TOPIC 5: PV Systems – Modelling, Design, Operation and Performance; 5.3 Operation, Performance and Maintenance of PV Systems

PERFORMANCE IMPUTATION TECHNIQUES FOR ASSESSING COSTS OF TECHNICAL FAILURES IN PV SYSTEMS

S. Lindig[1,2*], A. Louwen[1], M. Herz[3], J. Ascencio-Vásquez[2,4], D. Moser[1], and M. Topič[2]

[1]Institute for Renewable Energy, EURAC Research, Viale Druso 1, 39100 Bolzano, Italy
[2]University of Ljubljana, Faculty of Engineering, Tržaška cesta 25, 1000 Ljubljana, Slovenia
[3]TÜV Rheinland Energy GmbH, Am Grauen Stein, 51105 Cologne, Germany
[4]3E sa, Quai à la Chaux 6, 1000 Brussels, Belgium
*Phone: +39 0471 055 715; E-mail: sascha.lindig@eurac.edu

ABSTRACT: In the framework of the new H2020 project TRUST PV, the Cost Priority Number methodology, a cost-based failure modes & effect analysis method, is tailored for its application onto real-case studies during the operational phase of a PV plant. Thereby, it is possible to calculate the energetic and economic losses for each individual failure by combining actual technical characteristics of the failure tickets, the operational costs of fixing those failures, and the power loss due to the failure occurrence. In this work, different empirical, machine learning and univariate models have been tested to find the optimal model to impute power data for all possible scenarios of missing performance data. The parameter space comprises of the amount and type of missing data, the ratio between training and test set and the predictor availability. Overall, the most desirable setting for highly accurate data imputation is the availability of a neighbouring string/plant or measured climate data as predictor with a high training/test set ratio for small test sets and ratios between 80/20 to 50/50 for bigger test sets (longer than 12 hours). In general, machine learning imputation models and empirical power models perform similar for smaller test set sizes and machine learning models are preferrable for bigger data gaps. Univariate methods should be avoided if possible.

Keywords: PV systems, PV system performance, Data imputation, Data quality

1 INTRODUCTION

In recent years, the photovoltaic (PV) industry experiences considerable advancements in the operation and maintenance (O&M) of PV power plants. The aim is to increase plant performance and availability while decreasing costs. During the operational phase of a PV system, O&M concepts are key tools to achieve these goals through a decrease in the number and severity of performance impairing events. Advanced O&M routines consist of smart and site-specific preventive maintenance solutions and improved corrective maintenance, which is characterized by a reduced detection time of an appearing failure and consequently a reduced downtime of PV plant components in which the failure is remedied [1].

Possible risks, leading to a decrease in PV performance, are faster detected and counteracted against through a better understanding of typical PV system behaviour and advanced detection and prevention techniques. A method to map and understand the likeliness and severity technical failures may have on PV plant performance, together with the associated costs, is the Cost Priority Number (CPN) method. The initial CPN, a cost-based failure modes & effect analysis (FMEA) method, was developed in the H2020 project Solar Bankability [2] with the aim to assess the economic impact of failures in PV projects and is calculated in €/kWp/year. In the framework of the new H2020 project TRUST PV [3], the CPN is improved by adapting parameters for its application onto real-case studies. The first version of the CPN methodology was based on the development of failure scenarios. Instead, within TRUST PV, tickets of O&M activities coming from SCADA systems of operating PV plants are categorized and evaluated. The CPN is used to accurately calculate costs/benefits for each individual ticket in order to understand and further improve the effectiveness of operators and maintainers. To do so, the accurate estimation of energy losses due to recorded failures is essential. However, several data issues such as gaps, communication errors and high measurement uncertainties make this task quite challenging.

In this work, we test widely used data imputation models considering all possible scenarios of missing power/energy data, which include the following parameter:

- Type of missing data
- Amount of missing data
- Ratio training set / test set
- Predictor availability

For each scenario, the tested models have been benchmarked. Optimized imputation approaches for all considered cases were identified by minimizing the following statistical metrics: relative root-mean-square-error (rRMSE), relative mean bias error (rMBE), relative difference (rD) and the absolute difference (aD). The overall aim is to find the optimal imputation solution for each possible scenario in terms of available predictor data and missing data rate to automatize the imputation process to a high degree.

2 COST PRIORITY NUMBER METHODOLOGY

In the H2020 project Solar Bankability [2], the CPN methodology has been developed to quantify appearing degradation modes and other performance impairing effects in PV plants. The focus was on the assessment of risks connected to investments in PV projects [4]. Thereby, the methodology assesses the economic impact based on factors such as performance reduction and downtime. Within the project, the CPN was applied to theoretical scenarios to calculate extreme values for the CPN metric, expressed in €/kWp/year. All phases of a PV plant life cycle (from product testing to decommissioning) have been included and a nomenclature for possible failures has been created in form of a technical risk matrix [5]. Later, the CPN has been adapted towards the needs of a large O&M operator by assessing real tickets from the ticketing system of an operating PV plant manually [6]. In the H2020 project TRUST PV [3], the assessment of real tickets is automatized to not only study the economic impact of individual technical failures, but also to draw useful statistics in terms of failure appearance likeliness and severity. Therefore, the technical risk matrix has been updated. The TRUST PV risk matrix only includes failures from within the operational phase of a PV plant divided by component, subcomponent, and individual failure. Overall, 344 failures have been identified. Additionally, the following optional information on individual failures can be supplied: cause (what is the failure cause), accountability (which party is accountable), detection (how was the failure detected), origin (can the failure be attributed to a root-cause stemming from a specific phase of the PV plants life-cycle) and solution (how was the failure fixed). The workflow of assigning a cost to a failure and thereby calculating the CPN is the following:

1. Failure appearance in PV plant
2. Failure detection
3. Creation of ticket in SCADA system
4. Classification of failure according to TRUST PV's Risk Matrix
5. Resolution of failure
6. Statistical analysis of failure using the CPN methodology

The CPN is calculated by the sum of costs due to downtime and fixing costs due the failure appearance:

$$CPN\ [€/kWp/year] = C_{down} + C_{fix}$$

The two categories of costs are defined as:

a) Economic impact due to downtime (C_{down}):
   - failures that cause downtime or power loss
   - considers time from failure to repair/substitution (thereby including detection time, response time, repair time, shutdown time)
   - missing income related to the sale of electricity, defined by the feed in tariff, power-purchasing agreement or missing savings generated by PV plants installed on roofs/facades defined as retail cost of electricity
   - important to pay attention to component level of failure as other components might be affected

b) Economic impact due to repair (C_{fix}):
   - cost of detection
   - cost of labour
   - cost of transportation
   - cost of repair/substitution

The scope of this work is to reliably calculate the energy loss caused by an appearing failure to determine the economic impact due to downtime. A technical failure leads to a partial or complete reduction of the energy produced by a PV plant. With the help of this study, we determine how much energy a plant would produce without the appearance of a failure and the difference between the estimated energy production and the reduced actual production is the amount lost through the appearance of the failure.

3 PV PLANT

For this work, performance timeseries data of an experimental PV plant installed and operated at the Airport Bolzano Dolomiti (South Tyrol/Italy) have been used. The data are recorded and stored by EURAC Research. Longitude and latitude of the system are 46.4625°N and 11.3299°E respectively, and the plant is situated 240 m above sea-level. The system was installed in 2010 with a fixed tilt of 30° and an orientation of 8.5° west of south. The plant consists of 20 poly-crystalline silicon PV modules with an overall nominal power of 4.2 kWp. Additionally, climate data are recorded at a nearby meteorological station including plane-of-array irradiance and ambient temperature.

4 PARAMETER

Incoming performance data of PV systems can have different measured signals available. In order to cover all possible scenarios of missing data for calculating the energy loss, we created a parameter space including four different variables, namely the type and amount of missing data, the ratio between training and test set as well as the predictor availability.

4.1. Type of missing data

Missing data can usually be categorized as being randomly or continuously distributed. In the scope of this work, we only work with continuous missing data as this is the predominant type in case of failure appearance. Furthermore, data can be categorized based on pairwise or listwise missingness. In case of listwise missing data, the complete column (power/energy and all available measured predictors) is missing, whereas in case of pairwise missing data only the power/energy measurement is affected.

4.2. Amount of missing data

This parameter is expressed through downtime of the plant or plant component. It depends on the failure type in combination with the effectiveness of the O&M activities, driven by detection and response time on one hand and repair time paired with spare part management on the other hand. In order to simulate different grades of missingness, the tests set size (data to impute) was varied from 4 hours to 10 days in five steps.

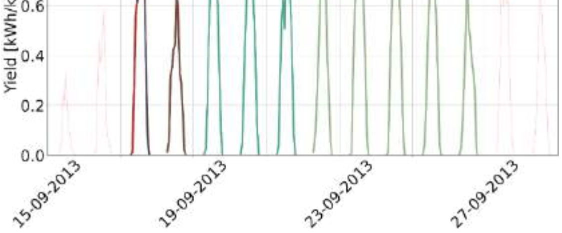

Figure 1: Yield time series of PV plant; test set intervals ranging from 4 hours to 10 days

In Figure 1, the intervals of the test sets are shown exemplarily. Thereby, the test set position is drawn randomly from the time series, which ranges from 2012 until 2019 spanning seven years of data.

4.3. Ratio training set / test set

After assigning the test set size and position, the training set is selected as the data preceding the test set in seven different training/test set splits. Ultimately, 35 different datasets are rendered for the imputation. The training set is thereby the dataset used to train a data imputation model and the test set the missing energy data to impute. Across all different settings the missing data rate ranges from 4 hours to 10 days. Different training/test set ratios are tested to find the amount of required training data to fit the imputation models sufficiently. The amount of training data is a trade-off between providing enough data for model fitting while capturing the instantaneous performance of the PV plant.

4.4. Predictor availability

It can be expected that different PV plants have different predictors available through varying data acquisition and monitoring concepts. That is why different predictors and predictor sources have been selected within this study:

Table 1: Predictor groups and predictors used for data imputation.

MEASURED PARAMETER	**SATELLITE DATA**	**NO PREDICTOR AVAILABLE**
Climate data	**ERA5** [7] **+ tilt & azimuth**	
$G_{POA_{meas}}$ $T_{amb_{meas}}$	$G_{POA_{ERA5}}$ $T_{amb_{ERA5}}$	
Neighbouring string/plant	**CAMS** [8] **+ tilt & azimuth**	
$Power$	$G_{POA_{Cams}}$	

Table 1 shows the different groups of predictors together with the specific parameters. The first group are measured climate data from a meteorological station on-site. If a neighbouring plant or a neighbouring string/inverter (in case of higher resolution monitoring) is available, the power signal of this neighbour is used. In this study, the neighbour is a poly-crystalline silicon PV plant installed on the same site as the test system having the same tilt and orientation while using modules from a different manufacturer. Both systems have the same configuration of front-glass/cell/backsheet while being installed on an open rack. This is important for the temperature evolution across the systems and thereby for the performance/efficiency evolution. For reanalysis/ satellite data, ERA5 data [7] as well as CAMS data [8] were selected as they are freely available in high resolution. In order to compute plane-of-array data [9, 10], tilt and azimuth of the PV plant in question must be known. Figure 2 shows the correlation [11] between the target variable (power of test system) and the tested predictors in hourly resolution. It is visible that the power signal of the neighbouring plant has the highest correlation with the target variable, rooting from the same installation conditions and the same PV technology. As expected, the measured plane-of-array irradiance (G_{POA}) also correlates highly, followed by irradiance values retrieved through CAMS and ERA5 data. Ambient temperature correlates to a lesser degree with the produced power.
The last predictor option is that no data are available for imputation. In this case, only univariate imputation methods can be applied and tested.

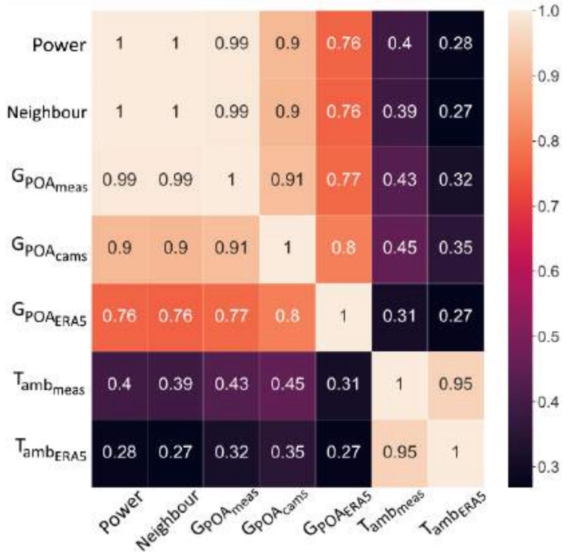

Figure 2: Heatmap showing the Pearson correlation **[11]** between the target variable (Power) and the tested predictors in hourly resolution

5 IMPUTATION MODELS

Overall, 14 different models have been tested. The applicability of the models always depends on the availability of predictors. The models are coming from three different families: empirical, regression & machine learning and univariate models.

5.1. Empirical power prediction models

Empirical models predict the power of a PV system as a function of climate inputs, usually irradiance and temperature.

For the application of most models the module temperature is required. The Nominal Operating Cell Temperature Model [12] has been used to calculate the module temperature. In this study, three different empirical models have been tested.

a) PVWatts model [13]

The PVWatts model relates the incoming irradiance and cell temperature to the PV system power with:

$$P = P_{nom}\frac{G_{POA}}{1000W/m^2}(1+\gamma(T_{mod}-25°C))$$

Here, γ is the temperature coefficient given by the module manufacturer. The model is provided by NREL as an online PV performance modelling application [14].

b) PVGIS model [15]

The PVGIS model combines regression of normalized irradiance and PV module temperature with six empirical coefficients and is a variant of King's model [16] (used in the PVGIS online tool [17]). In this work, modelled module temperature and measured/modelled plane-of-array irradiance is used without normalization.

$$P = G_{POA}(P_{nom} + k_1 \ln(G_{POA}) + k_2 \ln(G_{POA})^2 + T_{mod}(k_3 + k_4 \ln(G_{POA}) + k_5 \ln(G_{POA})^2) + k_6 {T_{mod}}^2)$$

k_1 to k_6 are the empirical coefficients which are fitted with non-linear least squares.

c) Three parameter model

This model was benchmarked among different heuristic models by Ding et al. [18] and proved to provide accurate results. It is described by:

$$P = \left(aG_{POA} + bG_{POA}^2 + cG_{POA} ln\left(\frac{G_{POA}}{1000\frac{W}{m^2}}\right)\right)\left(1 + \gamma(T_{mod} - 25°C)\right)$$

Here, a, b and c are fitting coefficients, again determined with non-linear least squares.

5.2. Regression & Machine Learning models

Several widely used regression and machine learning models have been tested as alternatives to more classical approaches of predicting power by using empirical models. All models have been implemented with the Python library scikit-learn [19]. For the models *b)* to *g)* hyperparameter tuning has been carried out with a pre-defined parameter space using an exhaustive searching approach.

a) Multivariate linear regression

The first model is an ordinary least-squares linear regression of the predictor/s versus power.

b) K-nearest neighbours

K-nearest neighbour regression creates data clusters based on training data. The predicted value will be approximated based on the predictors in the cluster space through weighted averaging. The size of the clusters is defined by the parameter K.

c) Decision tree

Decision trees design sets of logical decisions to divide data into nodes by minimizing the residual sum of squares within a node for a given tree depth.

d) Random forest

A random forest regressor constructs a number of decision trees and returns the average prediction of the individual trees to improve accuracy [20]. It fits the trees using bootstrapping and selecting an optimum split.

e) Extra tree

The extra tree regressor is very similar to random forest but uses the whole training sample for fitting individual decision trees and splits the tress randomly [21].

f) Gradient boosting regressor

When using gradient boosting regressors, trees are grown sequentially using information from previously grown trees. Thereby, additionally to the depth of the tree and the number of nodes the learning rate is variable [22].

g) Histogram based gradient boosting regressor

This model is implemented within scikit-learn based on the LightGBM framework [23]. It provides a novel gradient boosting regressor including new techniques to find the optimal split point of trees by reducing the number of data instances and reducing the feature space.

5.3. Univariate models

As mentioned before, the situation might arise where no training data are available to predict the lost energy. In this case, univariate imputation methods are the only option. Here, only energy data before and after the outage instance are used to train the models and to perform the imputation. The models have been implemented using the R packages imputeTS [24] as well as forecast [25, 26].

a) Average value

This method simply averages hourly values in the training set by the hour of the day. The imputed value is the average for the specific hour.

b) Random

Missing values are replaced by drawing a random value from the training dataset.

c) Kalman

This is a two step-process. First, the time series (training data) is transformed into a state space model using ARIMA via seasonal decomposition. Data imputation is performed via Kalman smoothing of the created model. The Kalman filter joins the probability distributions of the univariate data and estimates of unknown variables at each timeframe, allowing the prediction when no recorded data is available [27].

d) Interpolation

This model imputes data via linear interpolation. For seasonal data, such as performance time series, seasonal decomposition is first carried out.

5.4. Model evaluation metrics

The presented models have been evaluated based on the parameter introduced in Section 3 using four different statistical metrics. In this study, a full outage of the whole plant has been assumed. The following parameter were used to test the best imputation scenarios and to evaluate the accuracy of the different models:

$$rRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{N}(\tilde{y}_i - y_i)^2}{N}}}{\frac{1}{N}\sum_{i=1}^{N} y_i} * 100\%$$

$$rMBE = \frac{\sum_{i=1}^{N}(\tilde{y}_i - y_i)}{\sum_{i=1}^{n} y_i} * 100\%$$

$$aD = \left|\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \tilde{y}_i\right|$$

$$rD = \frac{\left|\sum_{i=1}^{N} y_i - \sum_{i=1}^{n} \tilde{y}_i\right|}{\sum_{i=1}^{n} y_i}$$

The relative root mean square error (rRMSE) helps to identify imputation scenarios with large individual errors between predicted value and target value. The relative mean bias error (rMBE) shows if a set of predictors (or a model) tends to over- or underestimate the target value. The absolute difference (aD) tells us by how many kWh our prediction deviates from the target. This parameter is very interesting as the outcome can be directly used to assess in economic impact of the imputation inaccuracy. Finally, the relative difference (rD) sets this parameter into relation to the overall lost energy to better intercompare the results on varying timescales.

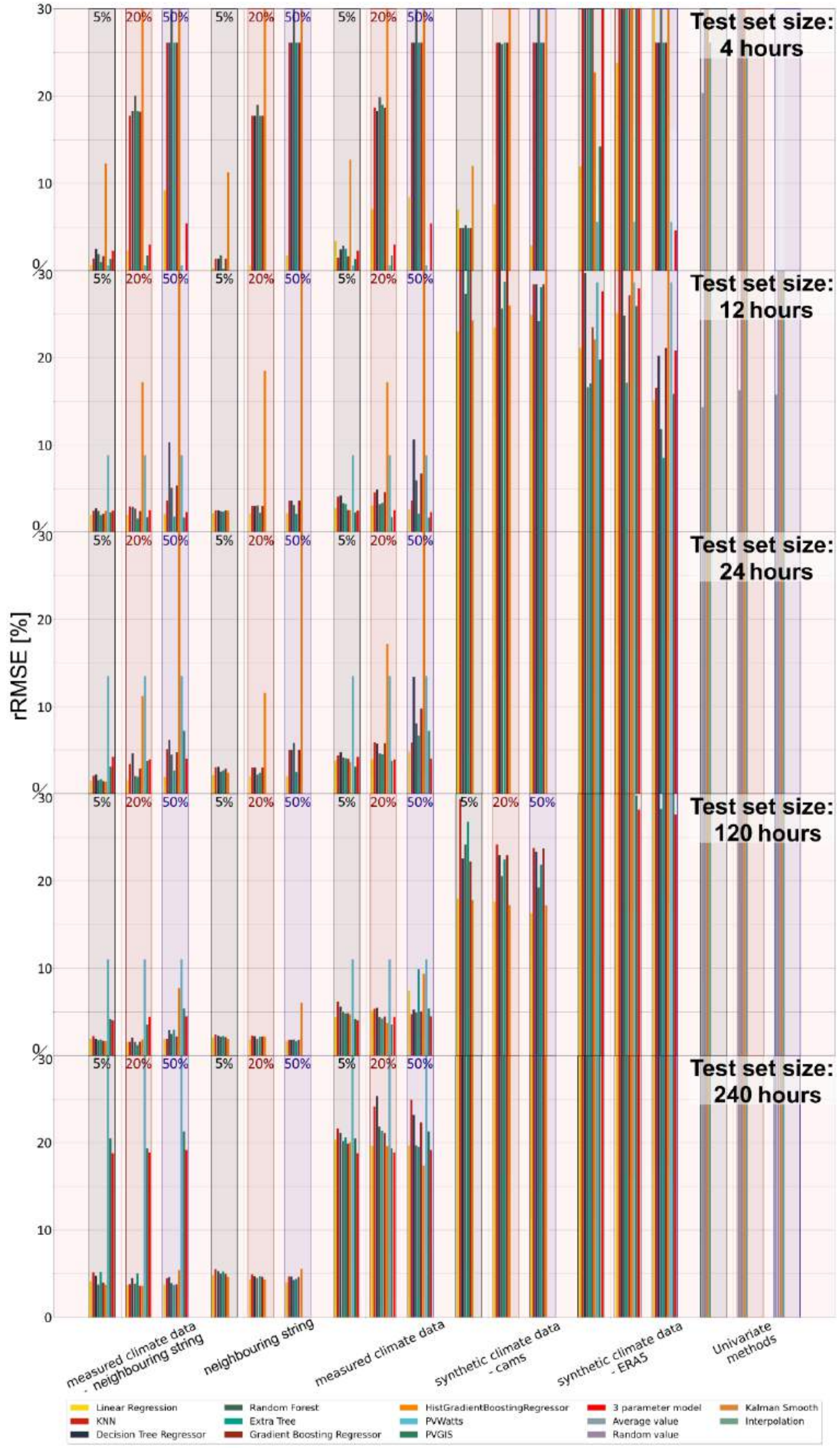

Figure 3: rRMSE [%] of individual predictors for training/test ratio of 95/5, 80/20 & 50/50 for test set sizes of 4 hours, 12 hours, 24 hours, 120 hours & 240 hours

6 RESULTS

The results of this study are shown in Figure 3 as well as the following figures in the Appendix. In the figures, the statistical metrics introduced in Section 5.4. are depicted. In the figures, the test set size increases from top to bottom and from left to right the different predictors are shown. For each predictor, the training/test set ratios of 95/5, 80/20 and 50/50 are included. For example, a 95/5 ratio for a 12-hour test set equals to 228 hours of training data versus 12 hours of test data, an 80/20 ratio equals 48 hours vs. 12 hours and a 50/50 ratio equals 12 hours vs. 12 hours.

Looking at the overall results, it is visible that a neighbouring string/plant is the most desirable predictor followed by measured climate data, indicated by the lowest prediction errors across all error metrics. Except for very short time series imputation (4 hours), lower training/test ratios (80/20 to even 50/50) perform as good and partially even better compared to very high training/test ratios of 95/5. The reason for the bad performance for a 80/20 or 50/50 split for a test set size of 4 hours is because the amount of training data is simply too low. It is assumed that 95/5 splits for longer imputation periods do not perform too well because a very long training dataset contains data from very different seasons of the year and temperature dependent performance deviations throughout the seasons might introduce a bias.

Looking at the predictors neighbouring string and/or measured climate data, it is visible that the rRMSE (Figure 3) is slightly increasing with a increasing test set size while the absolute and relative difference (Figure 4 & Figure 5) are almost constant for the individual best performing models. The bias (Figure 6) is decreasing with an increasing test size. In terms of imputation methods, both machine learning and empirical models with fitting coefficients perfom similarly well for small test set sizes while machine learning models are superior when being applied on larger test sets, especially if measured climate data are used.

PVWatts is the worst performing empirical model among the tested ones, probably due its nature of being generally applicable without any training of fitting coefficients. The three parameter model and the PVGIS model show a comparable accuracy. For measured climate data, impuation using these two models has a similar, partially even lower, bias compared to machine learning approaches.

For small test set sizes, multivariate regression shows also reliable results. In contrast, histogram based gradient boosting regression performs poor for absolute training set sizes below 60 hours, especially when looking at the rRMSE. Across the tested scenarios and machine learning models, the decision tree regressor and the extra tree regressor perform well.

If no measured predictors are available, two imputation options have been identified, either the usage of satellite climate data or univariate imputation. Univariate methods are not performing very good, they really should be a last resort if no other option is available. The average value method with a medium/low training/test ratio shows the best performance among the univariate methods. For this particular location, the accuracy using both satellite data sources is low compared to measured predictors. Looking at the rRMSE in Figure 3 as well as the absolute difference in Figure 4, CAMS seems to produce slightly better results compared to ERA5 reanalyis data. For CAMS data, the bias tends to be lower for a high training/test ratio for small test set sizes and for a lower training/test ratio for bigger data holes to impute. As CAMS does not provide ambient temperature data, only machine learning imputaton methods were tested. The best performing model changes depending on the test set size and the training/test ratio, but in general multivariate regression seems to perform well in comparison for small test set sizes, and histogram based gradient boosting regression for bigger test sets.

We expect to see lower imputation errors for satellite and reanalysis data for PV systems installed in areas with less complex climate conditions and less mountainous terrain. Bolzano is located at multiple climate borders. It is situated in a valley amid mountains at the foot of the Dolomites. Satellite data are by experience highly unreliable in this location. System performance of PV plants installed in more stable climatic conditions with a simpler terrain can usually be predicted with higher accuracy using satellite reanalysis data.

7 CONCLUSIONS & OUTLOOK

In this work, different imputation models for energy data of a PV plant, installed in Bolzano/Italy, have been studied. Thereby, empirical models, machine learning approaches as well as univariate methods have been tested. The imputation accuracy has been evaluated in dependency on the available predictors, the amount of data to be imputed and the ratio between training and test set size. This work is part of a larger study to estimate the costs and impact of an appearing failure in the field for operating PV systems (cost priority number – CPN). One of the primary costs is the energy lost due to the appearance of a failure. Reliable methods to accurately calculate the energy loss covering all possible scenarios in terms of predictor data availability and training/test set characteristics are required to estimate the economic impact of technical failures.

This study has shown that a neighbouring plant/string is the most desirable predictor to impute performance data, followed by measured climate data. In most parameter settings, the application of machine learning models delivers higher imputation accuracies. Short time series can be imputed with high accuracy using simple multivariate regression or empirical models including fitting coefficients. Results with higher accuracy for longer time series are calculated using decision tree or extra tree regression. Also, histogram based gradient boosting regression performed very well if the training set has more than 60 entries (tested on data with hourly resolution). Data holes exceeding 12 hours can reliably be imputed with lower training/test set ratios of 80/20 up to 50/50 splits. Lower training/test set ratios have two advantages:

- computation time is lower for model optimization & application
- long training sets tend to introduce a seasonal performance related bias in the models, which might also introduce a bias in the results

If no measured predictors are available, freely obtainable reanalysis/satellite data such as CAMS or ERA5 reanalysis data are a viable choice and preferred to univariate data imputation. The imputation using satellite data was subject to high uncertainties for the PV plant in question. That is because the plant is installed in complex terrain and at the border of several climate zones. It is expected to see more reliable imputation results for systems in flat terrain and clearly defined climate zones.

In the future, this study will be extended to several PV systems, installed in different climate and terrain conditions, to define general best practices as well as to prepare automatized scripts for the imputation of missing performance data. Furthermore, an adapted version of cross validation for time series will be used to generalize the results for individual plants by removing data artefacts. Finally, it is foreseen to carry out an economic sensitivity analysis across the variability of the results to understand the impact of imputation uncertainties on the calculated CPN value.

ACKNOWLEDGMENTS

This project has received funding from the Horizon 2020 programme of the European Union under GA. No. 952957 – Project TRUST PV.

[7] was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store. The results contain modified Copernicus Climate Change Service information 2020. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

REFERENCES

[1] National Renewable Energy Laboratory, Sandia National Laboratory, SunSpec Alliance, and the SunShot National Laboratory Multiyear Partnership (SuNLaMP) PV O&M Best Practices Working Group, “Best Practices for Operation and Maintenance of Photovoltaic and Energy Storage Systems; 3rd Editions,” Golden, CO: National Renewable Energy Laboratory, 2018.

[2] “H2020 Solar Bankability Project,” [Online]. Available: http://www.solarbankability.org/home.html. [Accessed 20 07 2021].

[3] “H2020 TRUST PV Project,” [Online]. Available: https://trust-pv.eu/. [Accessed 11 01 2021].

[4] D. Moser et al., “Identification of Technical Risks in the PV Value Chain and Quantification of the Economic Impact on the Business Model,” *Progress in Photovoltaics: Research and Applications,* vol. 25, no. 7, pp. 592-604, 2017.

[5] H2020 Project Solar Bankability, “Technical Risk Matrix,” 2017.

[6] G. Oviedo-Hernández et al., “Optimization of the Cost Priority Number (CPN) Methodology to the Needs of a Large O&M Operator,” in *36th EU PVSEC,* Marseille, 2019.

[7] Copernicus-Climate-Change-Service-ERA5, “Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS),” 2020. [Online]. Available: https://cds.climate.copernicus.eu/cdsapp#!/home . [Accessed 15 July 2020].

[8] “CAMS: Surface solar radiation data,” [Online]. Available: https://atmosphere.copernicus.eu. [Accessed 12 01 2021].

[9] W. Holmgren, C. Hansen and M. Mikofski, “Pvlib python: A python package for modeling solar energy systems,” *J. Open Source Softw.,* vol. 3, no. 29, p. 884, 2018.

[10] H. Hottel and B. Woertz, “The Performance of Flat Plate Solar-Heat Collectors,” *Trans. ASME,* vol. 64, pp. 64-91, 1942.

[11] J. Benesty et al., “Pearson correlation coefficient,” in *Noise reduction in speech processing,* Springer, 2009, pp. 37-40.

[12] R. G. Ross, “Flat-Plate Photovoltaic Array Design Optimization,” in *14th IEEE Photovoltaic Specialists Conference,* San Diego, CA, USA, 1980.

[13] NREL, PVWatts Version 5 Manual, National Renewable Energy Laboratory, 2014.

[14] NREL, “PVWatts Calculator,” 2019. [Online]. Available: https://pvwatts.nrel.gov/index.php. [Accessed 20 07 2021].

[15] T.A. Huld et al., “A power-rating model for crystalline silicon PV modules.,” *Solar Energy Mater. Solar Cells,* vol. 95, pp. 3359-3369, 2011.

[16] D. L. King, W. Boyson and J. Kratochvill, “Photovoltaic Array Performance Model,” Sandia National Laboratories, Alberquerque, 2004.

[17] European Commission, “Joint Research Centre Photovoltaic Geographical Information System (PVGIS),” [Online]. Available: https://ec.europa.eu/jrc/en/pvgis. [Accessed 20 07 2021].

[18] K. Ding, Z. Ye and T. Reindl, “Comparison of Parameterisation Models for the Estimation of the Maximum Power Output of PV Modules,” *Energy Procedia,* vol. 25, p. 101–107, 2012.

[19] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.,* vol. 12, p. 2825.2830, 2011.

[20] L. Breiman, “Random Forests,” *Machine Learning,* vol. 45, pp. 5-32, 2001.

[21] P. Geurts, D. Ernst and L. Wehenkel, “Extremely Randomized Trees,” *Machine Learning,* vol. 63, pp. 3-42, 2006.

[22] J. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Ann. Stat.,* vol. 29, pp. 1189-1232, 2001.

[23] G. Ke et al., “LightGBM: A Highly Efficient,” *Adv. Neural Inf. Process. Syst.,* vol. 2017, pp. 3149-3157, 2017.

[24] S. Moritz and T. Bartz-Beielstein, “imputeTS: Time Series Missing Value Imputation in R,” *The R Journal,* vol. 9, no. 1, pp. 207-218, 2017.

[25] R. Hyndman, Y. Khandakar, “Automatic time series forecasting: the forecast package for R,” *Journal of Statistical,* vol. 26, no. 3, pp. 1-22, 2008.

[26] R. Hyndman et al., “forecast: Forecasting functions for time series and linear models. R package version 8.11,” 2020. [Online]. Available: http://pkg.robjhyndman.com/forecast>.

[27] R. E. Kalman, “A New Approach to Linear Filtering and Prediction Problems,” *Journal of Basic Engineering,* pp. 35-46, 1960.

APPENDIX

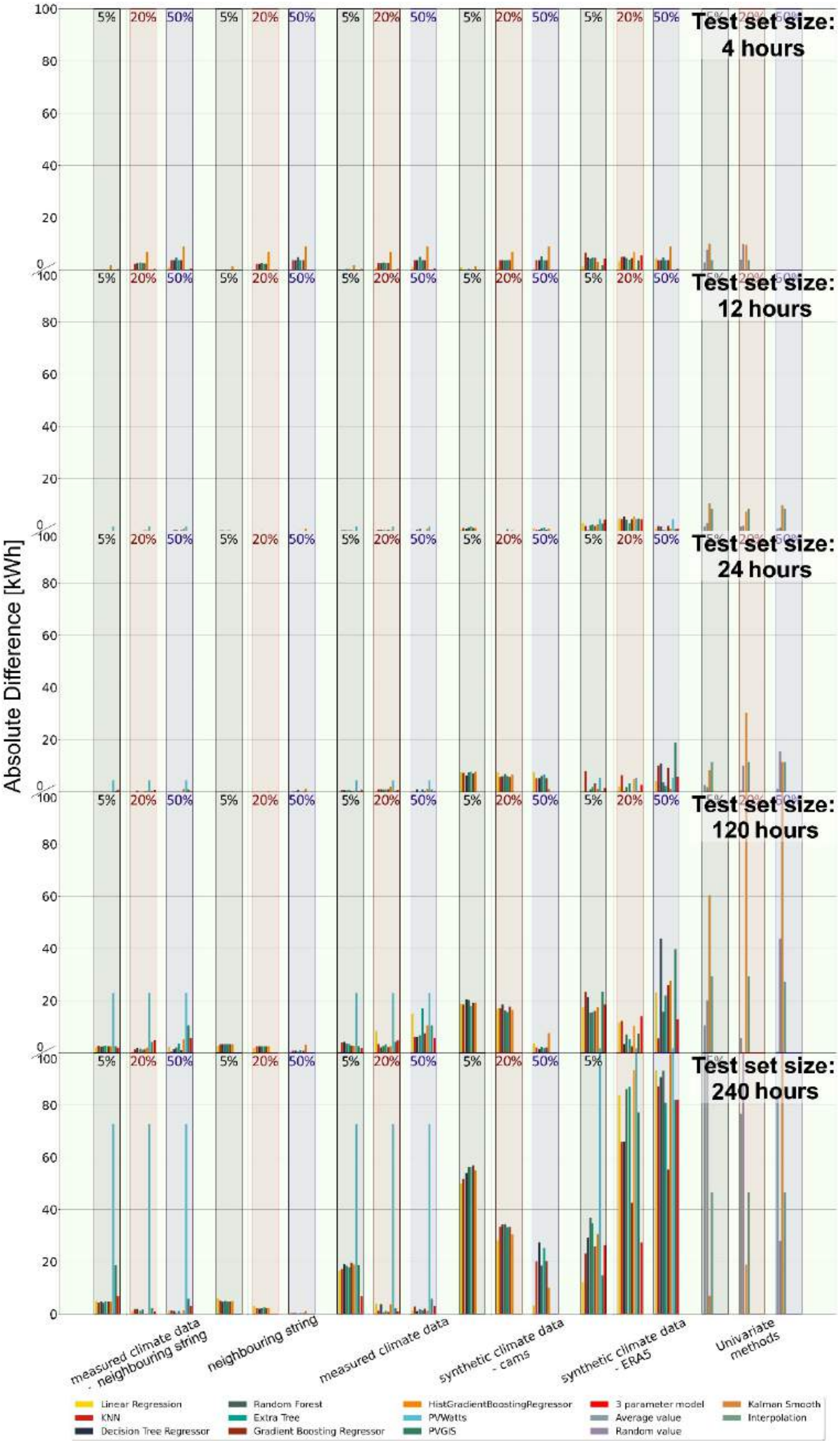

Figure 4: Absolute difference [kWh] of individual predictors for training/test ratio of 95/5, 80/20 & 50/50 for test set sizes of 4 hours, 12 hours, 24 hours, 120 hours & 240 hours

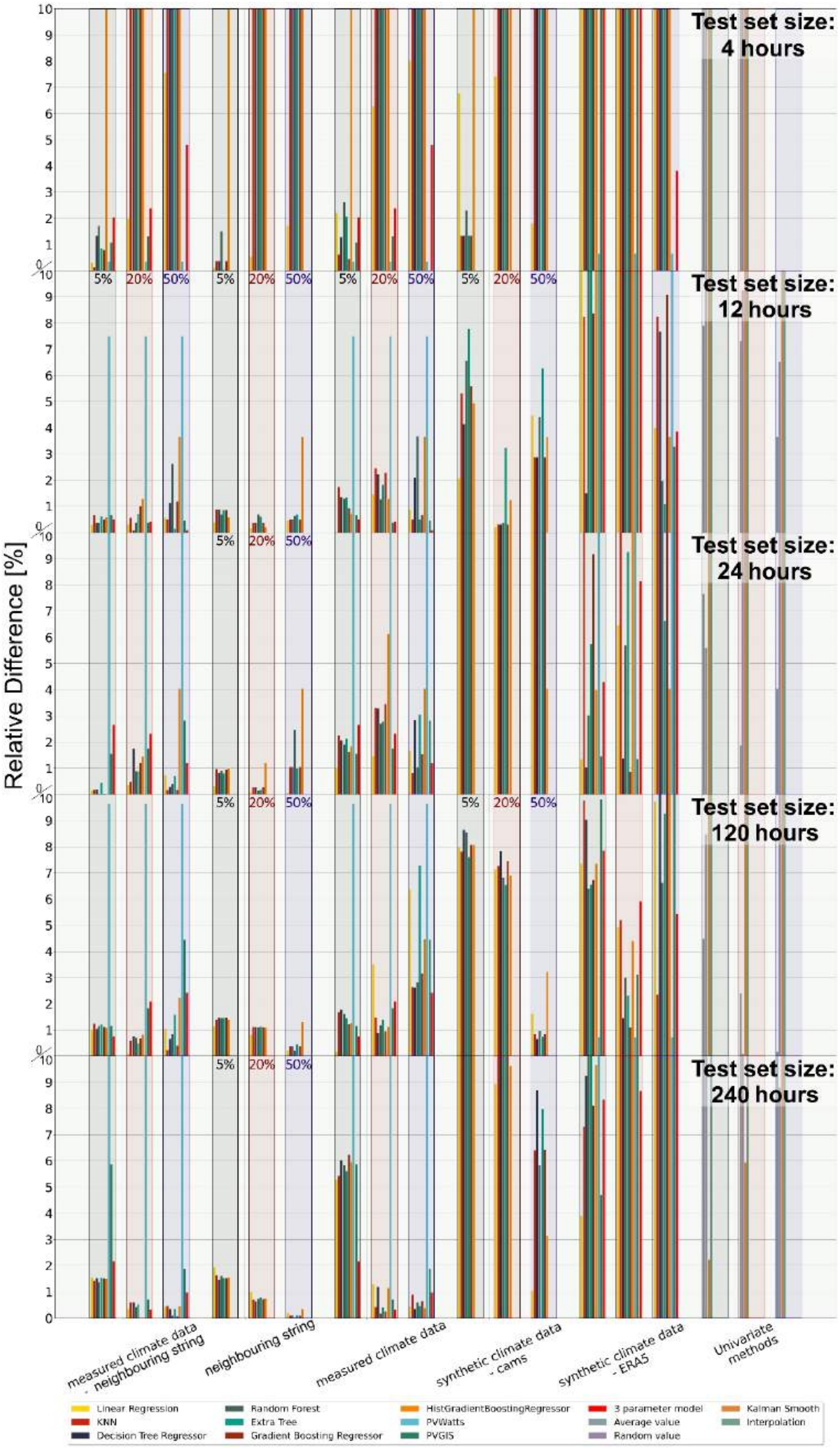

Figure 5: Relative difference [%] of individual predictors for training/test ratio of 95/5, 80/20 & 50/50 for test set sizes of 4 hours, 12 hours, 24 hours, 120 hours & 240 hours

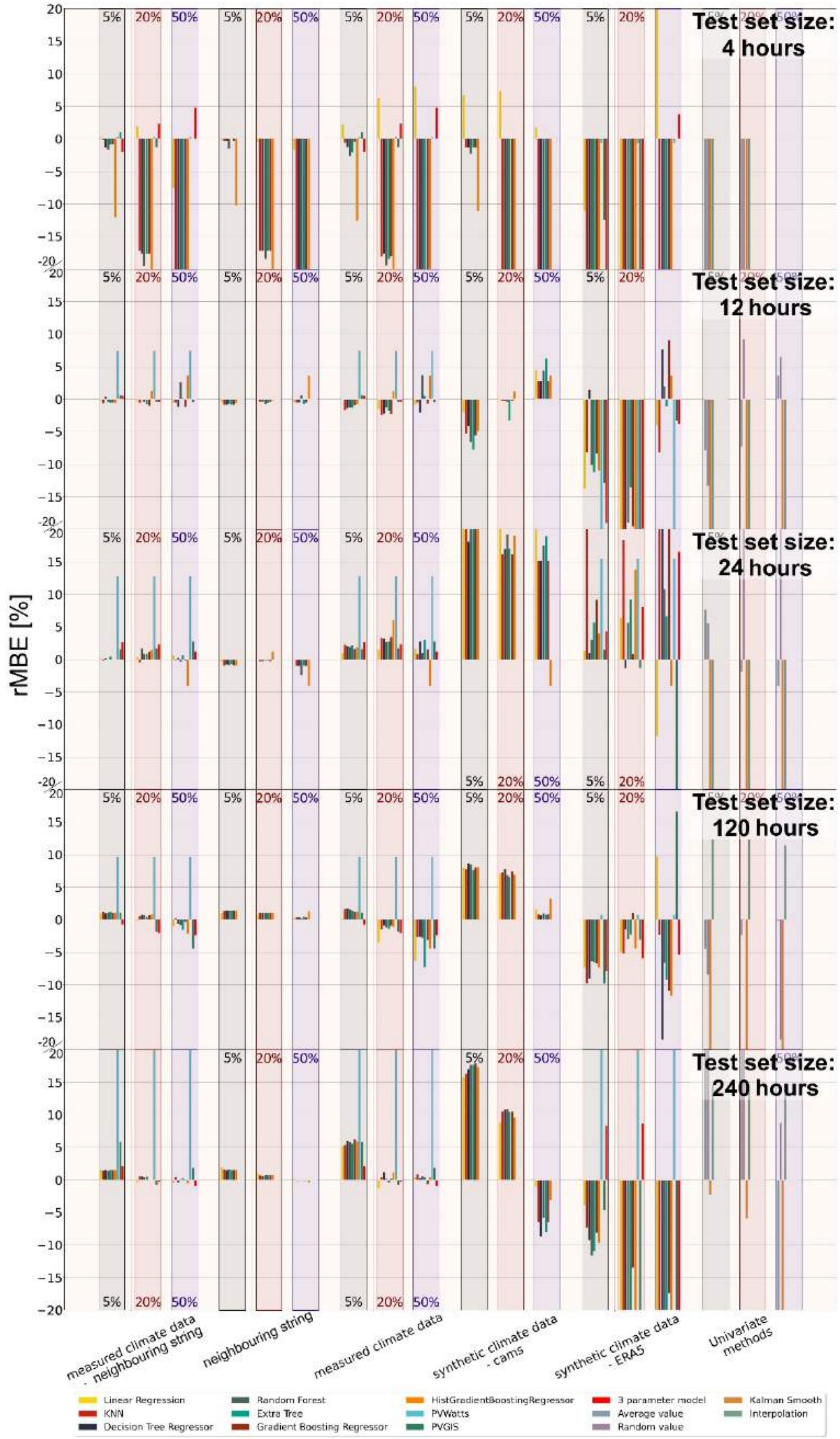

Figure 6: rMBE [%] of individual predictors for training/test ratio of 95/5, 80/20 & 50/50 for test set sizes of 4 hours, 12 hours, 24 hours, 120 hours & 240 hours