Smart4RES

Towards a generic seamless forecasting approach for multiple time scales

D3.2 Towards a generic seamless forecasting approach for multiple time scales

WP3, T3.2

Version V2.0

Authors:
Dennis van der Meer, ARMINES
Simon Camal, ARMINES
Akylas Stratigakos, ARMINES
Georges Kariniotakis, ARMINES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Disclaimer

The present document reflects only the author's view. The European Innovation and Networks Executive Agency (INEA) is not responsible for any use that may be made of the information it contains.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Technical references

Project Acronym	Smart4RES
Project Title	Next Generation Modelling and Forecasting of Variable Renewable Generation for Large-scale Integration in Energy Systems and Markets
Project Coordinator	ARMINES – MINES Paris
Project Duration	November 2019 – April 2023

Deliverable	D3.2 Towards a generic seamless forecasting approach for multiple time scales
Dissemination level[1]	PU
Nature[2]	R
Work Package	WP 3 – Data Science and the Future of RES Forecasting
Task	T 3.2 – Towards a generic seamless forecasting approach for multiple time scales
Lead beneficiary	ARMINES (01)
Contributing beneficiary(ies)	ARMINES (01)
Reviewers	Liyang Han (DTU), Jorge Lezaca (DLR)
Due date of deliverable	April 2022 (M)

1 PU = Public

PP = Restricted to other program participants (including the Commission Services)

RE = Restricted to a group specified by the consortium (including the Commission Services)

CO = Confidential, only for members of the consortium (including the Commission Services)

2 R = Report, P = Prototype, D = Demonstrator, O = Other

Document history		
V	Date	Description
0.1	08/04/2022	First version containing contributions from all partners
0.2	25/04/2022	Comments received from the reviewers
1.0	26/04/2022	Revised version submitted to Task leader and project coordinator
2.0	28/04/2022	Final version submitted to the European Commission

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Executive summary

This Deliverable Report presents the work developed by ARMINES in the frame of Task 3.2 ("Towards a generic seamless forecasting approach for multiple time scales") of the Smart4RES project. The main aim of this Task is to simplify the forecast model chain in light of increasing spatial footprint of virtual power plants (VPPs) and power system specifics such as missing data. The following presents a summary of the developed work and their main outcomes.

Automatic Feature Selection and Forecast Combination. The increasing spatial footprint of VPPs, i.e., a multitude of renewable energy sources (RESs) aggregated over space, requires a large set of input features. This induces collinearity between features and the curse of dimensionality. Filters are model agnostic feature selection methods that reduce the feature set to enhance accuracy and computational performance. In this work, we applied 6 filters that take into account nonlinear relationships and feature redundancy. Results on real-world data from mid-west France showed that the filters combined with an Analog Ensemble (AnEn) outperformed a Vanilla AnEn model that considered all features by **5.9%–15.6%** on average, while improving the computational performance by approximately **90%**. We employed forecast combination to improve forecast reliability. Linear and nonlinear probabilistic forecast combination effectively improve the reliability and sharpness, outperforming Vanilla AnEn by **16.0%–31.2%** on average.

Seamless trajectory forecasts. The increasing penetration of RESs, especially in electricity grids with few synchronous generators, requires accurate and autocorrelated forecasts at high temporal resolution to optimize storage control and maintain the power balance. The standard method to generate trajectory forecasts is to forecast each horizon and model the dependencies between the horizons via a covariance matrix. However, at high temporal resolution and large forecast horizons, such an approach quickly becomes cumbersome (5 min resolution at 48 h ahead requires 576 marginals). In this work, we proposed to simplify the model chain significantly by using a pattern matching model (PMM) that compares the current numerical weather prediction (NWP) forecast to a history of analog NWP forecasts. The results on real-world data of Rhodes, Greece, showed that there is **no statistical difference in the performance** of PMM and the state-of-the-art while increasing the computational performance by approximately **98%**.

Hierarchical forecasts with missing values. Power systems feature an inherent hierarchical structure. Ensuring forecast coherency across a hierarchy presents an emerging challenge in energy forecasting. Proposed reconciliation approaches assume coherent historical observations by construction; this is, however, often violated in practice due to equipment failures. We proposed an end-to-end learning approach that directly handles missing values. We described a conditional stochastic optimization approach based on prescriptive trees for end-to-end learning with missing values that fully utilizes the available data. We validated the proposed approach on real-world data from mid-west France comprising 60 wind turbines and 20 photovoltaic parks. The empirical results showed that end-to-end learning outperforms two-step reconciliation approaches by **2.0%–2.5%** on average while mitigating the adverse effect of missing data.

Key messages from the results presented in this Deliverable:

- Feature selection based on mutual information and forecast combination improve forecast accuracy by 5.9%–15.6% and 16.0%–31.2%, respectively.
- We found no statistical difference in performance between PMM and the state-of-the-art while PMM increases the computational performance by approximately 98%.
- End-to-end learning outperforms reconciliation approaches by 2.0%–2.5% on average while mitigating the adverse effect of missing data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Table of contents

I Introduction . . . **8**
I.1 Purpose and objectives of this Deliverable . . . 8
I.2 Context . . . 8
I.2.1 Electricity markets . . . 8
I.2.2 Characteristics of renewable energy generation . . . 9
I.3 Forecasting renewable power generation . . . 10
I.3.1 A brief introduction to stochastic processes . . . 11
I.3.2 General aspects of forecasting . . . 11
I.3.3 Point forecasts . . . 12
I.3.4 Probabilistic forecasts and scenarios . . . 13
I.3.5 Hierarchical forecasts and missing data . . . 14
I.4 State of the art . . . 16
I.4.1 Probabilistic forecasting and automatic feature selection . . . 16
I.4.2 Probabilistic forecasting and optimal forecast combination . . . 17
I.4.3 Seamless scenarios at very high temporal resolution . . . 18
I.4.4 Hierarchical forecasting with missing data . . . 19
I.5 Contributions . . . 19

II Seamless multi-source univariate and multivariate probabilistic forecasting . . . **21**
II.1 Introduction . . . 21
II.2 Managing high-dimensional data and forecast uncertainty . . . 21
II.2.1 Input features . . . 21
II.2.2 Automatic feature selection . . . 22
II.2.3 Probabilistic forecast combination . . . 23
II.3 Forecasting . . . 25
II.3.1 Probabilistic forecast generation . . . 25
II.3.2 Seamless trajectory forecasts . . . 26
II.3.3 Hierarchical forecasts with missing values . . . 26
II.4 Description of case studies . . . 29
II.4.1 Case study I . . . 29
II.4.2 Case study II . . . 31
II.4.3 Case study III . . . 32
II.4.4 Benchmarks . . . 32

III Results . . . **35**
III.1 Case study I . . . 35
III.1.1 VPP - PV . . . 35
III.1.2 VPP - Wind and PV . . . 39
III.2 Case study II . . . 43
III.3 Case study III . . . 46

IV Conclusions . . . **48**
IV.1 Summary . . . 48
IV.2 Dissemination . . . 49
IV.3 Future Work . . . 50

A Sample Average Approximation . . . **51**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

LIST OF TABLES

Table 1 The combination weights and SLP and BLP parameters for the VPP consisting solely of PV, expressed as "mean ± standard deviation" over all forecast horizons. The weights are optimized based on validation data and tested on the test data. . 36
Table 2 The combination weights and SLP and BLP parameters for the VPP consisting of wind and PV, expressed as "mean ± standard deviation" over all forecast horizons. The weights are optimized based on validation data and tested on the test data. 40
Table 3 Block bootstrapped loss differential presented as $\mu \pm \sigma$ $(2.5\% - 97.5\%)$. 45
Table 4 CRPS, ES and VS skill scores, relative to MuPEn. Note that we compute the mean and standard deviation ($\mu \pm \sigma$) over all forecast horizons for CRPS. 45
Table 5 Average SRMSE (± one standard deviation) per hierarchy level. The best-performing model is underlined in bold font. Bold font indicates that a result does not differ from the best-performing model at the 1% level (Welch's t-test). 47
Table 6 Refer to Fig. 6 for the KPIs of the project. 49

LIST OF FIGURES

Figure 1 In (a), time series of the normalized power output of 19 individual PV plants and the VPP and in (b), time series of the normalized power output of 19 individual wind turbines and the VPP (see Fig. 4 for a measure of the distances between the sites). 10
Figure 2 Example of a 3-level hierarchy (top) and the corresponding aggregation matrix (bottom) with $n = 8$, $n_a = 3$, and $n_b = 5$. 15
Figure 3 A flowchart of the pre-process, forecast, and post-process steps. 25
Figure 4 Overview of the satellite and NWP grid points, as well as the PV systems and wind turbines located in mid-west France. 30
Figure 5 Overview of the NWP grid points on and around the island of Rhodes (Greece) at which forecasts are available. 31
Figure 6 Overview of Key Performance Indicators of the Smart4RES project. 33
Figure 7 Forecast results on the validation set from the VPP consisting solely of PV. In (a), the numerical scores as a function of the forecast horizon. In (b), histograms of the PIT variables combined from all forecast horizons, including the variance. In (c), the proportion of the weights assigned to the feature groups. 35
Figure 8 The proportion of the weights assigned to the feature groups for the VPP consisting only of PV systems for the testing case. 36
Figure 9 The numerical scores as a function of the forecast horizon for the VPP consisting only of PV systems. The scores have been computed on data normalized using the installed capacity. 37
Figure 10 Histograms of the PIT variables combined from all forecast horizons, including their variance, for the VPP consisting only of PV systems. 37
Figure 11 Distribution of the CRPS conditioned on the binned deterministic forecast error of the Vanilla AnEn for 3 forecast horizons and for the VPP consisting only of PV systems. The points represent the average CRPS of each component model. 38
Figure 12 Forecast results on the validation set from the VPP consisting of wind and PV. In (a), the numerical scores as a function of the forecast horizon. In (b), histograms of the PIT variables combined from all forecast horizons, including the variance. In (c), the proportion of the weights assigned to the feature groups. 39
Figure 13 The proportion of the weights assigned to the feature groups for the VPP consisting of PV systems and wind turbines for the testing case. 40

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Figure 14 The numerical scores as a function of the forecast horizon for the VPP consisting of PV systems and wind turbines. The scores have been computed on data normalized using the installed capacity. 41
Figure 15 Histograms of the PIT variables combined from all forecast horizons, including their variance, for the VPP consisting of PV systems and wind turbines. 41
Figure 16 Distribution of the CRPS conditioned on the binned deterministic forecast error of the Vanilla AnEn for 3 forecast horizons and for the VPP consisting of PV systems and wind turbines. The points represent the average CRPS of each component model. 42
Figure 17 Histograms of the marginal PIT variables combined over all forecast horizons and testing instances where the zenith angle is smaller than 85°. 43
Figure 18 CRPS in percent of nominal capacity as a function of forecast horizon. The mean and standard deviation are computed across the 12 testing months. 43
Figure 19 ES and VS averaged over the entire testing set and on a monthly basis. Note that the scores are dimensionless. 44
Figure 20 Aggregated SRMSE for the hierarchy as a function of the number of sampled nodes and the percentage of missing values per node over 5 iterations. Bars correspond to one standard deviation. 46
Figure 21 Performance of `EtE-PF` for missing values at different timestamps. The lines indicate the number of malfunctioning nodes, the shaded areas show one standard deviation. 46

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Acronyms

AnEn Analog Ensemble.

BLP Beta-Transformed Linear Pool.

CDF Cumulative Distribution Function.

CRPS Continuous Ranked Probability Score.

ECMWF European Center for Medium-range Weather Forecasts.

ES Energy Score.

GCT Global Closure Time.

GHI Global Horizontal Irradiance.

GTI Global Tilted Irradiance.

MAE Mean Absolute Error.

MBE Mean Bias Error.

NRMSE Normalized Root Mean Squared Error.

NWP Numerical Weather Prediction.

OLP Opinion linear pool.

PDF Probability Density Function.

PIT Probability Integral Transform.

PMM Pattern Matching Model.

PV Photovoltaic.

QR Quantile Regression.

QRF QR Forests.

RES Renewable Energy Source.

RMSE Root Mean Squared Error.

SLP Spread-adjusted Linear Pool.

SRMSE Scaled RMSE.

TLP Traditional Linear Pool.

VPP Virtual Power Plant.

VS Variogram Score.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

I. Introduction

I.1 Purpose and objectives of this Deliverable

This deliverable proposes forecasting products that enhance the forecast accuracy and simplify the forecast model chain. Indeed, there generally exists a trade-off between forecast accuracy and model complexity where too low accuracy or too high complexity will not result in uptake by the industry. The objectives of the deliverable are the following:

1. Propose automatic feature selection techniques based on information theory to include high-dimensional data while discarding irrelevant or redundant features (e.g., adjacent satellite pixels).
2. Compare linear and nonlinear probabilistic forecast combination methods to enhance forecast calibration, i.e., the statistical similarity between forecasts and observations.
3. Propose to significantly simplify the forecast model chain for high temporal resolution (≤ 5 min) trajectory forecasts from 5 minutes ahead up to 48 hours ahead. The high temporal resolution aspect is relevant to new use cases such as inertia or frequency containment reserve forecasting.
4. Apply the aforementioned methods to real-world data made available by the project partners.

The document starts by defining the context of this work. Then the state of the art of similar forecasting methods is reviewed, and contributions from the proposed approaches are explicitly stated. After presentation of the case studies and the evaluation metrics, finally the results are presented and discussed.

I.2 Context

I.2.1 Electricity markets

A forecast does not have intrinsic value; instead, a forecast can offer value when it is used in a decision-making process. One particular decision-making process relevant to RES plant owners is that of offering energy in electricity markets. This section provides an overview of the organization of short-term electricity markets. However, new use cases such as inertia or frequency containment reserve forecasting emerge. These typically place additional requirements, often more stringent, on the forecasts.

Short-term electricity markets are most often organized in a day-ahead market and an intraday market. Most energy is traded on the day-ahead market, whereas the intraday market allows producers and sellers to adjust their offers considering new information such as the latest weather forecasts. Given that the underlying data generating process of RESs such as wind or solar is stochastic, the intraday market plays an important role for RES plant owners to improve their bids and subsequently maximize their revenue.

In either market, a producer (consumer) sells (buys) bids to (from) the market operator. A bid is defined by a volume of energy for a specific *Market Time Unit* (MTU) to which the producer or seller assigns a price (€/MWh). In the case of wind or solar energy, the assigned price is often set to 0 since the marginal cost to produce a unit of energy is 0. Bidding continues until the *Gate*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Closure Time (GCT), after which the market operator computes the demand and offer curves and determines the market clearing prices. Since the marginal cost of solar power and wind power are 0, the financial gain for a wind or solar power producer is the market clearing price.

Since a forecast is hardly ever correct, the bid placed by the RES plant owner most likely deviates from the actual production. The day-ahead bid can be adjusted by a bid on the intraday market. However, deviating bids on the intraday market have to be compensated on the balancing market by flexible energy producers that often feature high marginal cost. Compensating these imbalances effectively constitutes a penalty for the RES plant owner. Consequently, it can be said that a more accurate forecast results in lower penalties and therefore higher revenues for the RES plant owner.

Unforeseen outages or erroneous forecasts can create disturbances that can cause the frequency to deviate from its nominal operating point in the transmission grid or the voltage to deviate from its nominal operating point in distribution grids. In order to avoid a black-out, active power reserves need to be activated so that the system may return to its nominal operating point. Such a reserve is known as an ancillary service and needs to be available with high reliability and minimal delay, e.g., within 30 seconds for frequency containment reserve or within 15 minutes for automatic frequency restoration reserve. With increasing RES penetration levels, there is a pressing need to generate reliable and informative forecasts at much higher temporal resolution and update frequency in order to offer such ancillary services. Complementing an RES plant with, e.g., batteries, can improve the reliability further at additional costs.

I.2.2 Characteristics of renewable energy generation

Renewable energy generation is considered to be a stochastic process. For instance, the conversion of light to electrical energy depends on many aspects that are challenging to model accurately such as dust formation on solar panels, inverter characteristics, nonlinear temperature dependencies or shading events. Furthermore, PV power and wind power generation are both processes that are correlated in space and time, which is nicely captured by Tobler's first law of geography: "everything is related to everything else, but near things are more related than distance things". Figures 1a and 1b show that nearby PV plants and wind turbines, respectively, follow a highly similar pattern.

A forecast model is only as good as the data it is fed. It has been shown that spatially aggregating RES plants of various kinds, often referred to as a Virtual Power Plant (VPP), smooths the power generation profile (see the VPP subplots in Figs. 1a and 1b). Combining multiple RES plants as a VPP is of practical importance because a single RES plant might not have sufficient capacity to participate in electricity markets. An additional advantage is the aforementioned smoothing; since high variability is particularly challenging to forecast, VPP power output forecasts are often more accurate. The input to the forecast model of a VPP will span a wider area and will be more heterogeneous compared to the input to a single plant forecast model, e.g., a combination of numerous satellite pixels and numerical weather prediction (NWP) grid points. This increases the computational burden. Given emerging use cases that require substantially higher temporal resolution ($\leq$ 5 min), new forecast products need to be fast as well as reliable.

Generally speaking, a forecast horizon h larger than 6 h requires NWP forecasts as input features because the correlation between the current observation y_t and future observation Y_{t+h}[1] is low. However, the governing system of equations in NWP models put forth by meteorologists and atmospheric scientists comprise nonlinear differential equations that are sensitive to the initial conditions. For instance, even if it would be possible to fully model the atmosphere, oceans and land masses, it would still be impossible to observe all of it down to a molecular level. Given different initial conditions, integrating the differential equations forward in time would yield quickly

[1] Note that the future observation is still a random variable, hence the capitalized letter.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

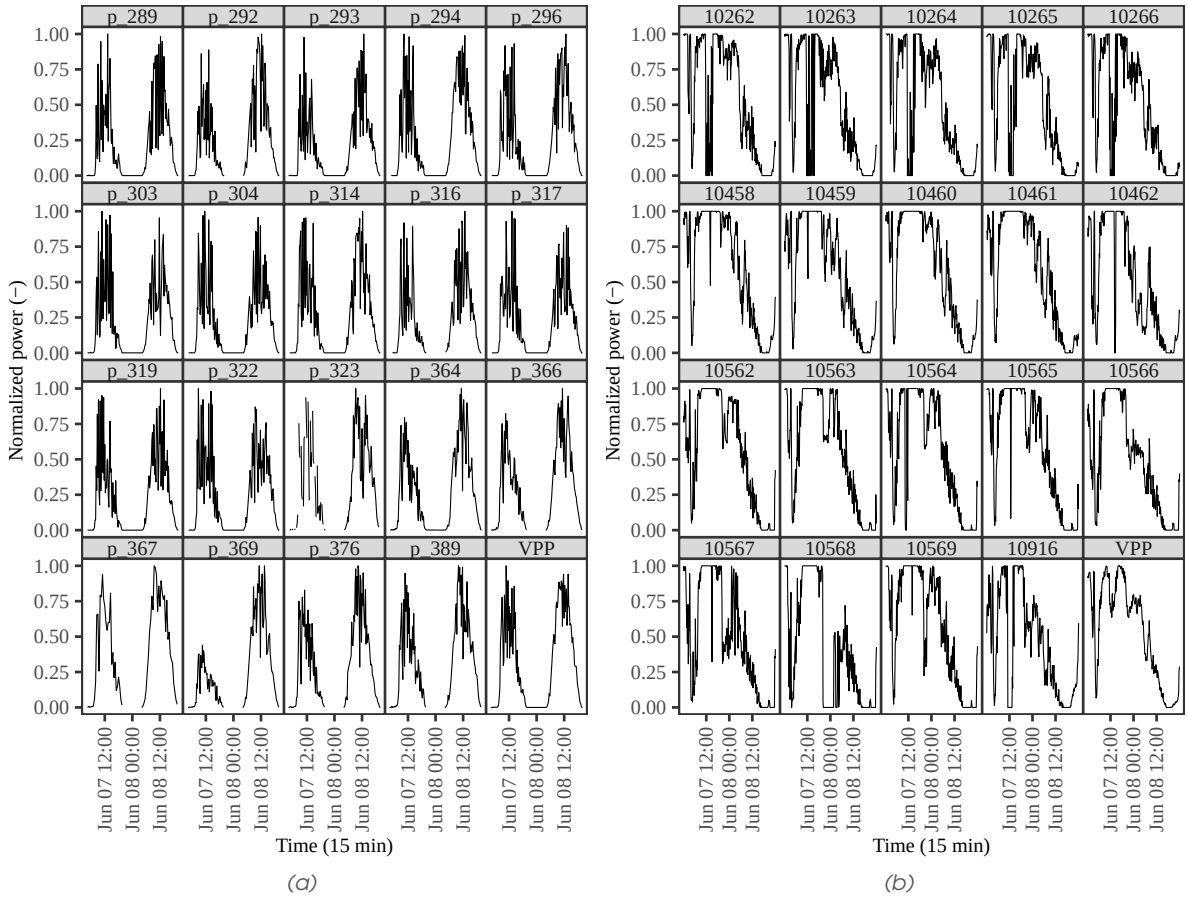


Figure 1 In (a), time series of the normalized power output of 19 individual PV plants and the VPP and in (b), time series of the normalized power output of 19 individual wind turbines and the VPP (see Fig. 4 for a measure of the distances between the sites).

diverging outcomes, which is commonly referred to as dynamical chaos. Given the stochastic nature of the renewable energy conversion process and the uncertainty related to the weather, probabilistic forecasts, as well as space-time scenarios, are typically preferred because it allows forecast providers to communicate the uncertainty in their estimates. The next section will go into more detail on stochastic processes and the various types of forecasts.

I.3 Forecasting renewable power generation

Renewable power generation is a stochastic process, as mentioned in Section I.2.2. Subsequently, a basic understanding of random variables and probability distributions is necessary and a brief introduction will be given in Section I.3.1. A description of general aspects of forecasting will be given in Section I.3.2 with a particular focus on what a good forecast is. Then, Sections I.3.3 and I.3.4 introduce point and probabilistic forecasts, respectively, along with metrics to determine their quality. We note that the following overview is mostly based on the book Morales et al. (2014). This section concludes with an introduction into hierarchical forecasting, which is a specific type of multivariate forecasting.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

I.3.1 A brief introduction to stochastic processes

To understand stochastic processes, it is helpful to first introduce random variables. The materialization of a random variable, often denoted by X, varies due to randomness. For instance, a discrete random variable is whether it will rain tomorrow or not. As such, this random variable takes on values in the set $X \in \{0, 1\}$, each with a probability defined by the probability mass function. In this report, we focus on continuous random variables, of which the probability distribution is defined by the probability density function (PDF) and the corresponding cumulative distribution function (CDF). While the discrete random variable that says whether or not it will rain tomorrow has a probability assigned to each outcome (e.g., $\mathrm{P}\{X = 1\} = 0.80$), the probability of a continuous random variable for any real value is 0, i.e., $\mathrm{P}\{X = x\} = 0 \; \forall \; x$. Consequently, continuous random variables are defined by their cumulative probability $\mathrm{P}\{X \leq x\}$, which is the cumulative distribution function:

$$F(x) = \mathrm{P}\{X \leq x\}, \qquad \forall x, \tag{1}$$

which is a nondecreasing function with a value between 0 and 1, respectively. The PDF is the derivative of the CDF and is therefore defined as:

$$f(x) = \frac{dF}{dx}, \qquad \forall x, \tag{2}$$

which is always positive and integrates to 1. We conclude this brief introduction to random variables by defining the quantile, since it is often used to describe nonparametric probabilistic forecasts. If $F(x)$ is strictly increasing, a quantile q with nominal probability level $\tau \in [0, 1]$ is the unique value x of random variable X:

$$\mathrm{P}\{X \leq x\} = \tau \qquad \Leftrightarrow \qquad q^{(\tau)} = F^{-1}(\tau). \tag{3}$$

A stochastic process is indicated using capitalized letters and it is common in forecasting to use Y to denote the stochastic process. Specifically, $\{Y_{r,s,t}\}$ is a multivariate stochastic process in the sense that it occurs over time $t = 1, \cdots, T$, locations $s = s_1, \cdots, s_n$ and renewable energy sources $r = r_1, \cdots, r_m$. Correspondingly, the observations of the multivariate stochastic process are denoted using lowercase letters: $\{y_{r,s,t}\}$. In this report, the focus is on stochastic processes of the type $\{Y_t\}$ since the types of RESs and all locations are aggregated to a single VPP. The stochastic process $\{Y_t\}$ is particularly relevant for RES plant owners when participating in electricity markets or problems involving optimal control. Furthermore, Y_t is equal to the power generation divided by the installed capacity. The consequence is that error metrics can be presented as percentages of installed capacity.

I.3.2 General aspects of forecasting

Section I.2.1 gave a brief introduction to electricity markets where bids needs to be placed before GCT. Consequently, the production of the RES plant is still a random variable and needs to be estimated before the plant owner can submit a bid. In case of thermal generators such as gas or coal fired generators, such forecasts are trivial since their power output can be controlled (barring failures). However, the production of an RES plant can vary substantially from day to day and from month to month, i.e., so-called seasonality. Exceptions for RES plants do exist, e.g., hydro power can make use of the fact that water can be stored in massive quantities, thus enhancing the flexibility of such a system. However, in this report the focus is mainly on solar and wind power, and the aggregation thereof. Notwithstanding, it is important to note that the methods presented here could be extended to include other RESs.

An important aspect of the forecasting process is forecast verification, i.e., the question whether a series of forecasts for a time period is any good. Murphy (1993) described three types of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

goodness that, when combined, define a good forecast: (i) consistency, which is to say that the forecasts correspond to the forecaster's best judgment; (ii) quality, i.e., the correspondence between the forecasts and the observations; and (iii) value, specifically the incremental value the forecasts bring to decision-makers.
Consistency has slightly different definitions for point and probabilistic forecasts. For the former, when the forecasts are the conditional mean of the predicted probability distribution, consistency is ensured by evaluating the forecasts using the root mean squared error (RMSE). For probabilistic forecasts, consistency is guaranteed by using strictly proper scoring rules that maximally reward forecasters in expectation only when their forecasts correspond to their judgments.
Quality is likely the most familiar aspect of forecast verification. However, there are several ways to evaluate the quality of forecasts. Most often, forecast quality is assessed using measures such as RMSE or the continuous ranked probability score (CRPS) in case of probabilistic forecasts. The CRPS is a strictly proper scoring rule that elicits consistency because it maximally rewards the forecaster if he or she quotes his or her best judgment as the forecast (Murphy, 1993). Besides numerical measures, it is also possible to evaluate time-independent forecast quality through a scatter plot or the probability integral transform histogram.
Value, as mentioned, is the benefit or cost that a decision-maker incurs by acting upon the information in a forecast. While forecasts are not evaluated by their value in this report, there is a growing body of research that develops so-called value-oriented forecasts, see, e.g., Stratigakos et al. (2022).

I.3.3 Point forecasts

A forecast is issued for a particular lead time, herein denoted with k. As such, the forecaster gives an estimate of the random variable $Y_{t+k|t}$, which is to say that the forecaster estimates the random variable at time $t+k$ given all information available at time t. A common way to summarize the uncertainty associated with $Y_{t+k|t}$ is the expectation:

$$\hat{y}_{t+k|t} = \mathbb{E}\left[Y_{t+k}|\hat{\theta}, g, \Omega_t\right], \tag{4}$$

where $\hat{\cdot}$ denotes an estimate, $\hat{\theta}$ represents the trained parameters of model g and Ω_t the available information at time t.

The error of forecast $\hat{y}_{t+k|t}$ can then be defined as:

$$\epsilon_{t+k|t} = \hat{y}_{t+k|t} - y_{t+k}, \tag{5}$$

where $\epsilon_{t+k|t} \in [-1, 1]$ if Y_t is normalized using the installed capacity[2]. Note that eq. (5) is defined such that if $\hat{y}_{t+k|t} > y_{t+k}$, it is an overprediction, whereas conversely it is an underprediction.

Using error $\epsilon_{t+k|t}$, it is possible to measure the accuracy—as a part of the quality—of the forecasts using a suite of measures. The most relevant are given here, starting with the mean bias error (MBE):

$$\text{MBE}(t+k) = \frac{1}{T}\sum_{t=1}^{T}\epsilon_{t+k|t}, \tag{6}$$

where T is the length of the testing set. In case of point forecasts, the parameters $\hat{\theta}$ have been learned from training data by minimizing the sum or mean of squared errors. In that scenario,

[2]Note that this can be exceeded for individual solar power systems due to cloud enhancement, i.e., the occasion where the edge of a cloud acts as a lens that momentarily increases the light intensity and consequently increases the power production above its rated power. However, this is unlikely to occur due to spatial and temporal smoothing.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

the RMSE is a consistent measure and it is defined as:

$$\text{RMSE}(t+k) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\epsilon_{t+k|t})^2}, \tag{7}$$

which is a score that assigns a larger penalty on large forecast errors. Alternatively, the mean absolute error (MAE) is a consistent score if the forecaster issues the median of Y_{t+k} as his or her estimate. It is defined as:

$$\text{MAE}(t+k) = \frac{1}{T}\sum_{t=1}^{T}|\epsilon_{t+k|t}|. \tag{8}$$

Point forecasts are not the main interest of this report and the reader is therefore referred to Morales et al. (2014) for a thorough discussion on this type of forecasts.

I.3.4 Probabilistic forecasts and scenarios

A notable issue with point forecasts is the lack of information concerning the uncertainty. For instance, a point forecast may be informative if tomorrow is expected to be a clear day. However, if broken clouds dominate tomorrow, then it would be helpful for the decision-maker to know what the probability is of generating a certain amount of energy during a time interval. In such a case, probabilistic forecasts—and scenarios as a multivariate extension thereof—are useful because they are designed to inform end-users of the uncertainty that the forecaster has about Y_{t+k}. In other words, instead of the point estimate $\hat{y}_{t+k|t}$ the decision maker is now given $\hat{F}_{t+k|t}$:

$$\hat{F}_{t+k|t}(y) = \text{P}\left\{Y_{t+k} \leq y|\hat{\theta}, g, \Omega_t\right\}, \qquad \forall y. \tag{9}$$

In this report, the forecast $\hat{F}_{t+k|t}$ consists of a sequence of quantile forecasts $\hat{q}_{t+k|t}$ where $\tau \in \{0.05, 0.10, \cdots, 0.95\}$. As mentioned above, consistency can be ensured by using strictly proper scoring rules such as the CRPS:

$$\text{CRPS}(t+k) = \frac{1}{T}\sum_{t=1}^{T}\int_x (\hat{F}_{t+k|t}(x) - H(x - y_{t+k}))^2 dx. \tag{10}$$

where $H(x)$ is the Heaviside step function, whose value is 1 when $x \geq y$ and 0 otherwise. At this point, it is important to note that the quality of probabilistic forecasts comprises three aspects: (i) reliability or calibration; (ii) resolution; and (iii) sharpness.
A series of forecasts is calibrated when it is statistically similar to the observations. Calibration can be evaluated using the probability integral transform over a testing set and allows the forecaster to uncover various kinds of miscalibration. When a series of forecasts is calibrated, the following holds:

$$\hat{F}_{t+k|t}(Y_{t+k}) \sim U\left[0, 1\right], \tag{11}$$

which is to say that calibrated forecasts result in a flat histogram bounded by 0 and 1. Miscalibration such as biases and erroneous dispersion—too wide or too narrow predictive distributions—can easily be gauged from such histograms.
The resolution of a forecast model is related to its capability to generate forecasts that deviate from the average observation. A climatological forecast has zero resolution since it is simply the distribution of all observations.
Sharpness is a measure of the informativeness of the predictive distributions. In the case of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

the climatological forecast, the sharpness is poor because the predictive distributions contains all observations—whereas it is perfectly calibrated. The paradigm in probabilistic forecasting is that one should maximize the sharpness of the predictive distributions subject to calibration. In this report, the root mean variance of probabilistic forecasts is used to measure the sharpness (Gneiting and Ranjan, 2013).

Probabilistic forecasts inform end-users of uncertainty but they do not provide an estimate of the dependencies across time, space or RES. Such information can be highly relevant, for instance when scheduling thermal generators that are constrained by ramping times or when performing a regional study into nodal voltage fluctuations. Then, the aim is to generate a multivariate probabilistic forecast $\hat{\boldsymbol{F}}$ for the entire stochastic process $\{Y_{r,s,t}\}$ or a subset thereof, e.g., along the temporal dimension as in this report.

Since conveying such a multivariate probabilistic predictive distribution is challenging, it is common to approximate this distribution using samples, which we refer to as trajectories or scenarios. Each of these trajectories is an equiprobable sample of the multivariate predictive distribution that, besides the quality aspects mentioned above, should reflect the dependence structure present in the observations.

The most prominent proper scores to evaluate trajectory forecasts are the energy score (ES) and variogram score (VS). Since there is no single forecast horizon k as previously, the multivariate forecasts are only indexed using their issue-time, i.e., $t|t$. ES is a generalization of CRPS and is defined as (Gneiting and Raftery, 2007):

$$\mathrm{ES}\left(\boldsymbol{F}_{t|t}, \boldsymbol{y}_t\right) = \mathbb{E}_{\boldsymbol{F}} \left\|\boldsymbol{X}_{t|t} - \boldsymbol{y}_t\right\| - \frac{1}{2}\mathbb{E}_{\boldsymbol{F}} \left\|\boldsymbol{X}_{t|t} - \boldsymbol{X}'_{t|t}\right\|, \tag{12}$$

where $\boldsymbol{X}_{t|t}$ and $\boldsymbol{X}'_{t|t}$ are independent random vectors sampled from $\boldsymbol{F}_{t|t}$ and $||\cdot||$ represents the Euclidean norm. An important result from the literature is that ES is is unable to discriminate between poorly or correctly specified dependence structures (Pinson and Girard, 2012). Therefore, VS is included to compare the quality of the dependence structure specification, which is defined as (Scheuerer and Hamill, 2015):

$$\mathrm{VS}_p\left(\boldsymbol{F}_{t|t}, \boldsymbol{y}_t\right) = \sum_{i,j=1}^{K} w_{ij} \left(|y_i - y_j|^p - \mathbb{E}_{\boldsymbol{F}} \left|\boldsymbol{x}_i - \boldsymbol{x}_j\right|^p\right)^2, \tag{13}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are components i and j of random vector $\boldsymbol{X}_{t|t}$, which is distributed according to $\boldsymbol{F}_{t|t}$. $\boldsymbol{F}_{t|t}$ is approximated by S K-dimensional trajectories $\begin{pmatrix}\boldsymbol{x}^{(1)} & \boldsymbol{x}^{(2)} & \cdots & \boldsymbol{x}^{(S)}\end{pmatrix}^\top$ and $\mathbb{E}_{\boldsymbol{F}} \left|\boldsymbol{x}_i - \boldsymbol{x}_j\right|^p$ can be approximated by Scheuerer and Hamill (2015):

$$\mathbb{E}_{\boldsymbol{F}} \left|\boldsymbol{x}_i - \boldsymbol{x}_j\right|^p \approx \frac{1}{S}\sum_{s=1}^{S} |x_i^{(s)} - x_j^{(s)}|^p, \qquad i,j = 1,\ldots,K, \tag{14}$$

where $x_i^{(s)}$ and $x_i^{(s)}$ are elements i and j of the s^{th} trajectory forecast. Weights w_{ij} can be used to add or reduce importance between certain forecast horizons. However, all weights are set to 1 and $p = 0.5$ in this report, as recommended by Scheuerer and Hamill (2015).

I.3.5 Hierarchical forecasts and missing data

The power system follows a natural hierarchy where traditionally high capacity thermal power generators serve customers by a complex system of transmission and distribution networks, i.e., a top-down approach. The energy transition implies a diversion from this because distributed power generators such a wind turbines and PV systems can be placed at different levels of the hierarchy. Nevertheless, the hierarchy remains in place because of the various voltage levels

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

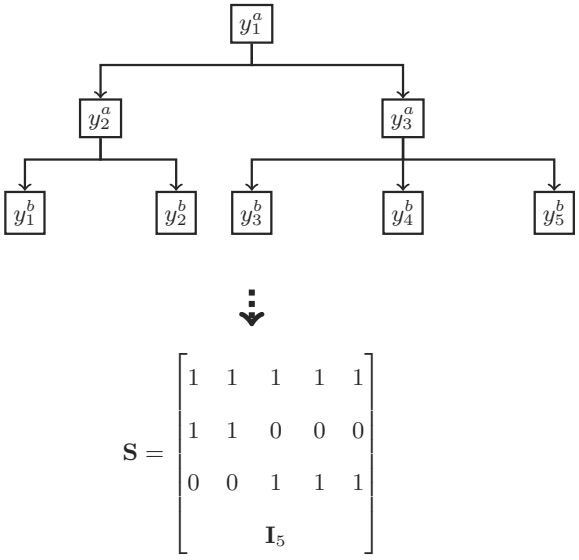


Figure 2 Example of a 3-level hierarchy (top) and the corresponding aggregation matrix (bottom) with $n = 8$, $n_a = 3$, and $n_b = 5$.

in the grid. An important attribute of such a hierarchy is that the power generated at child nodes must sum to the power recorded at the parent node (barring efficiency losses, which are assumed to be zero in this report). Figure 2 presents an example of a hierarchy with 3 levels and 5 bottom nodes. At each of these nodes power can be generated but missing values occur only at the bottom level while the top levels still record the aggregated power correctly. The methodology will be detailed in Section II.3.3. Mathematically, the hierarchy in Fig. 2 can be represented as follows:

$$\mathbf{y}_t = \mathbf{S}\mathbf{y}_t^b \iff \begin{bmatrix} \mathbf{y}_t^a \\ \mathbf{y}_t^b \end{bmatrix} = \begin{bmatrix} \mathbf{S}_a \\ \mathbf{I}_{n_b} \end{bmatrix} \mathbf{y}_t^b, \quad \forall t \in [T] \tag{15}$$

where $\mathbf{S}_a \in \{0,1\}^{n_a \times n_b}$ aggregates the bottom-level series $\mathbf{y}_t^b$, $\mathbf{I}_{n_b}$ is an n_b-size identity matrix, and $[T] := \{1, \ldots, T\}$. Following Rangapuram et al. (2021), a convenient way to represent (15) is given by

$$\mathbf{A}\mathbf{y}_t = \mathbf{0}, \quad \forall t \in [T], \tag{16}$$

where $\mathbf{A} = [\mathbf{I}_{n_a}, -\mathbf{S}_a]^\intercal$ and $\mathbf{I}_{n_a}$ is an n_a-size identity matrix.

In this context, so-called hierarchical forecasts need to behave similarly in the sense that the power forecast at the child nodes should sum to the power forecast at the parent node, which is referred to as a coherent forecast. Besides RMSE, the Scaled RMSE (SRMSE) has been proposed to evaluated hierarchical forecasts (Di Modica et al., 2021). SRMSE is computed by dividing by the total number of child notes. For a set of T forecasts, the SRMSE for the i-th series is defined as:

$$\text{SRMSE}_i = \left(\frac{1}{T} \sum_{t \in [T]} \left(\frac{y_{it} - \widehat{y}_{it}}{s_i} \right)^2 \right)^{\frac{1}{2}}, \tag{17}$$

where s_i the number of child nodes.

In the frame of decision-making, which is the main way to extract value from forecasts, incoherent forecasts can lead to inconsistent decisions across the hierarchy and result in sub-optimal

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

operation of the power system. However, these forecasts are often generated by different stakeholders that do not share data with each other and it is therefore uncommon that the forecasts are indeed coherent. To further complicate matters, it is common to encounter missing or erroneous data in power systems, especially at lower levels of the hierarchy, e.g., smart meters in the homes of consumers; this violates the assumption that the measured power throughout the hierarchy is coherent. Any forecast based on such a data set inevitably leads to incoherent forecasts and data completeness is often implicitly assumed in the literature. This part of the report proposes a forecast model that automatically generates hierarchical forecasts while considering missing values at the most disaggregated level. Note that hierarchical forecasts are in fact a special type of multivariate forecasts.

I.4 State of the art

The last part of this chapter introduces an overview of the state-of-the-art related to the present work. It is divided into subsections that are related to the papers that this report is based on, namely: (i) automatic feature selection; (ii) optimal forecast combination; (iii) seamless scenario forecasts; and (iv) hierarchical forecasts with missing values.

I.4.1 Probabilistic forecasting and automatic feature selection

Section I.2.2 discussed some of the most relevant characteristics of renewable energy generation. One of those characteristics is spatial smoothing, which is particularly relevant when aggregating RES plants as a VPP. However, such a VPP requires significantly more input features since it is likely to cover a much wider area and multiple resources. Two of the challenges that arise when one includes more features are: (i) features may be multicollinear, which is to say that one feature is a (near) linear combination of other features; and (ii) the curse of dimensionality, i.e., when there are simply not enough observations relative to the number of features for a model to learn a meaningful relationship. In other words, feature selection is important to discard redundant features so as to reduce time complexity and retain accuracy.

Feature selection can be categorized into three classes: (i) filtering methods; (ii) wrapper methods; and (iii) embedded methods (Guyon and Elisseeff, 2003). Filtering methods are model agnostic; features are ranked based on a score, e.g., Pearson correlation, and a subset is selected. Yang recently employed an ultra-fast similarity search algorithm to preselect relevant features (Yang, 2018). It is also possible to statistically determine a threshold distance beyond which information is excluded (Yang et al., 2014) or build regime-based models based on wind direction (Amaro e Silva et al., 2019).

Wrapper methods are easily the most computationally intensive out of the three classes; the aim is to recursively subset the feature set, learn a forecast model and compare its output to that of other forecast models on other feature subsets. Owing to their computational burden, wrapper methods are not common. One example is that of van der Meer et al. (2018), who attempted a brute-force search into the best performing subset of relevant endegenous features.

The embedded methods, such as Lasso regression (Tibshirani, 1996), embed feature selection. There is a long list of studies that employ embedded methods because they can be used as stand-alone forecast models or as a regularizer to constrain the model parameters. For instance, Agoua et al. (2018) combined Lasso with quantile regression, whereas Yang et al. (2022) first filtered features using a preselection algorithm after which Lasso-penalized quantile regression was used to generate forecasts.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

I.4.2 Probabilistic forecasting and optimal forecast combination

As mentioned above, the aim in probabilistic forecasting is to maximize sharpness subject to calibration. However, it is often the case that forecasts are not calibrated. For instance, the power may be overestimated or the predictive distributions are underdispersed. In that case, postprocessing techniques such as Ensemble Model Output Statistics (EMOS, (Gneiting et al., 2005)) or Bayesian Model Averaging (BMA, (Raftery et al., 2005)) can be used to remove miscalibration. EMOS takes poor man's ensemble members as input and outputs a parametric PDF. On the other hand, BMA dresses the members of a dynamical ensemble with parametric PDFs and linearly combines these using weights that represent the posterior model probability.[3] Many more methods have been proposed; for an overview the reader is referred to the book by Vannitsem et al. (2019).

This report focuses on the combination of predictive distributions to improve the calibration and potentially sharpness. An important advantage of forecast combination is to reduce the overall uncertainty that arises when selecting a model g, estimating parameters $\hat{\theta}$ or using the available information Ω_t. Heuristic, linear and nonlinear combination methods can be used to combine predictive distributions. A typical example of a heuristic combination method is the opinion linear pool (OLP) in which predictive distributions are linearly combined with equal weights assigned to the experts that produced the forecasts, a method that was introduced already in 1961 (Stone, 1961). Other heuristics exist as well, such as trimming the outer- or innermost predictive distributions; see Yang and van der Meer (2021) for an overview.

An extension of OLP is the traditional linear pool (TLP) in which weights $w_j \forall j = 1, \cdots, m$ are determined by optimizing a loss function (Gneiting and Ranjan, 2013) or based on the inverse of the error of each forecast model (Pauwels and Vasnev, 2016). The CRPS has been used as a loss function to optimize the combination weights of TLP in (Bracale et al., 2017). However, one challenge of such a "static" approach is that these optimal weights do not necessarily perform best on a testing set, something that has been coined the "forecast combination puzzle" and may be caused by variance induced by weight optimization on a small training set (Claeskens et al., 2016). To deal with structural breaks in time series, Thorey et al. (2018) developed an online optimization of the CRPS to linearly combine forecasts, which improved the calibration although underdispersion remained. Berrisch and Ziel (2021) went further and hypothesized that different experts may perform differently over time and within their forecast distributions and therefore developed the fully adaptive Bernstein online aggregation method for pointwise CRPS online learning.

Gneiting and Ranjan (2013) provided theoretical results that indeed TLP is limited in its flexibility. Therefore, they introduced nonlinear combination methods, specifically the spread-adjusted linear pool (SLP) and the beta-transformed linear pool (BLP). In their work, Gneiting and Ranjan (2013) showed that the probability integral transform variables of beta-transformed predictive distributions $\hat{F}_{t+k|t}(Y_{t+k})$ can attain any value in the interval $(0, \frac{1}{4})$, where $\frac{1}{12}$ indicates neutrally dispersed forecasts. SLP was used by Möller and Groß to calibrate the ensemble prediction system of the European Center for Medium-range Weather Forecasts (ECMWF) (Möller and Groß, 2020). BLP was used in the frame of GDP forecasting (Lahiri et al., 2015) and solar irradiance forecasting (Fatemi et al., 2018). As a conclusion, there is need for additional study into nonlinear combination methods applied to renewable energy power production.

[3] A poor man's ensemble is a collection of unperturbed numerical weather prediction (NWP) forecasts from different providers, whereas a dynamical ensemble is a collection of NWP forecasts from the same model but with slightly different initial conditions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

I.4.3 Seamless scenarios at very high temporal resolution

The forecaster communicates his or her uncertainty through the issued probabilistic forecasts. However, recall that renewable power generation is a stochastic process and that there exists a relationship over time and space. Therefore, a forecast error at $t+k$ is likely to propagate to $t+k+1$, which is important to consider when optimizing battery control (van der Meer et al., 2021) or offering automatic frequency restoration reserve (aFRR) and energy on the day-ahead market (Camal et al., 2019). A similar argument can be made for the case where spatial forecasts are issued. As a consequence, it can be valuable to issue so-called scenario or trajectory forecasts. In a purely temporal setting, the forecaster aims to approximate the multivariate distribution using S scenarios with maximum forecast horizon K: $\begin{pmatrix}\boldsymbol{x}^{(1)} & \boldsymbol{x}^{(2)} & \cdots & \boldsymbol{x}^{(S)}\end{pmatrix}^\top$, where $\boldsymbol{x}^{(s)} = \begin{pmatrix}x^{(s)}_{t+1|t} & x^{(s)}_{t+2|t} & \cdots & x^{(s)}_{t+K|t}\end{pmatrix}$.

Such trajectory forecasts can be generated using the combination of a forecast model for each forecast horizon $k = 1, \ldots, K$ and a covariance matrix that describes the temporal dependencies. Recall that a series of forecasts with horizon k is said to be calibrated when $P_k = \hat{F_{t+k|t}}(Y_{t+k}$ is uniformly distributed, i.e., $P_k \sim \mathcal{U}[0,1]$. It is then possible to transform P_k to a standard normal random variable $Z_k \sim \mathcal{N}(0,1)$ using the inverse Gaussian CDF:

$$z_{k,t} = \Phi^{-1}(p_{k,t}), \quad \forall t \tag{18}$$

where $p_{k,t}$ is an instant from P_k at time t. Given $k = 1, \ldots, K$ forecast horizons, the result is a multivariate normal distribution such that $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$. Here, $\boldsymbol{\mu}_0$ is a vector of zeros and $\boldsymbol{\Sigma}$ is a covariance matrix that describes the temporal dependencies with ones on its diagonal (Pinson et al., 2009). An unbiased estimate of $\boldsymbol{\Sigma}$ is (Pinson et al., 2009):

$$\boldsymbol{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N} \boldsymbol{Z}_i \boldsymbol{Z}_i^\top. \tag{19}$$

Using a multivariate normal random number generator with inputs $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$, it is possible to sample S K-dimensional vectors. These can be transformed to uniform variables P_k using the standard normal CDF Φ:

$$p_{s,k} = \Phi(z_{s,k}). \quad \forall s,k \tag{20}$$

These autocorrelated standard uniform samples can finally be used in combination with the predictive distributions $F^{-1}_{t+k|t}$ to generate trajectory forecasts $f_{s,t+k|t}$:

$$f_{s,t+k|t} = F^{-1}_{t+k|t}(p_{s,k}). \quad \forall s,k \tag{21}$$

This method is currently the de facto approach to generate trajectory forecasts, see, e.g., (Pinson, 2013; Camal et al., 2019, 2018; van der Meer et al., 2021). Even though the above-described method is computationally manageable for intra-day or even day-ahead trajectories at hourly resolution, it becomes more challenging when drastically increasing the temporal resolution. For instance, a forecast horizon of 24 hours with a temporal resolution of 1 minute requires 1,440 forecast models, whereas a forecast horizon of 48 hours with a temporal resolution of 5 minutes requires 576 forecast models.

Such high-resolution forecasts may become necessary to perform security-constrained unit commitment on islands featuring high RES penetration and without backup of large-scale synchronous generators. In such a setting, it is pertinent to mitigate critical disturbances such as fast ramps of RES production on both the intra-day and day-ahead horizon. In this part of the report, a seamless model is introduced that generates trajectory forecasts at a fraction of the time while delivering similar performance as the state-of-the-art.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

I.4.4 Hierarchical forecasting with missing data

Traditional approaches for hierarchical forecasting (Petropoulos et al., 2022), i.e., forecasting a group of time series that satisfy a set of linear aggregation constraints, include the bottom-up and the top-down. The bottom-up approach involves forecasting the bottom-level series and aggregating them; however, it usually performs poorly as the signal-to-noise ratio tends to be lower at the bottom-level series. Further, the top-down approach may introduce forecast bias. Thus, a significant body of research focuses on two-step methods, where each series is modeled independently, with individual (named *base*) forecasts being reconciled in a post-processing step. Wickramasuriya et al. (2019) reconciled unbiased base forecasts by minimizing the trace of the forecast error covariance matrix. Van Erven and Cugliari (2015) proposed reconciliation by weighted projection of base forecasts. A constrained multivariate regression framework is described by Di Modica et al. (2021), considering both batch and online learning, applied in wind power forecasting. Moving beyond point forecasts, Taieb et al. (2017) described a bottom-up approach for coherent probabilistic forecasts, while Yang (2020) proposed a block bootstrap method for probabilistic photovoltaic (PV) production forecasts. Recently, end-to-end learning, i.e., training a single model to predict all series in one-shot, has begun to attract attention, as it directly leverages dependencies across series. A deep learning model, with an internal projection step, is presented by Rangapuram et al. (2021) for coherent probabilistic forecasts. Lastly, a general framework for end-to-end forecasting of predictive quantiles is proposed by Han et al. (2021).

Regarding missing values, a recurring assumption in the literature is that training observations are coherent by construction. In practice, however, missing or erroneous values are commonplace due to communication failures or equipment malfunctions. Two main approaches are identified for dealing with missing values. The first is to ignore observations with missing values (i.e., complete case analysis), which is typically applied in the works mentioned above. However, in an end-to-end learning setting, disregarding observations leads to significant information loss. Further, if data are not missing at random, bias might be introduced. The second approach involves missing data imputation; in turn, this raises the problem of ensuring that imputed values are coherent. To this end, Liu et al. (2015) proposed an iterative algorithm to impute missing values while exploiting dependencies across series. The present work examines the case of missing values in the lower levels of the hierarchy, but accurate measurements in the upper levels, which is of practical interest in power system applications. For example, smart-meters might fail to transmit consumption data at a household level while the respective distribution feeder properly measures aggregated demand.

I.5 Contributions

The present report contributes to the state-of-the-art on the topic of probabilistic and multivariate forecasting of the output of RESs organized as a VPP. Specifically, the contributions can be summarized as follows:

- As more and more data become available, it becomes increasingly challenging to extract useful information due to multicollinearity of the features and the curse of dimensionality. To that end, filtering methods are applied to generate a feature subset with less redundancy in a model agnostic approach. Filters based on mutual information have the advantage that they can uncover nonlinear relationships and allow for automatic feature selection. As such, they can be used efficiently in an operational setting. This approach significantly improves the computational performance although the individual forecasts become underdispersive.
- Underdispersed forecasts stem from overconfident forecast models, which in this case is

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

caused by feature subsets that do not completely represent the original feature set. The combination of probabilistic forecasts is a powerful method to mitigate overconfident models. Here, linear and nonlinear combination methods are investigated that substantially improve the calibration of the probabilistic forecasts compared to the original ones. Furthermore, it leads to interesting perspectives for future works.

- Future power system operation requires probabilistic forecasts that are correlated in time and/or space at high temporal and spatial resolution. The current state-of-the-art then becomes computationally too intensive and alternative methods are required. To that end, a pattern matching algorithm is proposed that downscales NWP ensemble forecasts to multivariate probabilistic power forecasts at any temporal resolution lower than or equal to the native resolution. The proposed algorithm significantly reduces the computational burden and simplifies the forecast model chain, while the use of kd-tree ensures that the model can be applied to any RES power forecasting problem.
- Finally, missing values are commonplace in power systems. However, simply ignoring these may introduce biases if the source of the missing values is systematic. As of now, there is a knowledge gap on how to handle missing values in a hierarchical setting. Therefore, a decision tree algorithm for end-to-end forecasts with missing values is proposed that does not require imputation and fully utilizes the available training data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

II. Seamless multi-source univariate and multivariate probabilistic forecasting

II.1 Introduction

The methodology described in this chapter is based on two studies. The first study concerns high-dimensional data—data from various source such as satellite imagery or NWP forecasts—and how these can be used operationally. This study also concerns forecast combination as a means to reduce the uncertainty introduced when reducing the dimensionality of the data set. Here, the forecasting algorithm itself is not a novelty but based on the state-of-the-art. Section II.2 describes the methods employed in the first study.

Whereas the first study focuses on probabilistic forecasts, the second study focuses on a temporal sequence of probabilistic forecasts, i.e., trajectory forecasts. Instead of focusing mainly on the input features, here the focus lies on how to speed up the de facto method described in Section I.4.3, which is relevant when moving to very high temporal resolution and operational settings. Section II.3 details the forecast methods for both the first and the second studies, where the latter is a novel contribution as mentioned.

II.2 Managing high-dimensional data and forecast uncertainty

A particular challenge that arises when aggregating multiple RESs over space and time is that the number of input features sharply increases. However, when we remove features, are we certain that the "right" features have been removed? Sections II.2.2 and II.2.3 describe the methodologies that we propose to deal with the issues of data dimensionality and forecast uncertainty. First, Section II.2.1 gives a succinct description of the types of input features, which will later be expanded when describing the case studies of the deliverable.

II.2.1 Input features

Consider a single PV plant at a single forecast horizon. Then, a forecaster could consider as input features a forecast issued by the NWP forecast provider for the nearest grid point. These features could, for example, be temperature, wind speed and global horizontal irradiance. In case of a VPP and multiple forecast horizons, these inputs are distributed over space and time.

It is likely that only a subset of the available features is relevant at a particular point in time and the relevance of the features may vary over time. In addition, feature redundancy can cause computational issues for machine learning methods and unnecessarily increases the computational burden. However, there is no guarantee of selecting the optimal feature subset in a reasonable amount of time, which could become a problem in an operational setting.

In light of the aforementioned issues, this deliverable investigates filtering as a means to select relevant features. There is a multitude of filters, which will be introduced in the following section, and each will result in a different forecast. Section II.2.3 introduces the methods to optimally combine the forecasts that result from the various filters.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

II.2.2 Automatic feature selection

As mentioned in Section I.4.1, feature selection can be divided into filtering methods, wrapper methods and embedded methods. Here, the focus is on filtering methods because they are model agnostic and fast, which means they can be used in an operational setting. Essentially, a filter computes a score that ranks the importance of features. The aim is to find a feature subset $\mathcal{S}$ from the complete feature set $\mathcal{X}$ that represents the target variable Y with high accuracy and minimal residual uncertainty (Bommert et al., 2020).

In this deliverable, the filtering methods are based on mutual information, which originates from information theory and describes the amount of information that can be known about random variable Y when knowing random variable X (Bommert et al., 2020). First, it is helpful to introduce the entropy of a discretized random variable Y with univariate probability mass function p, i.e., a discrete version of a PDF (Bommert et al., 2020):

$$H(Y) = -\sum_{y} p(y)\log_2\left(p(y)\right), \tag{22}$$

which measures the uncertainty of Y. Furthermore, the conditional entropy of Y given X is defined as (Bommert et al., 2020):

$$H(Y|X) = \sum_{x} p(x)\left(-\sum_{y} p(y|x)\log_2\left(p(y|x)\right)\right). \tag{23}$$

The mutual information between Y and X is defined as $I(Y;X) = H(Y) - H(Y|X)$ and describes how the uncertainty of Y is lowered by knowing X (Bommert et al., 2020). Herein, the praznik package (Kursa, 2020) in the statistical software R is used. The praznik toolbox discretizes the range of the continuous features into $\max\{\min\{\frac{n}{3}, 10\}, 2\}$ equally spaced intervals, where n is the number of training samples (Kursa, 2020).

In total, 6 filtering methods based on mutual information are used. The first filter, mutual information maximization (MIM), computes the mutual information between feature i and target variable Y as:

$$J_{\text{MIM}}(X^{(i)}) = I(Y;X^{(i)}), \tag{24}$$

and returns a predetermined number of features that maximizes J (Kursa, 2020).

The remaining filters greedily add a feature to $\mathcal{S}$ that at each iteration maximizes the score $J(X^{(i)})$. The first of which is minimal conditional mutual information (CMIM), which uses the following score

$$J_{\text{CMIM}}(X^{(i)}) = \min\left\{I(Y;X^{(i)}), \min_{X^{(j)}\in\mathcal{S}} I(Y;X^{(i)}|X^{(j)})\right\}, \tag{25}$$

where $I(Y;X^{(i)}|X^{(j)}) = H(Y|X^{(j)}) - H(Y|X^{(i)},X^{(j)})$ Kursa (2020). In words, CMIM describes how knowing $X^{(i)}$ lowers the uncertainty of Y given that we already know $X^{(j)}$ and is therefore similar to MIM.

The third filter is conditional mutual information (CMI), which ranks features according to the score:

$$J_{\text{CMI}}(X^{(i)}) = I(Y;X^{(i)}|\mathcal{S}), \tag{26}$$

and therefore evaluates the added value of feature $X^{(i)}$ considering the already selected features.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

The fourth filter is double input symmetrical relevance (DISR), which uses the score (Kursa, 2020):

$$J_{\text{DISR}}(X^{(i)}) = \sum_{X^{(j)} \in \mathcal{S}} \frac{I(Y; X^{(i)}, X^{(j)})}{H(Y, X^{(i)}, X^{(j)})}, \tag{27}$$

where $I(Y; X^{(i)}, X^{(j)})$ evaluates the complementary information that $X^{(i)}$ and $X^{(j)}$ provide for Y, whereas the normalization term $H(Y, X^{(i)}, X^{(j)})$ reduces the possibility to choose highly variable features (Bommert et al., 2020).

Maximum relevance and minimum redundancy (MRMR) is the fifth filter and ranks the features according to the following score:

$$J_{\text{MRMR}}(X^{(i)}) = I(Y; X^{(i)}) - \frac{1}{|\mathcal{S}|} \sum_{X^{(j)} \in \mathcal{S}} I(X^{(i)}; X^{(j)}), \tag{28}$$

which ensures minimal redundancy between $X^{(i)}$ and $\mathcal{S}$ (second term) while providing maximal information about Y (Kursa, 2020).

The final filter is minimal normalized joint mutual information (NJMIM) and scores the features according to the following equation (Kursa, 2020):

$$J_{\text{NJMIM}}(X^{(i)}) = \min_{X^{(j)} \in \mathcal{S}} \left\{ \frac{I(Y; X^{(i)}, X^{(j)})}{H(Y, X^{(i)}, X^{(j)})} \right\}. \tag{29}$$

In essence, this is a modification of filter DISR that evaluates the minimal relative information between Y, $X^{(i)}$ and already selected features instead of the sum (Bommert et al., 2020).

The filters are applied to data that comprise the d previous days recorded during the same time stamp as the forecast issue time t (expressed as "HH:MM"). In this way, it is possible to include the most recent data in our filtering method. In order to account for temporal dependencies, one time instant prior to "HH:MM" and one time instant after "HH:MM" (except for $t+1$) are also included. As a result, there are a total of $3 \cdot d - 1$ time instances available to the filters.

These filters have been selected because they (i) can uncover nonlinear relationships; (ii) do not require the features to be on the same scale; (iii) perform decently as Bommert et al. (2020) concludes; (iv) tend to select a diverse set of features at time t (Bommert et al., 2020), which could benefit forecast diversity; and (v) except for MIM, take interactions between features into account.

II.2.3 Probabilistic forecast combination

As mentioned above, OLP is a heuristic method in which m predictive distributions F_j are linearly combined as $G = \sum_{j=1}^{m} w_j F_j$ $w_j = 1/m$. However, it has been shown that it is possible to improve upon OLP by the optimizing weights w_j using a scoring rule such as (10). In this deliverable the logarithmic score (IGN), as proposed by Gneiting and Ranjan (2013), is used to optimize the weights w_j and potential additional parameters on a validation set separate from the testing set. The following describes the three combination methods, while OLP is kept as a naive reference method.

II.2.3.1 Linear Combination The natural extension of OLP is the traditional linear pool (TLP), in which the weights are determined by optimizing the logarithmic score. The weights can be determined by evaluating $g_i(y) = \sum_{j=1}^{m} w_j f_{ij}(y)$, where i is an index for training data, and minimizing

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

over n fitting samples:

$$\text{IGN} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{j=1}^{m} w_j f_{ij}(y_i)\right), \tag{30}$$

which is equivalent to maximizing the log-likelihood (Yang and van der Meer, 2021). When the component forecasts are neutrally- or over-dispersed, the TLP will only be worse than the component forecasts because the linear combination always increases dispersion (cf. Theorem 3.1 in Gneiting and Ranjan (2013)).

II.2.3.2 Nonlinear combination Nonlinear combination methods allow more freedom and avoid introducing overdispersion. The first nonlinear combination method considered here is the spread-adjusted linear pool (SLP) that includes a strictly positive parameter c that adjusts the spread of the predictive distributions. The combined predictive distribution is defined as (Gneiting and Ranjan, 2013):

$$G_i^c(y) = \sum_{j=1}^{m} w_j F_{ij}^0\left(\frac{y-\mu_{ij}}{c}\right), \tag{31}$$

where $F_{ij} = F_{ij}^0(y-\mu_{ij})$ and μ_{ij} is the median of predictive distribution of F_{ij}. The parameters can be found by modifying (30) as follows:

$$\text{IGN} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{c}\sum_{j=1}^{m} w_j f_{ij}^0\left(\frac{y_i-\mu_{ij}^*}{c}\right)\right). \tag{32}$$

Equation (31) shows that SLP reduces to TLP when $c=1$. Setting $c<1$ tends to improve calibration when the individual predictive distributions are overdispersed or neutrally dispersed and setting $c \geq 1$ may benefit underdispersed predictive distributions. While SLP is more flexible than TLP, it is not flexible enough to be used when the component forecasts are overdispersed (Gneiting and Ranjan, 2013).

The beta-transformed linear pool is fully flexible. The CDF of the beta-transformed linear pool is defined as

$$G_i^{\alpha,\beta}(y) = B_{\alpha,\beta}\left(\sum_{j=1}^{m} w_j F_{ij}(y)\right), \tag{33}$$

where $B_{\alpha,\beta}$ represents the beta CDF with parameters $\alpha > 0$ and $\beta > 0$, and BLP reduces to TLP when $\alpha = \beta = 1$. When fixing the weights $w_1, \ldots, w_m$, it can be shown that the variance of the PIT random variable can attain any value in the open interval $\left(0, \frac{1}{4}\right)$ (Gneiting and Ranjan, 2013).[4] The following objective function is minimized to learn the optimal parameters (Gneiting and Ranjan, 2013):

$$\text{IGN} = -\frac{1}{n}\sum_{i=1}^{n}\left((\alpha-1)\log\left(\sum_{j=1}^{m} F_{ij}(y_i)\right) + (\beta+1)\log\left(1-\sum_{j=1}^{m} F_{ij}(y_i)\right)\right). \tag{34}$$

[4]Note that when the variance of the PIT variables is $\frac{1}{12}$, the probabilistic forecasts are neutrally dispersed.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

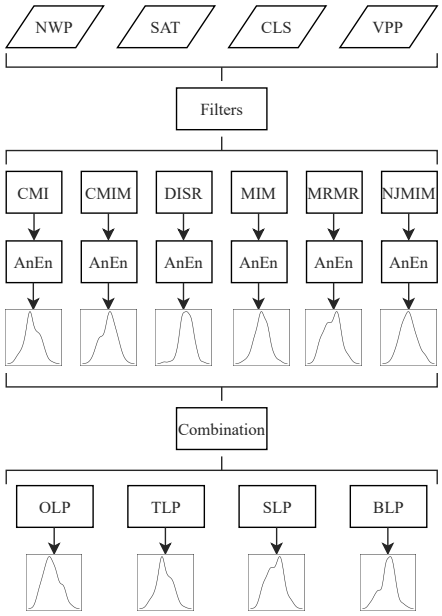


Figure 3 A flowchart of the pre-process, forecast, and post-process steps.

The methodology of feature selection and forecast combination to deal with high-dimensional data and forecast uncertainty is presented as a flowchart in Fig. 3. The Analog Ensemble (AnEn) probabilistic forecast model will be introduced in Section II.3.1. In essence, the filters introduced above are applied to all data sources, i.e., NWP, satellite (SAT), solar geometry (CLS) comprising the zenith angle and the solar azimuth, and measurements (VPP). The selected features are used in separate forecast models, which are then combined using any of the four combination methods described above.

II.3 Forecasting

This section describes the forecasting methods used in this deliverable. Note that the probabilistic forecasts described in Section II.3.1 are not novel; rather, the novelty there lies in the way features are automatically selected and forecasts are combined.

II.3.1 Probabilistic forecast generation

Ensemble NWP forecasts were already proposed by Epstein in 1969 (S., 1969), after Lorenz in 1963 noted that the lack of complete observation of the atmosphere as well as model uncertainty results in diverging outcomes between the observed and predicted state of the atmosphere (Lorenz, 1963). In order to generate an ensemble forecast, one can either combine multiple deterministic NWP forecasts (i.e., a poor man's ensemble) or perturb initial conditions slightly and run these concurrently (i.e., a dynamical ensemble). In either case, the effort is significant because a multitude of computational runs are necessary. Instead, Delle Monache et al. (2013) proposed AnEn as a means to generate probabilistic forecasts, which was a departure from previous uses of AnEn, e.g., related to calibration of NWP forecasts.

In essence, AnEn works by comparing the current NWP forecast, the "query", to a history of NWP forecasts, the "analogs". The query and analogs are compared based on similarity using

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

the Euclidean distance and the most similar analogs are selected. The final forecast is then an ensemble of the power measurements that correspond to the time stamps of the most similar analogs. In order to include additional data sources, the original similarity metric is modified (Carriere et al., 2020). The similarity metric is further modified such that the lags of the features are distinct features, in order to leverage the algorithmic efficiency of k-dimensional tree (kd-tree) (Yang, 2019; Yang and van der Meer, 2021). The similarity metric then becomes:

$$d(\boldsymbol{\mathcal{X}}_t^k, \boldsymbol{\mathcal{A}}_i^k) = \sqrt{\sum_{j=1}^{J} w_j^k \left(x_t^{(j)} - x_i^{(j)}\right)^2}, \tag{35}$$

where $J = N_v \cdot (2\tilde{t} + 1)$ is the total dimension, w_j^k is the j^{th} weight for the k^{th} forecast horizon and $x^{(j)}$ is the j^{th} feature. In (35), vector $\boldsymbol{\mathcal{X}}_t^k$ contains the most recent filtered features, whereas vector $\boldsymbol{\mathcal{A}}_i^k$ contains the same filtered features as $\boldsymbol{\mathcal{X}}_t^k$ except the historical ones. It is important to note that the features in $\boldsymbol{\mathcal{A}}_i^k$ are scaled and centered to ensure the distance metric does not favor features on a larger scale. These scaling factor are then used to scale and center $\boldsymbol{\mathcal{X}}_t^k$.

Finally, it is important to note that the weights w_j^k in (35) are taken directly from the scores that the filtering methods assign to the features. Subsequently, the weights are normalized by the sum of all feature scores to ensure that the weights sum to 1. This method affords a dynamical update of the feature weights rather than employing a costly wrapper.

II.3.2 Seamless trajectory forecasts

Section I.4.3 described the de facto method to generate trajectory forecasts. There, it was argued that the method becomes cumbersome when significantly increasing the temporal resolution of the forecast, e.g., up to 5 minutes or 1 minute. For a temporal resolution of 5 minutes, as in this deliverable, one would need to generate 576 univariate, i.e., marginal predictive distributions when considering a forecast horizon of 48 hours. In the following, we describe an alternative that simplifies the forecast modeling chain substantially.

The pattern matching model (PMM) proposed here is conceptually straightforward and is based on AnEn described in Section II.3.1. However, the innovation is to search for S analog trajectories to approximate the multivariate predictive CDF $\boldsymbol{F}_t$, rather than searching for analogs for a single forecast horizon to approximate $F_{t+k|t}$. Therefore, $\boldsymbol{\mathcal{A}}_i$ contains the analog NWP forecast issued at historical time i organized as one vector, instead of only extracting $i+k-1$, $i+k$ and $i+k+1$. Similarly, $\boldsymbol{\mathcal{X}}_t$ contains the query NWP forecast issued at testing time t. Then, using the algorithmic efficiency of kd-tree as mentioned above, it is possible to compute the distances using (35) as before. Note that $w_j = 1 \; \forall \; j$ for simplicity but this could be extended in future work, e.g., by using filters as in the previous section. Furthermore, all features are again scaled and centered using the historical analogs and these factors are then applied to the query.

The S most similar analog time stamps i are selected and the accompanying power measurements from time i up to time $i+K$ constitute the multivariate probabilistic forecast, such that $\boldsymbol{F}_t \in \mathbb{R}^{S \times K}$ and the corresponding observations are $\boldsymbol{y}_t \in \mathbb{R}^K$. In other words, PMM is a downscaling method in which time series at low resolution are compared, after which a high resolution time series can be extracted.

II.3.3 Hierarchical forecasts with missing values

II.3.3.1 Hierarchical forecasts Equation (16) from Section I.3.5 states that hierarchical observations are coherent if the difference between the top nodes and bottom nodes is zero. Fur-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

thermore, define $\mathcal{S} := \{\mathbf{y} \mid \mathbf{A}\mathbf{y} = \mathbf{0}\}$ to be the feasible set that satisfies the linear aggregation constraints. The following assumption is commonplace in the literature.

Assumption 1 *Historical observations $\mathbf{y}_t$ are coherent by construction, i.e., $\mathbf{y}_t \in \mathcal{S}\ \forall t \in [T]$.*

The following definition can be constructed for coherent forecasts, similar to that of the measurements.

Definition 1 *Forecasts $\widehat{\mathbf{y}}_{t+k}$ are said to be coherent if they satisfy $\mathbf{A}\widehat{\mathbf{y}}_{t+k} = \mathbf{0}$.*

As argued above, the base forecasts will likely not satisfy the coherency constraints and a post-processing step is thus required. The following shows that a class of non-parametric machine learning models directly provides coherent point forecasts. First, we state a standard result from convex analysis.

Proposition 1 *Any convex combination of historical observations $\mathbf{y}_t$ satisfies the coherency constraints.*

Proof This follows from convexity of $\mathcal{S}$ and Assumption 1. □

The above holds for additional convex constraints, e.g., non-negativity of forecasts. A corollary of Proposition 1 is that a class of machine learning models, including, among others, k-nearest neighbors, kernel regression, and decision trees, directly leads to coherent forecasts. These models are based on the idea of local averaging or smoothing of historical observations. For an out-of-sample observation $\mathbf{x}_{t+k}$, we derive a set of non-negative weights $\omega_t(\cdot)$, with $\sum_{t\in[T]} \omega_t(\mathbf{x}_{t+k}) = 1$, and the respective point forecast $\widehat{\mathbf{y}}_{t+k}$ is given by

$$\widehat{\mathbf{y}}_{t+k} = \sum_{t\in[T]} \omega_t(\mathbf{x}_{t+k})\mathbf{y}_t, \tag{36}$$

i.e., a convex combination of historical observations. From Proposition 1, we see that $\widehat{\mathbf{y}}_{t+k}$ are coherent. The practical implication is that off-the-shelf machine learning tools are readily applicable for end-to-end hierarchical forecasting. This result is somewhat trivial; nonetheless, it seems to have escaped the respective forecasting literature.

The above-mentioned models can also be employed for probabilistic hierarchical forecasting. For example, the selected neighbors $\mathbf{y}_t$ in a k-nearest neighbor model can be treated as (coherent) sample path realizations of the joint predictive density of all the series in the hierarchy. Similarly, one could treat the output of individual trees within an ensemble as realizations of the multivariate predictive density. Therefore, off-the-shelf machine learning tools are also applicable to probabilistic hierarchical forecasting, presenting a computationally cheaper alternative to the internal sampling and projection approach proposed by Rangapuram et al. (2021) and the bottom-up method described by Taieb et al. (2017).

II.3.3.2 Dealing with missing values Recall that we extend the hierarchical forecasts with the case when bottom-level series have missing values due to equipment failures, but aggregated series maintain correct measurements. Assume that missing values are set at 0, therefore $\mathbf{A}\mathbf{y}_t \neq \mathbf{0}$ and $\mathbf{y}_t \notin \mathcal{S}$. Without loss of generality, the term "missing" refers both to missing and erroneous measurements, as long as these are identified, e.g., by applying an outlier detection mechanism, and the missing values do not propagate through the hierarchy. This problem is examined

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

under a conditional stochastic optimization lens by integrating predictive and prescriptive analytics (Bertsimas and Kallus, 2020), and by formulating prescriptive trees for end-to-end hierarchical forecasting. Prescriptive trees (Stratigakos et al., 2022) refer to decision trees that output prescriptions rather than predictions. In our case, and with a slight abuse of terminology, the prescriptions correspond to hierarchical point forecasts, which must satisfy the coherency constraints (and possibly additional ones). Out-of-sample conditional prescriptions are derived via a weighted Sample Average Approximation (SAA) of the original stochastic optimization problem (Bertsimas and Kallus, 2020).

We follow the popular CART method (Breiman et al., 1984) by recursively partitioning the feature space with locally optimal splits. Mathematically, a node split separates a feature space $R \subseteq \mathbb{R}^{T\times p}$ of T observations and p features at feature j and point s into two disjoint partitions $R = R_l \cup R_r$, such that $R_l = \{t \in [T] \mid x_{tj} < s\}$ and $R_r = \{t \in [T] \mid x_{tj} \geq s\}$, with scalar x_{tj} denoting the t-th observation of the j-th feature[5]. The main idea of the node split is to partition the data such that similar observations are clustered into the partitions. Here, we minimize a generic cost function subject to a set of linear aggregation constraints. The problem of finding the locally optimal split is given by

$$\min_{j,s}\left[\min_{\mathbf{z}_l\in\mathcal{S}}\sum_{t\in R_l(j,s)} c(\mathbf{z}_l;\mathbf{x}_t)+\min_{\mathbf{z}_r\in\mathcal{S}}\sum_{t\in R_r(j,s)} c(\mathbf{z}_r;\mathbf{x}_t)\right], \tag{37}$$

with subscripts l, r referring to the left and right child node, $\mathbf{z}_{\{l,r\}} \in \mathbb{R}^n$ being the locally constant decisions (i.e., hierarchical forecasts), which satisfy the coherency constraints $\mathcal{S}$, $R_{\{l,r\}}$ being index sets, and $c(\cdot)$ being the cost function to be minimized. Thus, the main difference from the CART algorithm is the requirement for predictions to satisfy a set of constraints and the use of a generic, task-based loss function.

In a typical regression setting, $c(\cdot)$ would correspond to the squared ℓ_2 norm. Here, we want to ignore missing values, without disregarding any quality data points. To this end, we employ an indicator matrix $\mathbf{\Gamma} \in \{0,1\}^{n\times T}$ that checks whether historical observations are missing. A single entry γ of $\mathbf{\Gamma}$ is given by

$$\gamma_{it} = \begin{cases} 1, & \text{if } y_{it} \text{ is missing} \\ 0, & \text{otherwise} \end{cases} \quad \forall i \in [n], t \in [T], \tag{38}$$

and the cost of sample t is given by

$$c(\cdot) = \|\mathbf{y}_t - (\mathbf{1} - \boldsymbol{\gamma}_t) \odot \mathbf{z}\|_2^2, \tag{39}$$

with $\boldsymbol{\gamma}_t$ denoting the t-th column vector of $\mathbf{\Gamma}$ and $\odot$ denoting the elementwise multiplication. Note that missing values are effectively left out of the objective. This way the optimization process places more value/weight on series without any missing values during training. We consider this to be a desirable property, as these series correspond to more reliable nodes within the hierarchy.

As discussed, forecasts are required to satisfy the coherency constraints imposed by $\mathcal{S}$. Note that (37) involves two equality constrained quadratic sub-problems. An analytical solution is derived by solving a system of linear equations obtained from the Karush–Kuhn–Tucker (KKT) optimality conditions (see Appendix IV). Other possible constraints, e.g., non-negativity or monotonicity, can be readily included. In this case, a general-purpose convex solver can be called on to evaluate (37). For an out-of-sample observation $\mathbf{x}_{t+k}$ point forecasts are derived via a weighted SAA given by

$$\widehat{\mathbf{y}}_{t+k} = \arg\min_{\mathbf{z}\in\mathcal{S}} \sum_{t\in[T]} \omega_t(\mathbf{x}_{t+k})\|\mathbf{y}_t - (\mathbf{1} - \boldsymbol{\gamma}_t) \odot \mathbf{z}\|_2^2. \tag{40}$$

[5]For brevity of exposition we focus on quantitative features, although it is straightforward to also include categorical features

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

In general, decision trees are highly prone to overfitting. Randomization-based ensembles provide a remedy and lead to impressive predictive performance; these are readily applicable within the proposed framework, leading to a *prescriptive forest*. A single tree is fully compiled, with its leaves outputting coherent forecasts. The corresponding weights are given by

$$\omega_t(\mathbf{x}_{t+k}) = \frac{\mathbb{I}[R(\mathbf{x}_t) = R(\mathbf{x}_{t+k})]}{|R(\mathbf{x}_{t+k})|}, \tag{41}$$

where $R(\mathbf{x}_{t+k})$ is the leaf that out-of-sample observation $\mathbf{x}_{t+k}$ falls into, $|\cdot|$ the leaf cardinality, and $\mathbb{I}[\cdot]$ an indicator function that checks whether training observation $\mathbf{x}_t$ falls into $R(\mathbf{x}_{t+k})$. Lastly, for an ensemble of B trees the weights are obtained

$$\omega_t(\mathbf{x}_{t+k}) = \frac{1}{B}\sum_{b=1}^{B} \frac{\mathbb{I}[R^b(\mathbf{x}_t) = R^b(\mathbf{x}_{t+k})]}{|R^b(\mathbf{x}_{t+k})|}. \tag{42}$$

II.4 Description of case studies

The case studies presented in this section employ data made available by the partners. The case studies pertain to the three topics of this report: (i) intra-day probabilistic forecasting of the power output of a VPP located in mid-west France using a heterogeneous set of input data; (ii) intra-day and day-ahead probabilistic multivariate forecasts of the power output of a VPP located in Greece; and (iii) day-ahead hierarchical point forecasts with missing values of the same VPP as in (i). The following sections detail the case studies and the final section introduces the benchmark models.

II.4.1 Case study I

In the first case study, the aim is to include a heterogeneous set of input features for intra-day forecasting, i.e., up to 6 hours ahead, in mid-west France. The inputs comprise NWP forecasts, satellite derived GHI maps, astronomical data and measurements. Figure 4 provides an overview of the location of the PV systems, wind turbines and grid points of NWP and satellite features. More specifically, the European Center for Medium-range Weather Forecasts (ECMWF) provides NWP forecasts that are issued daily at 00:00 UTC with a spatial resolution of 0.1° $\times$ 0.1°. Although these forecasts comprise a multitude of features, here 5 features are extracted. Specifically: (i) surface solar radiation downwards (SSRD); (ii-iii) 100 m U- and V-wind speed (U100, V100); (iv) 2 m temperature (T2M); and (v) total cloud cover (TCC). GHI is computed from SSRD and the deterministic component of GHI, i.e., irradiance under clear-sky conditions, is removed by dividing GHI by the clear-sky irradiance[6] using the McClear clear-sky model (Lefèvre et al., 2013).

The satellite images have been recorded by the Meteosat Second Generation (MSG) satellite. Afterwards, GHI can be derived from these images using an improved version of the Heliosat-2 method (Zarzalejo et al., 2009) and these GHI maps are stored as time series for each grid point. Also in this case, the GHI is detrended using clear-sky irradiance such that the input feature is the clear-sky index.

The solar geometry (CLS) can be computed with very high accuracy at any point in the future and they inform the model as to when the sun rises and sets. Especially in the case when the VPP consists only of PV systems, this could be important information for the diurnal intermittency.

Finally, the power measurements are available from 2019-01-01 until 2020-09-30. The data from 2019-01-01 until 2019-09-30 is used as the historical analogs. The period from 2019-10-01 until

[6]This quantity is referred to as the clear-sky index.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

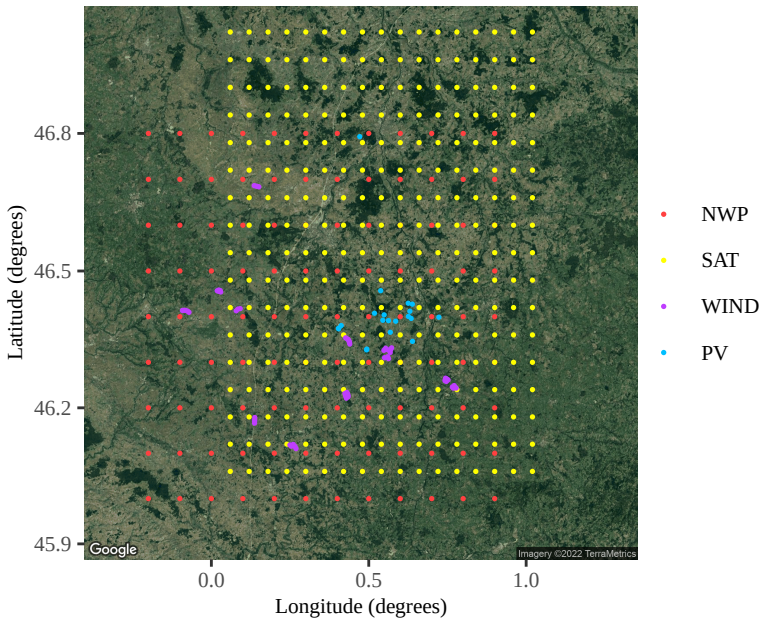


Figure 4 Overview of the satellite and NWP grid points, as well as the PV systems and wind turbines located in mid-west France.

2019-12-31 is used to optimize the combination weights, which is not necessarily a representative period but since we are dealing with time series, we want to maintain the chronological order of observations while keeping ample instances for testing. The remaining data from 2020-01-01 until 2020-09-30 is used for model testing. In the case of a PV only VPP, measurements recorded at a zenith angle larger than 85° are removed. The reason for removing measurements is that small errors in the clear-sky profile at high zenith angles can result in significant errors. Furthermore, a quality control similar to that proposed by Killinger et al. (2017) is used on the measurements. Briefly, the quality control: (i) ensures all time series are at the same temporal resolution; (ii) flags any instance higher than the physical limit based on the extraterrestrial irradiance or lower than zero, as well as power measurements when the zenith angle is greater than 95°; and (iii) compares measurements with overall daily variability (Killinger et al., 2017). In addition, the PV power is detrended using the clear-sky global tilted irradiance (GTI_{cs}) with one tilt and orientation because the systems are located near each other. Regarding the wind power measurements, these are checked whether they are nonnegative and do not exceed the installed capacity. The PV VPP has a total installed capacity of 4 MW, whereas the wind and PV VPP has a total installed capacity of 124 MW.

In this case study, filters are tested to reduce the dimensionality of feature set and speed up the computations. To cope with potentially too stringent filters, forecast combination is used as a tool to mitigate some of these uncertainties. These methods have been presented in Sections II.2.2 and II.2.3.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

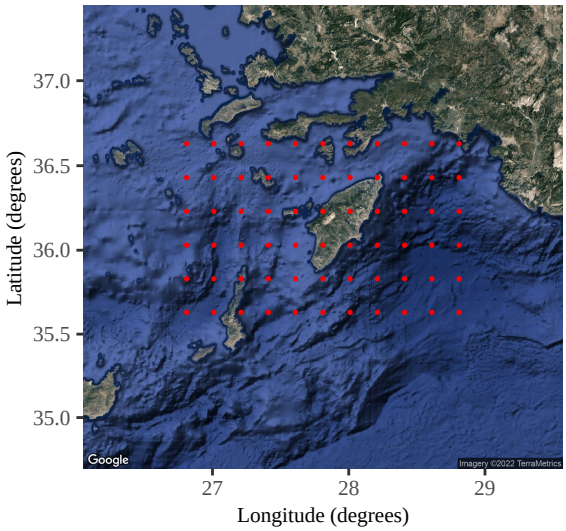


Figure 5 Overview of the NWP grid points on and around the island of Rhodes (Greece) at which forecasts are available.

II.4.2 Case study II

Unlike the previous case study, here the focus is on investigating whether a straightforward approach such as the one described in Section II.3.2 can improve upon, or at least equal, the state-of-the-art described in Section I.4.3. To isolate as much as possible the performance of the model itself, the focus is on one set of input features, namely NWP ensemble forecasts from ECMWF. This case study takes place on Rhodes, which is a Greek island. Figure 5 presents the 66 grid points at which 50 ensemble members have been collected (Leutbecher and Palmer, 2008) that predict the progression of 5 variables, initiated daily at 00:00 UTC. The 5 variables selected are SSRD converted to GHI and then clear-sky index, TCC, U10, V10 and T2M. A large number of grid points has been included to capture large scale weather patterns, while the ensemble members inform about the variability present in the forecasts. The query $\boldsymbol{x}_t$ therefore has a length of $48 \times 66 \times 50 \times 5 = 792,000$, which is reduced by summarizing the grid points and ensemble members by: (i) the mean and standard deviation of the ensemble members at each grid point, resulting in 31,680 features (referred to as case study MS); and (ii) the $0.01, 0.02, \ldots, 0.99$ quantiles over all ensemble members and grid points, resulting in 23,760 features (referred to as case study QS).

Both the aggregated PV power and wind power measurements are available at 1 minute resolution, which are then averaged to 5 minute resolution. The total installed capacity is 18,164 kW and 48,550 kW for PV power and wind power, respectively. When the VPP consists of PV alone, the measurements are detrended using clear-sky GHI generated by the McClear model (Lefèvre et al., 2013) for a tilt of 25° and azimuth of 180° due south. These are the optimal tilt and azimuth on Rhodes and the implicit assumption is therefore that most PV systems will have been installed with this orientation (Kambezidis and Psiloglou, 2021). However, the PV power measurements do not coincide perfectly during early morning and late afternoon, which is likely caused by systems orientated towards the east and west, respectively. Therefore, the detrended PV power measurements are therefore set to 0 when the zenith angle is larger than 90° or when the clear-sky GHI is lower than 1 W/m^2 as these values would otherwise be significantly higher than can be reasonably assumed. Both PV power and wind power are normalized with the installed capacity so that the errors can be presented as dimensionless numbers or as percentages.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

II.4.3 Case study III

The third and final case study concerns hierarchical point forecasting with missing values, as detailed in Section I.3.5. Since the case study involves energy production, the forecasts should be non-negative and therefore $\mathcal{S} := \{\mathbf{y}|\mathbf{A}\mathbf{y} = \mathbf{0}, \mathbf{y} \geq \mathbf{0}\}$; Proposition 1 holds as $\mathcal{S}$ remains convex. Only day-ahead forecasts are of interest with hourly resolution, i.e., the forecast horizon is 12-36 hours which simulates day-ahead market conditions. The data are the same as in case study I, except that VPPs of wind power and PV power are investigated separately. In both cases, a 3-level hierarchy is considered. Wind production data are naturally aggregated at park level (13 wind parks in total); for the PV production data we construct a fictitious hierarchy based on spatial k-means clustering. Figure 4 provides an overview of the geographical distribution of the power plants.

It is important to note that historical production lags are not considered as input features as these typically do not improve forecasts in the horizon of interest. Further, including historical lags would introduce missing values in $\mathbf{x}$, which is outside of our scope. For the i-th series, respective feature vector $\mathbf{x}_{it}$ comprises the NWPs from the closest grid point in terms of Euclidean distance; when forecasting a group of series in one-shot, respective features are concatenated in a single vector.

II.4.4 Benchmarks

Benchmark forecast models are important tools to gauge the performance of the model under investigation. Generally, benchmark models should either be naive to represent a lower bound to the achievable performance or they should be well-known state-of-the-art models. In this report, a multitude of benchmark models is used and the following gives a brief description for each of these as a function of case study.

In case study I, AnEn without any feature selection ("Vanilla AnEn") is used as a benchmark to evaluate what the value of feature selection and forecast combination is. The performance of Vanilla AnEn allows to compute a skill score ($s = 1 - \frac{\text{Score}_{\text{model}}}{\text{Score}_{\text{benchmark}}}$), which reflects the relative improvement over the benchmark model where a negative skill implies worse performance than the benchmark. Furthermore, Quantile Regression Forests (QRF) is used as a state-of-the-art benchmark. QRF is an extension of random forests (RF), in which the prediction is the weighted average of the observed response variables (Breiman, 2001). Instead in QRF, the output is the weighted distribution of the response variables (Meinshausen, 2006). During training, each tree is grown on a random sample of the training data, thus reducing the correlation between the trees. To further decrease the correlation between the trees, a random subset of the features is selected at each candidate split (Breiman, 2001). The QRF model is separately trained for each forecast horizon and takes as input all available features.

In case study II, again QRF is used to forecast the marginal predictive distributions. The Gaussian copula described in Section I.4.3 models the dependence structure between forecast horizons and is used to generate trajectory forecasts in combination with the probabilistic forecasts generated by QRF. The naive benchmark is the multivariate probabilistic ensemble (MuPEn) of which the marginal predictive CDFs are identical to those of the complete-history persistence ensemble (van der Meer, 2021). It can be constructed by gathering all N K-length vectors from the historical observations that start at the same time ("HH:MM") as the current forecast issue time t. The result is an $N \times K$ matrix from which we randomly sample—without replacement—S trajectories such that the result is a multivariate predictive distribution $\boldsymbol{F}_t \in \mathbb{R}^{S \times K}$.

Finally, in case study III, benchmarks comprise post-processing and end-to-end learning approaches. Note that for post-processing methods, any learning algorithm can be used to gen-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

KPI category	KPI index	KPI name	KPI baseline	KPI target
Weather Forecasting				
Project KPI	1.1.a	% absolute improvement Weather Forecasting score: 15 to 30 min ahead	Current operational solutions (AROME/ECMWF/GFS, DLR, EMSYS)	10-15% RMSE
Project KPI	1.1.b	% absolute improvement Weather Forecasting score: Few hours ahead	Whiffle forecast driven with ECMWF boundary conditions and without data-assimilation	10% RMSE
Project KPI	1.1.c	% absolute improvement Weather Forecasting score: From few hours to 96 hours ahead	Current operational solutions of MeteoFrance	10% RMSE, 4-6% CRPS (solar radiative) 5-10% CRPS (wind)
RES Forecasting				
Project KPI	1.2.a	% improvement RES Forecasting score: Up to 30 min ahead	Current operational solutions of EMSYS, EDP-R, errors from public datasets. Conditional evaluation on situations with highest forecasting errors.	Solar: 9-12% RMSE, 3-5% CRPS Wind: 7-9% RMSE, 2-4% CRPS
Project KPI	1.2.b	% improvement RES Forecasting score: Up to 96 h ahead		Solar: 16-20% RMSE, 4-6% CRPS Wind: 12-15% RMSE, 3-5% CRPS
Specific KPI	1.2.c	% improvement Variogram score for ensemble forecasts	State-of-the-art methods for RES ensemble forecasts	>= 0
Specific KPI	1.2.d	% improvement for seamless generic forecasts	Same as KPI 1.2.a, 1.2.b	Weighted combination of targets in KPI 1.2.a, 1.2.b over lead-times and RES sources

Figure 6 Overview of Key Performance Indicators of the Smart4RES project.

erate base forecasts. To allow for a fair comparison, similar base learners are considered in all cases, i.e., randomized tree ensembles based either on the Random Forest (Breiman, 2001) or the ExtraTrees (Geurts et al., 2006) algorithm. The following approaches are examined:

- `BASE`: Base forecasts with a Random Forest model for each series, without reconciliation. Typically, these will not be coherent.
- `BASE-BU`: Bottom-up reconciliation applied to the base forecasts of the bottom-level series.
- `BASE-PRJ`: Base forecasts post-processed with a Euclidean projection step. The reconciled forecasts are given by

$$\arg\min_{y\in\mathcal{S}} \|\mathbf{y} - \widehat{\mathbf{y}}_{t+k}\|_2, \tag{43}$$

  where $\widehat{\mathbf{y}}_{t+k}$ are the base forecasts at time t with horizon k. Alternative reconciliation methods, such as MinT (Wickramasuriya et al., 2019) and constrained multivariate least squares (Di Modica et al., 2021), were also examined. However, as these methods require additional training, results with missing values were not robust and thus are omitted.
- `EtE`: A single Random Forest model predicting the whole hierarchy, to examine the efficacy of end-to-end learning.
- `EtE-PF`: End-to-end learning with prescriptive forests to deal with missing values.

In all cases, except for `EtE-PF`, observations with missing values are disregarded prior to training. For the `EtE` approach, this means that observation $\mathbf{y}_t$ is disregarded if at least two of the n series have a missing value, thus this approach deals with the largest loss of information. A grid search is performed to tune the hyperparameters of the Random Forest models. For the `EtE-PF`, we employ random node splits to speed-up computations, following the ExtraTrees algorithm (Geurts et al., 2006), and similarly perform a grid search for tuning.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Finally, Fig. 6 presents the Key Performance Indicator (KPIs) of the project. Relevant to this deliverable are KPIs 1.2a - d, where the percentage improvement is computed using one of the benchmarks described above.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

III. Results

III.1 Case study I

This section presents the forecast results of the probabilistic forecasts of a VPP consisting of solely PV systems and a VPP consisting of wind turbines and PV systems, as described in Section II.4.1. The following subsections have been further divided to present the results of the forecast models on the validation data and testing data.

III.1.1 VPP - PV

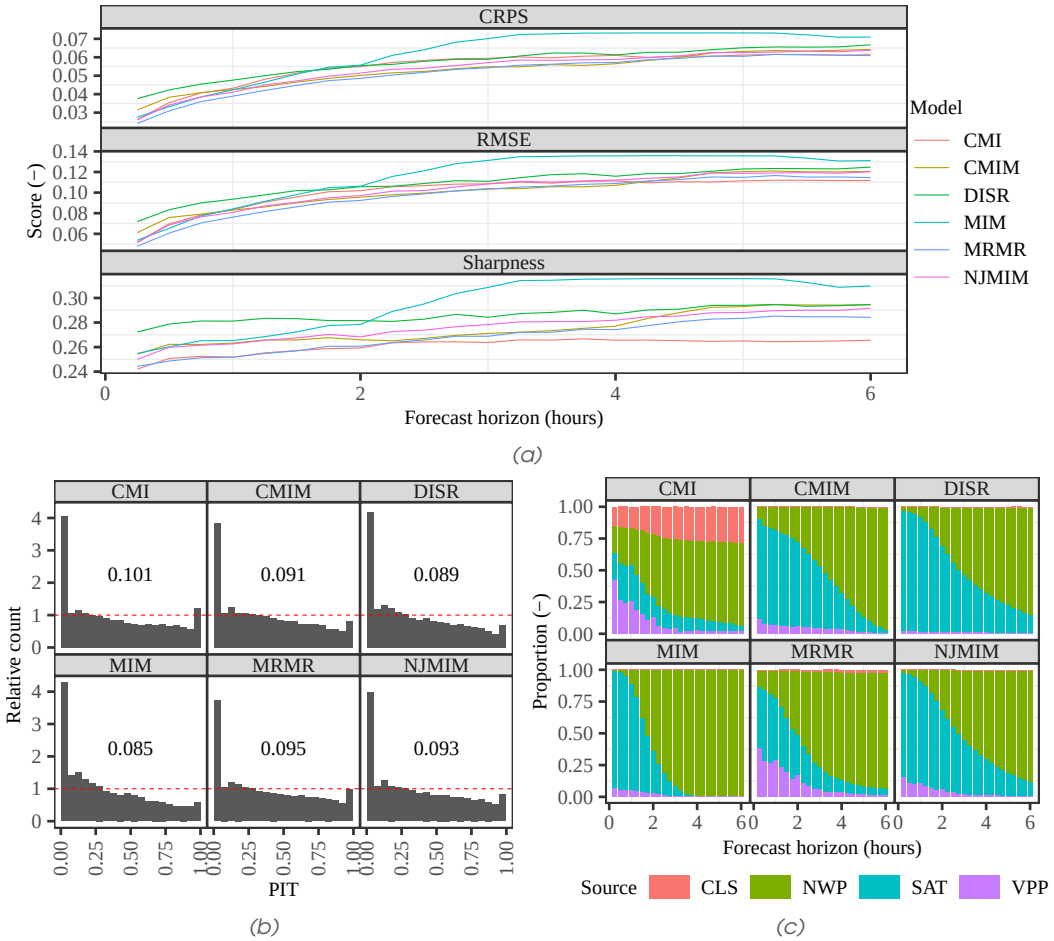


Figure 7 Forecast results on the validation set from the VPP consisting solely of PV. In (a), the numerical scores as a function of the forecast horizon. In (b), histograms of the PIT variables combined from all forecast horizons, including the variance. In (c), the proportion of the weights assigned to the feature groups.

III.1.1.1 Validation This section presents the results on the validation data of the VPP consisting solely of PV systems. The results of the validation set are relevant because the forecast combination weights are determined based on these results. Recall that data from 2019-01-01 until 2019-09-30 is used for the historical analogs while 2019-10-01 until 2019-12-31 is used for validation. Figure 7a presents the numerical scores of the probabilistic forecasts as a function of the forecast horizon, which is 15 minutes up to 6 hours. The scores are computed on the normalized data and therefore presented as such. It is clear that the forecasts resulting from the 6 filtering

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

methods perform quite similarly in terms of CRPS, RMSE and sharpness, except that filter MIM deteriorates after 2 hours compared to the other filters. This is likely caused by the fact that filter MIM does not consider interactions between features and therefore potentially selects redundant features. This can be seen in Fig. 7c, where the filter quickly selects only features from the NWP forecasts after the 2 hour forecast horizon.

In contrast, the other filters select a more diverse feature set over the forecasts horizons. There are some notable differences between the filters. For instance, filter CMI consistently selects CLS, i.e., solar geometric data, as an important input feature, while MRMR puts minor weight on CLS and the other filters ignore this information. In addition, DISR puts almost negligible weight on the most recent power measurement ("VPP") even though it is well known among forecasters that this information is relevant. Similarly, filters MIM and CMIM assign little weight to the most recent measurements.

Regardless, Fig. 7b shows that the probabilistic forecasts of all feature selection methods display positive bias, as indicated by the decreasing PIT histograms. This can be expected since the training and validation sets are disjoint, i.e., the seasons do not overlap. A full year of historical data would afford a more similar historical data set and therefore improved results.

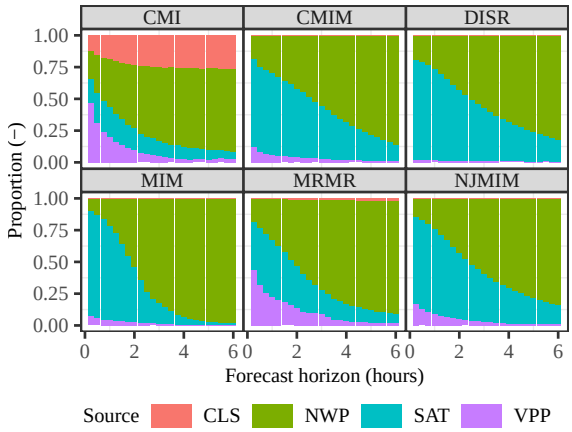


Figure 8 The proportion of the weights assigned to the feature groups for the VPP consisting only of PV systems for the testing case.

Table 1 The combination weights and SLP and BLP parameters for the VPP consisting solely of PV, expressed as "mean ± standard deviation" over all forecast horizons. The weights are optimized based on validation data and tested on the test data.

Model	CMI	CMIM	DISR	MIM	MRMR	NJMIM	c	α	β
TLP	0.249 ± 0.048	0.219 ± 0.086	0.081 ± 0.043	0.078 ± 0.08	0.254 ± 0.07	0.118 ± 0.039	(-)	(-)	(-)
SLP	0.277 ± 0.065	0.206 ± 0.099	0.087 ± 0.065	0.068 ± 0.078	0.253 ± 0.057	0.109 ± 0.046	0.931 ± 0.028	(-)	(-)
BLP	0.211 ± 0.021	0.206 ± 0.062	0.119 ± 0.049	0.166 ± 0.053	0.192 ± 0.058	0.106 ± 0.039	(-)	0.928 ± 0.048	1.506 ± 0.069

III.1.1.2 Testing Here the results on the testing data are presented. Recall that the testing data period runs from 2020-01-01 until 2020-09-30 while all data prior to this are used for model training or as analogs. Figure 8 presents the weights that the filters assign to the feature groups. The feature selection is very similar to that of the feature selection on the validation data and will therefore not be discussed further.

Figure 9 presents the numerical scores as a function of the forecast horizon including the combination methods and QRF, as well as the CRPS skill relative the Vanilla AnEn. In terms of the point forecasts, it is interesting to note that the RMSE is on average lower than during the validation period (cf. Fig. 7a). This could be due to the validation period being more challenging

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

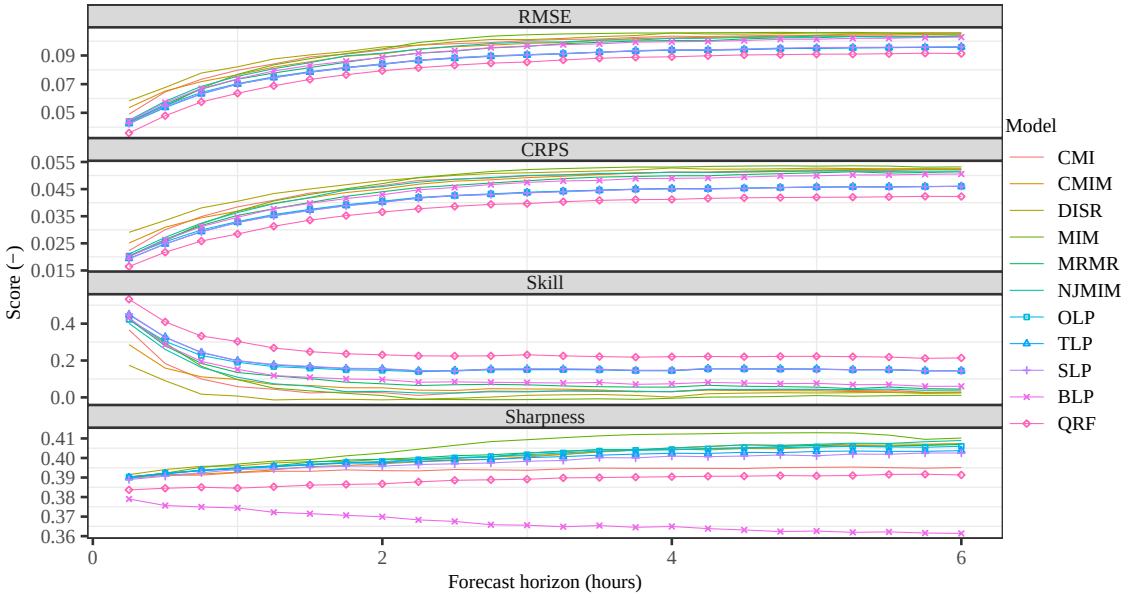


Figure 9 The numerical scores as a function of the forecast horizon for the VPP consisting only of PV systems. The scores have been computed on data normalized using the installed capacity.

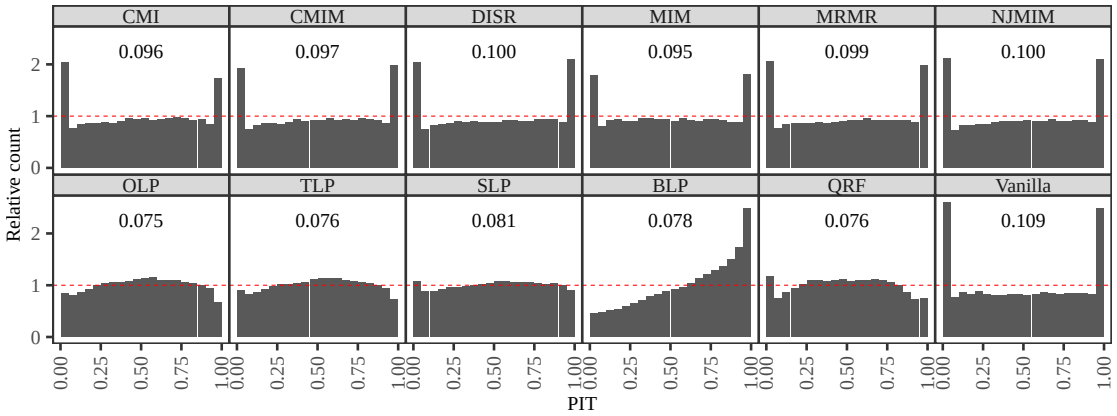


Figure 10 Histograms of the PIT variables combined from all forecast horizons, including their variance, for the VPP consisting only of PV systems.

in terms of weather patterns or that the validation period is substantially shorter than the testing period. Comparing the models during the testing data, it can be observed that the combination methods, except for BLP, outperform the component models and that QRF outperforms the combination methods in terms of RMSE and CRPS. In terms of skill, all models achieve positive skill and it can therefore be concluded that filtering methods do not deteriorate the predictive performance while significantly improving the computation time ($\approx 90\%$). It is worth noting that the KPI of Fig. 6 is achieved. Most notable from Fig. 9 is the poor performance of BLP, which theoretically should perform best. The poor performance of BLP is most likely caused by the biased component forecasts on the validation data (cf. Fig. 7b), which affects BLP more than the other combination methods. It is not certain what causes the bias increase although a likely explanation is that the parameters have been optimized to compensate for the negative bias on the training data, thus resulting in positive bias when combining the unbiased component forecasts on the testing data. The poor performance of BLP can also be observed from Table 1 where the large values of the α and β parameters of the Beta distribution attempt to correct the biases, which results in overcorrection. Regarding the weights that the combination methods assign to the component models, it is interesting to note that these are quite similar among the combi-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

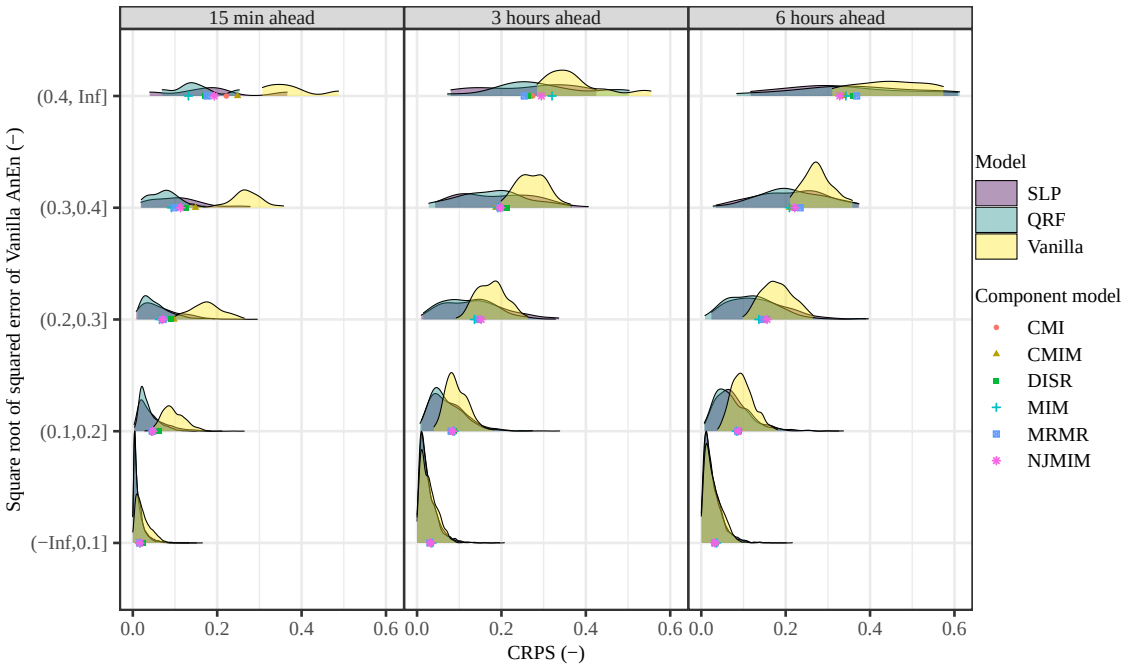


Figure 11 Distribution of the CRPS conditioned on the binned deterministic forecast error of the Vanilla AnEn for 3 forecast horizons and for the VPP consisting only of PV systems. The points represent the average CRPS of each component model.

nation methods. A notable difference is the weight assigned to MIM by BLP (0.166 on average) compared to 0.078 by TLP and 0.068 by SLP given that MIM performed worse on the validation data in terms of CRPS. This result highlights the need for a longer validation period.

Figure 10 shows that the component models (top row) show much better calibration in the main distribution than for the validation data. Nevertheless, the tails of the distribution are miscalibrated as indicated by the large peaks on either side. As a naive combination method, OLP performs well because the component models are too confident on average. However, both OLP and TLP are slightly overdispersed as also indicated by the variance of the PIT variables.[7] This is a theoretical result, which SLP and BLP mitigate by allowing for nonlinear forecast combination (Gneiting and Ranjan, 2013). The probabilistic forecasts combined with SLP are closest to neutral dispersion, whereas BLP introduces negative bias as discussed before. Finally, QRF forecasts are overdispersed whereas Vanilla AnEn forecasts are underdispersed more than the component models.

Finally, Fig. 11 presents a conditional error analysis of the 4 combination methods, QRF and the Vanilla AnEn. The vertical axis comprises 5 bins into which the square root of the squared error of the Vanilla AnEn has been categorized. The horizontal axis presents the continuous and contemporary CRPS. The points in the figure show the average CRPS of the component models for each bin and the coloured distributions present the CRPS distribution in each bin of the 3 models whose name is specified in the legend with the average of the distribution shown by the vertical line. Furthermore, the 3 subfigures show the results for 3 forecast horizons, specifically 15 min, 3 h and 6 h ahead. Several interesting observations can be made from the figure. First, all the models incur approximately similar CRPS in the lowest bin of the Vanilla AnEn deterministic error although the distribution of the Vanilla AnEn at the 15 min horizon is slightly wider. Second, the CRPS of Vanilla AnEn increases most with increasing point forecast error during the first forecast horizon, which indicates that the Vanilla AnEn has to consider too much unnecessary information—recall that NWP forecasts make up approximately 60% of the total number of

[7] Recall that the variance of the PIT variables computed from neutrally dispersed probabilistic forecasts should be close to 0.083.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

features—and this impedes the search for suitable analogs. The latter argument is supported by the CRPS distributions that increasingly converge with increasing forecast horizon, i.e., when NWP forecasts become increasingly important relative to the latest satellite image. Third, when comparing QRF and SLP directly, it is clear that the former performs better on the 15 min horizon in all the bins except for the lowest. However, at the 3 h forecast horizon in the highest deterministic forecast error bin it can be seen that SLP achieves lower absolute CRPS as well as on average. Similarly, the average CRPS achieved by SLP is slightly lower than that of QRF at the 6 h forecast horizon and at the highest deterministic forecast error bin although the difference is less pronounced. Finally, it is interesting to note that the component forecast models always outperform Vanilla AnEn at the first forecast horizon and with increasing deterministic forecast error, again highlighting that the computational burden can be significantly improved while also improving the accuracy.

III.1.2 VPP - Wind and PV

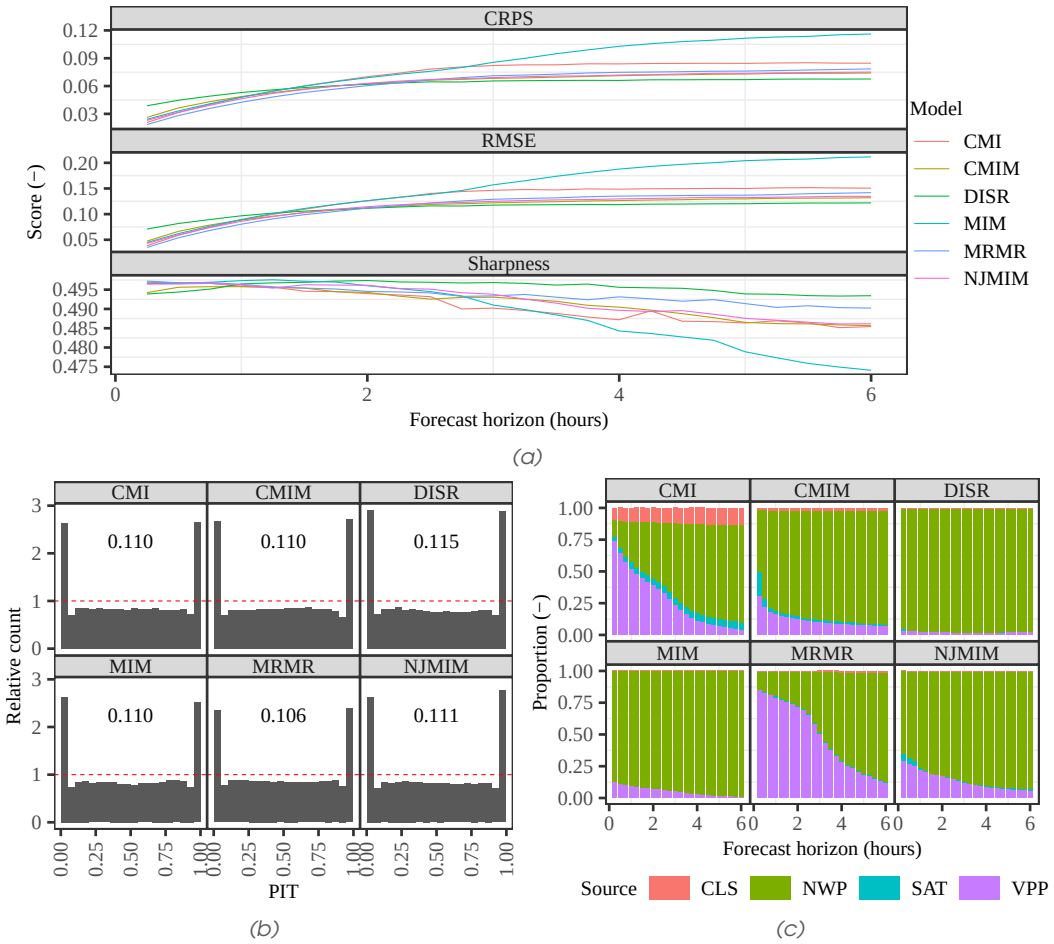


Figure 12 Forecast results on the validation set from the VPP consisting of wind and PV. In (a), the numerical scores as a function of the forecast horizon. In (b), histograms of the PIT variables combined from all forecast horizons, including the variance. In (c), the proportion of the weights assigned to the feature groups.

III.1.2.1 Validation Similar to the previous section, here the forecasts on the validation data are first analyzed. Recall that the VPP comprises 120 MW of installed wind power and approximately 4 MW of PV power. Fig. 12a presents the numerical scores on the validation data, which are slightly worse than the results of the VPP consisting solely of PV power (cf. Fig. 7a). Especially

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

filter MIM performs poorly; it diverges substantially in terms of CRPS, RMSE and sharpness from the other filters despite that the filter assigns more weight to the recent measurements than previously (cf. Fig. 12c). Noteworthy is the lack of sharpness compared to the VPP consisting of only PV, which is caused by the higher availability and variability of wind during the validation period (2019-10-01 until 2019-12-31).

In this case study where the share of wind power is significant, feature selection as a means to diversify the feature set becomes less pertinent. In essence, satellite imagery and solar geometry are mainly useful for PV power generation. This is reflected in Fig. 12c, where the filters mainly select features related to recent measurements of NWP forecasts. However, filter CMI still selects a diverse feature set and it can therefore be concluded that this filter selects the most diverse features, even when that is not necessary.

Interestingly, the probabilistic forecasts are calibrated in the main part of the distribution, as Fig. 12b shows. However, in all cases there are significant spikes at either sides of the distribution, which indicates that the models frequently either completely over- or underestimate the power generation. This indicates that the time series is variable and that the models do not fully capture the variability. This can be expected since the hourly NWP forecasts are linearly interpolated to the temporal resolution of the power measurements, which is 15 minutes. As a consequence, the NWP forecasts are smoothed and even though (35) includes time steps before and after the query time to capture variability, the forecasts are overconfident because some of the information concerning the variability is lost and the power measurements subsequently lack the appropriate variability.

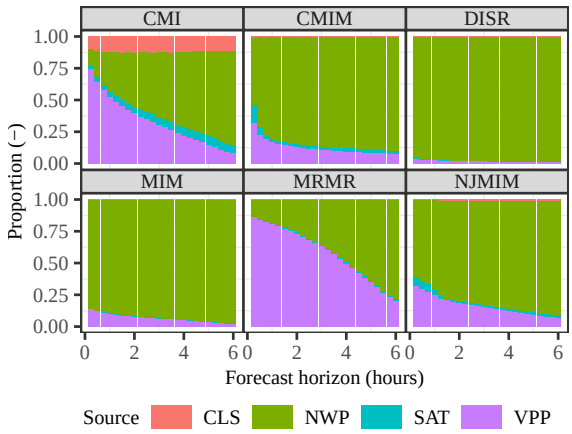


Figure 13 The proportion of the weights assigned to the feature groups for the VPP consisting of PV systems and wind turbines for the testing case.

Table 2 The combination weights and SLP and BLP parameters for the VPP consisting of wind and PV, expressed as "mean ± standard deviation" over all forecast horizons. The weights are optimized based on validation data and tested on the test data.

Model	CMI	CMIM	DISR	MIM	MRMR	NJMIM	c	α	β
TLP	0.11 ± 0.032	0.159 ± 0.061	0.242 ± 0.102	0.068 ± 0.024	0.268 ± 0.151	0.153 ± 0.029	(-)	(-)	(-)
SLP	0.146 ± 0.067	0.187 ± 0.081	0.19 ± 0.084	0.1 ± 0.052	0.239 ± 0.138	0.138 ± 0.058	0.882 ± 0.054	(-)	(-)
BLP	0.115 ± 0.039	0.16 ± 0.059	0.244 ± 0.098	0.07 ± 0.025	0.258 ± 0.131	0.152 ± 0.026	(-)	1.039 ± 0.058	1.029 ± 0.065

III.1.2.2 Testing Figure 13 presents the weights that the different filters assign to each feature group as a function of the forecast horizon. Similar to the case for the VPP consisting only of PV systems, the feature weights are very similar when comparing the validation (cf. Fig. 12c) and testing data. This indicates that the filters are consistent, as are the data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

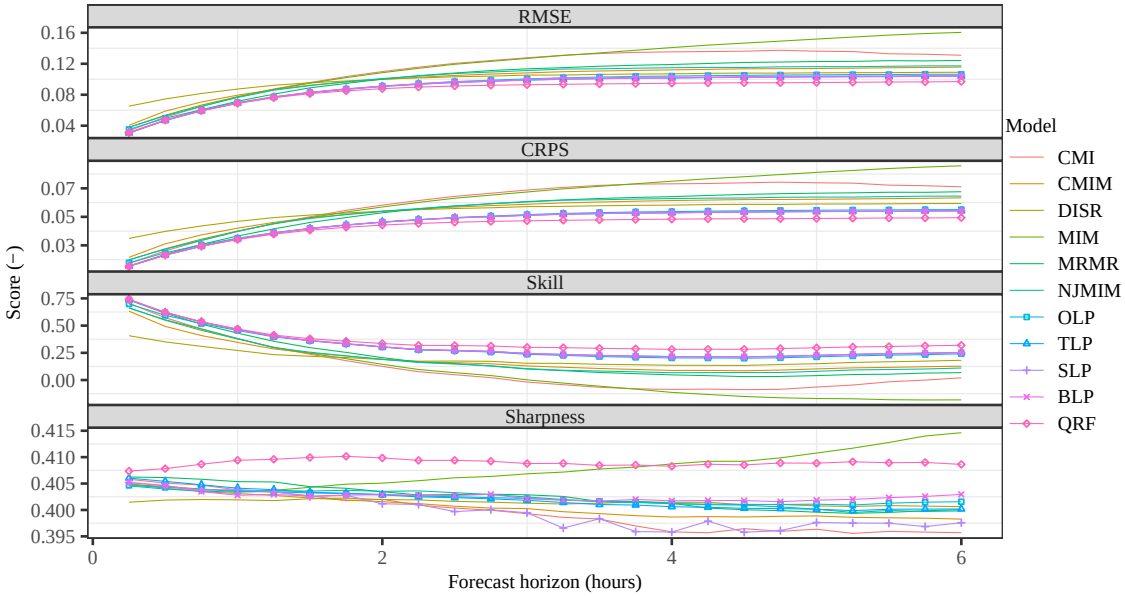


Figure 14 The numerical scores as a function of the forecast horizon for the VPP consisting of PV systems and wind turbines. The scores have been computed on data normalized using the installed capacity.

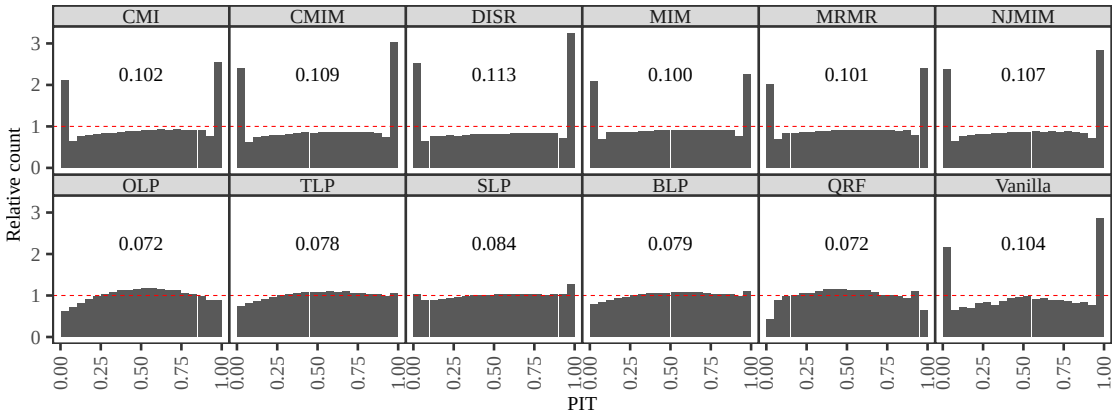


Figure 15 Histograms of the PIT variables combined from all forecast horizons, including their variance, for the VPP consisting of PV systems and wind turbines.

In terms of the numerical scores, Fig. 14 shows that these are generally lower on the testing data compared to the validation data. Again, the most likely explanation for this is the fact that the validation data run from 2019-10-01 until 2019-12-31, which is a challenging period to forecast. When comparing among the models, Fig. 14 indicates that the combination methods and QRF perform similarly during the first 2 forecast horizons. This is interesting because it departs from the results for the PV power forecasts and shows that static forecast combination can be competitive, while naive forecast combination should always be considered as a benchmark. In terms of skill, the difference is substantial compared to only forecasting PV power, which indicates that the inclusion of satellite data significantly deteriorates the performance of Vanilla AnEn. This is a logical consequence because satellite-derived irradiance maps are not useful when forecasting wind power while the additional features do increase the sparsity of the search space, in turn decreasing the quality of the analogs. The fact that 5 out of 6 filter methods result in positive skill confirms this observation. QRF produces the least sharp predictive distributions combined with MIM.

The latter observation becomes relevant when looking at the PIT histograms presented in Fig. 15.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

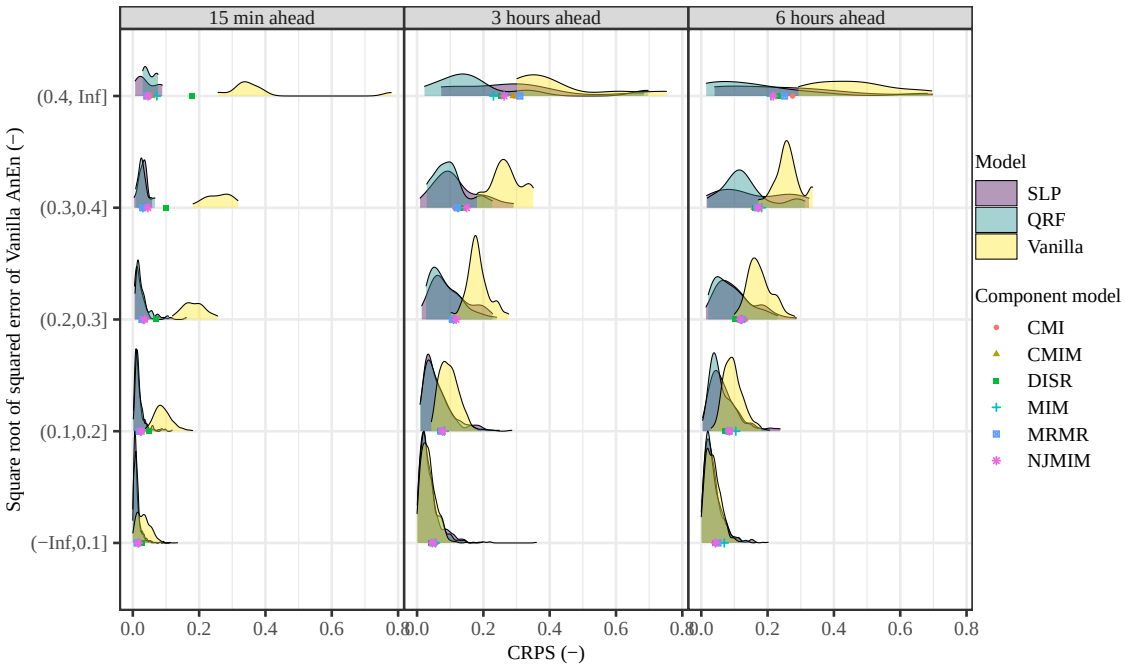


Figure 16 Distribution of the CRPS conditioned on the binned deterministic forecast error of the Vanilla AnEn for 3 forecast horizons and for the VPP consisting of PV systems and wind turbines. The points represent the average CRPS of each component model.

SLP and BLP forecasts appear better calibrated than those produced by QRF and the latter produces less sharp forecasts, which raises the question why the CRPS of QRF is lower than the CRPS of SLP and BLP. Upon closer inspection of the PIT histograms and computing the MBE (not shown here), it can be seen that SLP and BLP suffer from slight negative bias as indicated by the moderate positive slope of the PIT histograms. Interestingly, QRF shows better forecast resolution (not shown here), which relates to the CRPS as CRPS=*Reliability+Uncertainty-Resolution* and these values are averaged over the testing set (Hersbach, 2000). The uncertainty is inherent to the data set and cannot be improved upon; therefore, the forecaster aims to minimize the reliability and maximize the resolution in the aforementioned equation. The resolution is defined as a forecast model's capability to issue forecasts that depend on the prevalent conditions; as such, a climatological model by design has zero resolution because it always generates the same forecast (Lauret et al., 2019). This raises an important question of whether (probabilistic) forecast combination decreases the forecast resolution, which we intend to investigate in future work.

Figure 16 presents the binned square root of the squared error of the Vanilla AnEn model versus the CRPS. When comparing this with the same figure of the VPP consisting only of PV systems (cf. Fig 11), it is interesting to note that CRPS is generally lower when combining wind and PV in a VPP when considering the highest error bin on the vertical axis. This indicates that at high production levels, the forecast error for PV power can be significant, e.g., when broken clouds occur. Overall, however, the PV power forecast error is lower because of the lower capacity factor compared to wind power. Furthermore, Fig. 16 shows that QRF and SLP perform similarly across all point forecast error regimes for the first horizon whereas their performance diverges at the highest point forecast error regimens when the forecast horizon increases. This is a clear consequence of the lower forecast resolution of SLP; when the weather is most variable, a model with high resolution will likely be more accurate.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

III.2 Case study II

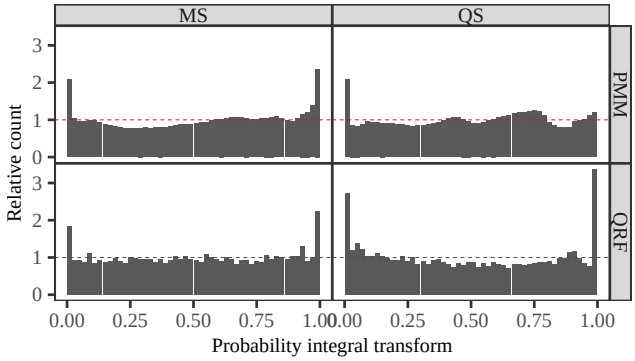


Figure 17 Histograms of the marginal PIT variables combined over all forecast horizons and testing instances where the zenith angle is smaller than 85°.

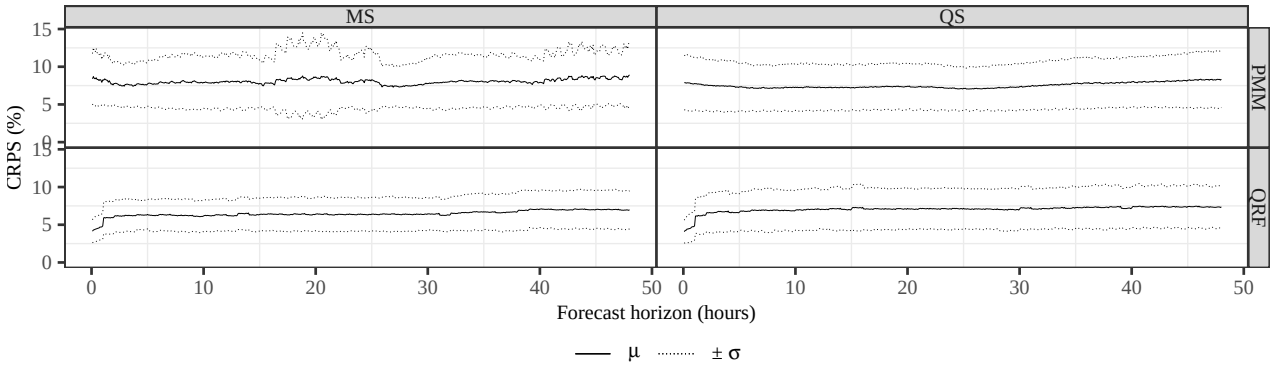


Figure 18 CRPS in percent of nominal capacity as a function of forecast horizon. The mean and standard deviation are computed across the 12 testing months.

This section presents the forecast results of the multivariate probabilistic forecasts of a VPP consisting of solely PV systems, as described in Section II.4.2. Unlike case study I, this case study does not require a validation set because the proposed model does not require training or tuning. Note that the QRF benchmark does require a validation set to estimate the covariance as described in (18) - (19), but we omit the analysis here.

Although the focus of this case study is on multivariate probabilistic forecasting, it is still important to evaluate the marginal predictive distributions. The main reasons are that the energy score and variogram score, or any numerical score for that matter, summarize all the information in a single value, whereas multivariate rank histograms can be challenging to interpret (Thorarinsdottir et al., 2016). Therefore, this section starts with an evaluation of the marginal predictive distributions, i.e., the probabilistic forecasts.

Figure 17 presents histograms of the PIT variables computed over the testing set. The figure shows that the PMM+MS and QRF+QS forecasts tend to be underdispersed in the main part of the distribution. In contrast, Fig. 17 shows that forecasts generated by PMM+QS and QRF+MS are calibrated better in the main part of the distribution. However, the lowest quantile deviates in case of PMM+QS, whereas the most extreme quantiles deviate in case of QRF+MS, which indicates consistent over- and underestimation similar to what was observed in Section III.1. In the case where the trajectory forecasts are generated in one-shot as with PMM, postprocessing the marginal distributions to enhance the calibration would alter the multivariate distribution and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

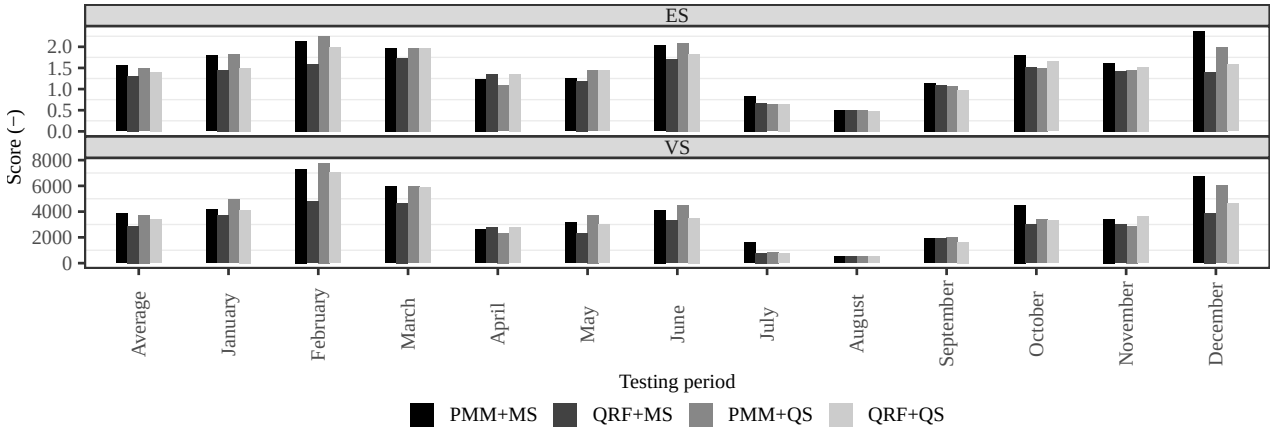


Figure 19 ES and VS averaged over the entire testing set and on a monthly basis. Note that the scores are dimensionless.

undermine the temporal dependence structure. Postprocessing multivariate forecasts is out of the scope of this report but the interested reader is referred to Schefzik and Möller (2018).

Furthermore, Fig. 18 presents the CRPS in percent of the nominal capacity as a function of the forecast horizon. The mean and standard deviations are computed across the 12 testing months and presented as solid and dashed lines, respectively. The figure shows that PMM+MS performs poorly, which is caused by forecasts in the test month December. The poor performance is likely caused by the combination of a lack of suitable ensemble members and the way the NWP ensemble information is summarized, since PMM+QS does not suffer from exceptionally poor performance in December.

Another important observation relates to the relative constancy of PMM+QS in terms of CRPS over the forecast horizons compared to that of QRF+MS and QRF+QS. In the latter cases, CRPS sharply increases during the first hours and then stabilizes. Stability in the forecast error variance over the entire forecast horizon is preferable because it reduces the bullwhip effect (Yang et al., 2019). An example of the bullwhip effect given underdispersive forecasts is that a control algorithm could plan an optimistic strategy that frequently requires the available storage to correct for the forecast errors. Nevertheless, QRF+MS outperforms the other models in terms of CRPS.

Next, we turn our attention to the multivariate forecasts. Figure 19 presents ES and VS as monthly averages in addition to the total averaged scores. The figure clearly shows that PMM+MS performs poorly in December. Besides December, the figure shows that QRF+MS performs substantially better in February than the other models. Overall, QRF+MS outperforms the others but given the limited number of forecast-verification pairs—recall that there are about 700 forecast-verification pairs per month—it is worthwhile to test the significance of the difference between the forecasts.

In hypothesis testing it is common to set the null hypothesis (H_0) such that there is no difference between the forecasts whereas the alternative hypothesis (H_a) is that there is a difference. The Diebold-Mariano (DM) test is commonly used to test H_0. The test requires computation of the forecast error loss differential $d_t = \ell(\boldsymbol{F}_{1,t}, \boldsymbol{y}_t) - \ell(\boldsymbol{F}_{2,t}, \boldsymbol{y}_t)$. However, the DM test is designed specifically for a forecast horizon and consequently, we are uncertain whether it applies here. Furthermore, the paired t-test is not recommended when temporal dependence and contemporaneous correlation are present (Gilleland et al., 2018). Instead, Gilleland et al. (2018) recommend the Hering-Genton (HG) test and circular block bootstrapping. We choose to use block bootstrapping because the HG test, like the DM test, assumes that the series $\boldsymbol{d}$ is covariance stationary, whereas circular block bootstrapping is relevant for small testing sets (Gilleland et al.,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

2018). The bootstrap is repeated 10,000 times with a block length of $\sqrt{T}$ and the results are presented in Table 3 as the mean plus-minus the standard deviation and the confidence interval (CI, significance level $\alpha = 5\%$) in parentheses. As the table shows, 0 always lies within the CI and therefore we cannot reject H_0.

Table 3 Block bootstrapped loss differential presented as $\mu \pm \sigma$ $(2.5\% - 97.5\%)$.

	ES			
	PMM+QS	QRF+QS	PMM+MS	QRF+MS
PMM+QS	0±0 (0—0)	0.06±0.31 (-0.5—0.83)	-0.08±0.21 (-0.67—0.31)	0.13±0.32 (-0.4—0.82)
QRF+QS	(-)	0±0 (0—0)	-0.11±0.29 (-0.78—0.46)	0.11±0.22 (-0.2—0.68)
PMM+MS	(-)	(-)	0±0 (0—0)	0.15±0.27 (-0.31—0.74)
QRF+MS	(-)	(-)	(-)	0±0 (0—0)
	VS			
	PMM+QS	QRF+QS	PMM+MS	QRF+MS
PMM+QS	0±0 (0—0)	397.02±1116.4 (-1557.52—3187.29)	-246.11±794.92 (-1996.46—1632.91)	722.43±1199.35 (-977.84—3611.67)
QRF+QS	(-)	0±0 (0—0)	-533.09±1108.93 (-3073.65—1598.25)	534.45±1119.49 (-657.42—4011.29)
PMM+MS	(-)	(-)	0±0 (0—0)	776.92±1091.41 (-770.52—3274.18)
QRF+MS	(-)	(-)	(-)	0±0 (0—0)

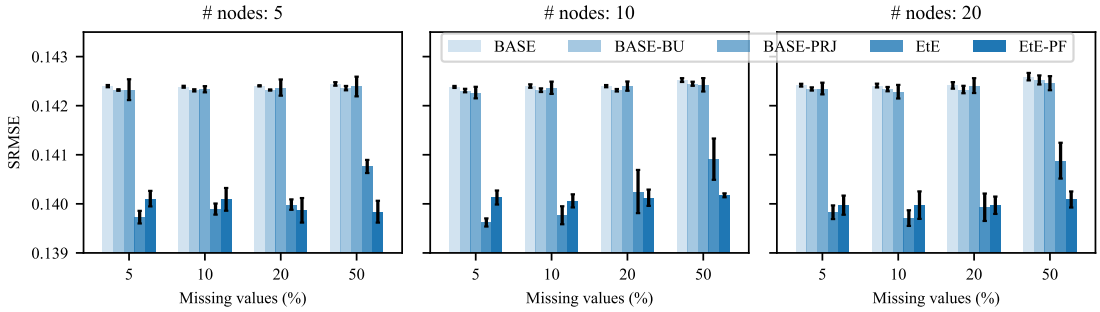
Finally, we report the skill scores relative to MuPEn in Table 4, computed as $1 - L_{\text{model}}/L_{\text{MuPEn}}$ and where L is the loss averaged over the testing set. Evidently, QRF+MS performs best on all scores. However, given that the marginal predictive CDFs of PMM+QS are calibrated slightly better, that no learning step but only proper data organization is required, and that computation time is reduced by approximately 98%, we argue that the PMM is at the very least a highly efficient and interpretable asset in a forecaster's toolbox.

Table 4 CRPS, ES and VS skill scores, relative to MuPEn. Note that we compute the mean and standard deviation ($\mu \pm \sigma$) over all forecast horizons for CRPS.

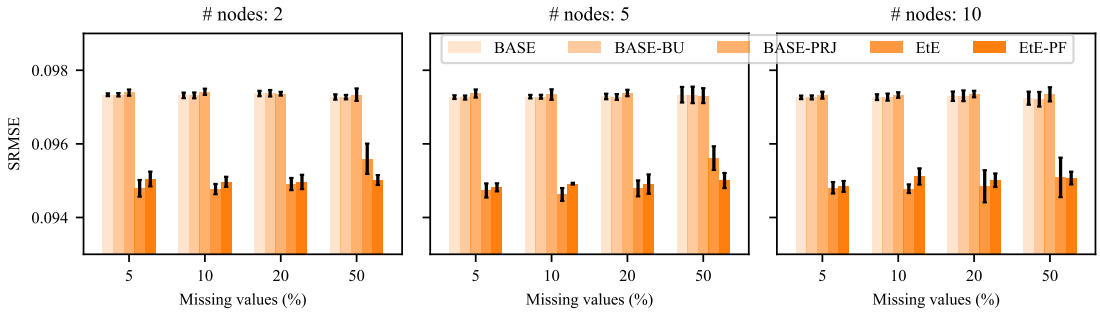
Model	CRPS	ES	VS
PMM+MS	0.188 ± 0.250	0.237	0.384
PMM+QS	0.241 ± 0.234	0.272	0.398
QRF+QS	0.289 ± 0.221	0.308	0.452
QRF+MS	0.348 ± 0.203	0.361	0.536

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

III.3 Case study III



(a) Wind production.



(b) PV production.

Figure 20 Aggregated SRMSE for the hierarchy as a function of the number of sampled nodes and the percentage of missing values per node over 5 iterations. Bars correspond to one standard deviation.

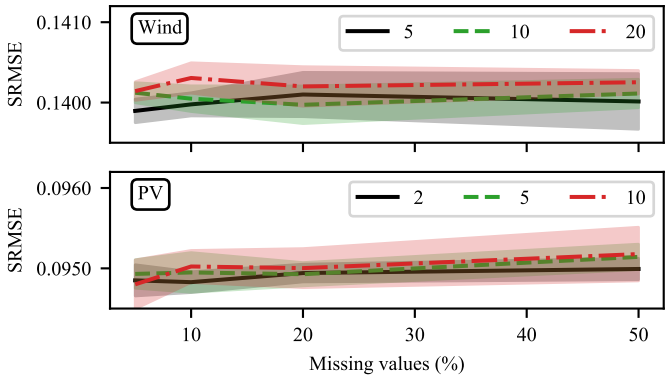


Figure 21 Performance of `EtE-PF` for missing values at different timestamps. The lines indicate the number of malfunctioning nodes, the shaded areas show one standard deviation.

The effect of missing values due to equipment malfunctions is simulated by sampling a subset of bottom-level nodes and setting a percentage of training observations to zero. Both the number of nodes and the percentage of missing values per node are varied and this experiment is repeated 5 times to derive aggregate statistics. For simplicity, we assume that missing values occur at the same timestamp for all nodes.

Figure 20 shows the aggregated SRMSE as a function of the number of sampled nodes and the percentage of missing values per node. Overall, the following can be observed: (i) end-to-end learning (`EtE`) outperforms post-processing methods but is also more heavily affected by missing values; (ii) post-processing methods are robust against missing values; and (iii) the proposed `EtE-PF` combines the best of both worlds. Regarding the wind data set (Fig. 20a), both `EtE`

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Table 5 Average SRMSE (± one standard deviation) per hierarchy level. The best-performing model is underlined in bold font. Bold font indicates that a result does not differ from the best-performing model at the 1% level (Welch's t-test).

Data set	Level (# nodes)	BASE	BASE-BU	BASE-PRJ	EtE	EtE-PF
Wind	1	0.0969 ± 0.0002	0.0977 ± 0.0001	**0.0968 ± 0.0002**	0.0972 ± 0.0005	0.0971 ± 0.0003
	2 (13)	0.1327 ± 0.0132	0.1322 ± 0.0132	0.1326 ± 0.0123	**0.1306 ± 0.0128**	**0.1305 ± 0.0127**
	3 (60)	0.1453 ± 0.0144	0.1453 ± 0.0144	0.1452 ± 0.0138	**0.1429 ± 0.0141**	**0.1428 ± 0.0141**
PV	1	0.0761 ± 0.0001	0.0762 ± 0.0001	**0.0759 ± 0.0001**	0.0765 ± 0.0005	0.0764 ± 0.0002
	2 (3)	**0.0812 ± 0.0040**	**0.0811 ± 0.0041**	**0.0811 ± 0.0037**	**0.0808 ± 0.0035**	**0.0807 ± 0.0035**
	3 (20)	0.1008 ± 0.0176	0.1008 ± 0.0176	0.1009 ± 0.0170	**0.0980 ± 0.0153**	**0.0980 ± 0.0153**

and EtE-PF show improved accuracy for lower percentage of missing values, with the performance of EtE gradually degrading as the percentage of missing values increases. Conversely, the EtE-PF proves to be robust, consistently outperforming the reconciliation methods. This result persists both for the case of an increased number of malfunctioning nodes and an increased percentage of missing values per node. Further, BASE-PRJ performs, albeit slightly, better than the BASE and BASE-BU, corroborating previous findings on the benefits of post-processing. Similar results are observed for the PV data set (Fig. 20b), which has a smaller sample size. Overall, the relative increase in average SRMSE for EtE from the smallest (5%) to the largest (50%) percentage of missing observations is 0.8% for the wind data set and 0.6% for the PV data set.

By examining the accuracy of end-to-end learning as a function of the number of selected nodes and the percentage of missing observations we observe that the former has a negligible effect in overall performance; this is partly attributed to the design of the experiment, as missing values occur at the same timestamp across all nodes. On the contrary, the percentage of missing values has a more pronounced effect. In order to present a comprehensive study, we repeat the above experiment only for EtE-PF with missing values occurring at different timestamps. The results presented in Fig. 21 are similar to the ones achieved before, with the forecast accuracy decreasing only slightly as the number of nodes increases. Thus, we conclude that the proposed EtE-PF successfully mitigates the adverse effects of missing values in the lower levels of the hierarchy.

Lastly, we examine performance for each level of the hierarchy. From Table 5 we observe that in all cases the SRMSE is lower for higher levels of aggregation due to the spatial smoothing effect. The effect is more pronounced for the wind production data, which we partly attribute to the larger number of wind production series examined. For both data sets, the EtE-PF leads to the best performance in the 2nd and 3rd (bottom) level, while BASE-PRJ leads to the best performance for the 1st (top) level, with the results being, generally, statistically significant. Overall, both EtE and EtE-PF consistently improve performance for the bottom-level nodes, highlighting the benefits of exploiting dependencies across time series in an end-to-end learning setting. Note that the results shown in Table 5 are obtained over all the iterations (for uniform timestamps); the difference between EtE-PF and EtE becomes statistically significant if we only examine adverse scenarios (higher percentage of missing values), as the performance of EtE declines.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

IV. Conclusions

This section summarizes the main contributions and findings from Task 3.2. The topics for future work are also identified.

IV.1 Summary

Accurate RES forecasts play an important role in the energy transition as they support decision-making that ensures the power system is operated efficiently and safely. However, variability and intermittency of the solar and wind resources make RES forecasting a challenging task. As the penetration of RESs in the power system increases, the requirements on the forecasts become more stringent as well (e.g., probabilistic, multivariate, high temporal/spatial resolution, etc.). Table 6 presents the KPIs that were achieved in the various case studies.

Task 3.2 provides the following contributions to the field of RES forecasting:

1. **Automatic feature selection.** As the amount of explanatory data increases, so does the need to ensure that redundant features are excluded. Of the feature selection methods, filters have high potential because they are model agnostic. When based on mutual information, these filters can additionally uncover nonlinear relationships between features. In addition, the scores assigned by the filters to the features can be used in an automated framework of feature weighting.

   The proposed method was tested on a real-world data set comprising PV systems and wind turbines that were aggregated as VPPs. In case of the VPP consisting solely of PV systems, the results showed that each filter method selected a feature subset that allowed the forecast model to improve upon the benchmark based on the same forecast model that considered all available features. Moreover, the computational speed improved by approximately 90%, which becomes increasingly important when the temporal resolution increases. In case of the VPP consisting of PV systems and wind turbines, feature selection was highly relevant because a large part of the features, i.e., the satellite-derived irradiance map, was nearly irrelevant due to the dominating share of wind power in the VPP.

2. **Probabilistic forecast combination.** In forecasting, and prediction in general, there are several sources of uncertainty related to, e.g., model selection, model parameters or feature selection. In the frame of automatic feature selection, such uncertainty is especially present because it is a priori uncertain whether the feature selection method captures the necessary information. Forecast combination is an effective method to hedge against uncertainties because the component models offer different perspectives.

   Linear and nonlinear combination methods for probabilistic forecasts were tested on the same VPPs as described above. When the VPP consisted solely of PV systems, the combination methods substantially improved upon the component models. Moreover, whereas the advanced benchmark (QRF) did not generate calibrated forecasts, combination method SLP did. Interestingly, even the naive combination method (OLP) performed quite well, which is often referred to as the "forecast combination puzzle" (Claeskens et al., 2016). On the other hand, BLP, as the theoretically most flexible combination method, performed poorly, which was most likely caused by the biased forecasts it was trained on. When tested on the VPP consisting of PV systems and wind turbines, BLP performed well and the forecasts it was trained on were unbiased.

3. **Seamless multivariate probabilistic forecasts at high temporal resolution.** For power systems that already feature high RES penetration levels, such as island grids, high-temporal resolution forecasts become increasingly important. Moreover, such forecasts should represent

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Table 6 Refer to Fig. 6 for the KPIs of the project.

KPI index	KPI name	KPI baseline	KPI target	KPI achieved
1.2a	% improvement RES forecasting score up to 30 min ahead	Vanilla AnEn	Solar: 3-5% CRPS. Wind: 2-4% CRPS	Solar: 38% CRPS. Wind: 62% CRPS.
1.2b	% improvement RES forecasting score up to 96 h ahead	Vanilla AnEn	Solar: 4-6% CRPS. Wind: 3-5% CRPS	Solar: 13% CRPS. Wind: 35% CRPS.
1.2c	% improvement variogram score for ensemble forecasts	QRF	≥ 0	Statistically insignificant difference
1.2d	% improvement for seamless generic forecasts	QRF	Weighted combination of 1.2a and 1.2b	37%

the correct autocorrelation in order to efficiently schedule thermal generators and control storage devices. However, the increasing temporal resolution brings a set of challenges, particularly related to the computational burden. Pattern matching and efficient similarity search algorithms can alleviate the aforementioned challenges, resulting in a forecast model that does not require training.

The results of the case study on the Greek island of Rhodes showed that the proposed model can significantly improve the computation time; specifically, with approximately 98%. Although the benchmark model (QRF) outperformed the proposed model in terms of the marginal predictive distributions, there was no statistical significant difference between the two models in terms of the multivariate forecasts. A longer data set at more geographical areas would allow for a definitive conclusion, as well as a value-oriented assessment of the forecasts in a decision-making framework.

4. **Hierarchical forecasts with missing values.** Hierarchical forecasts should be coherent to ensure consistent decision making throughout the power system. Usually, coherency is ensured by means of a post-processing step. However, there is a large class of non-parametric machine learning models that generates coherent hierarchical forecasts. The main practical implication is that off-the-shelf machine learning tools can be utilized for hierarchical forecasting, presenting an easy way to create benchmarks. In addition, missing values in the lower part of the hierarchy, e.g., smart-meters, are common and efficient methods are required to handle these instances.

   We proposed a prescriptive trees algorithm for end-to-end learning with missing values. Performance was evaluated in two case studies of wind and PV production point forecasting on a day-ahead horizon. Overall, end-to-end learning showed improved aggregate performance against two-step reconciliation approaches; for the bottom-level series this improvement was 1.7% and 2.8% for the wind and PV data, respectively. Conversely, reconciliation approaches proved to be more robust against the number of missing values. The proposed solution managed to combine the best of both worlds as it maintained improved performance while also mitigating the adverse effect of missing data for end-to-end learning.

IV.2 Dissemination

Each case study has one companion publication published in a peer-reviewed conference.

Automatic Feature Selection and Forecast Combination

Section II.2. D. van der Meer, S. Camal, G. Kariniotakis, "Generalizing Renewable Energy Forecasting Using Automatic Feature Selection and Combination," *17th International Conference on Probabilistic Methods Applied to Power Systems*, 2022, doi:10.5281/zenodo.6451891.

Seamless trajectory forecasts

Section II.3.2. D. van der Meer, S. Camal, G. Kariniotakis, "Seamless intra-day and day-ahead

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

multivariate probabilistic forecasts at high temporal resolution," *17th International Conference on Probabilistic Methods Applied to Power Systems*, 2022, doi:10.5281/zenodo.6451913.

Hierarchical forecasts with missing values

Section II.3.3. A. Stratigakos, D. van der Meer, S. Camal, G. Kariniotakis, "End-to-end Learning for Hierarchical Forecasting of Renewable Energy Production with Missing Values," *17th International Conference on Probabilistic Methods Applied to Power Systems*, 2022, url: https://hal.archives-ouvertes.fr/hal-03527644.

IV.3 Future Work

The following topics were identified for future work:

1. **Automatic Feature Selection and Forecast Combination.** The increasing spatial footprint of VPPs remains a challenge to forecast models, while forecast combination suffers from static parameter optimization. Future research should consider:
   (a) testing other nonlinear filter methods based on, e.g., Kendall's rank correlation coefficient;
   (b) lowering the correlation between the component models so as to improve the effect of combination;
   (c) converting the static combination parameter learning to an online setting (e.g., Thorey et al. (2018)).
2. **Seamless trajectory forecasts.** A challenge for search-based algorithms is the curse of dimensionality, which was addressed in the previous paper. Future research should consider:
   (a) applying feature selection to trajectory forecasts, which will be challenging considering the importance of the temporal and spatial correlation;
   (b) additional testing on multiple resources and, crucially, longer time periods to relieve the curse of dimensionality;
   (c) developing so-called consistency bands for multivariate rank histograms (not shown in this deliverable) to account for a test set of limited size in combination with high-dimensional forecasts.
3. **Hierarchical forecasts with missing values.** Missing data is common in the measurements but also occurs in input features and increases the uncertainty of the forecasts. Future research should consider:
   (a) probabilistic forecasting with missing values;
   (b) examining the impact of network losses on forecast coherency;
   (c) examining the effect of corrupted or missing values in the feature vector.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Appendices

A. Sample Average Approximation

This section describes analytical solutions for the SAA sub-problems in (37). For $t \in [T]$, the respective SAA is given by

$$\min_{\mathbf{z}} \left\{ \frac{1}{2} \sum_{t \in [T]} \|\mathbf{y}_t - (\mathbf{1} - \boldsymbol{\gamma}_t) \odot \mathbf{z}\|_2^2 \mid \mathbf{A}\mathbf{z} = \mathbf{0} \right\}. \tag{44}$$

For simplicity, the objective is scaled. The KKT optimality conditions for this problem can be written as

$$\sum_{t \in [T]} (\mathbf{y}_t - (\mathbf{1} - \boldsymbol{\gamma}_t) \odot \mathbf{z}) + \mathbf{A}^\intercal \mathbf{v} = \mathbf{0}, \mathbf{A}\mathbf{z} = \mathbf{0}, \tag{45}$$

where $\mathbf{v} \in \mathbb{R}^{n_a}$ denotes the dual variables. We write (45) as

$$\begin{bmatrix} \mathbf{P} & -\mathbf{A}^\intercal \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{z} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \sum_{t \in [T]} \mathbf{y}_t \\ \mathbf{0} \end{bmatrix}, \tag{46}$$

where $\mathbf{P} = \operatorname{diag}\left(\sum_{t \in [T]}(1 - \gamma_{1t}), \ldots, \sum_{t \in [T]}(1 - \gamma_{nt})\right)$ is an n-size diagonal matrix whose entries equal the number of non-missing values per series. Hence, we need to solve this set of $n + n_a$ linear equations in the $n + n_a$ variables. Lastly, note that it is possible for $\mathbf{P}$ to become singular; in this case, the least-squares solution of (46) can be used.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

References

Agoua, X. G., Girard, R., and Kariniotakis, G. Probabilistic Model for Spatio-Temporal Photovoltaic Power Forecasting. *IEEE Transactions on Sustainaible Energy*, -(-):1–9, 2018. ISSN 1949-3029. doi: 10.1109/TSTE.2018.2847558.

Amaro e Silva, R., Haupt, S. E., and Brito, M. C. A regime-based approach for integrating wind information in spatio-temporal solar forecasting models. *Journal of Renewable and Sustainable Energy*, 11(5):056102, 2019. doi: 10.1063/1.5098763.

Berrisch, J. and Ziel, F. Crps learning. *Journal of Econometrics*, 2021. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2021.11.008.

Bertsimas, D. and Kallus, N. From predictive to prescriptive analytics. *Management Science*, 66 (3):1025–1044, 2020.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., and Lang, M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 143:106839, 2020. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2019.106839.

Bracale, A., Carpinelli, G., and De Falco, P. A Probabilistic Competitive Ensemble Method for Short-Term Photovoltaic Power Forecasting. *IEEE Transactions on Sustainable Energy*, 8(2):551–560, apr 2017. ISSN 19493029. doi: https://doi.org/10.1109/TSTE.2016.2610523.

Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees*. CRC press, 1984.

Camal, S., Michiorri, A., and Kariniotakis, G. Optimal offer of automatic frequency restoration reserve from a combined pv/wind virtual power plant. *IEEE Transactions on Power Systems*, 33 (6):6155–6170, 2018. doi: 10.1109/TPWRS.2018.2847239.

Camal, S., Teng, F., Michiorri, A., Kariniotakis, G., and Badesa, L. Scenario generation of aggregated wind, photovoltaics and small hydro production for power systems applications. *Applied Energy*, 242:1396–1406, 2019. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2019.03.112.

Carriere, T., Vernay, C., Pitaval, S., and Kariniotakis, G. A novel approach for seamless probabilistic photovoltaic power forecasting covering multiple time frames. *IEEE Transactions on Smart Grid*, 11(3):2281–2292, 2020. doi: 10.1109/TSG.2019.2951288.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762, 2016. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2015.12.005.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., and Searight, K. Probabilistic Weather Prediction with an Analog Ensemble. *Monthly Weather Review*, 141:3498–3516, 2013. ISSN 0027-0644. doi: 10.1175/MWR-D-12-00281.1.

Di Modica, C., Pinson, P., and Taieb, S. B. Online forecast reconciliation in wind power prediction. *Electric Power Systems Research*, 190:106637, 2021.

Fatemi, S. A., Kuh, A., and Fripp, M. Parametric methods for probabilistic forecasting of solar irradiance. *Renewable Energy*, 129:666–676, 2018. ISSN 0960-1481. doi: https://doi.org/10.1016/j.renene.2018.06.022.

Geurts, P., Ernst, D., and Wehenkel, L. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Gilleland, E., Hering, A. S., Fowler, T. L., and Brown, B. G. Testing the tests: What are the impacts of incorrect assumptions when applying confidence intervals or hypothesis tests to compare competing forecasts? *Monthly Weather Review*, 146(6):1685 – 1703, 2018. doi: 10.1175/MWR-D-17-0295.1.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Gneiting, T. and Ranjan, R. Combining predictive distributions. *Electronic Journal of Statistics*, 7: 1747–1782, 2013. doi: https://doi.org/10.1214/13-EJS823.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005. doi: https://doi.org/10.1175/MWR2904.1.

Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003. ISSN 1533-7928.

Han, X., Dasgupta, S., and Ghosh, J. Simultaneously reconciled quantile forecasting of hierarchically related time series. In *International Conference on Artificial Intelligence and Statistics*, pages 190–198. PMLR, 2021.

Hersbach, H. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecasting*, 15(5):559–570, 2000. doi: doi.org/10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2.

Kambezidis, H. D. and Psiloglou, B. E. Estimation of the optimum energy received by solar energy flat-plate convertors in greece using typical meteorological years. part i: South-oriented tilt angles. *Applied Sciences*, 11(4), 2021. ISSN 2076-3417. doi: 10.3390/app11041547.

Killinger, S., Engerer, N., and Müller, B. Qcpv: A quality control algorithm for distributed photovoltaic array power output. *Solar Energy*, 143:120–131, 2017. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2016.12.053.

Kursa, M. B. *praznik: Tools for Information-Based Feature Selection*, 2020. URL `https://CRAN.R-project.org/package=praznik`. R package version 8.0.0.

Lahiri, K., Peng, H., and Zhao, Y. Testing the value of probability forecasts for calibrated combining. *International Journal of Forecasting*, 31(1):113–129, 2015. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2014.03.005.

Lauret, P., David, M., and Pinson, P. Verification of solar irradiance probabilistic forecasts. *Solar Energy*, 194:254 – 271, 2019. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2019.10.041. URL `http://www.sciencedirect.com/science/article/pii/S0038092X19310382`.

Lefèvre, M., Oumbe, A., Blanc, P., Espinar, B., Gschwind, B., Qu, Z., Wald, L., Schroedter-Homscheidt, M., Hoyer-Klick, C., Arola, A., Benedetti, A., Kaiser, J. W., and Morcrette, J.-J. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. *Atmospheric Measurement Techniques*, 6(9):2403–2418, 2013. doi: http://dx.doi.org/10.5194/amt-6-2403-2013.

Leutbecher, M. and Palmer, T. Ensemble forecasting. *Journal of Computational Physics*, 227 (7):3515–3539, 2008. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2007.02.014. Predicting weather, climate and extreme events.

Liu, Z., Yan, Y., Yang, J., and Hauskrecht, M. Missing value estimation for hierarchical time series: A study of hierarchical web traffic. In *2015 IEEE International Conference on Data Mining*, pages 895–900. IEEE, 2015.

Lorenz, E. N. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:130–141, 1963. ISSN 1520-0469. doi: doi.org/10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Meinshausen, N. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

Möller, A. and Groß, J. Probabilistic temperature forecasting with a heteroscedastic autoregressive ensemble postprocessing model. *Quarterly Journal of the Royal Meteorological Society*, 146(726):211–224, 2020. doi: https://doi.org/10.1002/qj.3667.

Morales, J., Conejo, A., Madsen, H., Pinson, P., and Zugno, M. *Integrating Renewables in Electricity Markets - Operational Problems*. Springer, 01 2014. ISBN 9781461494119.

Murphy, A. H. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting*, 8(2):281–293, 1993. doi: 10.1175/1520-0434(1993)008⟨0281:WIAGFA⟩2.0.CO;2.

Pauwels, L. L. and Vasnev, A. L. A note on the estimation of optimal weights for density forecast combinations. *International Journal of Forecasting*, 32(2):391–397, 2016. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2015.09.002.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., et al. Forecasting: theory and practice. *International Journal of Forecasting*, 2022. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.11.001.

Pinson, P. Wind energy: Forecasting challenges for its operational management. *Statistical Science*, 28(4):564–585, 11 2013. doi: https://doi.org/10.1214/13-STS445.

Pinson, P. and Girard, R. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96:12–20, 2012. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2011.11.004. Smart Grids.

Pinson, P., Madsen, H., Nielsen, H. A., Papaefthymiou, G., and Klöckl, B. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, 12(1):51–62, 2009. doi: 10.1002/we.284.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005. doi: https://doi.org/10.1175/MWR2906.1.

Rangapuram, S. S., Werner, L. D., Benidis, K., Mercado, P., Gasthaus, J., and Januschowski, T. End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In *International Conference on Machine Learning*, pages 8832–8843. PMLR, 2021.

S., E. E. Stochastic dynamic prediction. *Tellus*, 21(6):739–759, 1969. doi: 10.1111/j.2153-3490.1969.tb00483.x.

Schefzik, R. and Möller, A. Chapter 4 - Ensemble Postprocessing Methods Incorporating Dependence Structures. In *Statistical Postprocessing of Ensemble Forecasts*, pages 91 – 125. Elsevier, 2018. ISBN 978-0-12-812372-0. doi: https://doi.org/10.1016/B978-0-12-812372-0.00004-2.

Scheuerer, M. and Hamill, T. M. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321 – 1334, 2015. doi: 10.1175/MWR-D-14-00269.1.

Stone, M. The opinion pool. *Annals of Mathematical Statistics*, 32(4):1339–1342, 12 1961. doi: https://doi.org/10.1214/aoms/1177704873.

Stratigakos, A. C., Camal, S., Michiorri, A., and Kariniotakis, G. Prescriptive trees for integrated forecasting and optimization applied in trading of renewable energy. *IEEE Transactions on Power Systems*, pages 1–1, 2022. doi: 10.1109/TPWRS.2022.3152667.

Taieb, S. B., Taylor, J. W., and Hyndman, R. J. Coherent probabilistic forecasts for hierarchical time series. In *International Conference on Machine Learning*, pages 3348–3357. PMLR, 2017.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Thorarinsdottir, T. L., Scheuerer, M., and Heinz, C. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*, 25(1):105–122, 2016. doi: 10.1080/10618600.2014.977447.

Thorey, J., Chaussin, C., and Mallet, V. Ensemble forecast of photovoltaic power with online CRPS learning. *International Journal of Forecasting*, 34(4):762–773, 2018. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2018.05.007.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

van der Meer, D., Wang, G. C., and Munkhammar, J. An alternative optimal strategy for stochastic model predictive control of a residential battery energy management system with solar photovoltaic. *Applied Energy*, 283:116289, 2021. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2020.116289.

van der Meer, D. A benchmark for multivariate probabilistic solar irradiance forecasts. *Solar Energy*, 225:286–296, 2021. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2021.07.010.

van der Meer, D., Shepero, M., Svensson, A., Widén, J., and Munkhammar, J. Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using gaussian processes. *Applied Energy*, 213:195–207, 2018. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2017.12.104.

Van Erven, T. and Cugliari, J. Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts. In *Modeling and stochastic learning for forecasting in high dimensions*, pages 297–317. Springer, 2015.

Vannitsem, S., Wilks, D. S., and Messner, J. W., editors. *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, 2019. ISBN 978-0-12-812372-0. doi: https://doi.org/10.1016/B978-0-12-812372-0.09993-3.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019.

Yang, D. Ultra-fast preselection in lasso-type spatio-temporal solar forecasting problems. *Solar Energy*, 176:788 – 796, 2018. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2018.08.041.

Yang, D. and van der Meer, D. Post-processing in solar forecasting: Ten overarching thinking tools. *Renewable and Sustainable Energy Reviews*, 140:110735, 2021. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2021.110735.

Yang, D., Dong, Z., Reindl, T., Jirutitijaroen, P., and Walsh, W. M. Solar irradiance forecasting using spatio-temporal empirical kriging and vector autoregressive models with parameter shrinkage. *Solar Energy*, 103:550–562, 2014. ISSN 0038092X. doi: dx.doi.org/10.1016/j.solener.2014.01.024.

Yang, D., Yagli, G. M., and Srinivasan, D. Sub-minute probabilistic solar forecasting for real-time stochastic simulations. *Renewable and Sustainable Energy Reviews*, 153:111736, 2022. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2021.111736.

Yang, D. Ultra-fast analog ensemble using kd-tree. *Journal of Renewable and Sustainable Energy*, 11(5):053703, 2019. doi: 10.1063/1.5124711.

Yang, D. Reconciling solar forecasts: Probabilistic forecast reconciliation in a nonparametric framework. *Solar Energy*, 210:49–58, 2020.

Yang, D., Wu, E., and Kleissl, J. Operational solar forecasting for the real-time market. *International Journal of Forecasting*, 35(4):1499–1519, 2019. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.03.009.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Zarzalejo, L. F., Polo, J., Martín, L., Ramírez, L., and Espinar, B. A new statistical approach for deriving global solar radiation from satellite images. *Solar Energy*, 83(4):480–484, 2009. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2008.09.006.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337

Smart4RES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864337