computer sciences and mathematics forum

MDPI

Article

Deep Learning-Based Imputation of Irradiance Time Series Enhanced by Numerical Data †

Mónica Moreno Revelo [1,2,*], Juan Bernardo Gómez Mendoza [2], Javier Revelo Fuelagán [3] and Diego H. Peluffo Ordoñez [4]

[1] Departamento de Ingeniería Civil, Universidad Mariana, Pasto 520001, Colombia
[2] Departamento de Ingeniería Eléctrica, Electrónica y Computación, Universidad Nacional de Colombia, Manizales 170001, Colombia; jbgomezm@unal.edu.co
[3] Departamento de Electrónica, Universidad de Nariño, Pasto 520001, Colombia; javierrevelof@udenar.edu.co
[4] School of Mathematical and Computational Sciences, Yachay Tech University, Ibarra 100115, Ecuador; diego.peluffo@sdas-group.com
* Correspondence: momorenor@unal.edu.co
† Presented at the 12th International Conference on Time Series and Forecasting, Canaria, Spain, 15–17 July 2026.

Abstract

Performing imputation of gaps in irradiance time series plays a fundamental role in applications related to environmental monitoring and photovoltaic (PV) systems. Consequently, numerous studies have explored the use of artificial intelligence techniques to complete missing segments in time series. However, only a few works have examined the integration of irradiance data obtained from different sources to exploit their complementary strengths. In this context, our research proposes a methodology that compares two deep learning architectures-Bidirectional Long Short-Term Memory (BiLSTM) networks and Generative Adversarial Networks (GANs)—for imputing missing values in in situ irradiance measurements using external irradiance data as an additional input feature. The methodology is applied to datasets from Viento Libre and Alice Spring, and the results demonstrate that incorporating external data significantly enhances imputation performance. Specifically, for the Viento Libre dataset for example, the proposed approach achieves an improvement of 23.89%, 18.86% and 70.29% for Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Kullback–Leibler (KL) divergence respectively.

Keywords: imputation; artificial intelligence; solcast; neural networks

1. Introduction

Recording solar irradiance through meteorological stations is frequently affected by unavoidable factors such as sensor malfunctions, adverse weather conditions, and human errors [1]. Although numerical sources such as Solcast provide an alternative for obtaining irradiance and other environmental variables, their availability has traditionally been restricted to research applications [2]. Reliable irradiance measurements without missing values are essential for accurately assessing environmental conditions and play a key role in the design, operation, and performance evaluation of photovoltaic (PV) systems that contribute to the energy demand of buildings [3]. This is particularly relevant given that the building sector accounts for approximately 30–40% of global energy consumption and up to 40–50% in Europe [4]. Furthermore, high-quality irradiance data without gaps is crucial for downstream tasks such as forecasting [5].

Check for updates

Academic Editors: Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, Hector Pomares and Ignacio Rojas

Published: 25 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

The imputation of missing data has been addressed from multiple perspectives, ranging from traditional statistical techniques to advanced deep learning methods such as Bidirectional Long Short-Term Memory (BiLSTM) networks and Generative Adversarial Networks (GANs) [3,6]. For example, the authors in [7] compare nine traditional approaches—including linear and spline interpolation—for missing-value imputation aimed at improving a regression model for sunflower crop optimization. Similarly, ref. [1] applies linear interpolation, averaging, and the Hargreaves–Samani method to complete irradiance and irradiation data from the Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Likewise, Turrado [8] compares several traditional techniques for solar radiation imputation under varying sky and irradiance conditions.

Studies that rely on traditional imputation methods report that such approaches struggle to capture the intrinsic properties and temporal dynamics of time series. This issue is addressed in [9], where missing solar radiation data are reconstructed using Deep Neural Networks, achieving better performance than classical approaches such as mean and spline interpolation. Similarly, ref. [10] compares a BiLSTM with several traditional methods to complete pregnancy examination records collected between 2008 and 2018. The main challenges identified include the nonlinearity of the data and the risk of overfitting associated with the use of simple techniques such as oversampling. Likewise, the work in [11] employs a BiLSTM to fill gaps in half-hourly temperature records in China. To enhance accuracy, the authors introduce an optimization-based objective error function, concluding that new methods are still required to address long gaps in low-frequency time series.

Regarding GAN-based imputation, ref. [3] applies a GAN composed of an autoencoder-based generator and a recurrent discriminator combined with attention mechanisms to complete medical and aviation datasets. Their results show that incorporating attention improves GAN performance by better modeling the relationships among imputed values through an enhanced weighting matrix. Other works such as [5,12], which focus on meteorological data, highlight common issues in GAN-based imputation: generators often struggle to preserve local information, while discriminator losses may fail to provide stable probability estimates. The latter study addresses these limitations by integrating a gradient penalty mechanism.

The studies reported in [13,14] also demonstrate promising performance in time series imputation using more sophisticated architectures, including hybrid Transformer-based models for photovoltaic power data completion and Transformer-based approaches for general time series imputation. Finally, the works presented in [15,16] illustrate how imputation performance can be improved by incorporating additional variables beyond irradiance alone. In particular, ref. [15] applies Multivariate Imputation by Chained Equations (MICE) to solar radiation data. Similarly, ref. [16] incorporates climate reanalysis products as external sources for missing-data imputation in solar radiation, which is conceptually related to the use of external Solcast data in this work.

Although previous studies have explored multivariate time series for irradiance imputation, most of them have focused on using only in situ irradiance measurements rather than combining them with external irradiance data. Considering this limitation and those identified in previous studies, this work compares BiLSTM and GAN architectures and evaluates the improvement achieved when external irradiance samples are used as an additional feature to complete in situ data. To the best of our knowledge, few studies have explored the integration of external Solcast data with in situ measurements from Viento Libre in Nariño for irradiance imputation using BiLSTM and GAN models. The results obtained demonstrate that incorporating Solcast data significantly improves the

performance, reducing the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Kullback–Leibler (KL) divergence by 23.89%, 18.86% and 70.29% respectively.

2. Materials and Methods

2.1. Viento Libre Dataset

This is the name of a meteorological station operated by the "Instituto de Hidrología, Meteorología y Estudios Ambientales" (IDEAM). The station is located in Nariño, in the south of Colombia, with an average temperature of 20 °C and an annual precipitation of 844 mm. IDEAM records hourly data for several environmental variables across different regions of Colombia, including irradiance, temperature, and precipitation [1]. Depending on the station and the variable, the temporal coverage of the time series may vary. For the Viento Libre station, irradiance data are available from 3 April 2008 to 31 October 2016. Due to cloudiness, equipment failures, and other operational issues, the percentage of missing data in Viento Libre reaches nearly 28%.

2.2. Alice Spring Dataset

Alice is a region located in the north of Australia. It is situated in an arid desert environment, with maximum temperatures reaching up to 35 °C and an annual precipitation of 285.9 mm. The in situ dataset is provided by the Desert Knowledge Australia Solar Centre (DKASC). The center offers open access to data recorded every 5 min, covering the period from 12 September 2008 to 2 September 2022. Some of the variables measured include irradiance, active power, and wind speed. Due to issues related to meteorological stations and sensor maintenance, the percentage of missing data in this dataset is approximately 0.8%.

In this research, irradiance data from the in situ sources IDEAM and DKASC are used. Additionally, numerical irradiance data from Solcast are employed for both locations. Solcast is a data source that operates a monitoring system integrating near-real-time satellite imagery from 11 weather satellites, together with outputs from 7 numerical weather prediction (NWP) models [2]. Its spatial coverage is global, whereas its temporal coverage spans from January 2007 to date, with a sampling interval of five minutes.

2.3. Bidirectional Long Short-Term Memory (BiLSTM)

An LSTM is a recurrent network architecture designed to preserve long-term pattern information through its three gating mechanisms: forget, update, and output gates. BiLSTM is an extension of the LSTM that processes data in both forward and backward directions. This allows it to capture information from both past and future contexts [10].

2.4. Generative Adversarial Network (GAN)

A GAN analyses a collection of training samples to learn the underlying probability distribution that generated them [17]. GANs consist of two networks: the generator, which creates synthetic samples, and the discriminator, which distinguishes between real and fake samples.

2.5. Metrics

Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2) are used in this research. These metrics are suitable for models that aim to reproduce the exact target values such as LSTM [12]. However, they are not ideal for evaluating GAN-based imputation, since GANs focus on generating values that follow the underlying data distribution rather than reproducing each individual point exactly [18]. Therefore, Kullback–Leibler Divergence (KL) and Mutual Information (MI) are also used.

https://doi.org/10.3390/cmsf2026016011

KL divergence evaluates the difference between the probability distributions of the original and imputed series, whereas MI measures the statistical dependence between them. Higher MI values indicate stronger shared information content, whereas lower KL divergence values indicate greater similarity between distributions.

3. Experimental Setup

The experimental setup used to perform the imputation of in situ data using historical time series data obtained from Solcast platform is described in Figure 1.

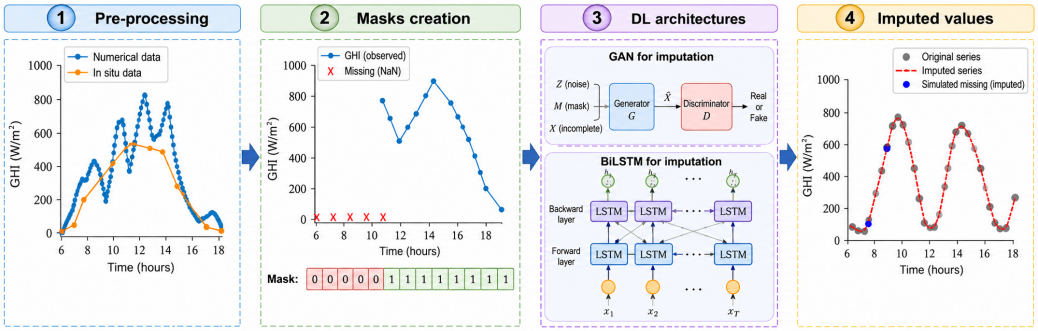


Figure 1. Methodology summarizing the main steps of the imputation process.

3.1. Preprocessing

The preprocessing applied is summarized in Table 1. First, nighttime data are removed from all datasets, retaining only the period between 06:00 and 18:00 h when irradiance takes non-zero values. Then, Solcast timestamps are adjusted to the corresponding Colombian and Australian local time zones to align them with the in situ observations from Viento Libre and Alice Spring.

Table 1. Preprocessing steps applied to the in situ and Solcast datasets. ✓ indicates that the preprocessing step is applied; — indicates that it is not applied.

Preprocessing Step	In Situ	Solcast
Night filtering (06–18 h)	✓	✓
Time-zone alignment	—	✓
Outlier removal (IQR)	✓	✓
Averaging hourly data (Alice Spring)	✓	—
Resampling to 1 h frequency	—	✓
Period matching	✓	✓
Normalization (0–1 MinMax)	✓	✓

Following this process, outliers are identified and removed using the Interquartile Range (IQR) criterion [19]. In order to compare our results with those reported in [20], the 5-minute in situ irradiance values from Alice Spring are converted to an hourly frequency by averaging all the values within each hour; this average is computed only when at least 50% of the samples are available; otherwise the hour is labeled as missing. The Solcast data for both regions are also resampled to a 1-hour resolution to match the in situ temporal frequency. Then, the data recording period is adjusted to ensure temporal overlap between the Solcast and in situ sources for each region. Finally, all variables are normalized to the range $[0, 1]$ using MinMax scaling.

3.2. Creation of Masks

Two masks are created for the in situ data of each region before applying the imputation process. The first mask, $M_o \in \mathbb{R}^{n\times 1}$, where n is the number of samples, marks with

zeros the positions in the time series where real missing values (NaN) occur, and with ones the positions where valid observations are available. The second mask, M_h, which has the same dimensions as M_o, is generated to simulate additional missing values for the evaluation of the proposed methodology. The simulated missingness follows the MCAR (Missing Completely At Random) mechanism, meaning that values are removed purely at random and independently of the variable itself or any other variable. Other missing-data mechanisms are not explored because the benchmark protocols used in this work are based on the MCAR assumption, allowing a fair comparison between methods.

To compare our methodology with the state of the art, the percentage of simulated missing data is set to 30% for Viento Libre and 15% for Alice Spring. Likewise, to assess the robustness of the models, all tests are performed using three different random seeds (123, 456, 789), and the results reported in Section 4 correspond to the mean (standard deviation).

3.3. BiLSTM Architecture and Parameters

The process of preparing the BiLSTM to perform imputation of in situ values begins by arranging the input feature matrix, which consists of two columns: the first contains the in situ irradiance values, and the second contains the Solcast irradiance values. Since the Solcast data do not contain missing values, this second feature is used as an auxiliary source during the imputation process.

Next, the feature matrix is reorganized into windows of a specific size. A target vector, composed of the in situ irradiance values, is also defined. Likewise, the previously described masks are taken into account to indicate which samples are valid and which are missing during training.

In the second stage, the BiLSTM architecture is defined following the recommendation in [21], where—after a preliminary evaluation—the authors conclude that the architecture illustrated in Figure 2 is the most suitable for modeling irradiance time series. Furthermore, the same structure was employed in our previous work, where it demonstrated strong performance for both temperature and irradiance forecasting [22].

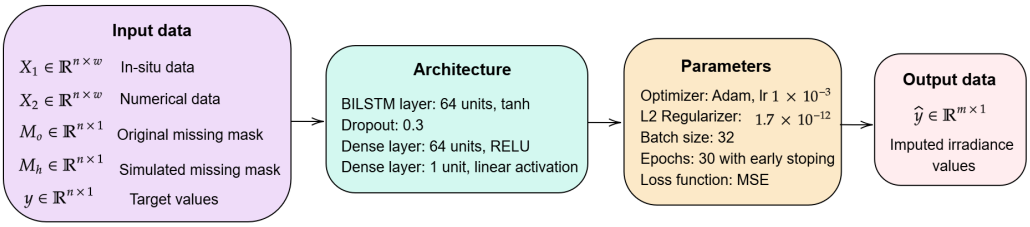


Figure 2. Architecture and parameters used for training the BiLSTM model. Here, n denotes the number of samples, w the window size, and m the number of imputed samples.

Subsequently, the hyperparameters of the BiLSTM are selected based on a methodology that consists of fixing a baseline configuration and varying one parameter at a time to isolate its effect on model performance. The parameters considered—listed below—were chosen according to the recommendations provided in [23]: window size, number of hidden units, learning rate, regularization rate, and batch size. For each parameter, two or three configurations are tested using only the Viento Libre dataset.

Finally, the trained model is used to complete both real and simulated missing values in the in situ data according to Algorithm 1. The simulated missing values are then employed to evaluate the performance of the model.

Algorithm 1: Sequential imputation using the BiLSTM model.

Input: Input data X, trained model, masks M_o, M_h
Output: Imputed series $\hat{y} \in \mathbb{R}^{m \times 1}$
Initialize $\hat{y} \leftarrow 1_m$;
for $i \leftarrow w$ **to** $n - 1$ **do**
  if $M_h(i) = 0$ ***or*** $M_o(i) = 0$ **then**
    Extract input window: $x_{input} \leftarrow \hat{X}_i \in \mathbb{R}^{1 \times w \times 2}$;
    Predict missing value: $\hat{y}_i \leftarrow \text{model}(x_{\text{input}})$;
    Update imputed series: $\hat{y}(i) \leftarrow \hat{y}_i$;
  end
end
return $\hat{y}$;

3.4. GAN Architecture and Parameters

The first stage required to prepare the GAN for imputation is almost identical to the preprocessing performed for the BiLSTM model. However, two key differences arise. First, a noise vector is generated, whose dimensions match those of the feature matrix. Second, no target vector $\hat{y}$ is defined, since, unlike the BiLSTM, the GAN produces a set of values (depending on the window size) rather than a single imputed sample.

In the second stage, the GAN structure and its parameters are defined by mirroring the BiLSTM architecture in the generator to maintain comparability between both models. Regarding the discriminator, ref. [24] employs the same architecture for both the generator and the discriminator in their GAN, which supports our decision to preserve the structural similarity between the two networks. We specifically restrict the discriminator to at most two LSTM layers with 64 units and moderate regularization to prevent it from either overpowering or underperforming relative to the generator.

The selection of discriminator hyperparameters is guided by the recommendations given in [24,25], where comprehensive hyperparameter searches emphasize the relevance of batch size, learning rate, regularization techniques, and the weighting of the generator loss. Since several of these settings are already fixed on the generator side in the proposed framework, the hyperparameter exploration in this study focuses primarily on the discriminator. Thus, the regularization strength and learning rate of the discriminator and the weight assigned to the generator loss are varied using a sequential one-factor-at-a-time strategy, following the same procedure adopted for the BiLSTM model.

For each configuration, experiments are conducted using in situ and external Viento Libre samples, and the optimal configuration—summarized in Figure 3—is selected based on the average performance across three independent runs with different random seeds.

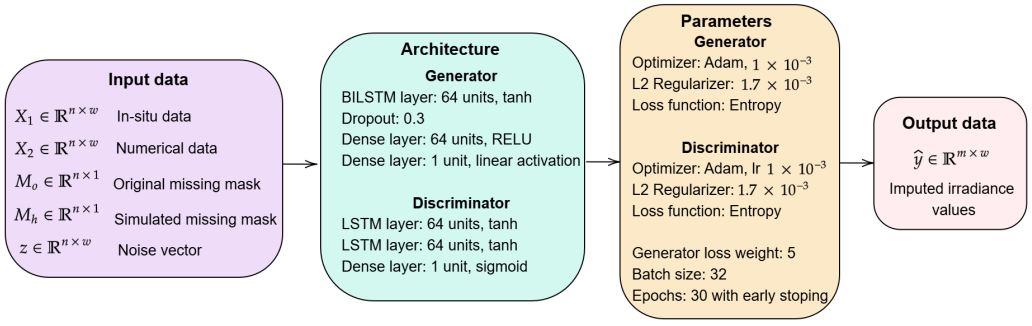


Figure 3. Architecture and parameters used for training the GAN model. Here, n denotes the number of samples, w the window size, and m the number of imputed samples.

In the third stage corresponding to training, the generator and discriminator are optimized in an alternating manner according to Algorithm 2. Since the GAN outputs a sequence of values (depending on the window size) instead of a single imputation, the last element of the generated sequence is selected as the final imputed value according to Algorithm 3. This value is evaluated only at the positions marked as 1 by the mask M_h, which correspond to the simulated missing entries.

Algorithm 2: GAN training procedure for time series imputation.

Input: Input data X, Masks M_o and M_h, noise z, generator G, discriminator D
Output: Trained generator G
Compute total mask: $M_t \leftarrow M_o \odot M_h$;
Apply windowing to M_t to match the input tensor X, obtaining $M \in \mathbb{R}^{n \times w}$;
for *each epoch* **do**
  for *each minibatch* **do**
    Extract batch X_b, mask M_b, and noise z_b;
    Generator input: $G_{in} \leftarrow \text{concat}(X_b, M_b, Z_b)$;
    Generate imputed values: $\hat{X_b} \leftarrow G(G_{in})$;
    Construct reconstructed data: $X_R \leftarrow M_b \cdot X_b + (1 - M_b) \cdot \hat{X_b}$;
    Discriminator input: $D_{in} \leftarrow \text{concat}(X_R, M_b)$;
    Compute discriminator output: $D_{prob} \leftarrow D(D_{in})$;
    Compute discriminator loss:
    $D_{loss} = -\mathbb{E}\Big[M_b \log(D_{prob}) + (1 - M_b) \log(1 - D_{prob})\Big]$;
    Compute generator adversarial loss: $G_{adv} = -\mathbb{E}\Big[(1 - M_b) \log(D_{prob})\Big]$;
    Compute reconstruction loss: $G_{rec} = \mathbb{E}\big[M_b \cdot (X_b - X_R)^2\big]$;
    Total generator loss: $G_{loss} = G_{adv} + \lambda \cdot G_{rec}$;
    Update generator parameters using ∇G_{loss};
    Update discriminator parameters using ∇D_{loss};
  end
end
return G;

Algorithm 3: Sequential imputation using the trained GAN generator.

Input: Input data X, trained generator G, window size w, masks M_o and M_h
Output: Imputed series $\hat{y} \in \mathbb{R}^{m \times 1}$
Initialize $\hat{y} \leftarrow 1_m$;
for $i \leftarrow w$ **to** $n - 1$ **do**
  if $M_h(i) = 0$ ***or*** $M_o(i) = 0$ **then**
    Extract temporal windows: $x_{input} \leftarrow \hat{X}_i \in \mathbb{R}^{1 \times w \times f}$, $m_{input} \leftarrow \hat{M}_i \in \mathbb{R}^{1 \times w}$;
    Generate random noise: $z_{input} \sim \mathcal{N}(0, 1)$, with $Z_{input} \in \mathbb{R}^{1 \times w}$;
    Concatenate generator input: $G_{in} \leftarrow \text{concat}(x_{input}, m_{input}, z_{input})$;
    Generate imputed sequence: $\hat{X} \leftarrow G(G_{in})$;
    Update imputed value: $\hat{y}(i) \leftarrow \hat{X}(0, w - 1, 0)$;
  end
end
return $\hat{y}$;

https://doi.org/10.3390/cmsf2026016011

4. Results and Discussion

As highlighted in the previous section, several candidate values are evaluated for selected BiLSTM and GAN hyperparameters. However, this section only presents the results obtained using the best configurations identified during the evaluation process.

The models are evaluated not only on the Viento Libre dataset but also on Alice Spring. Although the datasets differ in the number of samples and percentage of missing values, the same configurations are applied in order to assess their generalization capability. Additional experiments excluding Solcast as an external data source are also conducted to quantify its contribution to the imputation task.

4.1. Results Obtained Using BiLSTM

Table 2 presents the performance metrics obtained for the Viento Libre and Alice Spring datasets. The metrics consistently show a remarkable improvement when Solcast data are incorporated. For instance, the MAE improves by 23.89% in Viento Libre and by 17% in Alice Spring. The distributional metrics in Table 2 also indicate that the real and imputed samples are highly aligned, with KL values close to zero and MI values higher than one in almost all cases.

The overall performance is better for Alice Spring across all metrics. This may be partially attributed to the lower elimination rate used for this dataset (half of that applied to Viento Libre). Another contributing factor could be the higher temporal resolution of the Alice Spring irradiance data, which were originally recorded every 5 min, allowing for a more reliable hourly resampling through averaging.

Table 2. Performance metrics obtained with the BiLSTM model for Viento Libre and Alice Spring.

Dataset		MAE (W/m^2)	RMSE (W/m^2)	R^2	KL	MI
With Solcast	Viento Libre	70.98 (0.47)	105.60 (0.62)	0.86 (0.02)	0.30 (0.23)	1.23 (0.01)
	Alice Spring	32.70 (0.35)	62.00 (1.26)	0.97 (0.13)	0.15 (0.05)	0.96 (0.04)
Without Solcast	Viento Libre	93.27 (2.23)	130.15 (2.90)	0.79 (0.01)	1.01 (0.03)	2.04 (0.01)
	Alice Spring	49.90 (6.38)	82.55 (3.54)	0.95 (0.02)	0.15 (0.01)	1.84 (0.01)

4.2. Results Obtained Using GAN

According to the results reported in Table 3, a noticeable increase in MAE and RMSE is observed, particularly in the configuration where Solcast data are not incorporated. Regarding the coefficient of determination, R^2 decreases for both datasets, and when Solcast data are excluded, negative values are obtained, indicating that the model fails to explain the variability in the irradiance series.

In terms of distribution-based metrics, KL divergence and MI show improved values for the GAN model when Solcast samples are included. However, both metrics deteriorate significantly in the absence of this external source. These results indicate that Solcast data play a more critical role in the GAN-based imputation framework than in the BiLSTM model. In particular, without Solcast, the GAN is unable to adequately capture the underlying data distribution. This behavior can be attributed to the high proportion of missing values in the in situ dataset, which causes the model to rely predominantly on noise during training.

Table 3. Performance metrics obtained with the GAN model for Viento Libre and Alice Spring.

Dataset		MAE (W/m^2)	RMSE (W/m^2)	R^2	KL	MI
With Solcast	Viento Libre	97.43 (0.86)	132.4 (0.84)	0.78 (0.02)	0.09 (0.01)	1.05 (0.01)
	Alice Spring	85.95 (2.04)	121.6 (1.76)	0.89 (0.01)	0.06 (0.03)	1.46 (0.01)
Without Solcast	Viento Libre	449.2 (121.88)	536.7 (106.57)	−2.68 (1.49)	18.15 (0.12)	0.23 (0.13)
	Alice Spring	473.4 (0.74)	594.2 (1.39)	−1.74 (0.01)	17.21 (0.09)	0.38 (0.35)

4.3. Graphical Examples Using BiLSTM and Viento Libre

Taking into account that the best results are achieved with the BiLSTM model, additional analyses are performed using this architecture and Solcast as an external source on the Viento Libre dataset. First, the sensitivity of the proposed approach to random seed initialization is analyzed. A total of 10 different random seeds are evaluated, and the corresponding results are presented in Figure 4.

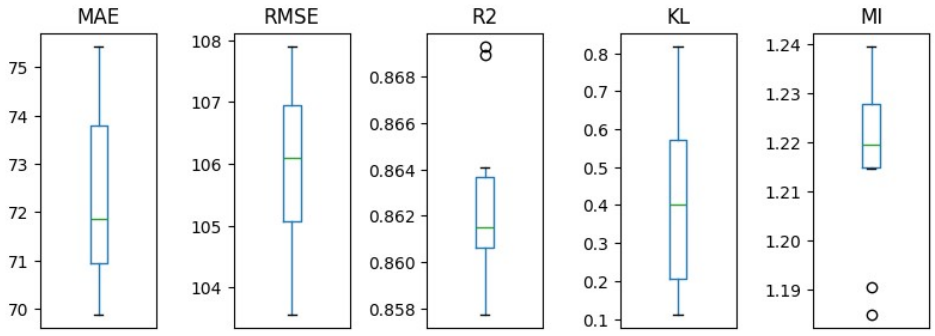


Figure 4. Boxplots of performance metrics obtained using the BiLSTM model across 10 random seeds.

Overall, there is no significant variation in the metrics across different random seeds. This indicates that the model is robust to the initialization of the indices used for simulating missing values. Thus, the model exhibits stable and reproducible performance.

On the other hand, Figure 5 shows a segment of the imputed Global Horizontal Irradiance (GHI). Both real and simulated missing points are displayed in the plot. The simulated points closely reconstruct the real values, and the imputed samples follow the overall behavior of the time series with high fidelity.

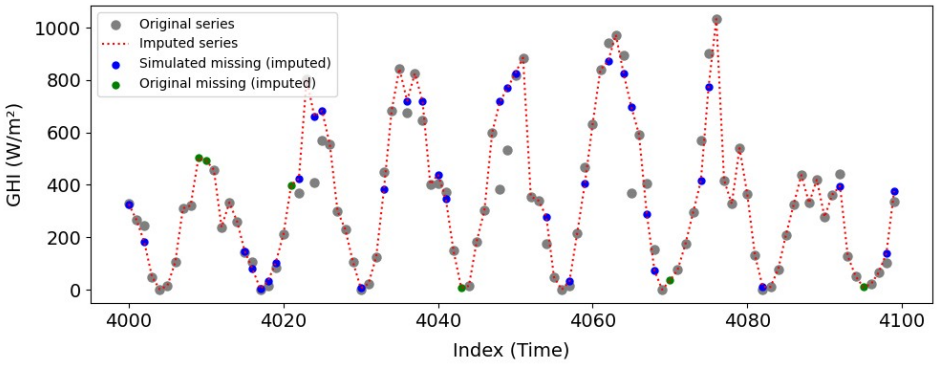


Figure 5. Imputed irradiance in the 4000–4100 interval for the Viento Libre dataset using the BiLSTM model.

Finally, Figure 6 shows the relationship between all true and imputed irradiance values obtained on the simulated positions. The scatter plot confirms that most points align closely

with the 1:1 diagonal, with greater dispersion occurring primarily at higher irradiance levels—an expected behavior in solar radiation analyses.

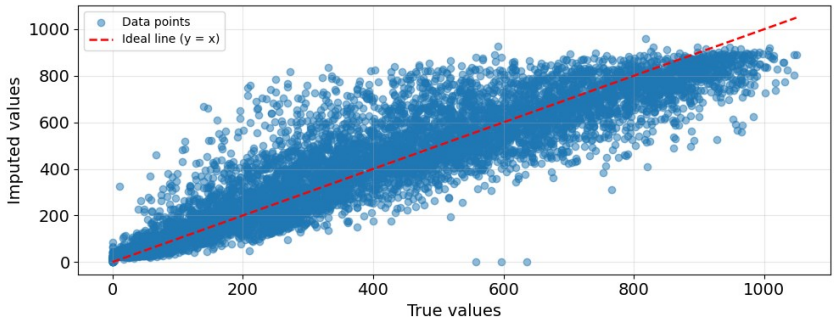


Figure 6. Scatter plot between true and imputed irradiance values obtained using the BiLSTM model.

4.4. Comparison with the State of the Art

Table 4 compares the results obtained with the proposed BiLSTM model using Solcast as an external data source. For the Viento Libre dataset, the only related benchmark identified in the literature is the study presented in [1], which performs irradiance imputation on several stations in Nariño, including Viento Libre, using traditional techniques such as interpolation. In contrast, only a limited number of studies have addressed irradiance imputation on the Alice Spring dataset using more sophisticated approaches. One representative example is the work described in [20], which evaluates methods based on Transformers and recurrent neural networks, reporting the best performance with the Transformer-based architecture. The proposed BiLSTM model outperforms the benchmark reported for Viento Libre and achieves results comparable to those obtained with recurrent neural network approaches on the Alice Spring dataset.

Table 4. Comparison with performance metrics reported in the state of the art.

Methodology	MAE (W/m^2)	RMSE (W/m^2)	R^2
Proposed	70.98	105.60	0.86
Interpolation method [1]	80.41	118.04	–
Proposed	32.70	62.00	0.97
LSTM method [20]	27.46	52.47	0.98

5. Conclusions

This work compared two deep learning approaches across two datasets and analyzed the impact of incorporating external weather data during training. The obtained results are comparable to those reported in the state of the art.

Despite the effectiveness of the proposed models, future work will explore strategies to reduce computational cost, such as the use of pre-trained models combined with Solcast data for irradiance imputation. In addition, the impact of the imputed series on forecasting models and the influence of other missingness mechanisms, such as Missing at Random (MAR) and Missing Not at Random (MNAR), will also be analyzed.

Author Contributions: M.M.R. conducted the experiments and wrote the manuscript; J.B.G.M., J.R.F. and D.H.P.O. contributed to review, supervision, and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this research can be found in https://dkasolarcentre.com.au/ (accessed on 15 May 2025) and https://www.ideam.gov.co/ (accessed on 15 May 2025).

Acknowledgments: The authors acknowledge the use of ChatGPT Plus 5.6 for assistance with English grammar revision and language polishing, as well as for improving the graphical presentation of Figure 1. The scientific content, methodology, and data represented in the figure were defined and verified by the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hoyos-Gómez, L.S.; Ruiz-Muñoz, J.F.; Ruiz-Mendoza, B.J. Short-term forecasting of global solar irradiance in tropical environments with incomplete data. *Appl. Energy* **2022**, *307*, 118192. [CrossRef]
2. Almarshoud, A.F. Validation of satellite-derived solar irradiance datasets: A case study in Saudi Arabia. *Future Sustain.* **2024**, *2*, 1–7. [CrossRef]
3. Zhao, J.; Rong, C.; Lin, C.; Dang, X. Multivariate time series data imputation using attention-based mechanism. *Neurocomputing* **2023**, *542*, 126238. [CrossRef]
4. Wang, Z.; Liu, X.; Huang, Y.; Zhang, P.; Fu, Y. A multivariate time series graph neural network for district heat load forecasting. *Energy* **2023**, *278*, 127911. [CrossRef]
5. Qin, R.; Wang, Y. ImputeGAN: Generative adversarial network for multivariate time series imputation. *Entropy* **2023**, *25*, 137. [CrossRef]
6. Xu, L.; Xu, L.; Yu, J. Time series imputation with GAN inversion and decay connection. *Inf. Sci.* **2023**, *643*, 119234. [CrossRef]
7. Călin, A.D.; Coroiu, A.M.; Mureşan, H.B. Analysis of preprocessing techniques for missing data in the prediction of sunflower yield in response to the effects of climate change. *Appl. Sci.* **2023**, *13*, 7415. [CrossRef]
8. Turrado, C.C.; López, M.d.C.M.; Lasheras, F.S.; Gómez, B.A.R.; Rollé, J.L.C.; de Cos Juez, F.J. Missing data imputation of solar radiation data under different atmospheric conditions. *Sensors* **2014**, *14*, 20382–20399. [CrossRef]
9. Başakın, E.E.; Özger, M. Missing Data Imputation for Solar Radiatıon by Deep Neural Network. *Eur. J. Sci. Technol.* **2022**, *35*, 548–555. [CrossRef]
10. Lu, X.; Yuan, L.; Li, R.; Xing, Z.; Yao, N.; Yu, Y. An improved Bi-LSTM-based missing value imputation approach for pregnancy examination data. *Algorithms* **2022**, *16*, 12. [CrossRef]
11. Xie, C.; Huang, C.; Zhang, D.; He, W. BiLSTM-I: A deep learning-based long interval gap-filling method for meteorological observation data. *Int. J. Environ. Res. Public Health* **2021**, *18*, 10321. [CrossRef]
12. Wang, Y.; Xu, X.; Hu, L.; Fan, J.; Han, M. A time series continuous missing values imputation method based on generative adversarial networks. *Knowl.-Based Syst.* **2024**, *283*, 111215. [CrossRef]
13. Zhang, C.; Fu, X.; Qiu, D.; Badihi, H.; Gu, H. Robust imputation of missing photovoltaic power data using a weather-and context-aware hybrid transformer framework. *Renew. Energy* **2025**, *256*, 124576. [CrossRef]
14. Nie, T.; Qin, G.; Ma, W.; Mei, Y.; Sun, J. ImputeFormer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 2260–2271.
15. Akbar, G.; Prajitno, P.; Ariffudin; Ananda, N. Multivariate imputation chained equation on solar radiation in automatic weather station. *J. Penelit. Pendidik. IPA* **2024**, *10*, 3633–3639. [CrossRef]
16. Araujo, M.A.d.S.G.; Oliveira, F.L.C.; Souza, R.C. Missing data imputation using climate data reanalysis in solar energy generation time series. *Energy Convers. Manag. X* **2026**, *30*, 101653. [CrossRef]
17. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
18. Richter, A.; Ijaradar, J.; Wetzker, U.; Jain, V.; Frotzscher, A. A Survey on Multivariate Time Series Imputation using Adversarial Learning. *IEEE Access* **2024**, *12*, 148167–148189. [CrossRef]
19. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, USA, 1977.
20. Zhang, H.; Li, B.; Su, S.F.; Yang, W.; Xie, L. A novel hybrid transformer-based framework for solar irradiance forecasting under incomplete data scenarios. *IEEE Trans. Ind. Inform.* **2024**, *20*, 8605–8615. [CrossRef]
21. Etxegarai, G.; López, A.; Aginako, N.; Rodríguez, F. An analysis of different deep learning neural networks for intra-hour solar irradiation forecasting to compute solar photovoltaic generators' energy production. *Energy Sustain. Dev.* **2022**, *68*, 1–17. [CrossRef]

22. Revelo, M.Y.M.; Mendoza, J.B.G.; Fuelagán, J.R. Hourly Prediction of Irradiance and Temperature Using Recurrent Neural Networks and Gaussian Process Models. *Tecnura* **2025**, *29*, 1–16. [CrossRef]
23. Almuhaini, S.H.; Sultana, N. Bayesian-optimization-based long short-term memory (LSTM) super learner approach for modeling long-term electricity consumption. *Sustainability* **2023**, *15*, 13409. [CrossRef]
24. Klidbary, S.H.; Babaei, A.; Ghorbani, R. Evolutionary Approach to GAN Hyperparameter Tuning: Minimizing Discriminator and Generator Loss Functions. In *Proceedings of the 2024 14th International Conference on Computer and Knowledge Engineering (ICCKE)*; IEEE: Piscataway, NJ, USA, 2024; pp. 99–104.
25. Pinheiro, D.N.; Gonzalez, J.C.; Corso, G.; Gebre, M.G.; da Costa, C.A.; Xavier-de Souza, S.; Barros, T. Hyperparameter determination for GAN-based seismic interpolator with variable neighborhood search. *Comput. Geosci.* **2024**, *192*, 105689. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.