Environmental Data Science (2025), 4: e45, 1–28
doi:10.1017/eds.2025.10021

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Toward accurate forecasting of renewable energy: Building datasets and benchmarking machine learning models for solar and wind power in France

Eloi Lindas[1,2], Yannig Goude[3,4] and Philippe Ciais[1]

[1]Laboratoire des Sciences du Climat et de l'Environnement (LSCE), IPSL, CEA/CNRS/UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France
[2]Atos Inno'Lab TS Bezons, Atos, Bezons, France
[3]Laboratoire de Mathématiques d'Orsay (LMO), Faculté des Sciences d'Orsay, CNRS, Université Paris-Saclay, Orsay, France
[4]EDF R&D Lab, OSIRIS, EDF, Palaiseau, France
Corresponding author: Eloi Lindas; Email: eloi.lindas@lsce.ipsl.fr

Received: 10 December 2024; **Revised:** 09 July 2025; **Accepted:** 18 August 2025

Keywords: climate; electricity supply; forecasting; machine learning; renewable sources

Abstract

Accurate prediction of nondispatchable renewable energy sources is essential for grid stability and price prediction. Regional power supply forecasts are usually indirect through a bottom-up approach of plant-level forecasts, incorporate lagged power values, and do not use the potential of spatially resolved data. This study presented a comprehensive methodology for predicting solar and wind power production at a country scale in France using machine learning models trained with spatially explicit weather data combined with spatial information about production sites' capacity. A dataset is built spanning from 2012 to 2023, using daily power production data from Réseau de Transport d'Electricité (the national grid operator) as the target variable, with daily weather data from ECMWF Re-Analysis v5, production sites capacity and location, and electricity prices as input features. Three modeling approaches are explored to handle spatially resolved weather data: spatial averaging over the country, dimension reduction through principal component analysis, and a computer vision architecture to exploit complex spatial relationships. The study benchmarks state-of-the-art machine learning models as well as hyperparameter tuning approaches based on cross-validation methods on daily power production data. Results indicate that cross-validation tailored to time series is best suited to reach low error. We found that neural networks tend to outperform traditional tree-based models, which face challenges in extrapolation due to the increasing renewable capacity over time. Model performance ranges from 4% to 10% in normalized root-mean-squared error for midterm horizon, achieving similar error metrics to local models established at a single-plant level, highlighting the potential of these methods for regional power supply forecasting.

Impact Statement

Accurate power production forecasts, particularly for solar and wind power which are sensitive to weather conditions, are critical for grid stability, optimizing renewable energy integration, and supporting the transition to cleaner energy. We predict national power output in France by taking advantage of time-varying images of weather and power generation units' capacity as input data for different machine learning models. The key

This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

finding is that image-based models outperform time series-based models. The results of this research provide a practical model benchmark usable for practitioners and policymakers.

1. Introduction

To meet the 2050 net-zero scenario (United Nations Convention on Climate Change, 2015) of the European Union (EU) reinforced by the European Green Deal, which aims at decreasing net greenhouse gas emissions by 55% by 2030 (European Commission, 2019). Sustainable energy sources have become key to clean power production and reduced emissions from the energy sector in Europe. As power demand increases, however, fossil reliance is still high, accounting for 68% of the global primary energy consumed in 2023 and 40% of the electricity produced in the EU (British Petroleum [BP], 2024; Ritchie and Rosado, 2020). Electrification, coupled with more renewable and other low-carbon power supplies, is needed to reduce dependence on fossil fuels. To meet the CO_2 emissions goals of the EU, solar and wind power generation need to double their capacity by 2030 to produce 48% of Europe's energy share (International Renewable Energy Agency (IRENA), 2020b).

France has set a reduction of 33% of its emissions by 2030 compared to 1990, and pledged to reach greenhouse gas neutrality in 2050 (Ministère de la Transition Ecologique, 2020). This involves an increase in renewable power capacity installed throughout the country. The capacity of solar and wind power plants has tripled since 2012, and this growth is expected to accelerate with the capacity being planned to double from 2017 to 2028 (Ministère de la Transition Ecologique, 2019). Increasing renewable capacity comes with grid distribution challenges to prevent gaps between supply and demand, especially during the day when production may exceed consumption (Liu et al., 2023a). Accurate forecasts of power generation can improve the stability, reliability, quality, and penetration level of renewable energy (International Renewable Energy Agency (IRENA), 2020a). Solar and wind power sources depend on environmental and climate variables such as temperature, solar radiation, and wind speed, making their load highly variable (Engeland et al., 2017; Wang et al., 2019b). This variability leads to obstacles for grid operators as they need to constantly balance the demand with the supply. This is one of the reasons why specific models for understanding and predicting day-to-day renewable power generation have motivated interest from researchers and practitioners.

Many studies addressed the problem of short- (10 min–1 h) to medium-term (3 h–3 days) forecasting of renewable power using weather data from stations or numerical weather prediction (NWP). The impact of weather data and variable importance on forecasting energy supply, photovoltaic (PV), and wind power was studied thoroughly (Vladislavleva et al., 2013; De Giorgi et al., 2014; Zhong and Wu, 2020; Liu et al., 2023b). At the local scale, Malvoni et al. (2016) used solar radiation and temperature to predict the generation of a Mediterranean PV plant. The effect of various climates throughout the planet on hourly PV production was also investigated by Alcaniz et al. (2023). Other works such as Ahmad and Hossain (2020) made use of weather forecasts to maximize hydropower generated from dams while Couto and Estanqueiro (2022) who examined model-based predictive features for wind power predictions. Frequently, the availability of accurate weather observation is a bottleneck when working with a dedicated local area, not to mention their inherent sparsity and noise level, leading to NWP being preferred by researchers. Yet, when both types of weather data are available, they can be combined (Sharma et al., 2011; López Gómez et al., 2020).

Recent advances in forecasting variable renewable energy generation have seen statistical, machine learning, and deep learning models gain popularity among practitioners (Wang et al., 2019a; Iheanetu, 2022; Krechowicz et al., 2022; Tsai et al., 2023). Thanks to the increase in weather and power data availability and quality, models have proven to be useful in revealing driving factors and learning from complex patterns (Sweeney et al., 2020). Depending on the spatial and temporal scale, statistical models can outperform traditional physics-based models, which motivated the development of hybrid models (Bellinguer et al., 2020; Castillo-Rojas et al., 2023; Gijón et al., 2023). The link function between weather conditions and PV panels or wind turbines power output has been thoroughly investigated through

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

different types of models (Dolara et al., 2015; Mayer and Gróf, 2021; Zhou et al., 2022; Bilendo et al., 2023). Still, challenges remain when developing models for a large region or country.

Statistical data-driven models such as auto-regressive moving average (ARMA) and their variants (ARIMA, ARIMAX, SARIMA, and SARIMAX) have demonstrated reasonable performance, as shown in recent work (Chen and Folly, 2018; Ryu et al., 2022). Support vector machine, k-Nearest Neighbors, Generalized Additive Model (GAM), and tree-based and boosted models also gave good performance in forecasting power output from weather data (Kim et al., 2019; Condemi et al., 2021). Current trends have seen the use of artificial neural networks, computer vision (CV), and natural language processing models. Their application in renewable power forecasting shows promising performance. Multilayered perceptron (MLP), convolutionnal neural network (CNN), vision transformers (ViT) (Lim et al., 2022; Keisler and Naour, 2025), and sequence architectures such as recurrent neural network or long–short term memory deep learning models were also applied in various renewable energy forecasting frameworks (solar and wind) (Elsaraiti and Merabet, 2022; Abdul Baseer et al., 2023). A key advantage is their flexibility and ability to combine several data sources to make predictions, not to mention the different ways they can exploit complex spatiotemporal data.

Research on statistical models is not limited to model architectures. Data preprocessing techniques are also important to improve forecast performance. Principal component analysis (PCA), wavelet decomposition, time series detrending, and exponential smoothing can be applied to extract relevant features, reduce dimension, remove noise, or reveal pertinent phenomena from the data (Liu and Chen, 2019; Iheanetu, 2022). These techniques are mainly used as a first step to improve the robustness and performance of a model. It is important to point out that such techniques can be applied regardless of the type of data at hand, whether it is time series or gridded data over a region, albeit the second option being less explored.

Besides the methodology and models used for forecasting, differences between studies arise from the input and output data. Depending on the purpose and the availability of the data, the time and space resolution as well as temporal and spatial ranges differ between studies (Engeland et al., 2017). Research works encompass scales from short-term single plant forecasts with a time resolution of 5–10 minutes (Malvoni et al., 2017; Ryu et al., 2022; Gijón et al., 2023) to medium-term daily forecasts of a region (Kim et al., 2017). However, due to the lack of available good quality data, regional forecasts are often made out of single plant forecasts aggregated to the desired region. This means an indirect prediction of the regional power supply. Moreover, the temporal scale rarely exceeds a few years' worth of data (Chen and Folly, 2018; Iheanetu, 2022). Thus, gaps exist between short to medium term and regional forecasts, leading to difficulties in comparing results between studies and improving modeling performance.

Most prior studies have used a bottom-up approach based on single-plant models, which neglects the integration of spatial information for prediction. Additionally, many existing models enhanced their performance by incorporating lagged data of the target time series itself, such as power supply from the previous day or hour. To overcome these limitations, in this study, we use supervised machine learning models and test the impact of using spatially resolved data as model inputs. We also decided to exclude the use of lagged inputs from the time series themselves as model inputs. The first goal is to assess the influence of the model calibration procedure, especially the cross-validation protocol, on time series-based model error estimation. The second goal is to compare models ingesting explicit weather "images" against averaged variables as inputs.

We first explain how we build input datasets for wind and PV production integrating spatially resolved weather data and generation units' capacity and locations. These input images span the period from January 1, 2012, to December 31, 2023, at hourly resolution as presented in Section 2. Second, we present three different modeling approaches to handle the weather-gridded data to forecast daily wind and PV power production in Section 3.1. Finally, we explore cross-validation and hyperparameter optimization procedures in Section 3.3 to give insights and recommendations for model calibration before benchmarking widespread state-of-the-art machine learning models on our different modeling approaches in Section 4.

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

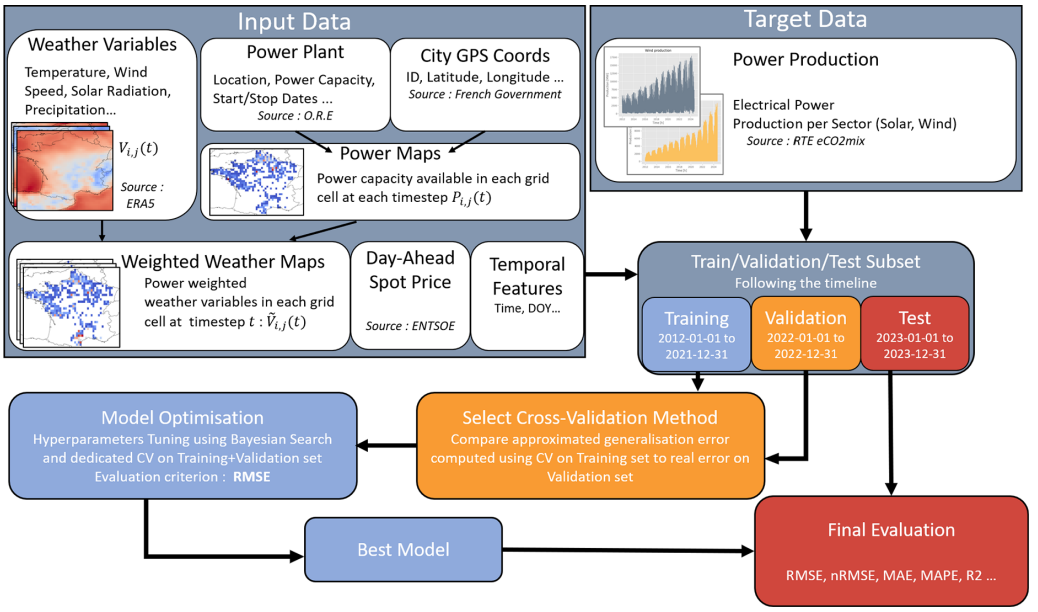

Figure 1. *Global framework of this study represented schematically.*

2. Data

In this section, we describe the target power supply data, the input weather data and power units data, and other input data sources, with the processing workflow to prepare them as input for supervised learning approaches. Figure 1 presents the overall approach, with more details given in the following sections.

2.1. Target data

We used as target wind and solar power from the RTE eCO$_2$mix database. RTE is the public French national Transmission System Operator (TSO) managing the whole electrical grid. RTE provides near-real-time data on electrical consumption, production, flows, and CO_2 emissions within the eCO$_2$mix application.[1] Electricity production data from RTE covers eight sectors: coal, oil, gas, nuclear, hydro, solar, wind, and bioenergy. We recovered production data for nondispatchable renewable wind and solar power. Solar refers to photovoltaic solar panels and wind to both onshore and offshore turbines.

Time-wise, data are available since January 1, 2012, and were retrieved until December 31, 2023. Resolution is half-hourly from January 1, 2012, to January 31, 2023, and quarter-hourly from February 1, 2023, to December 31, 2023.[2] We aggregated the data to an hourly resolution to be consistent with the time resolution of our inputs (see Section 2.2). Data being available at the country (NUTS0) or regional (NUTS1) scale, we chose to work directly with country-scale data. This dataset excluded Corsica and other French islands or overseas territories, which are considered self-sufficient in electricity.

France is part of the EU electricity market and the EU grid interconnection. In this work, we aimed to model the electrical power produced using solar and wind from France only, without taking into account any connection with neighboring countries. Therefore, we did not integrate imports and exports into our power supply target and retained only the production data, presented in Figure 2.

[1] RTE eCO$_2$mix website, available at https://www.rte-france.com/en/eco2mix (accessed 19 September 2024).
[2] Resolutions might change for 2023 in future releases. Current resolutions and types of data are given for September 2024 release.

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

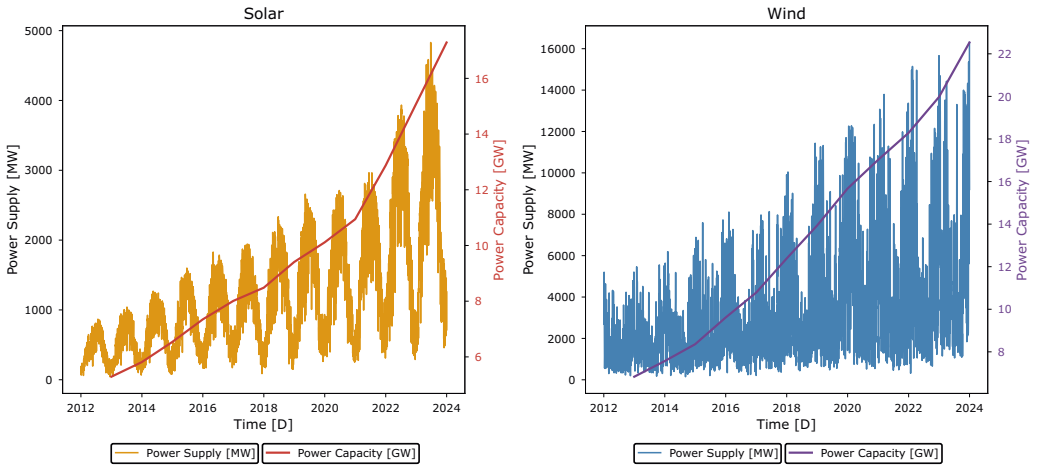

Figure 2. *Power supply and capacity time series for wind and solar in France for the period of interest. The power capacity curves have been smoothed to a yearly resolution.*

2.2. Input data

Our input data are based on gridded weather data weighted by the power capacity available at the given time and location, electricity day-ahead spot price, and other temporal features such as time or day of the year. We combined several different high-quality open-access databases from French governmental or government-affiliated organizations to create coherent inputs.

2.2.1. Weather data

We recovered hourly weather data from the ERA5 reanalysis (Hersbach et al., 2020) on single levels for the period of interest from January 1, 2012, to December 31, 2023. We used the domain bounded by 51° North, 42.5° South, −4.55° West, and 7.95° East which covers France, re-interpolating the original spatial grid of 0.25° × 0.25° or 30 km × 30 km. The weather variables we selected are those usually used for renewable power prediction: temperature at 2 m, Northward and Eastward wind speed at 10 and 100 m, instantaneous wind gust speed at 10 m, surface solar radiation downwards, total precipitation, evaporation, and runoff (Table A1). To select the variables relevant to wind and solar power, we used the mutual information between weather variables and power supply targets (Kraskov et al., 2004). We normalized the mutual information to one and kept only variables that had a score higher than 20%. This leads to hourly maps with 35 latitude and 51 longitude points for each considered variable in netCDF files.

2.2.2. Power units location, capacity, and activity

To get information on the location of facilities with installed solar panels or wind turbines, we used yearly released data from the Opérateurs Réseaux Energies (ORE)[3] agency database of all electrical facilities used for producing or storing electricity in France. The inventory published on December 31, 2023, contained around 84,000 electricity-producing units, among which 2,183 are wind facilities and 72,703 are PV farms. Rooftop PV panels dedicated to autoconsumption are not included. Because the ORE dataset did not provide the exact location of each facility, we merged it with the French governmental city database[4] using City ID, to allocate each facility to a 30-km grid cell of our weather maps. A city refers to an NUTS4 entity. City ID is a unique identifier provided to every French city by Institut National de la Statistique et des Etudes Economiques. Facilities' city IDs that were missing in ORE accounted for less

[3] Dataset used can be retrieved from ORE website, available at https://opendata.agenceore.fr/pages/accueil/.

[4] This database can be found on the French government Open-Data platform, available at https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-referentiel-geographique/export/.

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

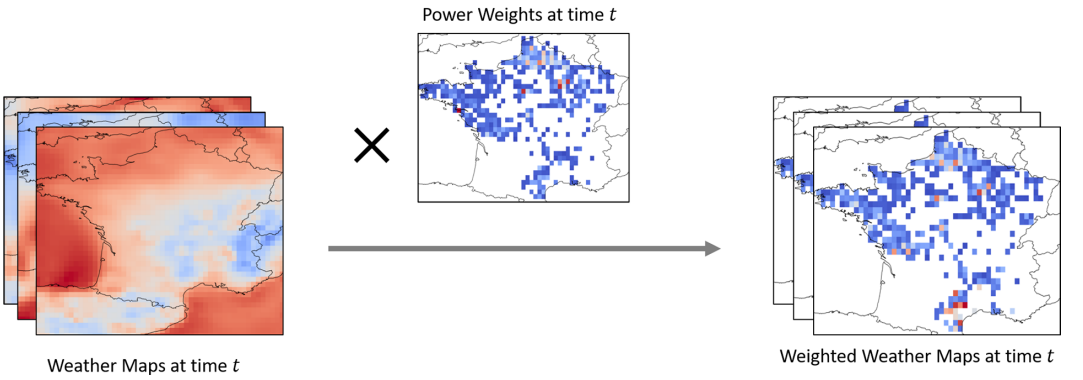

Figure 3. *Illustration of power-weighted weather maps creation for wind.*

than 2% of the data and were discarded. We assigned facilities to their corresponding wind or solar sector, keeping only PV panels for solar and including both offshore and onshore turbines for wind. The maximum power that can be produced by each facility in MW provided by ORE was used as its capacity. Some power capacity data were missing, representing 0.25% fo the data and thus were discarded. To account for the activity period of each facility, we added its start and stop dates. If the stop date was not given in the ORE inventory, we assumed that the facility was still in activity. For the start date, we used the start-up date or the date the plant was connected to the grid. We verified that those two starting dates were close to each other for facilities where both were reported. After latitude, longitude, sector, power capacity, and start/stop dates for each facility were added, we only dropped 4.4% of the initial ORE dataset. Most of those discarded plants are located overseas or in Corsica.

2.2.3. Power-weighted weather maps

We generated power capacity-weighted weather maps, by assigning each power facility to the nearest grid cell in the gridded hourly weather data. The weather parameters are thus multiplied by the power capacity weights defined as:

$$w_{i,j}^{t} = \frac{P_{i,j}^{t}}{\sum_{t}\sum_{i,j} P_{i,j}^{t}} \tag{2.1}$$

with the power capacity $P_{i,j}^{t}$ at time t and latitude, longitude i,j in MW. We use a spatiotemporal normalization of the weights to account for the fact that nondispatchable renewable energy sources have seen their available production capacity increase in the last few years (see Figure 2). Because this behavior is expected to carry on, it is important to account for it in the model's input. Figure 3 recaps the weighted weather map creation schematically.

2.2.4. Additional input features

To ensure that models could grasp all of the seasonality and trend, we added two temporal features as it is usually done in the electricity forecasting literature (Chatfield, 1986; Taylor, 2010; Goude et al., 2014). The time step converted to a numerical integer, and the day of the year encoded using a cosine: $doy_{cos} = \cos\left(\frac{2\pi doy_{int}}{365}\right)$, where doy_{int} is the day of the year encoded as an integer between 1 and 365. We used those two temporal features for the wind and solar sectors. However, to be more consistent with the physical process of producing electricity with PV panels, we replaced doy_{cos} for solar by the sunshine duration of the day. This duration was computed from sunrise and sunset times. We did it for every grid cell and timestep.

Even though PV and wind power supply to the grid are related to weather conditions, they are also dependent on the demand that electricity providers need to meet. The last few years have seen negative

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

electricity prices on the market soar as the electrical demand was low, and the available renewable power was in oversupply. This led to a new practice from electricity providers called curtailment, which consists of deliberately restricting the electricity generation from renewable energy sources to prevent negative prices (De Vita et al., 2020; Biber et al., 2022; Yasuda et al., 2022). Thus, we added as input the electricity spot price for France at hourly resolution from ENTSO-E.[5] There are different ways participants trade electricity on the market and therefore different electricity prices. We chose to use the auction day-ahead spot price as it is the only one that can be freely retrieved through ENTSO-E. Auction day-ahead spot price is the price of an MWh^{-1}, which was decided the day before delivery through an auction.

The above-described data processing methodology and workflow allowed us to have input and target datasets for Solar and Wind power, designed for a supervised learning approach, and consisting of a set of (X, Y) observations. X refers to hourly weather maps gridded over France for each selected weather variable, weighted by the power capacity of plants located in the corresponding cells. It also includes day-ahead spot price and temporal features such as the time and day of the year or sunshine duration. Y refers to the corresponding electrical power produced during this hour.

3. Models and calibration

This section describes the models we tested to predict electricity power production from weather variables. It also includes a discussion on model calibration techniques.

3.1. Modeling choices and approaches

As we aimed to develop models able to predict the power production of PV and wind for a day, given the weather conditions, day-ahead price, and temporal features of that same day, we aggregated all input data from hourly to daily resolution. Aggregation also helped to increase the signal-to-noise ratio and prevent overfitting when predicting daily power from hourly data. This leads to a day-to-day prediction approach without utilizing values of the previous days. In operation, real forecasts could then be easily obtained with our model by plugging daily weather forecasts from numerical weather prediction models.

3.1.1. Model architectures

We chose to test three modeling architectures of increasing complexity, as summarized in Figure 4: first using power-weighted weather images averaged over the whole French territory, second applying to power-weighted weather a dimension reduction method, and third applying a vision or image-based technique.

Models using spatially averaged images as input. The first approach is to train models on spatially averaged input data, to have a time series-to-time series regression framework. After averaging, weather time series are combined with price and temporal features series to leverage one-to-one models (models using one input point to predict the corresponding target point). In this family of models, we tested linear regressions, generalized additive models, tree-based models, boosting or artificial neural network, all proven to be capable of reaching state-of-the-art performance (Wood et al., 2014; Gaillard et al., 2016; Krechowicz et al., 2022; Chen et al., 2023; Liu et al., 2023b).

Models using dimensionally reduced input images. The second approach is to use dimension reduction techniques to extract key features from our high-dimensional input power-weighted weather maps before combining them with price and other time features for training a model (Teste et al., 2024). Several dimension reduction methods exist, ranging from empirical orthogonal functions, widely used in the earth sciences community, to autoencoder based on deep network architectures. These methods enable us to reduce the dimension of the input space while providing rich features. In this work, we focused on PCA

[5] Transparency Platfor Transparency Platform, available at https://transparency.entsoe.eu/.

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

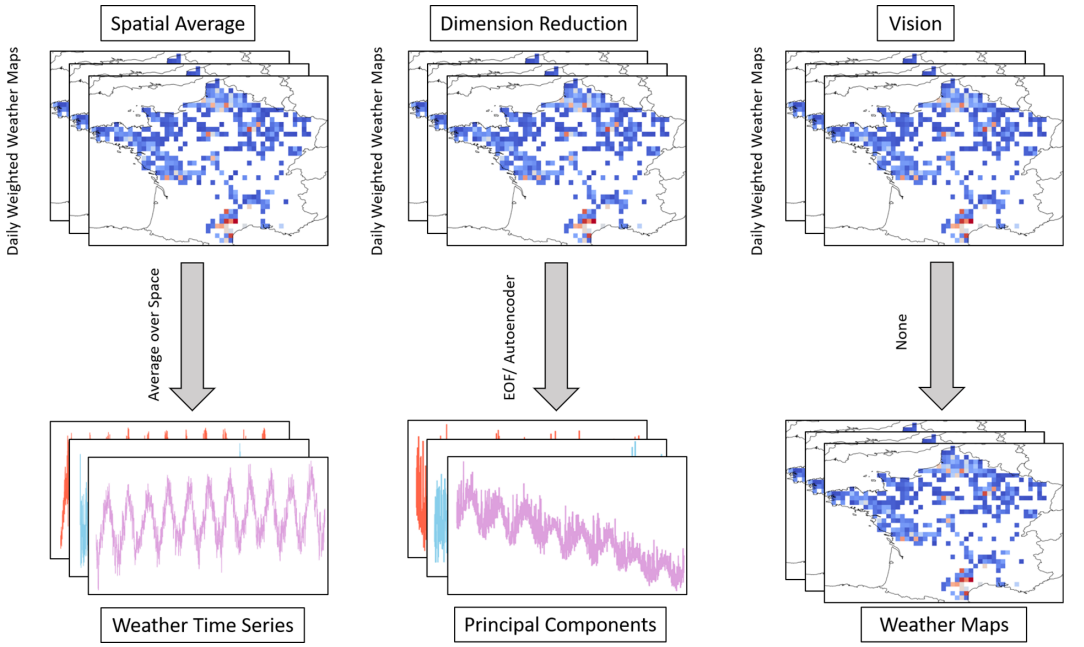

Figure 4. *Representation of the three modeling approaches used in this work to make use of weather maps.*

and optimized the number of principal components as any other model hyperparameter. After obtaining the principal components that behave as time series, we applied the same models as for the spatial average: tree-based models, GAM, and NN.

Models using images as input. The third approach consists of building models capable of directly ingesting the power-weighted weather maps alongside price and temporal features. Here, we used a CNN architecture, previously shown to be capable in image classification, segmentation, or regression tasks, even though they are now slowly being replaced by better performing ViT (Keisler and Naour, 2025).

3.2. Train, validation, and test subsets

We split our dataset into a training and a test subset for the evaluation of model performance. As our data is time-dependent, power production changed throughout the years, mainly due to openings of new facilities. We chose the period from January 1, 2012, to December 31, 2022, to be the train set and January 1, 2023, to December 31, 2023, to be the test set. Nonetheless, hyperparameter tuning is a key step of model development as it often makes the difference between poor and high-performing models. To perform hyperparameter optimization (HPO) we can use different CV methods as well as different optimization frameworks. To ensure the robustness of our model selection procedure, we chose to keep a validation set dedicated to the investigation of cross-validation and optimization methods. This validation set spans the period from January 1, 2022, to December 31, 2022. After choosing a proper model selection and HPO procedure, it is included in the train set for final HPO and model calibration before evaluation on the test set, as described later.

3.3. Cross-validation and HPO

Cross-validation is used to approximate the generalization error, that is, the error of the trained model exposed to new unseen data (Hyndman and Athanasopoulos, 2018). Different techniques are used for

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

splitting the training set into a new training set to train the model and a new left-out test set to evaluate its performance for computing the approximated generalization error. This step is usually combined with HPO to select the best set of hyperparameters for a given model architecture. Selecting the best-suited calibration procedure is a complicated process (Arlot and Celisse, 2009; Bergstra et al., 2011), and we explain later the proposed optimization scheme.

3.3.1. Procedures inspected

Our data are time-dependent because our target is a power supply time series. Different studies investigated which cross-validation procedure was best suited in this case (Tashman, 2000; Bergmeir and Benítez, 2012; Cerqueira et al., 2019). However, the scope of those studies was mainly synthetic and stationary, not to mention small, that is, a few hundred points, time series. Another major limitation is that even though real datasets were used, those modeling approaches involved lagged values of the target time series as predictors, which were excluded in our case. Therefore, we chose to study different cross-validation procedures and HPO algorithms to guide the choices for the calibration of our models. We did these experiments using only the models based on spatial averages of input weather images. The following cross-validation procedures were used:

- **Hold-out**: Split the training set into a train set and a test set.
- **K-fold**: Split the training set into K folds. At each iteration, a fold is chosen to be the test set while the $K-1$ others form the train set. Iterate until all folds were used as test once. After all the iterations, the approximated generalization error is taken to be the average of the error made on each test fold.
- **Expanding**: Split the training set into K folds following the order of the samples. During the i^{th} iteration, the first i folds are used as the train set and the $i+1$ fold is used as the test. Repeat until the entire training set has been used. After all the iterations, the approximated generalization error is taken to be the average of the errors made on each test fold.
- **Sliding**: Split the training set into K folds following the order of the samples. During the ith iteration, the i fold is used as the train set, and the $i+1$ fold is used as the test. Repeat until the entire training set has been used. After all the iterations, the approximated generalization error is taken to be the average of the errors made on each test fold.
- **Blocking**: Choose a block length l based on the temporal structure to conserve most of the correlation between neighboring samples. Split the training set into blocks of length l. Attribute blocks to the train or test set at random (inspired by Wood, 2024).

Figure 5 shows the scheme of these five cross-validation methods. We split the data into a 1-year test set for the Hold-Out method, 10 splits to get yearly folds for every method using folds and blocks of 7 days for the blocking method. The block size was chosen to keep most of the temporal structure using autocorrelation and partial autocorrelation analysis. We also considered the shuffling variants of the K-fold and hold-out methods, which involve randomly shuffling the samples before the folds or subset attributions.

Regarding hyperparameter optimization, we compared two optimization algorithms: Random search and Bayesian search using Gaussian Processes (Bergstra et al., 2011; Bischl et al., 2023).

To assess the impacts of cross-validation and HPO for different model architectures, we repeated the experiments using three models: a random forest, a tree-based boosting scheme (XGBoost), and a feed-forward neural network or MLP. In total, this led to 7 cross-validations × 2 HPO × 3 models estimators of the generalization error. At first glance, one might think that cross-validation procedures that respect the temporal order of the data are best suited to our approach. Still, we wanted to make an informed decision by doing the experiments. Our final goal is to choose the pairs of cross-validation techniques and HPO algorithms that give the "best" estimator of the generalization error. Here, best refers to different criteria ranging from the precision of the generalization error estimate to the computational resource usage.

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

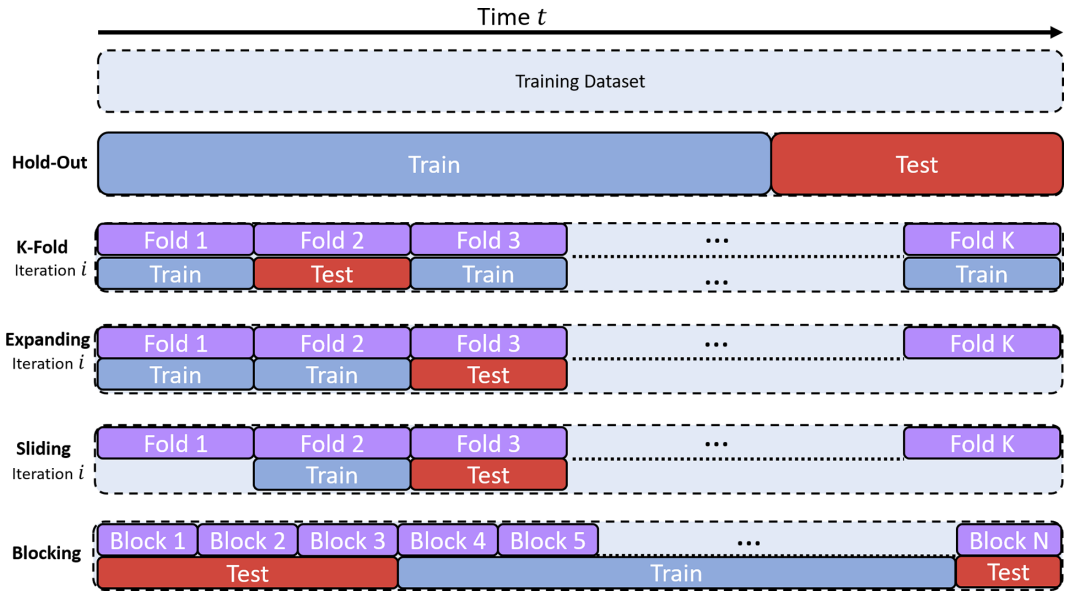

Figure 5. *Different cross-validation procedures considered in this work represented schematically. For Hold-Out and K-Fold, only the method without prior random shuffling is represented.*

3.3.2. Cross-validation experiments

As cross-validation's main goal is to obtain an approximation of the generalization error $\widehat{\varepsilon}$, we monitored how far the estimate was from the real error. To do so, we recorded for each of the 100 optimization iterations the test error made during cross-validation on the training part of the data for a given set of hyperparameters. Then, we compared it to the real generalization error ε made on the validation set. Here, the training and validation part refers to the one visible in Figure 1. Since we are dealing with a regression task, the error ε was taken to be the root-mean-squared error (RMSE) of the modeled and observed daily power production. See Appendix B for metrics definition. Our target being a power production daily time series, the unit of RMSE is MW. Given the real generalization error ε and its estimate $\widehat{\varepsilon}$ from cross-validation, for each procedure, we computed the difference between the two quantities as $\Delta\varepsilon = \varepsilon - \widehat{\varepsilon}$ and analyzed the average $\overline{\Delta\varepsilon}$ and its standard deviation $\sigma(\Delta\varepsilon)$ across the HPs. We also determined the optimum value of $\widehat{\varepsilon}$ reach after optimization and compared it with the real error in $\Delta\varepsilon_{min}$.

During the experiments, we monitored the time taken to perform one iteration and the permutation feature importance of each feature obtained during cross-validation compared to the one obtained on the validation set. These times of computation tell us how costly each error estimation method was. The feature importance tells us if the cross-validation technique impacted the interpretability of the model. Last, we experimented with different dataset sizes to inspect the influence of data size on cross-validation methods since the literature only deals with small sample sizes. As the dataset size increases, older and older data are utilized for training. Computation times can be found in Table 1 and results for random forest on solar are presented in Figures 6 and 7. Results for other models on solar are in Appendix C and on wind in Appendix D, Figures D1–D6. Results about permutation feature importance showed that despite the different cross-validation methods, the ranking of the features stayed the same for the different hyperparameter combinations explored, meaning that the method does not impact the model interpretability.

On the radar chart of Figure 6, we can see that $\Delta\varepsilon$ is positive on average and for the optimum. This means that our generalization error estimates $\widehat{\varepsilon}$ is lower than the real error ε. In other words, the cross-validation tends to overestimate the model performance leading to overconfidence in the model. We can also see that methods that do not preserve the chronological order or shuffling perform worse than those

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

Table 1. *Average and standard deviation of computing times for 1 iteration for each cross-validation method in seconds*

Model	Sector	Hold-Out	Hold-Out (S)	K-Fold	K-Fold (S)	Expanding	Sliding	Blocking
Forest	Solar	2.3 ± 1.6	2.1 ± 1.8	22.7 ± 11.8	14.1 ± 14	15.4 ± 9.3	19.8 ± 1.5	2.3 ± 1.7
Boosting	Solar	3.0 ± 3.9	3.9 ± 4.5	21.2 ± 17.8	53.2 ± 41.6	33.4 ± 27.7	1.7 ± 1.4	4.6 ± 5.7
Neural network	Solar	2.7 ± 2.3	2.8 ± 2.4	27.6 ± 24.1	27.8 ± 23.9	16.3 ± 14.0	4.3 ± 3.3	2.7 ± 2.3
Forest	Wind	2.4 ± 2.4	3.8 ± 2.5	19.6 ± 19.1	37.0 ± 23.2	9.9 ± 10.9	1.9 ± 1.0	1.8 ± 1.8
Boosting	Wind	3.7 ± 2.6	4.5 ± 3.9	61.2 ± 78.7	122.6 ± 92.2	78.4 ± 75.5	6.9 ± 3.9	6.2 ± 3.7
Neural network	Wind	2.8 ± 2.4	2.8 ± 2.4	28.1 ± 24.4	57.5 ± 53.3	33.0 ± 30.0	8.7 ± 7.4	2.8 ± 2.4

Note. The (S) indicates the shuffling variant of the method. Medals indicate the top three fastest methods for each model and dataset.

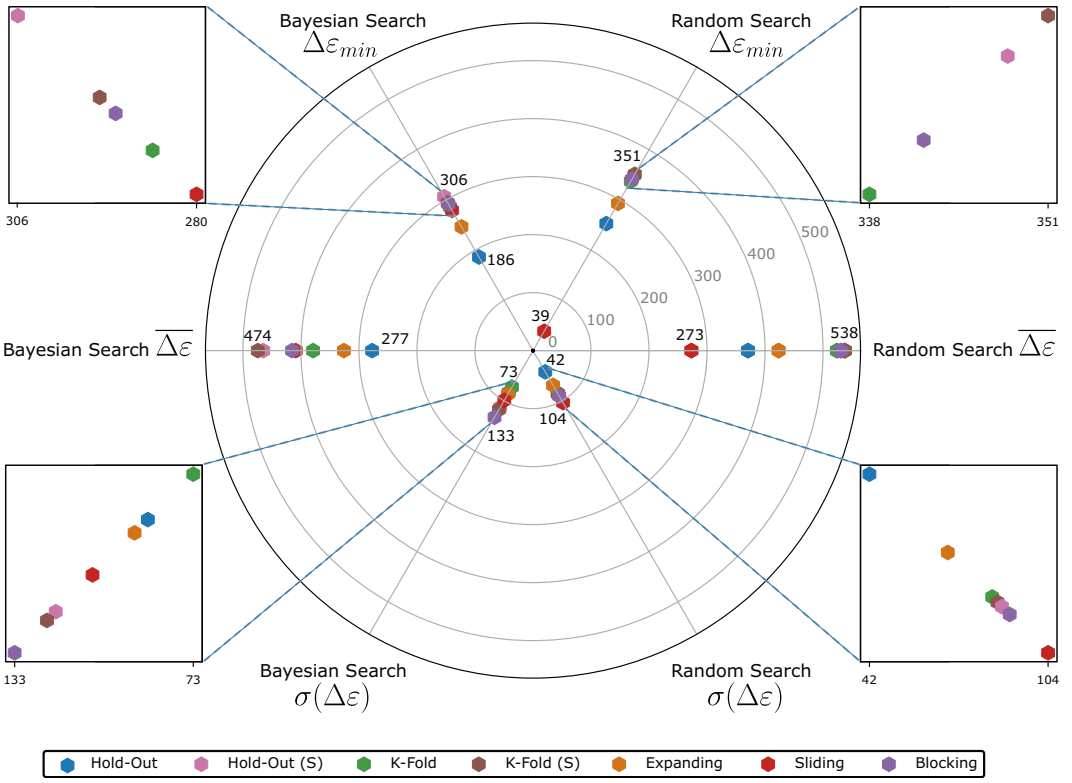

Figure 6. *Results of different cross-validation techniques for random forest on solar. Each axis represents a monitored quantity for a given HPO optimization procedure. The values for each method are plotted as points, and only the worst and best values for each axis are printed. The (S) indicates the shuffling variant of the method.*

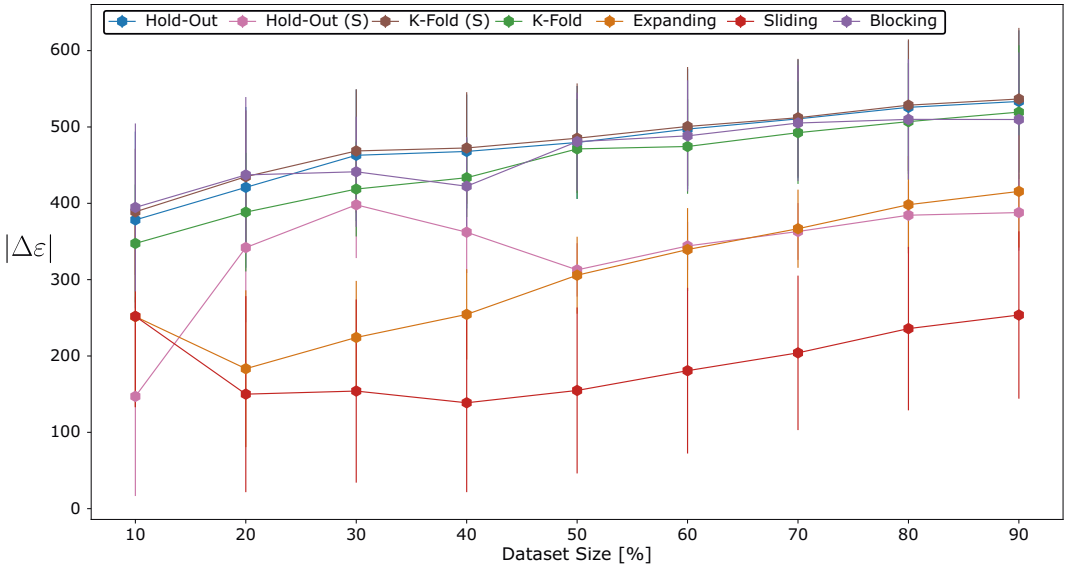

Figure 7. *Robustness of cross-validation procedure regarding the dataset size for random forest on solar. The marker indicates the average |Δε|, while the error bars display the standard deviation. The (S) indicates the shuffling variant of the method.*

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

that do. Specifically, hold-out, expanding, and sliding lead to the closest estimate on average and the optimum for both searches. However, sliding is the most sensitive to the set of hyperparameters as its variability $\sigma(\Delta\varepsilon)$ is the highest. This might stem from its small training set size which never exceeds 1 year of data. This is also confirmed by the error bars of Figure 7. This same figure shows that increasing the dataset size by appending older and older data leads to a slight increase in $|\Delta\varepsilon|$ meaning that our generalization error estimate is moving away from the real one. This is because older data such as 2012 carry less meaningful information than more recent data such as 2020 for predicting the validation set which is the year 2022. This behavior also explains why some methods display an inflection point for a certain dataset size meaning that there is an optimum past period of time to consider to make better predictions on the validation set.

The same conclusions hold for boosting and feed-forward neural networks on the solar dataset (see Figures C1–C4). It is worth mentioning that the neural network shows a high variability and a high $\Delta\varepsilon$ for the Bayesian search HPO, suggesting that this algorithm might not be the best for optimizing neural network hyperparameters. For the Wind dataset (see Appendix D, Figures D1–D6), hold-out, sliding, and expanding methods are the best methods to estimate the generalization error for all three model architectures. Yet, we can see for the random forest and boosting models that increasing the dataset size with older data does help better approximate the generalization error with the expanding and sliding methods. This means that in the wind dataset, older data still carry meaningful information for predicting the most recent validation set, even if there is a pronounced annual trend in the wind power production time series (see Figure 2).

Finally, Table 1 shows that cross-validation procedures involving folds are more computationally intensive per iteration, as one can expect. Combined with the previous graphs we can conclude that the longer computing times arising from the use of K-fold methods are not worth it since hold-out and sliding are better performers and between 5 and 10 times faster to compute per iteration.

From the result of those experiments testing different cross-validations, with different HPO and different model architectures we were able to make recommendations on how to choose a model selection procedure when dealing with time series to time series forecasting from covariates. We found that dedicated procedures that keep the chronological order during cross-validation perform better than standard K-fold or shuffled hold-out. Depending on the model architecture and the underlying data, some techniques tend to overestimate or underestimate model performance leading to underconfidence or overconfidence in our model. This systematic work could be extended to deep learning models that directly ingest images as inputs, to also get recommendations to push their performance even further.

4. Benchmark results and discussion

In this section, we present the results of our calibrated models on the training + validation set and evaluated on the test set. The best hyperparameters for each model were selected from the best generalization error, based on experiments from the previous section, that is, using Bayesian search with either an expanding or hold-out cross-validation method, depending on the model complexity, to save computing time. Expanding was preferred over sliding cross-validation due to the high sensitivity of sliding to hyperparameter sets. We assessed the performance of the model using the RMSE, mean absolute error (MAE), mean absolute percentage error (MAPE), normalized root-mean-squared error (nRMSE), and R^2 score (R^2). The definitions of these metrics are given in Appendix B. Table 2 contains all our results on the solar dataset, while results for wind can be found in Appendix E, Table E1.

As nondispatchable renewables capacity increased throughout our study period, solar and wind power production time series have an increasing trend from 2012 to 2023 as highlighted by Figure 2. This trend requires the models to be able to extrapolate on the test set. Despite reaching state-of-the-art performance in many tasks, tree-based models such as random forest and boosting are known to face difficulties when it comes to extrapolation outside of the training domain (Hengl et al., 2018; Malistov and Trushin, 2019). Our case makes no exception, despite low errors on the train set, random forest, and boosting models errors soared on the test set (see Tables 2 and E1). To address this issue, many research works propose

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

Table 2. *Benchmark results for different models using three different modeling approaches on the solar dataset*

Metrics			MAE		MAPE (%)		RMSE		nRMSE (%)		R^2	
Approach	Model	Detrend	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Average	Linear Regression		106	350	14.0	15.8	140	423	3.59	9.57	0.96	0.86
Average	Random Forest		50.7	300	6.12	13.9	69.3	375	1.78	8.5	0.99	0.89
Average	Random Forest	✓	57.4	179	6.95	9.83	82.3	279	2.12	6.33	0.99	0.94
Average	Linear Forest		78.0	300	9.54	13.9	109	374	2.80	8.45	0.98	0.89
Average	Tree Boosting		47.3	253	6.26	13.9	63.2	331	1.63	7.48	0.99	0.91
Average	Tree Boosting	✓	56.9	176	7.10	9.71	80.7	271	2.07	6.14	0.99	0.94
Average	Linear Tree Boosting		106	352	14.0	15.8	140	425	3.59	9.62	0.96	0.86
Average	GAM		82.0	321	10.3	16.0	113	401	2.91	9.10	0.98	0.87
Average	MLP		123	229	16.4	11.8	164	310	4.21	7.01	0.95	0.93
PCA	Linear Regression		100	310	9.09	12.7	135	394	3.47	8.92	0.97	0.88
PCA	Random Forest		62.1	349	7.49	16.9	86.9	436	2.23	9.86	0.99	0.85
PCA	Linear Forest		79.1	282	9.84	13.3	109	364	2.79	8.23	0.98	0.90
PCA	Tree Boosting		53.7	268	7.43	14.4	70.3	381	1.81	8.63	0.99	0.89
PCA	Linear Tree Boosting		96.9	319	12.3	14.5	133	403	3.40	9.12	0.97	0.87
PCA	GAM		83.3	434	10.9	20.4	112	501	2.87	11.3	0.98	0.80
PCA	MLP		85.6	195	9.52	10.7	129	294	3.33	6.65	0.97	0.93
Vision	CNN		147	182	15.2	10.1	200	277	5.1	6.30	0.93	0.94

Note. Medals indicate the top three best-performing models on the test set for each metric.

alternatives such as stochastic or linear trees (Gama and Brazdil, 1999; Zhang et al., 2019; Numata and Tanaka, 2020; Ilic et al., 2021; Raymaekers et al., 2024). We chose to apply two different methods to try to solve this extrapolation problem: linear trees and detrending of the time series.

Our detrending scheme consisted of applying a trend estimation method, such as seasonal trend decomposition using loess, on the entire dataset. Once the trend is estimated, we remove it from the data. The transformed data were thus passed to the model for calibration. The predictions were obtained by reconstruction from the model's output and trend estimate. The detrending was done on both weather input and power output data, as the weighting scheme introduced trends in the covariates.

Linear trees did not seem to be a silver bullet on the solar dataset as their performance was only marginally better for the forest and worse in the case of boosting. In contrast, for the wind dataset, they prove to be useful in enhancing the extrapolation performance. However, their performance was still far from the tree-based models predicting detrended power supply from detrended weather averages before reconstructing the proper production time series. Despite the error induced by the trend estimation and reconstruction step, this method displays some of the best results on both solar and wind within the spatial average method and even outside. Such behavior could be expected because the trend is estimated on the whole dataset. The extrapolation problem is weaker for GAM and MLP as they manage to better grasp the trend, achieving better performance on the test set.

Compared with the spatial input averaging approach, using tree-based models with PCA did not achieve better performance due to the extracted principal components exhibiting the same trend as the spatial averages. This time, we only applied linear trees, as detrending principal components was more challenging. They exhibited a small improvement on the solar dataset but a bigger decrease in performance when used to predict wind power supply. Combining PCA with GAM does not seem to improve performance on both datasets. For MLP, it depends on the sector, but one thing that we noticed after our calibration is that networks combined with PCA are deeper than networks without it, meaning that it requires more layers to extract meaningful information from the principal components.

Although the increase in complexity between dimension reduction and spatial average approach did not lead to clear improvements in model performance for every model architecture, leveraging the entire weather maps with a more complex computer vision architecture, such as a CNN clearly did. This phenomenom stems from the unsupervised nature of the PCA compared to the supervised CNN. In fact, the CNN is the best-performing model on the wind dataset and the second-best on the solar dataset. By utilizing our spatiotemporal weighting scheme, the CNN has a better grasp of the trends in renewables implementation, as highlighted in Figure 8, and avoids extrapolation difficulties. Combined with the MLP results, it highlights the versatility and suitability of neural network-based models for predicting power production from renewable sources.

Tables 2 and E1 illustrate the challenges that tree-based models face with extrapolation. Without the detrending scheme, these models would not rank among the top three performers. Instead, neural networks would dominate the podium, with the rankings reflecting the increasing complexity of the modeling approaches. Specifically, as models incorporate more spatially explicit data, their performance improves, with vision models outperforming MLPs combined with PCA, which in turn surpass MLPs on time series. Therefore, we recommend that practitioners incorporate spatial information when designing forecasting models.

The work conducted on cross-validation procedures and HPO schemes allowed us to push state-of-the-art machine learning architecture to their best performance. However, such a study could be extended to include deep learning models such as CNN to improve their performance. As deep convolutional neural networks are already amongst the best models for both solar and wind, we did not pursue this path. However, it is worth mentioning that a systematic study would benefit deep learning models and strengthen their edge.

5. Conclusion

This study presented datasets and tested a modeling framework based on machine learning and climate as well as facility locations as an input for predicting daily solar and wind supply at the country level in

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

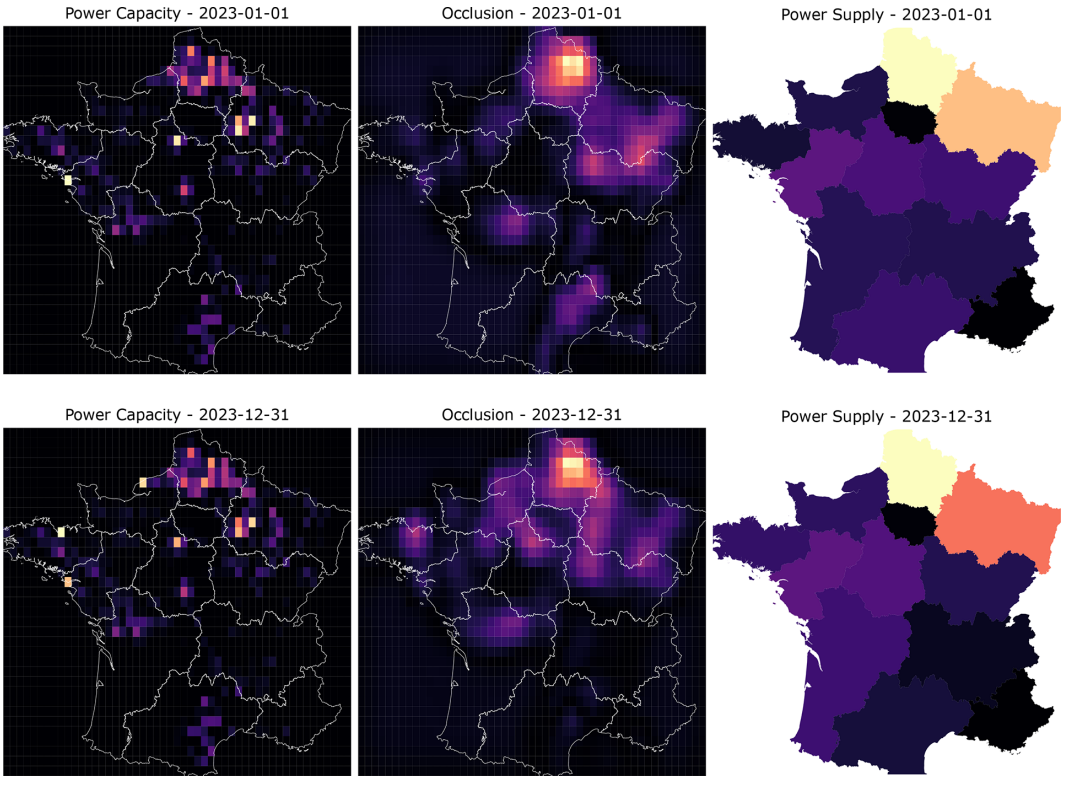

Figure 8. *Power capacity, occlusion attribution, and regional realized power supply for early and late 2023 for Wind. Occlusion is an interpretation method that hides part of the input and sees how it impacts the CNN prediction. The higher the impact is, the higher the hidden part's importance (Zeiler and Fergus, 2014). Power supply data are obtained from RTE for all of France's regions (NUTS1).*

France. Several different machine learning models with different complexities were applied to create a benchmark. Attention was paid to the methods used for calibrating the model to avoid displaying overconfident metrics. The method proposed was applied over France and could be extended to any other country or region.

Our model calibration experiments showed that there is no "silver bullet" model, as it is dependent on the data and the model at hand. Under- or overconfidence can arise depending on the calibration, leading to desillusions if the model is chosen to be run in operations based on the calibration results. Thus, a thorough validation procedure and analysis are required to avoid such phenomena and improve the production launch. Still, some general recommendations can be made towards preferring cross-validation methods, keeping the temporal structure of the data intact, as they are both more computationally efficient and less biased, leading to more robust models.

Trying to model renewable power supply from weather inputs without including the power capacity at facility locations in the inputs is pointless, as some state-of-the-art already failed to correctly model the trend with this added information. Models that are able to ingest the entire high-dimensional weather input can learn from spatial patterns to achieve better predictions, improving the forecasts. This means that being spatially explicit in both the data curation and preparation, as well as in the modelling process, is key to achieving good predictions. Therefore, we encourage other practitioners to include geospatial data in their framework. However, one must bear in mind that power capacity inventories are not available everywhere and can be of different quality depending on the data source.

In summary, geospatial weather information is key for renewable energy forecasting. By providing an open dataset and benchmark, we hope to foster research and improve comparison between studies.

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.10021.

Author contribution. Conceptualization: E.L., Y.G., P.C.; Methodology: E.L., Y.G.; Data curation: E.L. Data visualization: E.L. Supervision: Y.G, P.C. Writing original draft: E.L.; Writing review & editing: E.L., Y.G., P.C. All authors approved the final submitted draft.

Competing interests. The authors declare none.

Data availability statement. The datasets built for this work can be accessed https://doi.org/10.5281/zenodo.14287949.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This research was supported by a grant from the Association Nationale de la Recherche et de la Technologie (ANRT) No. 2024/0010.

References

Abdul Baseer M, **Almunif A**, **Alsaduni I and Tazeen N** (2023) Electrical power generation forecasting from renewable energy systems using artificial intelligence techniques. *Energies 16*, 6414.

Ahmad SK and Hossain F (2020) Maximizing energy production from hydropower dams using short-term weather forecasts. *Renewable Energy 146*, 1560–1577.

Alcañiz A, **Lindfors AV**, **Zeman M**, **Ziar H and Isabella O** (2023). Effect of climate on photovoltaic yield prediction using machine learning models. *Global Challenges*, *7*(1):2200166.

Arlot S and Celisse A (2009) A survey of cross validation procedures for model selection. *Statistics Surveys 4*, 40–79.

Bellinguer K, **Mahler V**, **Camal S and Kariniotakis, G.** (2020). Probabilistic Forecasting of Regional Wind Power Generation for the EEM20 Competition: a Physics-oriented Machine Learning Approach. In *17th European Energy Market Conference, EEM 2020*. Stockholm, Sweden: KTH, IEEE.

Bergmeir C and Benítez JM (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*: 192–213.

Bergstra J, **Bardenet R**, **Bengio Y and Kegl B** (2011) Algorithms for hyper-parameter optimization. *Neural Information Processing Systems Proceedings (NeurIPS) 24*, 1–9.

Biber A, **Felder M**, **Wieland C and Spliethoff H** (2022). Negative price spiral caused by renewables ? Electricity price prediction on the german market for 2030. *The Electricity Journal*, *35*(8):107188.

Bilendo F, **Meyer A**, **Badihi H**, **Lu N**, **Cambron P and Jiang B** (2023) Applications and modeling techniques of wind turbine power curve for wind farms: A review. *Energies 16*(1), 180.

Bischl B, **Binder M**, **Lang M**, **Pielok T**, **Richter J**, **Coors S**, **Thomas J**, **Ullmann T**, **Becker M**, **Boulesteix A**, **Deng D and Lindauer M** (2023) Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery 13*(2), e1484.

British Petroleum (BP) (2024). Energy Outlook 2024: Exploring the key trends and uncertainties surrounding the energy transition. Available at https://www.bp.com/en/global/corporate/energy-economics/energy-outlook.html (accessed 26 August 2024).

Castillo-Rojas W, **Medina Quispe F and Hernández C** (2023) Photovoltaic energy forecast using weather data through a hybrid model of recurrent and shallow neural networks. *Energies 16*(13), 5093.

Cerqueira V, **Torgo L and Mozetic I** (2019) Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning 109*, 1997–2028.

Chatfield C (1986) Comparative models for electrical load forecasting. *Journal of the Royal Statistical Society. Series A (General) 149*(3), 272–272.

Chen Q and Folly K (2018) Wind power forecasting. *IFAC-PapersOnLine 51*(28), 414–419.

Chen G, **Hu Q**, **Wang J**, **Wang X and Zhu Y** (2023) Machine-learning-based electric power forecasting. *Sustainability 15*(14).

Condemi C, **Casillas-Pérez D**, **Mastroeni L**, **Jiménez-Fernández S and Salcedo-Sanz S** (2021) Hydro-power production capacity prediction based on machine learning regression techniques. *Knowledge-Based Systems*, *222*:107012.

Couto A and Estanqueiro A (2022) Enhancing wind power forecast accuracy using the weather research and forecasting numerical model-based features and artificial neuronal networks. *Renewable Energy 201*, 1076–1085.

De Giorgi MG, **Congedo PM and Malvoni M** (2014) Photovoltaic power forecasting using statistical methods: impact of weather data. *IET Science, Measurement & Technology 8*(3), 90–97.

De Vita A, **Capros P**, **Evangelopoulou S**, **Kannavou M**, **Siskos P**, **Zazias G**, **Boeve S**, **Bons M**, **Winkel R**, **Cilhar J**, **De Vos L**, **Leemput N and Mandatova P** (2020) Sectoral integration: Long-term perspective in the EU energy system, European Commission Directorate-General for Energy, E3 Modelling, Ecofys, Tractebel, Publications Office. https://data.europa.eu/doi/10.2833/347937.

Dolara A, **Leva S and Manzolini G** (2015) Comparison of different physical models for pv power output prediction. *Solar Energy 119*, 83–99.

Elsaraiti M and Merabet A (2022) Solar power forecasting using deep learning techniques. *IEEE Access 10*, 31692–31698.

Engeland K, **Borga M**, **Creutin J-D**, **François B**, **Ramos M-H and Vidal J-P** (2017) Space-time variability of climate variables and intermittent renewable electricity production – a review. *Renewable and Sustainable Energy Reviews 79*, 600–617.

European Commission (2019) The European Green Deal: Striving to be the first climate-neutral continent. Available at https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en (accessed 26 August 2024).

Gaillard P, **Goude Y and Nedellec R** (2016) Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting 32*(3), 1038–1050.

Gama J and Brazdil P (1999) Linear tree. *Intelligent Data Analysis 3*(1), 1–22.

Gijón A, **Pujana-Goitia A**, **Perea E**, **Molina-Solana M and Gómez-Romero J** (2023). Prediction of wind turbines power with physics-informed neural networks and evidential uncertainty quantification, arXiv 2307.14675, https://arxiv.org/abs/2307.14675.

Goude Y, **Nedellec R and Kong N** (2014) Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE Transactions on Smart Grid 5*(1), 440–446.

Hengl T, **Nussbaum M**, **Wright M**, **Heuvelink G and Graeler B** (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, https://doi.org/10.7287/peerj.preprints.26693v2.

Hersbach H, **Bell B**, **Berrisford P**, **Hirahara S**, **Horányi A**, **Muñoz-Sabater J**, **Nicolas J**, **Peubey C**, **Radu R**, **Schepers D**, **Simmons A**, **Soci C**, **Abdalla S**, **Abellan X**, **Balsamo G**, **Bechtold P**, **Biavati G**, **Bidlot J**, **Bonavita M**, **De Chiara G**, **Dahlgren P**, **Dee D**, **Diamantakis M**, **Dragani R**, **Flemming J**, **Forbes R**, **Fuentes M**, **Geer A**, **Haimberger L**, **Healy S**, **Hogan RJ**, **Hólm E**, **Janisková M**, **Keeley S**, **Laloyaux P**, **Lopez P**, **Lupu C**, **Radnoti G**, **De Rosnay P**, **Rozum I**, **Vamborg F**, **Villaume S and Thépaut J** (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049.

Hyndman R and Athanasopoulos G (2018) *Forecasting: Principles and Practice*. 2nd Edn. Australia: OTexts.

Iheanetu KJ (2022) Solar photovoltaic power forecasting: A review. *Sustainability 14*(24), 17005.

Ilic I, **Görgülü B**, **Cevik M and Baydoğan MG** (2021) Explainable boosted linear regression for time series forecasting. *Pattern Recognition*, *120*:108144.

International Renewable Energy Agency (IRENA) (2020a) Advanced forecasting of variable renewable power generation: Innovation landscape brief. Available at https://www.researchgate.net/profile/Kien-Vu-6/post/Which_techniques_can_be_used_for_renewable_energy_prediction/attachment/6078ac650f39c7000141ebc3/AS%3A1012965440487428%40161852118 9613/download/IRENA_Advanced_weather_forecasting_2020.pdf (accessed 26 August 2024).

International Renewable Energy Agency (IRENA) (2020b). Renewable Energy Prospects for Central and South-Eastern Europe Energy Connectivity (CESEC). Available at https://www.irena.org/Publications/2020/Oct/Renewable-Energy-Prospects-for-Central-and-South-Eastern-Europe-Energy-Connectivity-CESEC (accessed 26 August 2024).

Keisler J and Naour EL (2025) WindDragon: Automated Deep Learning for Regional Wind Power Forecasting, *Environmental Data Science 4*, e19, https://doi.org/10.1017/eds.2025.10.

Kim S-G, **Jung J-Y and Sim M** (2019) A two-step approach to solar power generation prediction based on weather data using machine learning. *Sustainability 11*(5), 1501.

Kim J, **Kim D**, **Yoo W**, **Lee J and Kim YB** (2017) Daily prediction of solar power generation based on weather forecast information in korea. *IET Renewable Power Generation 11*(10), 1268–1273.

Kraskov A, **Stögbauer H and Grassberger P** (2004) Estimating mutual information. *Physical Review E*, *69*:066138.

Krechowicz A, **Krechowicz M and Poczeta K** (2022) Machine learning approaches to predict electricity production from renewable energy sources. *Energies 15*(23), 9146.

Lim S-C, **Huh J-H**, **Hong S-H**, **Park C-Y and Kim J-C** (2022) Solar power forecasting using CNN-LSTM hybrid model. *Energies 15*(21), 8233.

Liu H and Chen C (2019) Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Applied Energy 249*, 392–408.

Liu L, **He G**, **Wu M**, **Liu G**, **Zhang H**, **Chen Y**, **Shen J and Li S** (2023a) Climate change impacts on planned supply–demand match in global wind and solar energy systems. *Nature Energy 8*(8), 870–880.

Liu L, **He G**, **Wu M**, **Liu G**, **Zhang H**, **Chen Y**, **Shen J and Li S** (2023b) Climate change impacts on planned supply–demand match in global wind and solar energy systems. *Nature Energy 8*, 1–11.

López Gómez J, **Ogando Martínez A**, **Troncoso Pastoriza F**, **Febrero Garrido L**, **Granada Álvarez E and Orosa García JA** (2020) Photovoltaic power prediction using artificial neural networks and numerical weather data. *Sustainability 12*(24), 10295.

Malistov A and Trushin A (2019) Gradient boosted trees with extrapolation. *18th IEEE International Conference On Machine Learning And Applications (ICMLA) 2019*, 783–789.

Malvoni M, **De Giorgi M and Congedo P** (2016) Data on photovoltaic power forecasting models for mediterranean climate. *Data in Brief 7*, 1639–1642.

Malvoni M, **De Giorgi MG and Congedo PM** (2017) Forecasting of PV power generation using weather input data-preprocessing techniques. *Energy Procedia 126*, 651–658.

Mayer MJ and Gróf G (2021) Extensive comparison of physical models for photovoltaic power forecasting. *Applied Energy*, *283*: 116239.

Ministère de la Transition Ecologique (2019) Programmations pluriannuelles de l'énergie (PPE). Available at https://www.ecologie.gouv.fr/politiques-publiques/programmations-pluriannuelles-lenergie-ppe (accessed 26 August 2024).

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

Ministère de la Transition Ecologique (2020) Stratégie nationale bas-carbone (SNBC). Available at https://www.ecologie.gouv.fr/politiques-publiques/strategie-nationale-bas-carbone-snbc (accessed 26 August 2024).

Numata K and Tanaka K (2020) *Stochastic threshold model trees: A tree-based ensemble method for dealing with extrapolation*. ArXiv, arXiv:2009.09171

Raymaekers J, **Rousseeuw P**, **Verdonck T and Yao R** (2024) Fast linear model trees by pilot. *Machine Learning 113*, 1–50.

Ritchie H and Rosado P (2020) Electricity mix. *Our World in Data*. Available at https://ourworldindata.org/electricity-mix (accessed 26 August 2024).

Ryu J-Y, **Lee B**, **Park S**, **Hwang S**, **Park H**, **Lee C and Kwon D** (2022) Evaluation of weather information for short-term wind power forecasting with various types of models. *Energies 15*(24), 9403.

Sharma, N., **Sharma, P.**, **Irwin, D., and Shenoy, P.** (2011). Predicting solar generation from weather forecasts using machine learning. In *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, IEEE. pp. 528–533.

Sweeney C, **Bessa RJ**, **Browell J and Pinson P** (2020) The future of forecasting for renewable energy. *WIREs Energy and Environment 9*(2), 365.

Tashman LJ (2000) Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting 16*(4), 437–450 The M3- Competition.

Taylor JW (2010) Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research 204*(1), 139–152.

Teste F, **Makowski D**, **Bazzi H and Ciais P** (2024) Early forecasting of corn yield and price variations using satellite vegetation products. *Computers and Electronics in Agriculture*, *221*:108962.

Tsai W-C, **Hong C-M**, **Tu C-S**, **Lin W-M and Chen C-H** (2023) A review of modern wind power generation forecasting technologies. *Sustainability 15*(14), 10757.

United Nations Convention on Climate Change (2015) Paris Agreement: Climate Change Conference (COP21). Available at https://unfccc.int/documents/184656 (accessed 26 August 2024).

Vladislavleva K, **Friedrich T**, **Neumann F and Wagner M** (2013) Predicting the energy output of wind farms based on weather data: Important variables and their correlation, *Renewable Energy*, *50*, 236–243.

Wang H, **Lei Z**, **Zhang X**, **Zhou B and Peng J** (2019a) A review of deep learning for renewable energy forecasting. *Energy Conversion and Management*, *198*:111799.

Wang J, **Zhong H**, **Lai X**, **Xia Q**, **Wang Y and Kang C** (2019b) Exploring key weather factors from analytical modeling toward improved solar power forecasting. *IEEE Transactions on Smart Grid 10*(2), 1417–1427.

Wood SN (2024) On neighbourhood cross validation, ArXiv, https://arxiv.org/abs/2404.16490.

Wood SN, **Goude Y and Shaw S** (2014) Generalized Additive Models for Large Data Sets. *Journal of the Royal Statistical Society Series C: Applied Statistics 64*(1), 139–155.

Yasuda Y, **Bird L**, **Carlini EM**, **Eriksen PB**, **Estanqueiro A**, **Flynn D**, **Fraile D**, **Gómez Lázaro E**, **Martín-Martínez S**, **Hayashi D**, **Holttinen H**, **Lew D**, **McCam J**, **Menemenlis N**, **Miranda R**, **Orths A**, **Smith JC**, **Taibi E and Vrana TK** (2022) C-e (curtailment – energy share) map: An objective and quantitative measure to evaluate wind and solar curtailment. *Renewable and Sustainable Energy Reviews*, *160*:112212.

Zeiler MD and Fergus R (2014) Visualizing and understanding convolutional networks, *Computer Vision: ECCV 2014, ECCV 2014, Lecture Notes in Computer Science*, *vol 8689*, 818–833.

Zhang H, **Nettleton D and Zhu Z** (2019) Regression-enhanced random forests, arXiv:1904.10416.

Zhong Y-J and Wu Y-K (2020) Short-term solar power forecasts considering various weather variables. In 2020 *International Symposium on Computer, Consumer and Control (IS3C)*. IEEE, pp. 432–435.

Zhou H, **Qiu Y**, **Feng Y and Liu J** (2022) Power prediction of wind turbine in the wake using hybrid physical process and machine learning models. *Renewable Energy 198*, 568–586.

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

A. Appendix A: Weather variables

Table A1. *Description of climate variables*

Variable full name	Variable abbreviation	Unit	Description	Sector
2-m temperature	t2m	K	Temperature of air at 2 m above the surface	Solar, Wind
Surface solar radiation downward	ssrd	Jm^{-2}	Amount of solar radiation (direct and diffuse) reaching a horizontal plane at the surface of the Earth	Solar
10-m U wind component	u10	ms^{-1}	Northward component of the wind speed at 10m	Wind
10-m V wind component	v10	ms^{-1}	Eastward component of the wind speed at 10m	Wind
100-m U wind component	u100	ms^{-1}	Northward component of the wind speed at 100m	Wind
100-m V wind component	v100	ms^{-1}	Eastward component of the wind speed at 100m	Wind
Instantaneous 10-m wind gust	i10fg	ms^{-1}	Maximum wind gust speed at 10m	Solar, Wind
Total precipitation	tp	m	Accumulated liquid and frozen water that falls to the Earth's surface	Wind
Evaporation	e	m	Accumulated amount of water that has evaporated from the Earth's surface	Solar
Runoff	ro	m	Water from rainfall, snow melt or deep soil that drains away over the surface or under the ground	Wind

Source: ERA5 Documentation (https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation).
Note. There are 110,808 hourly weather observations spanning 4383 days with a 35 × 51 grid for each time step.

B. Appendix B: Metrics Definition

$$\mathrm{MAE} = \frac{1}{N}\sum_{i}^{N}|y_i - \widehat{y}_i| \tag{B.1}$$

$$\mathrm{MAPE} = 100 \times \frac{1}{N}\sum_{i}^{N}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{B.2}$$

$$\mathrm{nRMSE} = 100 \times \frac{\sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - \widehat{y}_i)^2}}{y_{max} - y_{min}} \tag{B.3}$$

$$\mathrm{R2} = 1 - \frac{\sum_{i}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i}^{N}(y_i - \bar{y})^2} \tag{B.4}$$

where y_{max}, y_{min}, and $\bar{y}$ are the maximum, minimum, and the average of the true target y, respectively.

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

C. Appendix C: Cross-validation experiment results for solar

C.1. Boosting

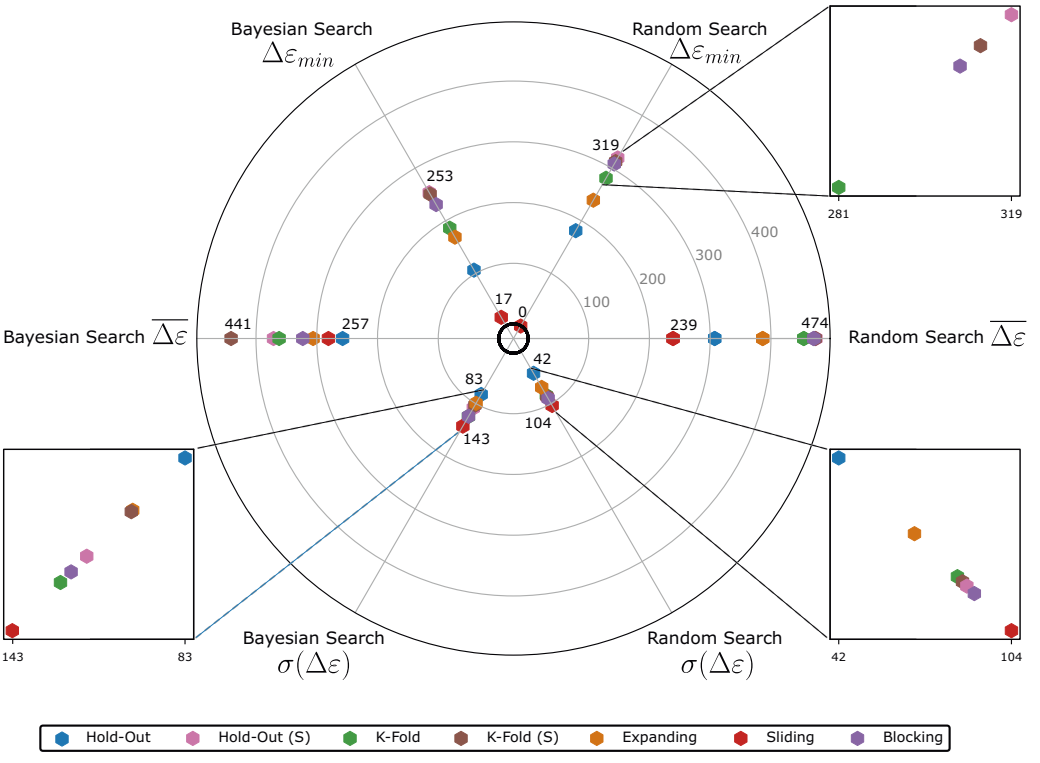

Figure C1. *Results of different cross-validation techniques for boosted trees on solar. Only the worst and best values for each axis are printed. The (S) indicates the shuffling variant of the method.*

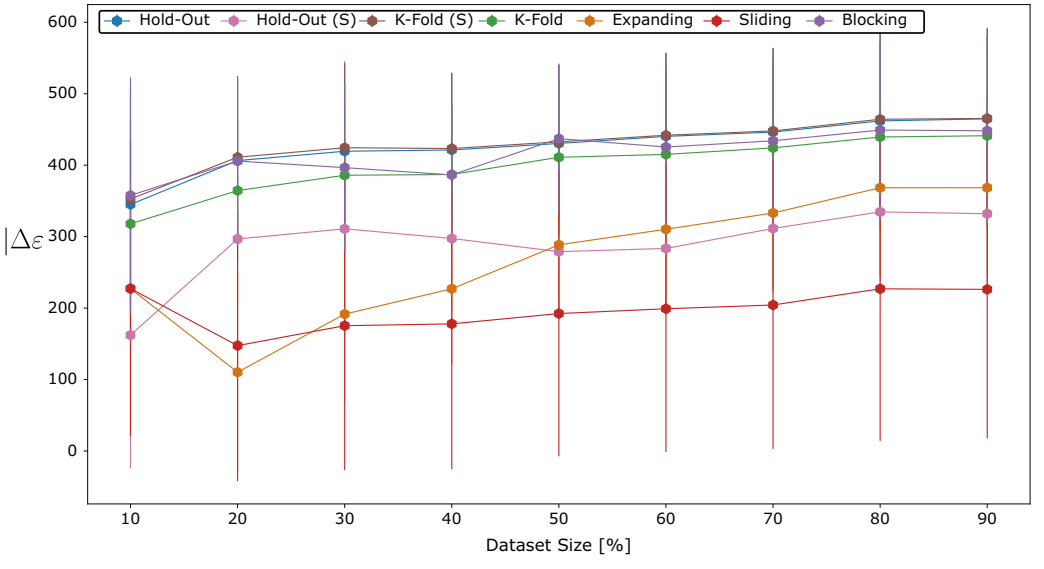

Figure C2. *Robustness of cross-validation procedure regarding dataset size for boosted tress on Solar. The marker indicates the average $|\Delta\varepsilon|$, while the error bars display the standard deviation. The (S) indicates the shuffling variant of the method.*

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

C.2. Feed-forward neural network (MLP)

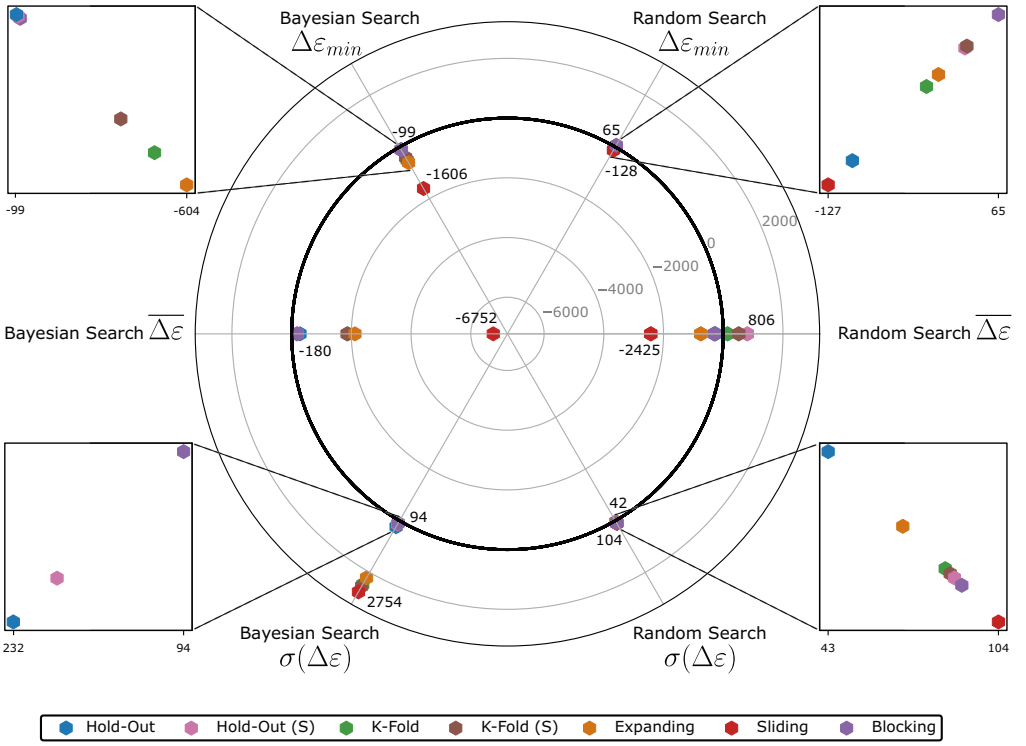

Figure C3. *Results of different cross-validation techniques for feed-forward neural network on solar. Only the worst and best values for each axis are printed. The (S) indicates the shuffling variant of the method.*

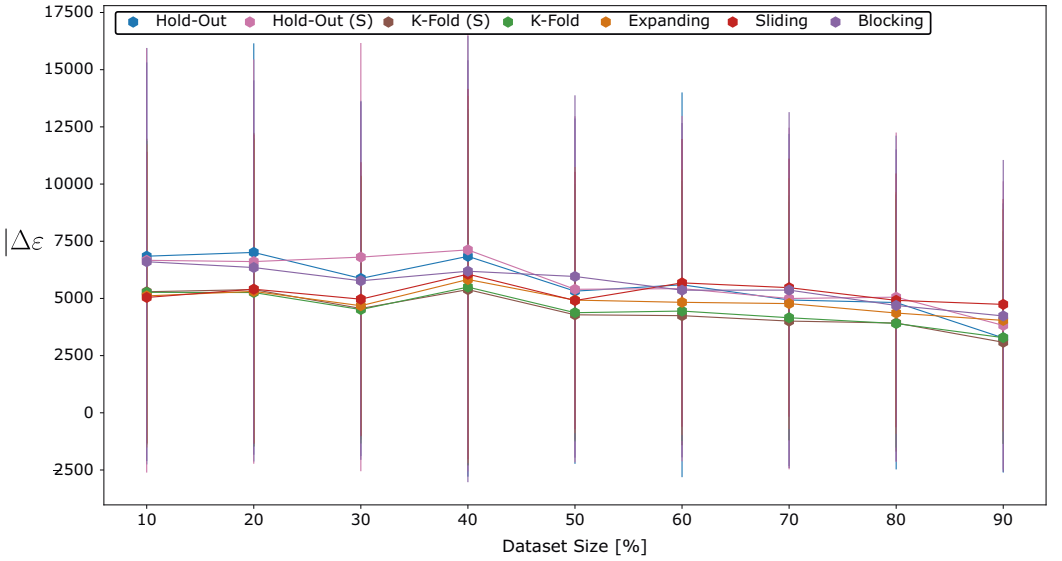

Figure C4. *Robustness of cross-validation procedure regarding dataset size for feed-forward neural network on solar. The marker indicates the average |Δε|, while the error bars display the standard deviation. The (S) indicates the shuffling variant of the method.*

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

D. Appendix D: Cross-validation experiment results for wind

D.1. Random forest

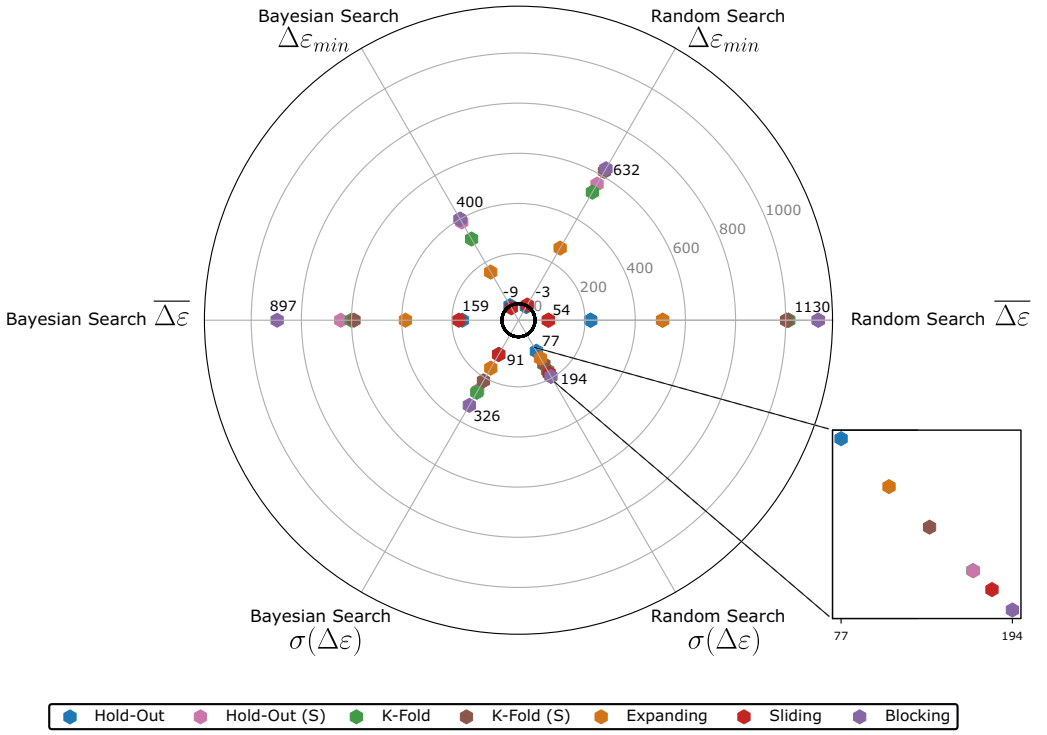

Figure D1. *Results of different cross-validation techniques for random forest on wind. Only the worst and best values for each axis are printed. The (S) indicates the shuffling variant of the method.*

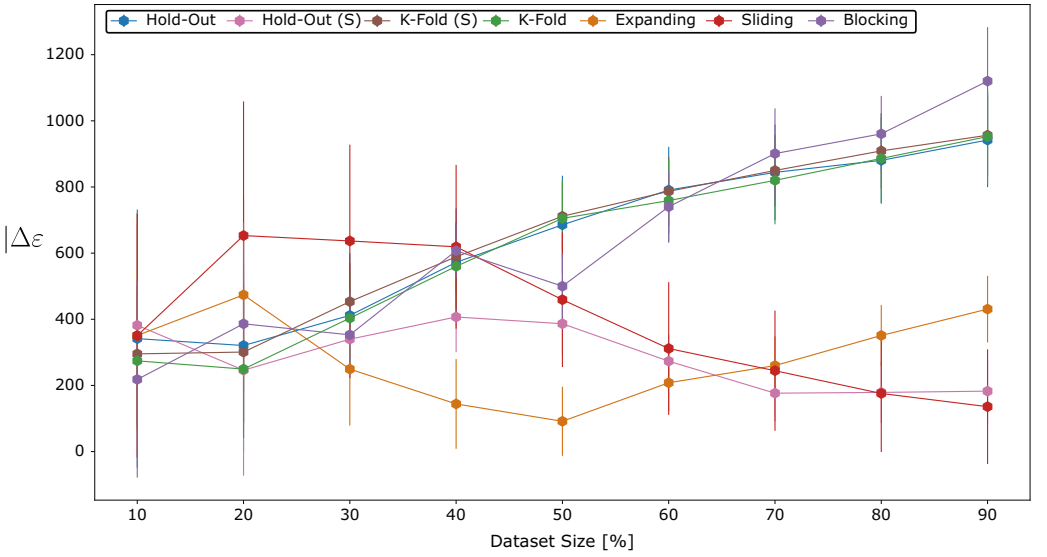

Figure D2. *Robustness of cross-validation procedure regarding dataset size for random forest on wind. The marker indicates the average $|\Delta\varepsilon|$, while the error bars display the standard deviation. The (S) indicates the shuffling variant of the method.*

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

D.2. Boosting

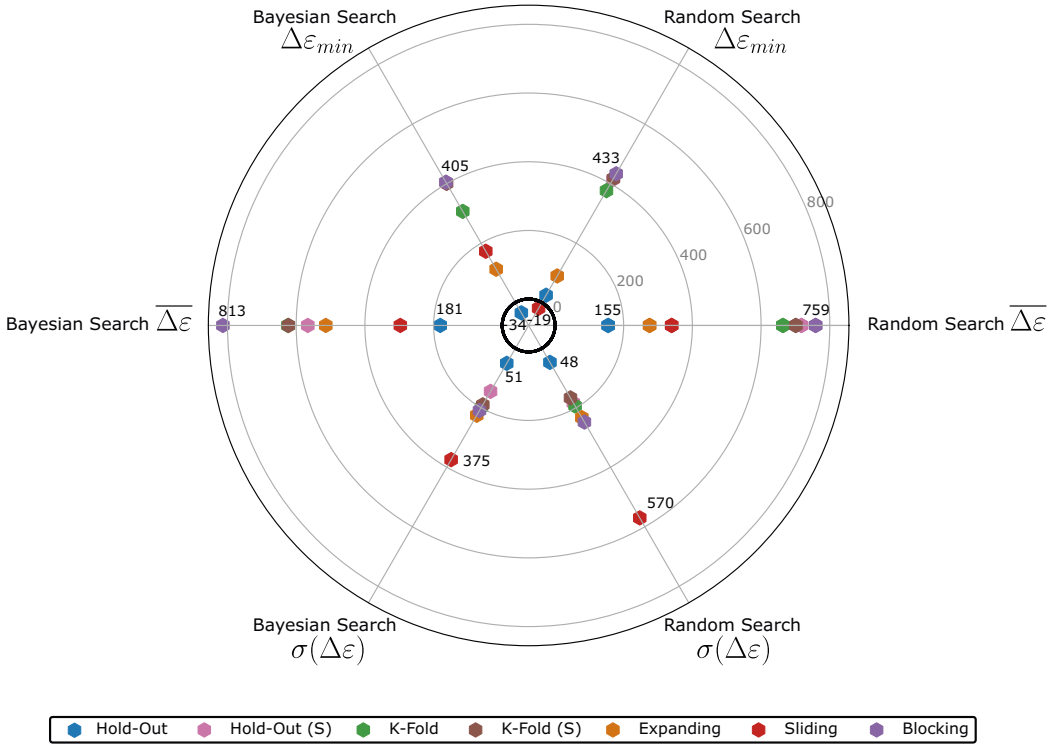

Figure D3. *Results of different cross-validation techniques for boosted trees on wind. Only the worst and best values for each axis are printed. The (S) indicates the shuffling variant of the method.*

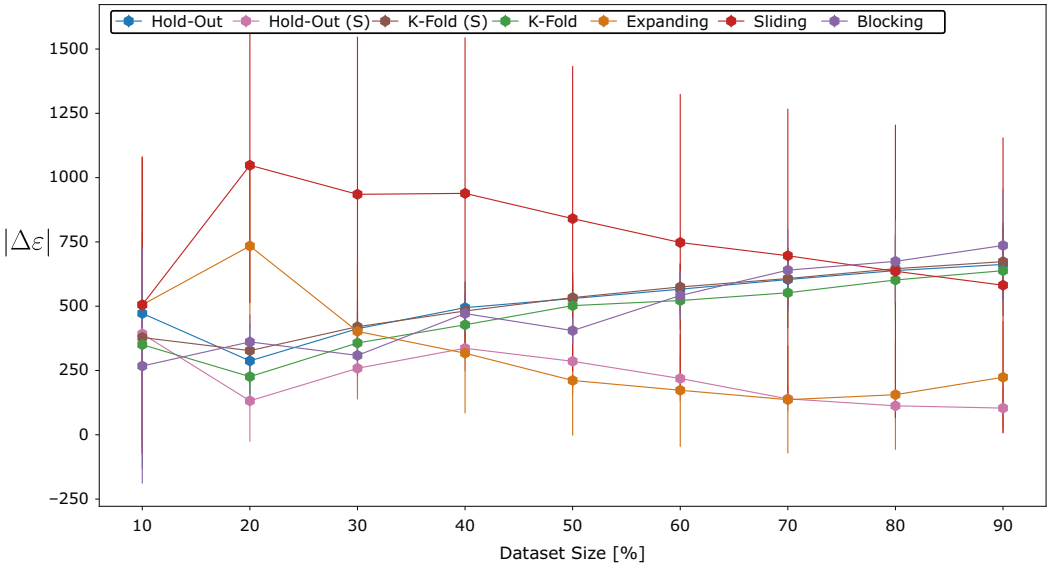

Figure D4. *Robustness of cross-validation procedure regarding dataset size for boosted trees on wind. The marker indicates the average $|\Delta\varepsilon|$, while the error bars display the standard deviation. The (S) indicates the shuffling variant of the method.*

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

D.3. Feed-forward neural network (MLP)

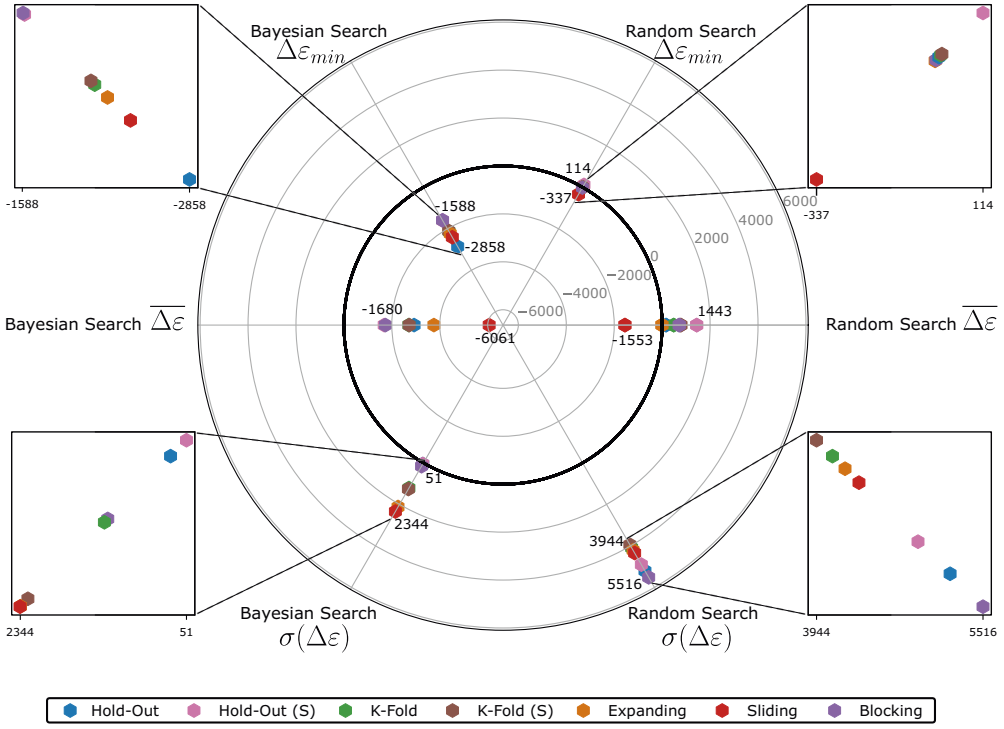

Figure D5. *Results of different cross-validation techniques for feed-forward neural network on wind. Only the worst and best values for each axis are printed. The (S) indicates the shuffling variant of the method.*

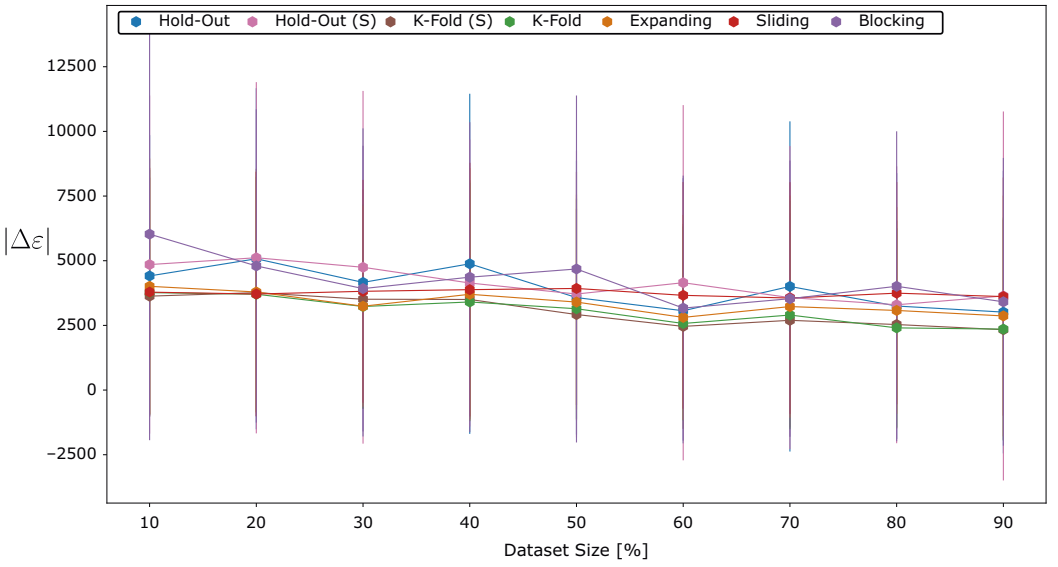

Figure D6. *Robustness of cross-validation procedure regarding dataset size for feed-forward neural network on wind. The marker indicates the average $|\Delta\varepsilon|$, while the error bars display the standard deviation. The (S) indicates the shuffling variant of the method.*

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

E. Appendix E: Benchmark results for wind

Table E1. *Benchmark results for different models using three different modeling approaches on the wind dataset*

Metrics			MAE		MAPE (%)		RMSE		nRMSE (%)		R^2	
Approach	Model	Detrend	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Average	Linear Regression		313	834	15.8	19.5	424	1,104	2.73	7.24	0.97	0.90
Average	Random Forest		99.2	1,180	4.27	25.6	140	1,464	0.90	9.60	1.0	0.83
Average	Random Forest	✓	112	483	4.80	11.3	159	650	1.03	4.26	1.0	0.97
Average	Linear Forest		260	707	12.2	19.8	354	967	2.28	6.35	0.98	0.93
Average	Tree Boosting		92.6	901	2.90	19.1	145	1,173	0.93	7.70	1.0	0.89
Average	Tree Boosting	✓	181	496	8.04	11.9	250	650	1.61	4.26	0.99	0.97
Average	Linear Tree Boosting		361	748	17.2	19.8	484	1,042	3.12	6.84	0.96	0.92
Average	GAM		243	628	11.8	17.4	332	807	2.14	5.29	0.98	0.95
Average	MLP		267	438	10.8	9.56	383	608	2.47	3.99	0.98	0.97
PCA	Linear Regression		484	799	22.6	19.1	654	1,089	4.21	7.14	0.93	0.91
PCA	Random Forest		177	1,168	8.27	26.1	252	1,500	1.63	9.84	0.99	0.82
PCA	Linear Forest		159	1,233	7.24	30.3	240	1,551	1.55	10.2	0.99	0.81
PCA	Tree Boosting		156	803	8.14	17.5	204	1,057	1.32	6.93	0.99	0.91
PCA	Linear Tree Boosting		423	801	19.7	19.0	574	1,066	3.70	6.99	0.94	0.91
PCA	GAM		328	750	16.3	20.2	438	929	2.82	6.10	0.97	0.93
PCA	MLP		283	508	12.6	11.5	373	693	2.40	4.54	0.98	0.96
Vision	CNN		240	417	9.98	9.12	340	575	2.19	3.77	0.98	0.97

Note. Medals indicate the top three best-performing models on the test set for each metric.

Table F1. *Comparison of ENTSO-E day ahead renewable Forecast performance for France with our model forecast performance in 2023 (test set)*

Sector	ENTSO-E Day-Ahead Forecast RMSE (MW)	CNN Forecast RMSE (MW)	Relative Improvement (%)
Solar	337	276	18.1
Wind	717	575	19.8

Note. The hourly ENTSO-E forecasts were aggregated to daily to match our work's granularity.

F. Appendix F: Comparison with ENTSO-E day-ahead forecasts

In the literature of renewable energy forecasting, most of the studies use numerical weather predictions, that is, forecasted weather, as inputs to the models, and mainly focus on a local scale, such as a single solar or wind farm. In this work, we used re-analysis ERA5 data as the weather inputs, which do not account for the weather forecasting error, and we directly predict the supply at the regional scale without any lags. These aspects make the comparison with other work difficult. However, we provide in Table F1 a comparison of the spatially explicit CNN results with the ENTSO-E day-ahead forecasts for wind and solar generation in France.[6] The day-ahead forecasts available on ENTSO-E are sourced from each TSO, and since they are run operationally, they must use numerical weather forecasts. Since the available forecast data granularity is hourly, we aggregated it to daily forecasts for the sake of comparison. We can see that our approach, combined with the use of re-analysis data, improved the forecasts by 18% on solar and around 20% on wind.

G. Appendix G: Sensitivity of CNN model to Gaussian noise applied to the weather inputs

Since this study's weather aspect is based on re-analysis data and not forecasted data, we study the degradation of the CNN model performance when mimicking weather forecasts as inputs. To do so, a Gaussian white noise without any correlation between the different weather variables is added to each weather map. The noise level is controlled, and the results of the performance degradation are reported in Table G1. It is worth mentioning that adding the same noise level to all the weather predictors does not

Table G1. *Comparison of our model performance when adding Gaussian noise to the weather inputs to mimic weather forecast data*

			RMSE (MW)	RMSE 95% empirical confidence interval (MW) (%)		Relative change (%)	Relative change 95% empirical confidence interval (%)	
Sector	Model	Noise level (%)	Mean	Lower bound	Upper bound	Mean	Lower bound	Upper bound
Solar	CNN	0	276.5					
Solar	CNN	5	276.8	274.8	279.1	0.11	−0.61	0.94
Solar	CNN	10	278.8	274.8	282.7	0.83	−0.61	2.24
Solar	CNN	15	284.7	278.3	290.5	2.90	0.65	5.10
Solar	CNN	20	295.6	287.4	302.7	6.90	3.90	9.50
Solar	CNN	30	336.4	324.6	347.1	21.6	17.4	25.5
Wind	CNN	0	575.1					
Wind	CNN	5	636.3	624.8	647.5	10.6	8.6	12.6
Wind	CNN	10	800.5	782.0	813.7	39.2	36.0	41.5
Wind	CNN	15	1,009.2	983.2	1,032.1	75.5	71.0	79.5
Wind	CNN	20	1,227.3	1,194.6	1,252.6	113.4	107.7	117.8
Wind	CNN	30	1,664.1	1,625.6	1,701.6	189.4	182.7	195.9

Note. The RMSE is computed on 2023 (test set). The relative change compares the metric with the noise to the metric without. Negative values mean improvement. The experiment was repeated 100 times before computing the mean and a 95% empirical confidence interval.

[6] Generation Day-Ahead Forecasts for wind and solar can be accessed at: https://transparency.entsoe.eu/.

https://doi.org/10.1017/eds.2025.10021 Published online by Cambridge University Press

translate into the same error for every weather variable. Even though this analysis is simple, we can notice that the solar model is less sensitive to the noise added than the Wind model. When looking at the score/metrics given by European Center for Medium-range Weather Forecast for their forecasts in their reference document,[7] we can see in Figure 26 that the RMSE for wind at 10 m is less than $0.5\,\mathrm{m\,s^{-1}}$ (around $0.25\,\mathrm{m\,s^{-1}}$) for 60- and 72-hour ahead forecasts. Our range of wind speed values is between -14 and $14\,\mathrm{m\,s^{-1}}$, with an average of 3–$4\,\mathrm{m\,s^{-1}}$, where the wind turbines are located. This would mean an error on forecast variables of around 5-10%, which would lead to a decrease of 10–40% of our predictions for a "fake" 3-day-ahead forecast.

[7] ECMWF Reference Document 2021 Release, available at https://www.ecmwf.int/en/elibrary/81235-evaluation-ecmwf-forecasts-including-2021-upgrade.

Cite this article: Lindas E, Goude Y and Ciais P (2025). Toward accurate forecasting of renewable energy: Building datasets and benchmarking machine learning models for solar and wind power in France. *Environmental Data Science*, 4: e45. doi:10.1017/eds.2025.10021