**Proceedings of the ASME 2024&
Energy Sustainability
ES2024
July 15–17, 2024, Anaheim, CA, USA**

ES2024-130057

MACHINE LEARNING FOR FORECASTING SOLAR IRRADIANCE USING SATELLITE AND LIMITED GROUND DATA

Jocellyn Luna, Alex Chancusig, Jose Cordova-Garcia, Guillermo Soriano
Escuela Superior Politécnica del Litoral - ESPOL University
Guayaquil, Ecuador
Email: {jmluna, abchancu, jecordov, gsorian}@espol.edu.ec

ABSTRACT

We propose a machine learning-based methodology suitable for use in regions where full availability of meteorological data is lacking. In response to the growing global energy demand and the need to reduce the carbon footprint generated by fossil fuels, solar energy has become the central renewable energy source through small and large-scale photovoltaic systems. Solar energy production depends on the available amount of solar irradiation in a given area, considering the influence of external factors such as environmental conditions, seasons, geographic location, and others. Many regions in the global south do not maintain an updated solar irradiance database, limiting an efficient solar potential analysis. Meteorological stations can provide high-precision ground measurements. However, such stations cover specific locations that cause a spatial data availability problem. This problem can be solved using satellite data, which provides zone-wise spatial information. However, the measurements made by satellites are not as accurate as those obtained on the ground. In this work, we opted for tree-based regression models to map satellite to meteorological stations' Global Horizontal Irradiation (GHI). Then, we propose using these data generated through regression for GHI forecasting to facilitate applications such as sizing and operating photovoltaic and solar thermal systems. We illustrate this methodology through a case study where we used the generated dataset along with an LSTM neural network showing better performance in forecasting short-term irradiance when compared to an statistical baseline.

Keywords: AI for Sustainability, Renewable Energy

NOMENCLATURE

GHI Global Horizontal Irradiance.
LSTM Long short-term memory.
DMQ Distrito Metropolitano de Quito. Municipality in charge of weather stations.

1 INTRODUCTION

Solar energy, as a renewable source with great potential, plays a fundamental role in developing a sustainable energy system free of carbon emissions [1–3]. The efficiency of photovoltaic and solar thermal systems depends on the availability of solar radiation incident on the Earth's surface. Technological progress in recent decades has contributed to the increase in the energy efficiency of solar panels, added to the low production costs, and has made photovoltaic energy an essential commitment of many countries to produce electricity economically [1, 3]. According to the Global Market Outlook for Solar Power 2021 – 2025 report [4], 39% of the net power installed capacity in 2020 corresponds to the solar component in countries such as China (33%), the United States (12%), Japan (9%) and Germany (7%), showing significant steps in promoting environmental awareness through the use of renewable energy. Therefore, the ability to accurately predict solar irradiance is essential for the efficient planning, operation, and management of solar-based energy systems [5]. Recent efforts have showcased the importance of analyzing solar irradiation through interactive tools. These tools often focus on a specific country or region, such as

the Photovoltaic Geographical Information System in European territory [6] or the "Explorador de Energía Solar" in Chile [7]. These types of tools rely on the availability of data at the location of the intended operation of photovoltaic systems.

Meteorological stations can provide high-precision ground measurements, facilitating data-driven analysis and predictions of solar energy systems. Specifically, global horizontal irradiance forecasting is crucial for planning and operating solar-based renewable energy systems. However, such stations cover specific, often sparse and scarce, locations that, when not available for a given zone of interest, cause a spatial data availability problem. Esposito et al. [8] compared irradiance data from ground-based sources and satellites, obtaining a maximum Mean Relative Error (MRE) of $<7\%$ for hourly averaged Global Horizontal Irradiance (GHI) and $<9\%$ for daily averaged GHI. These results confirmed data reliability for hourly and daily intervals in Copernicus Satellite. Thus, satellite data can be leveraged in areas where ground observations are not available, providing zone-wise spatial information at a specific resolution. Nevertheless, the measurements made by satellites are not as accurate as those obtained on the ground.

Therefore, several studies in the forecast of solar irradiation [9–11] are based on data recorded locally through meteorological stations distributed in different locations and integrated with the satellite data. In this work, we propose a two-stage workflow to incorporate satellite measurements to not only improve the accuracy of the data but also to overcome data availability limitations of scarce meteorological stations. Specifically, we showcase the practical implications of the workflow through an illustrative case study centered in a region of the global south where the problem of data availability from meteorological stations is often more prevalent.

In the following sections, we describe the proposed workflow. The first stage includes data selection from weather stations and satellites, its temporality and spatiality, and data cleaning of outliers values using a machine learning method shown in Section 2.1. Then, in Section 2.2, we describe how to interpolate the weather station data with the satellite data using tree-based regression models to leverage the precision of ground measurements and overcome the spatial availability problem. Then, the second stage in Section 2.3 describes the usage of statistical and deep learning methods to produce the solar radiation forecast. Finally, we illustrate this methodology through a case study in Section 3 and discuss its effectiveness.

2 TWO STAGE MIXED-DATA GHI FORECASTING

This work seeks to develop a methodology that allows data integration of highly accurate local information sources, such as meteorological stations, with the global availability of solar irradiation records from satellite sources through machine learning techniques. Thus, this methodology enables the availability of an updated, continuous, and accurate dataset for improved forecasting, even in locations where ground measurements are scarce. Here, we present the proposed GHI data generation and forecasting stages along with a pre-processing step before diving into details through the implementation of the proposed methodology in a motivating case study in the next section.

2.1 Dataset preprocessing

For satellite data, we are going to focus on open sources such as the National Solar Radiation Database (NSRDB) [12] in the Americas by the National Renewable Energy Laboratory (NREL). This dataset includes data on longitude, latitude, Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance (DHI), Global Horizontal Irradiance (GHI), temperature, wind speed and direction, atmospheric pressure, humidity, and precipitation. The variables DNI and GHI are mainly used for predicting solar energy generation in specific regions. However, GHI is essential for overall system performance, while DNI exhibits better accuracy during clear-sky conditions [13]. Lopes et al. [14] concluded that short-term forecast models perform accurately using GHI, yet DNI predictions were overestimated mainly due to cloud representation issues. Thus, we will use all available variables from the satellite data to provide our models with different options to extract predictive power. These variables are provided globally for 4x4 km^2 area resolution.

Following the motivation described in the previous section, we use meteorological variables to complement the satellite data. Commonly used variables in forecasting studies include temperature, global horizontal irradiance, relative humidity, wind direction, and speed. Due to its precision, the global horizontal irradiance (GHI) from meteorological stations is the target variable to be used in solar installations [15–20].

Beyond availability, a common issue in the global south's meteorological stations, particularly in South America, is that data from these stations have been affected by various external factors, such as dirt accumulation, lack of maintenance, and sensor failures. Thus, defining a preprocessing step is necessary before performing any predictive analysis. We guide the definition of preprocessing steps by analyzing data that contain such possible issues.

Specifically, we analyzed meteorological data obtained from weather stations of the Environmental Secretary of Quito Municipality (DMQ) [21] in the province of Pichincha, Ecuador. This meteorological network consists of 26 weather stations recording data on relative humidity, atmospheric pressure, Diffuse Horizontal Irradiance (DHI), wind speed, direction, and the target variable of Global Horizontal Irradiance (GHI) at a time resolution of 60 minutes.

Besides traditional treatment of missing values [22], our analysis revealed that the key preprocessing step for the success of the forecasting workflow was to identify outliers. Thus, Isola-

tion Forest, a method based on classification and regression trees, is proposed [23] to identify anomalies in the meteorological stations' time series.

After identifying the outliers, we remove them and then use a spline interpolation with second-degree polynomials to replace them. Thus, we treat the outliers while maintaining the natural behavior of solar radiation without drastically altering the irradiance peaks. Next, we described how this preprocessed dataset is integrated with satellite data.

2.2 Stage 1: Integrated dataset generation

Despite their high precision, as explained before, a forecast model relying on the data from meteorological stations limits its applicability as these data are often not available due to a lack of geographical coverage. Moreover, we also identified that even when the weather station is available, the data often exhibit considerable amounts of incomplete records that cannot be treated through traditional imputation methods that could be included in the preprocessing step. In contrast, the availability of satellite data is limited by its precision. To take advantage of each data source, we propose the integration of both datasets using regression techniques.

The data integration process required a margin error analysis between them. Therefore, we match both datasets according to their time interval and location. As the satellite data has a resolution of 4 km, we matched each satellite grid reference location to the closest point to each meteorological station. Thus, the offset between the weather station data and the satellite data has a maximum distance of 2 km.

Due to their ability to represent non-linear relationships in data, we opted for two tree-based regression models to map the data from satellite to meteorological stations' GHIs: Random Forest and Gradient Boosting Machines. These types of ensemble tree methods are commonly used in the machine learning literature and are often the top-performing model families across many tasks [24]. Moreover, boosted trees can handle data issues such as missing samples by construction. Training a regression model for each available station enables the generation of meteorological-like GHI datapoints at locations distant from them. Thus, using these data allows us to create a grid of GHI values that aligns with the temporal and spatial coverage of the satellite data while maintaining the accuracy of meteorological data, *extending* the coverage of the latter.

In the previous step, a model is obtained for each weather station. We then proceed to evaluate the performance of each model at randomly chosen locations on the region of interest to identify the best-performing models.

Finally, an ensemble model is created by selecting the best models and combining their predictions. This ensemble approach enhances overall performance and generalization capabilities. With this approach, an integrated GHI dataset is generated by combining the accuracy of measurements from weather stations with the spatial information obtained from satellite data.

2.3 Stage 2: Regression-based forecasting

After obtaining the integrated dataset of solar irradiation in the area of interest, we can use the generated meteorological-like GHI in the forecasts that will be used for the planning and operation of photovoltaic systems. Hence, this section describes suitable techniques for predicting solar irradiation.

In the case of forecasting GHI, it helps to consider light cycles and their changes over the days. We propose to use an LSTM-based model from the machine learning literature. LSTM (Long Short-Term Memory) [25] is a type of recurrent neural network (RNN) architecture specifically designed to address the limitations of traditional RNNs. Traditional RNNs are limited when attempting to learn dependencies from long sequential data. In contrast, LSTM incorporates gates that enable the retention of information over extended sequences.

Furthermore, our workflow includes the use of a statistical approach, the Holt-Winters algorithm [26], as a benchmarking method to be compared with the machine learning model. The Holt-Winters method is used widely for its ability to capture the trend behavior of a time series depending on the importance given to the age of the data and the speed of change in the trend of the series.

For the LSTM model, it is necessary to sequentially split the dataset by choosing a specific sequence length or window. The goal is for the model, given a GHI sequence spanning several consecutive days, to predict the irradiance for one or more days in the future. To enhance the model's performance, we vary the window size, the number of neurons, and the prediction horizon. This approach helps identify the optimal hyperparameter configuration, as well as the limitations of the implemented model.

Finally, we proposed combining several individual models to create an ensemble model for irradiance forecasting, aiming for improved generalization. Since individual models were employed to forecast specific points within an irradiance grid, concerns may occur regarding the generalization of proposed regression models. Ensemble models address this by combining the predictive power of possibly high variance individual models, helping generalization. [27].

3 CASE STUDY

To better detail the methodology presented in the previous section, we describe the application of the data integration and forecasting workflow for a particular region in Ecuador. This region illustrates the problem of meteorological data availability that motivates this work. Moreover, despite the country's geographical position and high potential for solar energy generation, Ecuador exemplifies a common problem of lack of standardized

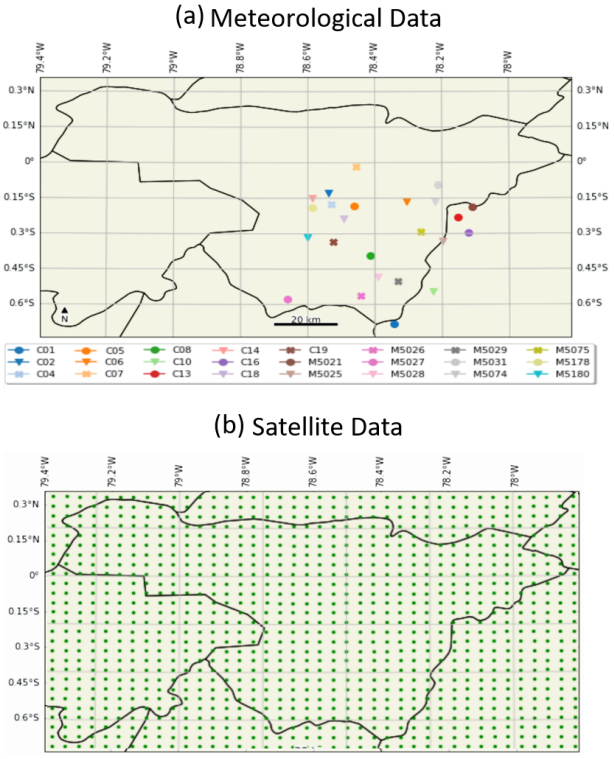

FIGURE 1. DATASETS SELECTED: (A) SATELLITE DATA, (B) METEOROLOGICAL DATA

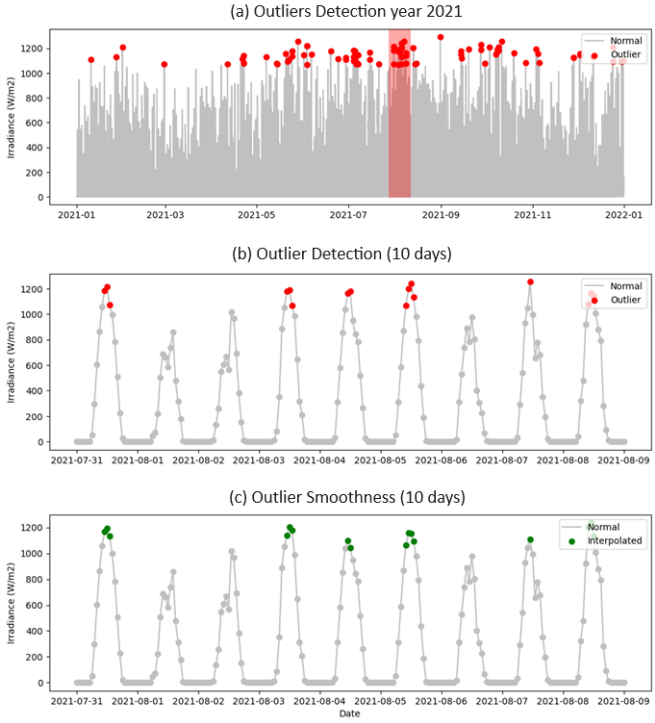

FIGURE 2. OUTLIERS DETECTION USING ISOLATION FOREST (M5029 WEATHER STATION): (A) OUTLIER DETECTION FROM YEAR 2021, (B) OUTLIER DETECTION FROM 10 DAYS, (C) OUTLIER INTERPOLATION SMOOTHNESS FROM 10 DAYS

and centralized data on solar data [28–30] in the global south. This lack of information has represented a significant challenge in the planning process of solar energy projects [31].

The selected region corresponds to the province of Pichincha. The DMQ dataset [21] has 24 weather stations located throughout the province (see Fig. 1.a). The figure shows that the weather station network does not cover the entirety of the province outlined in black. Moreover, the southeast of the province shows a relatively dense deployment of stations, while most of the province is not being monitored.

To better understand the spatial coverage of these stations, we will contrast the deployment resolution with the reference locations in the records of satellite data. As the satellite dataset has a 4 km resolution, the extension of the province of Pichincha was divided into equidistant points every 4 km, resulting in a matrix comprised of 1080 coordinate points constructed from the combination of 27 latitudes and 40 longitudes. The limits considered in the area of interest are the North limit (0.3595), South limit (-0.7391), East limit (-77.8176), and West limit (-79.3953), as is shown in Fig. 1.b. Fig. 1 allows us to contrast the availability and coverage of meteorological and satellite data. When compared to satellite data, the figure shows that, besides being concentrated in the southeast of the province, the 24 meteorological stations fall short of providing measurements for the province.

Then, we obtained data from the satellite and meteorological stations within this area of interest for the year 2021. This year provided the most complete data records of ground measurements, allowing us to focus on addressing the data availability problem due to the scarcity of weather stations and not major sensor failures or disconnection of the stations. Then, as detailed in Section 2.1, we used the Isolation Forest method to detect outliers within the time series corresponding to solar radiation for each station. This method

To illustrate this preprocessing step, Figure 2 shows an example of the detection carried out on the data corresponding to the *El Carmen station (M5029)*. Fig. 2.a shows the complete year of data and the outliers identified in the whole dataset. In the middle and bottom panes (Fig. 2.b and 2.c), we zoomed in a few days to highlight the interpolation used to smooth the identified outliers. Next, we will describe how the data generation stage integrates satellite data and the data from the weather stations.

3.1 Generating province coverage from limited weather stations

After the preprocessing step, we generated an integrated dataset based on ground (weather/meteorological stations) and satellite data, following Stage 1 described in 2.2. For the tree-based methods we used the libraries: Skforecast and LightGBM for the Random Forest and Gradient Boosting Trees implementations, respectively.

Skforecast [32] is a library based on Scikit-Learn [32], developed specifically to use regression models applied to forecast models, with the particularity of incorporating tools for preprocessing the time series before its training. Within the library, different types of cross-validation of the time series before training are available. The most representative being readjustment with fixed origin, readjustment with fixed training size, and validation without readjustment.

LightGBM [33], another widely used model for time series prediction stands out for its precision and rapid training. It uses two concepts: GBDT (Gradient Boosting Decision Trees) and GOSS (Gradient-based onesided sampling). The latter is responsible for selecting gradients in each instance that improve the model's performance and discarding those gradients that are less representative. This process allows the acquisition of models with good performance without the necessity for deep and complex trees.

Then, for the first stage, the predicted variable was the GHI, measured by meteorological stations, using nine features of the satellite data as the regression predictors: GHI, DHI, DNI, temperature, pressure, humidity, precipitation, wind speed and direction.Even though GHI can be calculated using DHI and DNI measurements, tree-based algorithms typically are not sensitive to correlated features. Therefore, all three measurements were used as predictors.

Satellite features were matched with the ground data corresponding to the closest point to each meteorological station. Thus, for evaluating the regression models, we only use the available ground location and its corresponding satellite area reference location. The segmentation used was 80% data for training and the remaining 20% to evaluate the proposed models.

We inspected the results of each ground GHI prediction corresponding to each meteorological station. In Tab. 1, we show an example of the regression results for a ground station. Both regression models show satisfactory results in modeling the GHI ground data using satellite data, with LGBM being the best-performing model. Then, we generate 24 LGBM regression models for each DMQ meteorological station. Next, the challenge is to use each model to obtain GHI values for the 1080 satellite data points that correspond to the specified area of interest, resulting in the integrated dataset for the province.

From the initial GHI 1080 point grid obtained from the satellite data (see Fig. 1.b), we selected 64 points uniformly distributed to validate how the method regression models are used,

TABLE 1. TREE-BASED REGRESSION FOR GROUND GHI PREDICTION (M5075 WEATHER STATION)

Model	RSME	MAE	R2
RF	63.77	32.29	0.89
LGBM	58.14	29.71	0.91

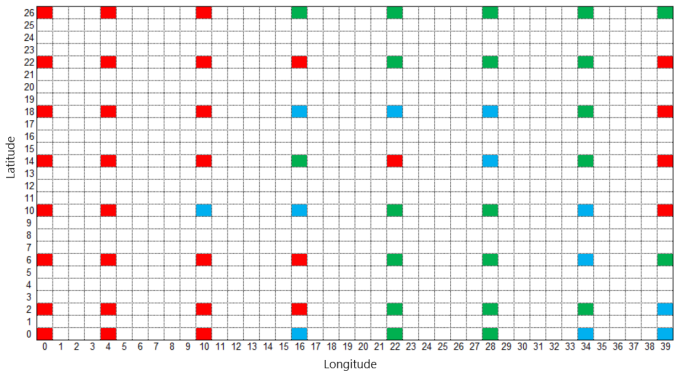

FIGURE 3. INITIAL EVALUATION OF REGRESSION AT LOCATIONS WITHOUT GROUND DATA

particularly for regions far from meteorological station locations.

It is crucial to highlight that not all the selected data points feature ground measurements. Consequently, domain experts conducted a manual classification process, considering various factors such as identifying outliers using the Isolation Forest method, detecting inconsistencies in sunlight distribution at unexpected periods, and identifying irradiance forecasting of zero during nocturnal cycles. The manual classification process provided a robust analysis of the forecasting quality, which is not easily characterized using a single metric, and allowed to keep domain experts involved in the proposed data-driven procedure.

Figure 3 represents the points where we analyzed the behavior of the regression. After inspection, we rated performance and color-coded it: good (green), acceptable (blue), and inconsistent (red), for the results for visualization purposes.

Examining the representation of the regression results, the majority of the points rated as *good* correspond to locations close to the actual location of the 24 meteorological stations considered. As expected, as the distance from a meteorological station increases, performance in an area where we want to generalize the ground measurement decreases.

To overcome this challenge, we examined the potential of combining the information available in more than one weather station to enhance the prediction of ground GHI in locations where performance is not adequate. First, we evaluated each model's effectiveness using irradiance measurements weather stations with the most complete data, which correspond to 9 locations depicted by their DMQ code in the y-axis of Fig. 4.

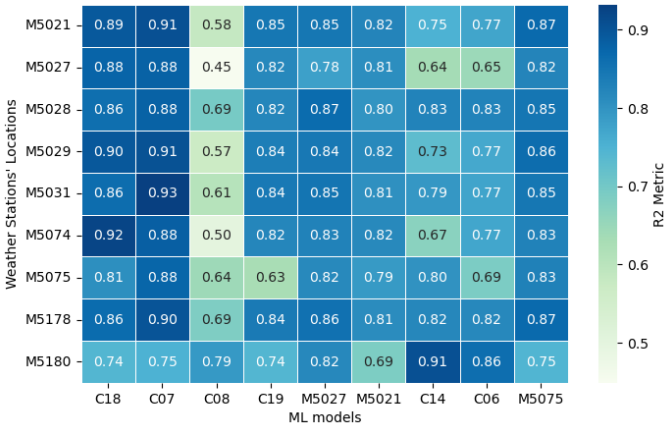

FIGURE 4. IDENTIFYING TOP PREDICTING MODELS

Despite our methodology having a data preprocessing step and the ability of the selected LightGBM to inheritely handle missing data, we want to analyze the learning capabilities of the 24 stations. Thus, by selecting these stations, we can analyze the predictive power of the other 24 stations without them having to overcome the noise induced by data issues. We analyze how the 24 regression models perform when predicting the ground GHI of the 9 *clean data* locations.

This strategy aims to discover regression models with high generalization capabilities. Among the 24 regression models, the top-performing models are shown in the figure's x-axis using their station code. Specifically, the models associated with stations C07, C18, M5075, M5027, and C19 demonstrated a satisfactory capacity to generalize irradiance data in the test areas. Following this result, we combined the selected models and averaged their regressions to create a more robust ensemble model to be used to generate province-wide coverage through ground GHI predictions.

3.2 GHI forecast results

After generating the GHI province dataset through the ensemble model, we evaluated the irradiance forecast of the proposed models in Section 2.3: Holt-Winters and LSTM. The dataset consisted of 1080 points located every 4 km, represented in a 40x27 matrix, where we selected the model whose station was closest to each point in the matrix. The segmentation used 80% data for training and the remaining 20% testing. In the forecast models, the night cycle data was excluded due to its potential to introduce bias. Consequently, the dataset used considered data from 8:00 a.m. to 5:00 p.m., providing 10 hours of GHI per day. Thus, the following tests aim to forecast approximately 40 days in the future. Then, we discuss representative results of a randomly selected location. A graphical representation of the forecast is shown in Fig. 5.

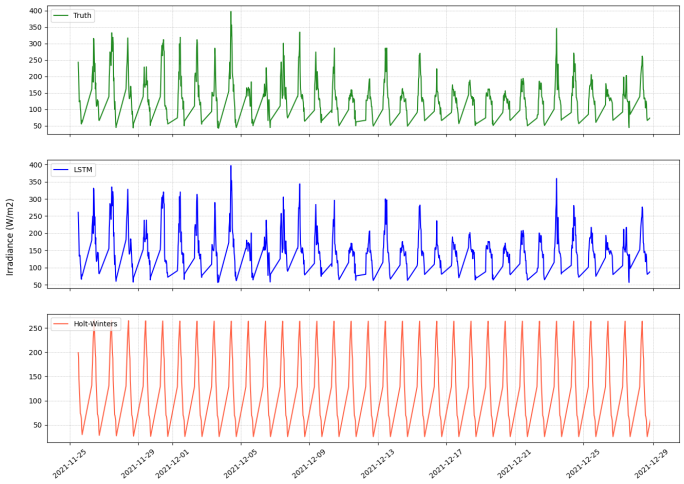

FIGURE 5. GHI FORECASTING

TABLE 2. IRRADIANCE FORECAST MODELS METRICS

Model	RSME	MAE	R^2
Holt-Winters	76.100	39.310	0.836
LSTM	12.453	10.082	0.985

The Holt-Winters algorithm was configured using the hyperparameters: additive seasonality, variable damping, and 48-hour seasonality analysis periods. The LSTM architecture used two LSTM layers with ten neurons in each layer and a dropout of 0.1 between them, where the output forecast uses one step and is subsequently concatenated to generate the output series. We selected a reference point in the test set to illustrate the results in Fig. 5. We can identify the difficulty of the Holt-Winters algorithm to model the behavior of solar GHI as the data gets older in time by comparing the top and bottom panes. While the seasonality of the GHI cycles seems to have been captured by the model, the different trends present in the series are not present in the forecast. In contrast, the middle pane illustrates that the forecast made by the LSTM model can satisfactorily represent the real behavior of the irradiance. Furthermore, we inspect the performance on the test set by evaluating the forecasting metrics listed in Tab. 2. Consistent with the graphical results the metrics show a notable reduction in the RSME when using the proposed LSTM.

4 CONCLUSIONS

We described a methodology for improving irradiance data forecasting by combining ground-based and satellite datasets. The SKforest model played a crucial role in reducing the occurrence of outlier values in the ground dataset without compromising the training data for the models. The GBM model demon-

strated strong performance in learning a predictive relationship of satellite data to ground data, creating a grid of meteorological-like irradiance data. This approach maintains both the temporal and spatial coverage provided by satellite data while preserving the accuracy of meteorological data. Additionally, the ensemble of stations model improved the capability of generating GHI data when meteorological stations are limited, enabling the usage of ground-like irradiance values to be used in the forecasting models. Finally, in the forecasting phase, the LSTM model produced results that closely align with the actual values experienced in the short term.

For future work, we propose to use transformer models capable of handling both data types, satellite and weather station, facilitating the forecasting of irradiance without the need for separate stages. This approach aims to identify concise models that can accelerate the outcome of results within the proposed methodology.

REFERENCES

[1] Ugli, T. J. T., 2019. "The importance of alternative solar energy sources and the advantages and disadvantages of using solar panels in this process". *International Journal of Engineering and Information Systems (IJEAIS)*.

[2] Guangul, F. M., and Chala, G. T., 2019. "Solar energy as renewable energy source: Swot analysis". In 2019 4th MEC international conference on big data and smart city (ICBDSC), IEEE, pp. 1–5.

[3] Ghadami, N., Gheibi, M., Kian, Z., Faramarz, M. G., Naghedi, R., Eftekhari, M., Fathollahi-Fard, A. M., Dulebenets, M. A., and Tian, G., 2021. "Implementation of solar energy in smart cities using an integration of artificial neural network, photovoltaic system and classical delphi methods". *Sustainable Cities and Society,* ***74***, p. 103149.

[4] Europe, S., 2021. "Eu market outlook for solar power 2021–2025". *Solar Power Europe: Brussels, Belgium*.

[5] Al-Shahri, O. A., Ismail, F. B., Hannan, M., Lipu, M. H., Al-Shetwi, A. Q., Begum, R., Al-Muhsen, N. F., and Soujeri, E., 2021. "Solar photovoltaic energy optimization methods, challenges and issues: A comprehensive review". *Journal of Cleaner Production,* ***284***, p. 125465.

[6] European Commission, J., 2022. Photovoltaic geographical information system.

[7] Sandoval Vilches, M. J. L., 2021. "Introducción de la energía solar térmica para reemplazar el uso de combustibles fósiles en procesos industriales en chile". *Universidad de Chile*.

[8] Esposito, E., Leanza, G., and Di Francia, G., 2024. "Comparative analysis of ground-based solar irradiance measurements and copernicus satellite observations". *Energies,* ***17***(7), p. 1579.

[9] Alkhayat, G., and Mehmood, R., 2021. "A review and taxonomy of wind and solar energy forecasting methods based on deep learning". *Energy and AI,* ***4***, p. 100060.

[10] Jebli, I., Belouadha, F.-Z., Kabbaj, M. I., and Tilioua, A., 2021. "Prediction of solar energy guided by pearson correlation using machine learning". *Energy,* ***224***, p. 120109.

[11] Atique, S., Noureen, S., Roy, V., Subburaj, V., Bayne, S., and Macfie, J., 2019. "Forecasting of total daily solar energy generation using arima: A case study". In 2019 IEEE 9th annual computing and communication workshop and conference (CCWC), IEEE, pp. 0114–0119.

[12] Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., and Shelby, J., 2018. "The national solar radiation data base (nsrdb)". *Renewable and sustainable energy reviews,* ***89***, pp. 51–60.

[13] Ramírez, L., and Vindel, J., 2017. "Forecasting and nowcasting of dni for concentrating solar thermal systems". *Advances in concentrating solar thermal research and technology*, pp. 293–310.

[14] Lopes, F. M., Silva, H. G., Salgado, R., Cavaco, A., Canhoto, P., and Collares-Pereira, M., 2018. "Short-term forecasts of ghi and dni for solar energy systems operation: assessment of the ecmwf integrated forecasting system in southern portugal". *Solar Energy,* ***170***, pp. 14–30.

[15] Narvaez, G., Giraldo, L. F., Bressan, M., and Pantoja, A., 2021. "Machine learning for site-adaptation and solar radiation forecasting". *Renewable Energy,* ***167***, pp. 333–342.

[16] Ordoñez-Palacios, L.-E., Bucheli-Guerrero Ph, V.-A., Ordoñez-Eraso Ph, H.-A., et al., 2020. "Predicción de radiación solar en sistemas fotovoltaicos utilizando técnicas de aprendizaje automático". *Revista Facultad de Ingeniería,* ***29***(54).

[17] Pasion, C., Wagner, T., Koschnick, C., Schuldt, S., Williams, J., and Hallinan, K., 2020. "Machine learning modeling of horizontal photovoltaics using weather and location data". *Energies,* ***13***(10), p. 2570.

[18] Urraca, R., Antonanzas, J., Alia-Martinez, M., Martinez-de Pison, F., and Antonanzas-Torres, F., 2016. "Smart baseline models for solar irradiation forecasting". *Energy conversion and management,* ***108***, pp. 539–548.

[19] de Freitas Viscondi, G., and Alves-Souza, S. N., 2021. "Solar irradiance prediction with machine learning algorithms: a brazilian case study on photovoltaic electricity generation". *Energies,* ***14***(18), p. 5657.

[20] Kayri, M., Kayri, I., and Gencoglu, M. T., 2017. "The performance comparison of multiple linear regression, random forest and artificial neural network by using photovoltaic and atmospheric data". In 2017 14th International Conference on Engineering of Modern Electric Systems (EMES), IEEE, pp. 1–4.

[21] FONAG, E. ., 2021. Anuario hidrometeorológico 2021: Red integrada de monitoreo hidrometeorológico epmaps - fonag.

[22] Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., and Stork, J., 2015. "Comparison of different methods for univariate time series imputation in r". *arXiv preprint arXiv:1510.03924*.

[23] Liu, F. T., Ting, K. M., and Zhou, Z.-H., 2008. "Isolation forest". In 2008 eighth ieee international conference on data mining, IEEE, pp. 413–422.

[24] Grinsztajn, L., Oyallon, E., and Varoquaux, G., 2022. "Why do tree-based models still outperform deep learning on typical tabular data?". *Advances in Neural Information Processing Systems,* ***35***, pp. 507–520.

[25] Yu, Y., Si, X., Hu, C., and Zhang, J., 2019. "A review of recurrent neural networks: Lstm cells and network architectures". *Neural computation,* ***31***(7), pp. 1235–1270.

[26] Gürel, A. E., Ağbulut, Ü., and Biçen, Y., 2020. "Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation". *Journal of Cleaner Production,* ***277***, p. 122353.

[27] Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., and Suganthan, P. N., 2022. "Ensemble deep learning: A review". *Engineering Applications of Artificial Intelligence,* ***115***, p. 105151.

[28] Romero, D. H., Icaza, D., and González, J., 2019. "Technical-economic study for the implementation of solar energy in the presence of biomass and micro hydraulic generation, for sectors that do not have electricity supply in the province of bolívar-ecuador". In 2019 7th International Conference on Smart Grid (icSmartGrid), IEEE, pp. 149–154.

[29] Icaza, D., Jurado, F., and Galindo, S. P., 2020. "What is of interest that the buildings of the public electrical companies are also provided with solar energy? case study" empresa eléctrica centro sur ca" in cuenca-ecuador". In 2020 9th International Conference on Renewable Energy Research and Application (ICRERA), IEEE, pp. 377–383.

[30] Echegaray-Aveiga, R. C., Masabanda, M., Rodríguez, F., Toulkeridis, T., and Mato, F., 2018. "Solar energy potential in ecuador". In 2018 International Conference on eDemocracy & eGovernment (ICEDEG), IEEE, pp. 46–51.

[31] García, J. L., Jurado, F., and Larco, V., 2019. "Review and resource assessment, solar energy in different region in ecuador". In E3S Web of Conferences, Vol. 80, EDP Sciences, p. 01003.

[32] Rodrigo, J. J. A., and Ortiz, J., 2022. Skforecast: time series forecasting with python and scikit-learn.

[33] Aziz, R. M., Baluch, M. F., Patel, S., and Ganie, A. H., 2022. "Lgbm: a machine learning approach for ethereum fraud detection". *International Journal of Information Technology,* ***14***(7), pp. 3321–3331.