Machine Learning and Data Augmentation Techniques to Cope With Solar Data Scarcity to Simulate PV Generation in Mountain Environments

Aleksandr Gevorgian – Free University of Bozen-Bolzano, Italy – aleksandr.gevorgian@student.unibz.it
Giovanni Pernigotto – Free University of Bozen-Bolzano, Italy – giovanni.pernigotto@unibz.it
Andrea Gasparella – Free University of Bozen-Bolzano, Italy – andrea.gasparella@unibz.it

Abstract

Accurate prediction of Global Horizontal Irradiance (GHI) is crucial for optimizing solar power generation systems, especially in mountainous regions characterized by complex topography and specific microclimates. These areas face significant challenges due to limited availability of reliable data and accuracy issues stemming from the dynamic nature of the atmosphere and local weather conditions. This scarcity of precise GHI measurements impedes the development of accurate solar energy prediction models, affecting both economic and environmental aspects.
In this framework, this paper proposes a novel methodology to address data scarcity challenges in solar energy prediction, particularly focusing on Alpine regions. We employ machine learning techniques such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) regressors, in conjunction with synthetic data generation, to predict GHI. To assess our approach's accuracy, we selected Bolzano as a case study and modelled the PV AC power outputs before and after optimizing GHI data.

1. Introduction

Solar energy stands as a pivotal pillar of sustainable development, as underscored by the International Energy Agency (IEA, 2020). Consequently, accurate solar irradiance prediction plays a central role in harnessing this renewable resource efficiently, in particular when it comes to the Global Horizontal Irradiance (GHI). However, obtaining accurate GHI measurements presents formidable challenges, rooted in the dynamic nature of the atmosphere and the variability of weather conditions (Qazi et al., 2019). Furthermore, assembling a comprehensive and precise dataset of GHI measurements is often a resource-intensive endeavour, demanding expensive equipment and periodic maintenance (Kalogirou, 2009). Indeed, inaccuracies in GHI measurements can reverberate through the entire solar energy prediction process, carrying ramifications that ripple through economic and environmental aspects (Kosmopoulos et al., 2015).
Machine learning (ML) techniques have gained traction in the field of solar energy prediction, holding significant promise for technological advancements (Javed et al., 2019). However, a major concern arises: the effectiveness of these ML algorithms relies heavily on the quantity and quality of the training dataset (Javed et al., 2019). Frequently, the scarcity of accessible data becomes a bottleneck, hindering the creation of accurate solar energy prediction models. Therefore, there is a pressing demand for cost-effective approaches capable of efficiently obtaining and utilizing GHI data to enhance the accuracy of solar energy prediction models.
In response to this pressing challenge, our study introduces a new approach designed to enhance the accuracy of GHI predictions, even when confronted with limited datasets. Our method capitalizes on the power of machine learning, specifically the Random Forest (RF) regressor (Breiman, 2001), to identify the optimal distribution of training data based on cloud opacity values — a pivotal factor in GHI measurements. Subsequently, we harness the same RF regressor to construct a new RF model, which generates synthetic data points. These synthetic data points undergo augmentation via techniques such as flipping, rotating, scaling, and the introduction of random noise (Maharana et al., 2022). This augmentation strategy enriches dataset variability, enhanc-

Part of
Pernigotto, G., Ballarini, I., Patuzzi, F., Prada, A., Corrado, V., & Gasparella, A. (Eds.). 2025. Building simulation applications BSA 2024. bu,press.
https://doi.org/10.13124/9788860462022

ing model robustness. In the last step of the proposed approach, we trained and tested the Extreme Gradient Boosting (XGBoost) regressor (Chen & Guestrin, 2016) on the combined structured dataset, which integrates the original and synthetic data via data augmentation techniques.

2. Methodology

2.1 Data Collection

In our study, we started by collecting hourly values of various meteorological and atmospheric quantities and Global Horizontal Irradiance (GHI) during the years 2019 and 2021. We selected four distinct Alpine locations (Fig. 1), with intricate topography and unique microclimates (Ohler et al., 2020): Bolzano (46.50° N, 11.35° E), Aosta Valley (45.75° N, 7.34° E), Locarno (46.16° N, 8.88° E), and Esine (45.92° N, 10.25° E).

Our data collection process relied on two primary sources: local weather stations for GHI data and satellite imagery for other predictor variables. Specifically, for Bolzano, we obtained GHI data from the weather station situated at the Free University of Bozen-Bolzano campus. For the remaining locations, we collected GHI data from weather stations located in close proximity to each respective site. Additionally, we acquired the Actual Meteorological Year (AMY) dataset from Solcast.com (Solcast, n.d.). This dataset encompasses seven crucial meteorological and atmospheric parameters, including air temperature, cloud opacity, precipitable water, relative humidity, surface pressure, wind direction, and wind speed. We also incorporated solar geometry variables, as well as time-related information such as azimuth and zenith angles, year, month, day, and hour, to serve as predictors in our analysis.

2.2 Data Preprocessing

Before analyzing the dataset, data preprocessing was conducted according to (Nugroho et al., 2021). This preprocessing phase included essential data cleansing steps, with a primary focus on the removal of missing values and a thorough identification and treatment of outliers.

Fig. 1 – Map of selected alpine locations
(source: https://en-gb.topographic-map.com/)

2.3 Data Splitting and Optimization

To ensure robust model training and evaluation, we partitioned the dataset into training (1 %) and testing (99 %) subsets, simulating a scenario with limited data available for training, using data from the years 2019 and 2021 for the training and testing subsets, respectively. Subsequently, we embarked on the task of optimizing the training dataset. This optimization process was fueled by the pursuit of the most effective distribution of training data, with the primary goal of maximizing the model's performance in terms of metrics such as the coefficient of determination (R^2) score, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Bias Error (MBE). In essence, we sought to find the ideal arrangement of training data that would yield the best predictive accuracy. To accomplish this, we leveraged the Random Forest (RF) regression model, which allowed us to discern the distribution that corresponded to specific cloud opacity values, a fundamental factor influencing GHI measurements.

2.4 Synthetic Data Generation and Augmentation

The process of generating synthetic data revolved around a Python-based algorithm designed to harness the capabilities of the Random Forest (RF) regressor. The objective was to create synthetic variables of predictors known as input features and generate new predicted GHI data points that closely mimicked real-world conditions while significantly expanding the size of our training dataset.

We created a grid of input feature values that covered the full range of possible values for each feature, utilizing the same feature ranges and resolu-

tions as the original, limited training dataset. This enabled us to construct a more diverse and expansive dataset than the original one. Subsequently, we passed the input features through the trained Random Forest (RF) model to predict the corresponding GHI values. This approach enabled us to create supplementary data points, thus improving the precision of our ML algorithm for GHI prediction. This strategy significantly enriched our training dataset, effectively expanding its size by a factor of up to 200 times its original magnitude.

To further improve the diversity of the dataset, we implemented a range of established data augmentation techniques (Maharana et al., 2022):

- Flipping: mirroring existing data points to introduce variations that capture inverted scenarios, such as changes in solar angles.
- Rotating: applying rotations to data points to simulate different solar angles and azimuths, thereby expanding the dataset's coverage of potential conditions.
- Scaling: introducing scaling factors to data points to represent varying magnitudes of meteorological and atmospheric parameters, effectively diversifying the dataset.
- Introducing Random Noise: injecting controlled random noise into the synthetic data to mimic the inherent variability in real-world atmospheric conditions.

2.5 Model Testing and Evaluation

In the next stages of our methodology, we employed the Extreme Gradient Boosting (XGBoost) regressor (Chen & Guestrin, 2016) as our primary machine learning model. This model was trained using a structured dataset from 2019, which we created by combining the original dataset with the synthetic and augmented data. To evaluate the model's performance thoroughly, we used a testing dataset from 2021, ensuring the robustness of our approach by testing the model with data from a different year.

2.6 PV System Modelling

To understand the impact of our methodology in increasing the accuracy of predicted GHI, we utilized the PVLib library (PVLIB, 2020) to model the AC power output of a photovoltaic (PV) system, focusing on Bolzano as a case study. PVLib is an open-source library that provides a set of tools for simulating the performance of PV energy systems. We modeled the AC power output using three different sets of GHI data: measured GHI, GHI predicted from augmented data, and GHI predicted from scarce data. The GHI data served as the primary input, the impact of which we analyzed before decomposing it into Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI) components for our PV system simulation. The analysis of the impact of decomposition on the PV system performance was not implemented as it is outside the scope of this research. This approach assumes the common scenario where only GHI data is known.

For the estimation of the diffuse horizontal irradiance (DHI) from the predicted GHI, the model by Erbs (1982) was adopted. Furthermore, the Perez model, described in (Perez et al., 1987; Ineichen & Perez, 2002), was implemented to estimate beam and diffuse components on tilted surfaces.

PV System Configuration

The PV system was configured using the following parameters from the PVLib library (PVLIB, 2020):

- *Temperature Parameters*: the open-rack glass-glass temperature model parameters.
- *Module*: the Trina Solar TSM-300DEG5C-07 II module, with efficiency of 18.19 %.
- *Inverter Specifications*: the ABB MICRO-0.25-I-OUTD-US-208 inverter.

To evaluate the accuracy of the modeled AC power output, we employed four statistical metrics: Mean Absolute Deviation (MAD), Root Mean Squared Deviation (RMSD), Coefficient of Variation (CV), and Autocorrelation Function (ACF). These metrics were chosen to assess how well each model captured the smooth transitions in AC power output typically observed in real-world PV systems.

Mean Absolute Deviation (MAD):

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}| \quad (1)$$

Root Mean Squared Deviation (RMSD):

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2]} \quad (2)$$

Coefficient of Variation (CV):

$$\text{CV} = \frac{\sigma}{\bar{y}} \quad (3)$$

Autocorrelation Function (ACF):

$$\text{ACF}(k) = \frac{\sum_{i=1}^{n-k}(y_i - \bar{y})(y_{i+k} - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

Where y_i is the hourly AC power output, $\bar{y}$ is the mean AC power output, σ is the standard deviation and k is the lag.

The primary purpose of using these statistical metrics is to identify which model best captures the gradual changes in AC power output over consecutive hours, thereby minimizing outliers. A model with lower MAD, RMSD, and CV values, combined with a higher ACF, suggests fewer unrealistic fluctuations and smoother transitions in the predicted AC power output.

To gain deeper insight into the accuracy of modeled AC power output under varying GHI conditions, we conducted an analysis of how cloud opacity influences PV AC output.

3. Results and Discussion

3.1 Data Splitting and Optimization

Fig. 2 reports the minimum number of hours with favorable sky conditions necessary to train our machine learning model, enhancing its accuracy in predicting GHI (Sarmas et al., 2022). As can be seen in the figure, various locations across the Alps exhibit notable similarities in terms of hours with corresponding cloud opacity values. This insight provides valuable guidance on pinpointing specific days of the year that require attention for instrument inspection, maintenance, and the collection of training data for ML models to make precise predictions. Furthermore, this knowledge aids in the more accurate calibration of instruments, obviating the need for year-round inspections (Lester & Myers, 2006; Santiago, 2023).

In addition, our approach offers several advantages in terms of data collection. It streamlines data gathering by concentrating resources on days with specific sky conditions crucial for precise predictions (Zellweger et al., 2023). This approach hints at potential optimizations in the allocation of time and resources, which could result in a more cost-effective process. This is especially relevant in remote or hard-to-access locations where data collection can be resource-intensive (Ohler et al., 2020). Moreover, improved data quality arises from the reduced influence of confounding variables like cloud cover or extreme weather conditions, which can introduce inaccuracies into the dataset (Krishnan et al., 2023). Data collected under favorable sky conditions is assumed to yield more consistent and reliable measurements.

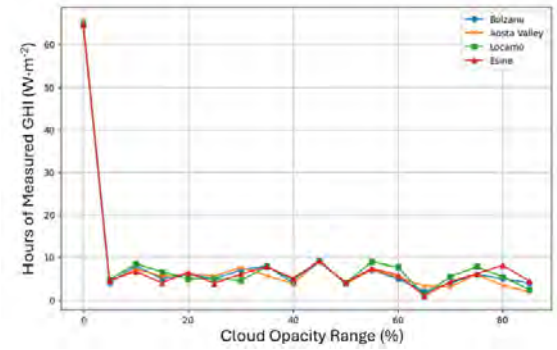

Fig. 2 – Cloud opacity range and corresponding hours of measured GHI

3.2 Models Accuracy and Reliability

Our model has achieved a R^2 score ranging from 0.91 to 0.97 when evaluated against the 2021 testing dataset. This outstanding performance signifies a robust correlation between the predicted and actual GHI values. Importantly, this high level of accuracy has been consistently observed across multiple Alpine locations, as evidenced in Table 1.

Furthermore, the low RMSE, MAE, and MBE values provide strong evidence of the quality of prediction accuracy. Particularly noteworthy is the substantial decrease in RMSE and MAE values when synthetic data augmentation techniques were applied, indicating that our model's predictions closely align with actual GHI values.

To provide a more tangible demonstration of our approach's impact, we included Scatter Plots in Fig. 3, that illustrate the model's performance before and after the application of synthetic data generation and augmentation techniques. The inclusion of synthetic data generation and augmentation techniques has not only improved the model's accuracy but also strengthened its overall reliability. By significantly enlarging our initially limited training dataset, our model now exhibits enhanced abilities to make precise predictions that can generalize effectively. The synthetic data generation process, covering a wide range of meteorological and atmospheric conditions, has equipped the model to adapt to various scenarios and comprehended the intricate patterns of GHI in mountainous regions.

Table 1– Performance metrics for GHI prediction with and without synthetic data generation and augmentation techniques

Method	R^2	RMSE [W m^{-2}]	MAE [W m^{-2}]	MBE [W m^{-2}]
Aosta Valley				
Data Scarcity	0.84	99.75	71.36	39.03
Augmented data	0.97	42.64	20.33	3.13
Bolzano				
Data Scarcity	0.79	114.93	78.61	38.58
Augmented data	0.91	74.61	41.82	3.06
Esine				
Data Scarcity	0.80	110.58	73.93	38.80
Augmented data	0.93	66.40	38.62	3.12
Locarno				
Data Scarcity	0.80	111.75	77.77	38.73
Augmented data	0.92	69.97	38.42	3.06

3.3 AC Power Output Result Analysis

The study aims also to evaluate the smoothness of transitions in AC power output from one hour to the next, which serves as an indicator of the model's accuracy in capturing gradual changes in solar irradiance. This evaluation spans four seasons and considers three scenarios: using measured GHI, predicted GHI from augmented data, and predicted GHI from limited data, as depicted in Fig. 4.

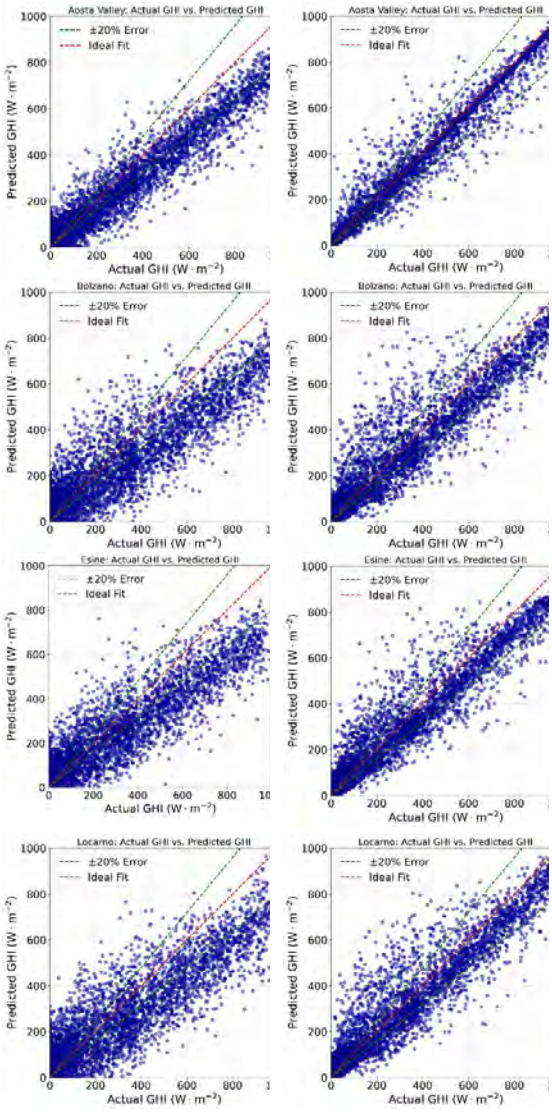

Fig. 3 – Performance comparison scatterplots

Across all seasons, the AC power output modeled with measured GHI consistently exhibits the smoothest transitions with minimal fluctuations. In contrast, using predicted GHI from augmented data shows moderate fluctuations, while using predicted GHI from scarce data exhibits more pronounced irregularities, suggesting lower accuracy.

Table 2 provides a quantitative assessment of each model's performance using statistical metrics (MAD, RMSD, CV, and ACF). We can observe that:

- *AC power output modelled with measured GHI* consistently shows the lowest MAD and RMSD values. It also has the lowest CV values, reflecting stable power output predictions, and the highest ACF values, suggesting smoother transitions.

- *AC power output modelled with augmented GHI* shows slightly higher MAD and RMSD values than measured GHI, suggesting moderate accuracy. CV values are slightly higher but remain relatively stable. ACF values are lower than measured GHI but still indicate relatively smooth transitions.
- *AC power output modelled with scarce GHI* exhibits the highest MAD and RMSD values, indicating higher deviations and less accuracy. CV values are the highest, particularly in winter, indicating the most variability. Low ACF values suggest more abrupt changes.

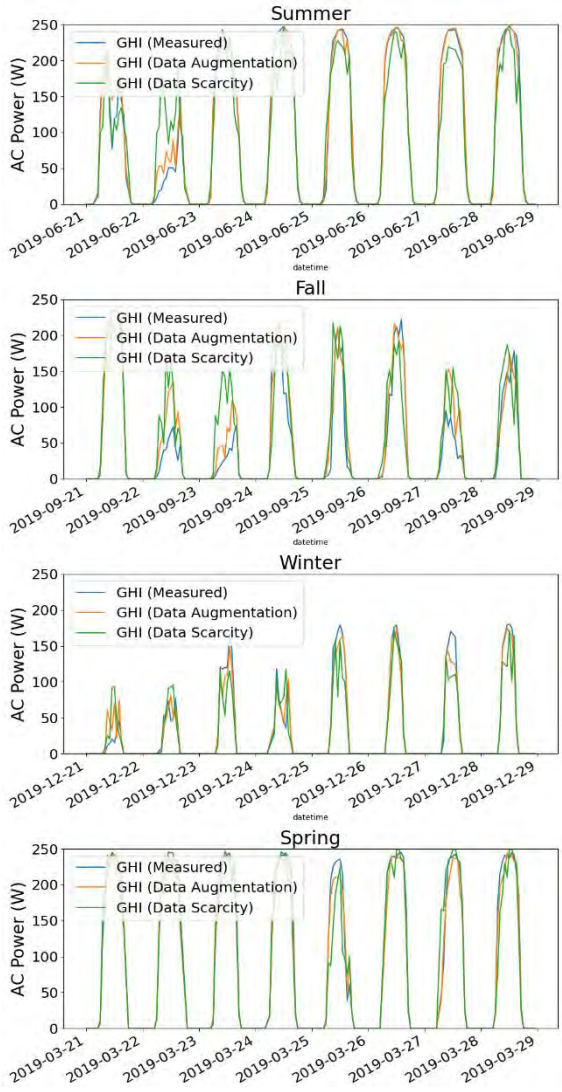

Fig. 4 – Modeled AC Power Output by seasons and GHI models

Table 2– Performance metrics for modeled AC power output with GHI data inputs

GHI data	Season	MAD [W]	RMSD [W]	CV [·]	ACF [·]
Measured	Summer	16.53	30.17	0.30	0.94
Augmented	Summer	18.83	34.35	0.33	0.93
Scarced	Summer	22.07	37	0.36	0.92
Measured	Fall	15.12	28.06	0.46	0.92
Augmented	Fall	16.76	30.46	0.51	0.9
Scarced	Fall	21.36	37.29	0.58	0.86
Measured	Winter	10.12	20.74	0.71	0.89
Augmented	Winter	12.46	25.93	0.79	0.86
Scarced	Winter	14.97	30.93	0.97	0.77
Measured	Spring	18.67	38.71	0.40	0.92
Augmented	Spring	20.16	42.02	0.46	0.91
Scarced	Spring	23.48	47.06	0.49	0.90

Fig. 5 illustrates the relationship between cloud opacity and PV AC output, highlighting the impact of cloud optical properties on power generation. The correlation plot shows a clear inverse relationship between cloud opacity and AC power output. Higher cloud opacity results in lower power output, reflecting reduced solar irradiance.

The model with measured GHI exhibits a more linear and consistent correlation, indicating its effectiveness in capturing the impact of cloud opacity on power output. This consistency further supports its superior performance in modeling smooth transitions. The models with GHI predicted from augmented and scarce data show greater variability in correlation, especially pronounced in the model with GHI predicted from scarce data. This suggests a less accurate representation of cloud effects on power generation. This scatter indicates potential inaccuracies and greater fluctuations in predicted power output.

Analysis of Figs. 4 and 5, along with the statistical metrics in Table 2, highlights that, when using measured GHI, the model consistently produces smoother transitions in AC power output between consecutive hours. This is reflected in its lower MAD, RMSD, and CV values, and higher ACF values, indicating more accurate and stable power output predictions. The model with augmented GHI shows moderate fluctuations and variability, indicating acceptable accuracy, though it is slightly less

accurate than the model with measured one. In contrast, the model with scarce GHI exhibits the highest fluctuations and variability, particularly in less predictable seasons like winter, suggesting significant inaccuracies that may lead to over-/underestimation of power output.

Accurate representation of cloud opacity is therefore crucial for reliable PV power output modeling. Inadequate representation of solar irradiance variability can lead to unrealistic jumps in power output, which are unlikely in real-world scenarios. The augmented model offers a reasonable alternative, while the scarce model's high variability and error rates make it less realistic.

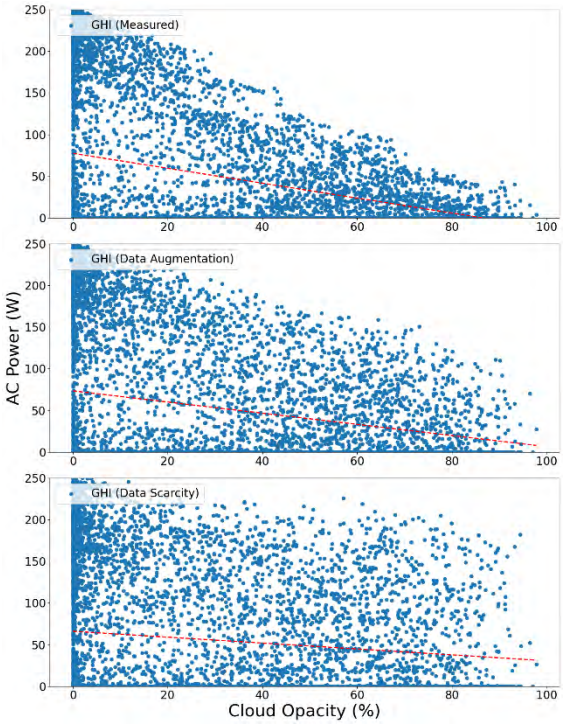

Fig. 5 – Correlation between Cloud Opacity and PV AC output

4. Conclusion

This study addresses the critical challenge of data scarcity in solar energy prediction, particularly in Alpine regions characterized by complex topography and microclimates. Accurate predictions of Global Horizontal Irradiance (GHI) are paramount for optimizing solar power generation. However, limited data availability in such regions poses a significant hurdle to achieving precise predictions. To overcome this challenge, our approach combines machine learning techniques, training data distribution, synthetic data generation, and augmentation.

Integration of synthetic data generation and augmentation techniques to expand the training dataset enhanced the model's ability to generalize and make accurate predictions. Machine learning models achieved high accuracies, with an R^2 score ranging from 0.91 to 0.97 and substantial reductions in RMSE, MAE, and MBE values across various Alpine locations. Findings also suggest that optimizing the distribution of training data based on cloud opacity values can identify specific days with favourable sky conditions for accurate GHI measurements.

The results of PV AC power output modelling suggest that when using GHI data, decomposed into DNI and DHI, predicted with the use of synthetic and augmentation techniques, the model shows moderate fluctuations and acceptable accuracy. In contrast, the use of GHI predicted from scarce data exhibited the highest fluctuations and variability, indicating significant inaccuracies. This highlights the importance of using our approach to increase the accuracy of PV system modelling in alpine and mountainous regions.

However, potential applications of this approach extend beyond traditional design and performance assessment of solar systems. The methodology could improve data collection efficiency, reduce costs, enhance data quality, and aid in instrument calibration. It may also optimize maintenance schedules, reduce downtime, lower maintenance costs, and extend equipment lifespan.

Future research will be dedicated to refining synthetic data generation processes, optimizing the integration of additional meteorological and environmental parameters, and extending the methodology to other regions.

Reproducibility Statement

The authors are committed to ensuring reproducibility and facilitating research by readily sharing source codes upon reasonable request.

Acknowledgement

This research was funded by the internal project of the Free University of Bozen-Bolzano "SOMNE - Bolzano Solar Irradiance Monitoring Network" (CUP: I56C18000930005; CRC Call 2018).

References

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Chen, T., and C. Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Erbs, D. G., S. A. Klein, and J. A. Duffie. 1982. "Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation." *Solar Energy* 28(4): 293–302.

Ineichen, P., and R. Perez. 2002. "A New Airmass Independent Formulation for the Linke Turbidity Coefficient." *Solar Energy* 73: 151-157.

International Energy Agency (IEA). 2020. *Renewables 2020*.

Javed, A., B. K. Kasi, and F. A. Khan. 2019. "Predicting Solar Irradiance Using Machine Learning Techniques." In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, 1458-1462. IEEE.

Kalogirou, S. A. 2009. *Solar Energy Engineering: Processes and Systems*. Academic Press.

Kosmopoulos, P. G., S. Kazadzis, K. Lagouvardos, V. Kotroni, and A. Bais. 2015. "Solar Energy Prediction and Verification Using Operational Model Forecasts and Ground-Based Solar Measurements." *Energy* 93 (Part 2): 1918-1930.

Krishnan, N., K. R. Kumar, and C. S. Inda. 2023. "How Solar Radiation Forecasting Impacts the Utilization of Solar Energy: A Critical Review." *Journal of Cleaner Production* 388: 135860.

Lester, A., and D. R. Myers. 2006. "A Method for Improving Global Pyranometer Measurements by Modeling Responsivity Functions." *Solar Energy* 80(3): 322-331.

Maharana, K., S. Mondal, and B. Nemade. 2022. "A Review: Data Pre-processing and Data Augmentation Techniques." *Global Transitions Proceedings* 3(1): 91-99.

Nugroho, H., N. P. Utama, and K. S. Surendro. 2021. "Normalization and Outlier Removal in Class Center-Based Firefly Algorithm for Missing Value Imputation." *Journal of Big Data* 8: 129.

Ohler, L. M., M. Lechleitner, and R. R. Junker. 2020. "Microclimatic Effects on Alpine Plant Communities and Flower-Visitor Interactions." *Scientific Reports* 10(1): 1366.

Perez, R., R. Seals, P. Ineichen, R. Stewart, and D. Menicucci. 1987. "A new simplified version of the Perez diffuse irradiance model for tilted surfaces." *Solar Energy* 39(3): 221–232.

PVLIB Python. 2020. "v0.7.2." PVLIB Python. Accessed May 11, 2024. https://github.com/pvlib/pvlib-python

Santiago, O. 2023. "Mastering Weather Predictions: Unleash the Power of AI with LSTM Deep Learning Models for Accurate Temperature Forecasts." *Towards Data Science*.

Sarmas, E., N. Dimitropoulos, V. Marinakis, Zoi Mylona & Haris Doukas. 2022. "Transfer Learning Strategies for Solar Power Forecasting under Data Scarcity." *Scientific Reports* 12: 14643.

Qazi, A., et al. 2019. "Towards Sustainable Energy: A Systematic Review of Renewable Energy Sources, Technologies, and Public Opinions." *IEEE Access* 7: 63837-63851.

Zellweger, F., Sulmoni, E., Malle, J. T., Baltensweiler, A., Jonas, T., Zimmermann, N. E., Ginzler, C., Karger, D. N., De Frenne, P., Frey, D., and Webster, C. 2023. "Microclimate Mapping Using Novel Radiative Transfer Modeling." EGUsphere Preprint. https://doi.org/10.5194/egusphere-2023-1549