five hours.

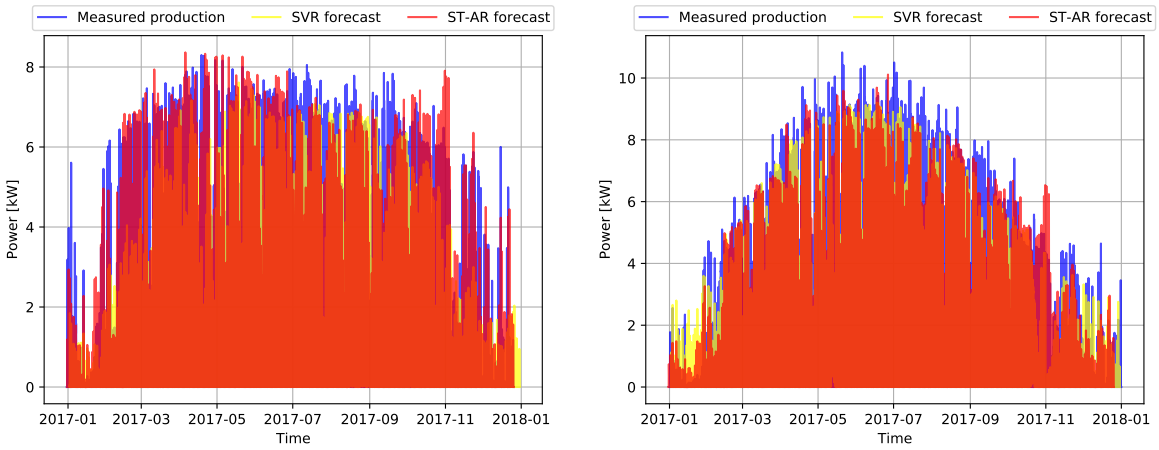


Figure 9. Illustration of the annual PV production for the selected sites. (**a**) Measured and forecasted production for Bern. (**b**) Measured and forecasted production for Bätterkinden. The forecasts are done every six hours with a horizon of six hours.

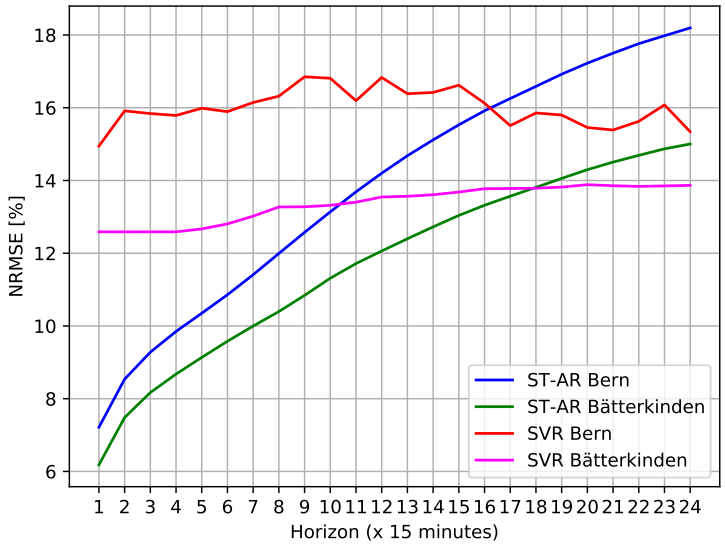
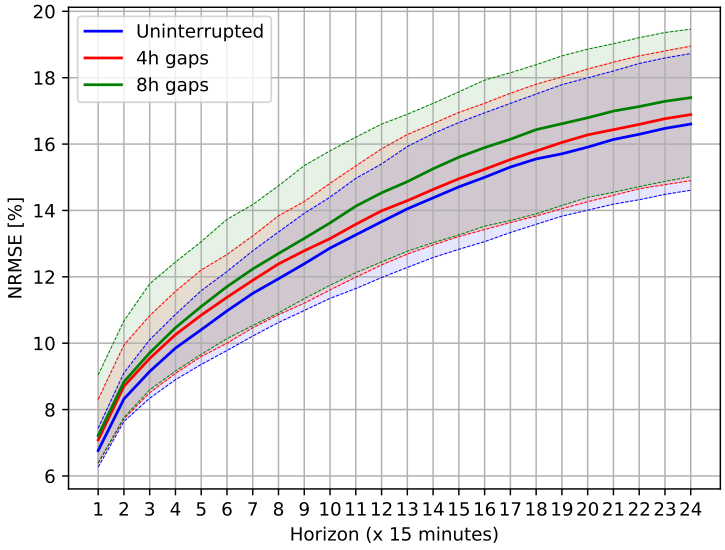


Figure 10. Forecast error (NRMSE) comparison between ST-AR and SVR with numerical weather predictions (NWP) for a single site. The forecast horizon is six hours in steps of 15 min.

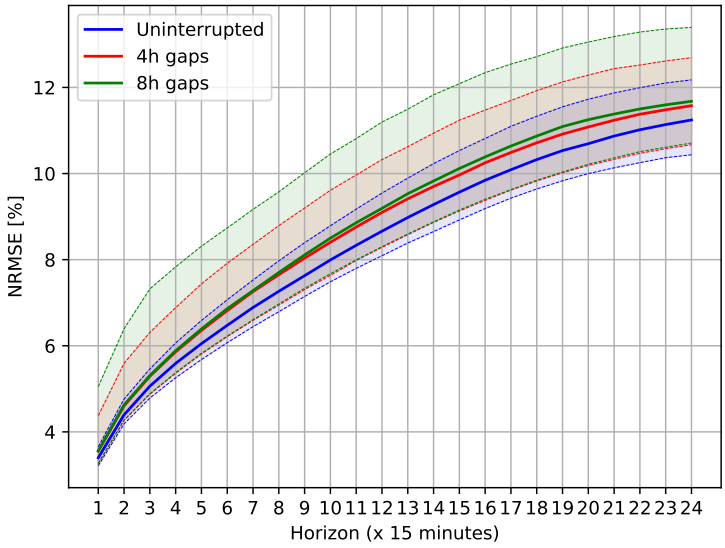
3.4. Forecasting Results Using Incomplete Data

Finally, we evaluated the complete forecast framework, which receives incomplete data, fills the gaps using the graph-based reconstruction method and then uses the cleaned data for forecasting. We evaluated the proposed approach using two different scenarios: (1) gaps with 4-h duration on average and (2) gaps with 8-h duration on average. The model learning phase of the ST-AR method, i.e., solving Equation (9), was performed using the reconstructed data. Figure 11a,b show the forecasting error results for the real and synthetic data sets, respectively, comparing against the results obtained using uninterrupted data. For each prediction step the median value (solid line) and the interquartile distance (shadow bounds) over all sites are shown.

The results show the robustness of the proposed approach with the median NRMSE for both 4- and 8-h gaps staying slightly above the median error from uninterrupted data. However, the interquartile distance of the NMRSE is increased for both datasets and scenarios. Moreover, in the case of the real dataset, the interquartile distance of the errors is closer to the one obtained using uninterrupted data for the first scenario. The average daytime NMRSE over the entire 6 h ahead horizon is 13.8% and 9%, for the real and synthetic databases, respectively, for the first scenario, and 14.5% and 10%, for the real and synthetic databases, respectively, for the second scenario. In comparison, the average NRMSE using uninterrupted data is 13.5% for the real data set and 8.5% for the synthetic data set, showing the effectiveness of the proposed framework to deal with incomplete data.



(**a**)



(**b**)

Figure 11. Forecast error comparison between ST-AR with uninterrupted data and ST-AR with data with gaps. (**a**) Forecast NRMSE for the real dataset. (**b**) Forecast NRMSE for the synthetic dataset. The forecast horizon is six hours in steps of 15 min. Solid lines show the median error while the shaded areas show the inter-quantile distance of the errors.

4. Conclusions

We proposed a robust graph-based framework that exploits the spatio-temporal correlation between different systems to forecast the local PV production based on imperfect production measurements. The proposed method achieves an average daytime NRMSE (over all sites) of 13.5% for the real data set and 8.5% for the synthetic data set for forecasting horizon of 6 h ahead (15 min resolution). It outperforms methods based on numerical weather forecasts on time horizons between 0 h and 4 h, confirming the intuition that PV systems can effectively be used as distributed weather stations for forecasting purposes.

The proposed graph-based method proved very effective in reconstructing missing or faulty data. The NRMSE of this reconstruction is well below 20% for gaps of up to 4 h on the real dataset. Not only is the uncertainty on the reconstruction lower with graph-based methods than with conventional, linear interpolation, its increase with an increasing duration in the gaps is also slower. As a result, the full forecasting framework proved very robust against faulty data: the median NRMSE across all the systems with the ST-AR method is 17% for a 6 h horizon over one year and increases by less than one percentage points with gaps in data of up to 8 h.

Finally, as compared to the state of the art, this work has significantly increased the confidence in the investigated algorithms since they have been tested on entire years and on uniquely large datasets: more than 300 real PV systems spread over Switzerland, and 1000 synthetic ones that reproduce the statistical distribution of installed PV in the country in terms of size, orientation and location. Future work will investigate more advanced graph-based machine learning models such as graph neural networks to learn models that overcome the non-stationarity of PV production signals. We will also investigate the generalization of the approach to other weather-dependent signals in the power system such as wind power production.

Author Contributions: Conceptualization, R.E.C. and P.-J.A.; methodology, R.E.C., M.L. and P.-J.A.; software, M.L., B.S., R.L. and T.C.; validation, R.E.C., B.S. and R.L.; formal analysis, R.E.C., B.S. and P.-J.A.; investigation, R.E.C., M.L. and B.S.; resources, C.T.; data curation, M.L., R.L. and C.T.; writing—original draft preparation, R.E.C. and P.-J.A.; writing—review and editing, R.E.C., B.S., C.T. and P.-J.A.; visualization, R.E.C. and B.S.; supervision, R.E.C.; project administration, P.-J.A.; funding acquisition, P.-J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out on behalf of and with the support of the Swiss Federal Office of Energy (research contract SI/501803-01). The authors are solely responsible for the content and conclusions of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alet, P.J.; Efthymiou, V.; Graditi, G.; Henze, N.; Juel, M.; Moser, D.; Nemac, F.; Pierro, M.; Rikos, E.; Tselepis, S.; et al. Forecasting and Observability: Critical Technologies for System Operations with High PV Penetration. In Proceedings of the 32nd European Photovoltaic Solar Energy Conference and Exhibition (EUPVSEC), Munich, Germany, 20–24 June 2016.
2. Antonanzas, J.; Osorio, N.; Escobar, R.; Urraca, R.; Martinez-de Pison, F.J.; Antonanzas-Torres, F. Review of photovoltaic power forecasting. *Sol. Energy* **2016**, *136*, 78–111. [CrossRef]
3. Visser, L.; AlSkaif, T.; Sark, W.V. Benchmark analysis of day-ahead solar power forecasting techniques using weather predictions. In *Proceedings of the 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*; IEEE: Chicago, IL, USA, 2019; pp. 2111–2116. [CrossRef]
4. Pierro, M.; De Felice, M.; Maggioni, E.; Moser, D.; Perotto, A.; Spada, F.; Cornaro, C. A New Approach for Regional Photovoltaic Power Estimation and Forecast. In Proceedings of the 33rd European Photovoltaic Solar Energy Conference and Exhibition (EUPVSEC), Amsterdam, Netherlands, 25–29 September 2017.
5. Chow, C.W.; Urquhart, B.; Lave, M.; Dominguez, A.; Kleissl, J.; Shields, J.; Washom, B. Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed. *Sol. Energy* **2011**, *85*, 2881–2893. [CrossRef]
6. Marquez, R.; Coimbra, C.F.M. Intra-hour DNI forecasting based on cloud tracking image analysis. *Sol. Energy* **2013**, *91*, 327–336. [CrossRef]
7. Jang, H.S.; Bae, K.Y.; Park, H.S.; Sung, D.K. Solar Power Prediction Based on Satellite Images and Support Vector Machine. *IEEE Trans. Sustain. Energy* **2016**, *7*, 1255–1263. [CrossRef]
8. Perez, R.; Kivalov, S.; Schlemmer, J.; Hemker, K.; Renné, D.; Hoff, T.E. Validation of short and medium term operational solar radiation forecasts in the US. *Sol. Energy* **2010**, *84*, 2161–2172. [CrossRef]

9. Sharma, N.; Sharma, P.; Irwin, D.; Shenoy, P. Predicting solar generation from weather forecasts using machine learning. In Proceedings of the 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm), Brussels, Belgium, 17–20 October 2011; pp. 528–533.
10. Bae, K.Y.; Jang, H.S.; Sung, D.K. Hourly Solar Irradiance Prediction Based on Support Vector Machine and Its Error Analysis. *IEEE Trans. Power Syst.* **2017**, *32*, 935–945. [CrossRef]
11. Yang, C.; Thatte, A.A.; Xie, L. Multitime-Scale Data-Driven Spatio-Temporal Forecast of Photovoltaic Generation. *IEEE Trans. Sustain. Energy* **2015**, *6*, 104–112. [CrossRef]
12. Agoua, X.G.; Girard, R.; Kariniotakis, G. Short-Term Spatio-Temporal Forecasting of Photovoltaic Power Production. *IEEE Trans. Sustain. Energy* **2018**, *9*, 538–546. [CrossRef]
13. Xu, J.; Yoo, S.; Heiser, J.; Kalb, P. Sensor network based solar forecasting using a local vector autoregressive ridge framework. In Proceedings of the 31st Annual ACM Symposium on Applied Computing, SAC '16, Association for Computing Machinery, Pisa, Italy, 4–8 April 2016; pp. 2113–2118.
14. Kashyap, Y.; Bansal, A.; Sao, A.K. Spatial Approach of Artificial Neural Network for Solar Radiation Forecasting: Modeling Issues. *J. Sol. Energy* **2015**, *2015*, 410684. [CrossRef]
15. Ghaderi, A.; Sanandaji, B.M.; Ghaderi, F. Deep Forecast: Deep Learning-based Spatio-Temporal Forecasting. *arXiv* **2017**, arXiv:1707.08110.
16. Lee, J.I.; Lee, I.W.; Kim, S.H. Multi-site photovoltaic power generation forecasts based on deep-learning algorithm. In Proceedings of the 2017 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, 18–20 October 2017; pp. 1118–1120.
17. Zhu, Q.; Chen, J.; Zhu, L.; Duan, X.; Liu, Y. Wind Speed Prediction with Spatio–Temporal Correlation: A Deep Learning Approach. *Energies* **2018**, *11*, 705. [CrossRef]
18. Jeong, J.; Kim, H. Multi-Site Photovoltaic Forecasting Exploiting Space-Time Convolutional Neural Network. *Energies* **2019**, *12*, 4490. [CrossRef]
19. Khodayar, M.; Mohammadi, S.; Khodayar, M.E.; Wang, J.; Liu, G. Convolutional Graph Autoencoder: A Generative Deep Neural Network for Probabilistic Spatio-Temporal Solar Irradiance Forecasting. *IEEE Trans. Sustain. Energy* **2020**, *11*, 571–583. [CrossRef]
20. Lauret, P.; Voyant, C.; Soubdhan, T.; David, M.; Poggi, P. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. *Sol. Energy* **2015**, *112*, 446–457. [CrossRef]
21. Lee, W.; Kim, K.; Park, J.; Kim, J.; Kim, Y. Forecasting Solar Power Using Long-Short Term Memor and Convolutional Neural Networks. *IEEE Access* **2018**, *6*, 73068–73080. [CrossRef]
22. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
23. Lai, G.; Chang, W.C.; Yang, Y.; Liu, H. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Ann Arbor, MI, USA, 8–12 July 2018; pp. 95–104.
24. Killinger, S.; Engerer, N.; Müller, B. QCPV: A quality control algorithm for distributed photovoltaic array power output. *Sol. Energy* **2017**, *143*, 120–131. [CrossRef]
25. González Ordiano, J.; Waczowicz, S.; Reischl, M.; Mikut, R.; Hagenmeyer, V. Photovoltaic power forecasting using simple data-driven models without weather data. *Comput. Sci. Res. Dev.* **2017**, *32*, 237–246. [CrossRef]
26. Heidari Kapourchali, M.; Sepehry, M.; Aravinthan, V. Multivariate spatio-temporal solar generation forecasting: A unified Approach to deal with communication failure and invisible sites. *IEEE Syst. J.* **2019**, *13*, 1804–1812. [CrossRef]
27. Qiu, K.; Mao, X.; Shen, X.; Wang, X.; Li, T.; Gu, Y. Time-Varying Graph Signal Reconstruction. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 870–883. [CrossRef]
28. Shuman, D.; Narang, S.; Frossard, P.; Ortega, A.; Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* **2013**, *30*, 83–98. [CrossRef]
29. Combettes, P.L.; Pesquet, J.C. Proximal Splitting Methods in Signal Processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*; Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H., Eds.; Springer: New York, NY, USA, 2011; pp. 185–212.

30. Stein, J.S.; Holmgren, W.F.; Forbess, J.; Hansen, C.W. PVLIB: Open source photovoltaic performance modeling functions for Matlab and Python. In Proceedings of the 2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC), Portland, OR, USA, 5–10 June 2016; pp. 3425–3430.
31. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2006**, *68*, 49–67. [CrossRef]
32. Dong, X.; Thanou, D.; Rabbat, M.; Frossard, P. Learning graphs from data: A signal representation perspective. *IEEE Signal Process. Mag.* **2019**, *36*, 44–63. [CrossRef]
33. Boegli, M.; Pierro, M.; Moser, D.; Alet, P.J. Machine learning techniques for forecasting single-site PV production. In Proceedings of the 34th European Photovoltaic Solar Energy Conference and Exhibition (EUPVSEC), Brussels, Belgium, 24–27 September 2018.