ASADES

Avances en Energías Renovables y Medio Ambiente - AVERMA
Vol. 29, pp. 484-499 2025, ISSN 2796-8111

RECONSTRUCCIÓN DE SERIES DE IRRADIANCIA GLOBAL HORIZONTAL CON HUECOS SINTÉTICOS MEDIANTE MODELOS DE MACHINE LEARNING Y DATOS SATELITALES. CASO DE ESTUDIO: EL ROSAL, SALTA

Constanza B. López Ruiz[1,2], **Rubén D. Ledesma**[1,2], **Germán A. Salazar**[1,2], **Janis Galdiño**[3]
[1] Instituto de Investigaciones en Energía no Convencional (INENCO). Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).
[2] Departamento de Física, Facultad de Ciencias Exactas, Universidad Nacional de Salta
[3] Centro de Energias Renováveis, Universidade Federal de Pernambuco (Av. da Arquitetura, 857 -Cidade Universitária, Recife - PE, 50740-550
E-mail: constanza.lopezruiz@conicet.gov.ar

RESUMEN: La radiación solar global sobre plano horizontal (GHI) es clave para el diseño y optimización de sistemas de energía solar, pero sus registros suelen presentar datos faltantes originados en fallas de instrumentación o mantenimiento. Este trabajo compara tres estrategias basadas en Machine Learning —regresión lineal simple (SLR), regresión lineal múltiple (MLR) y perceptrón multicapa (MLP)— para la imputación de datos faltantes. También se emplean productos satelitales (CAMS), de reanálisis (ERA5) y estimaciones del modelo ARGP2 para mejorar la precisión de las predicciones. Para evaluar el desempeño de los modelos, se generaron huecos sintéticos de manera semi-aleatoria, en una serie de GHI de la localidad El Rosal, Salta, Argentina. En todos los casos el MLP logró los menores errores (rRMSD entre 18 % y 21 %), seguido por MLR (rRMSD entre 20 % y 24 %) y SLR (rRMSD entre 21 % y 26 %).

Palabras clave: Irradiancia solar, machine learning, CAMS, ERA5, series temporales, gap filling.

INTRODUCCIÓN

La Irradiancia global horizontal sobre plano horizontal (GHI) es un parámetro fundamental para la evaluación y optimización de sistemas de energía solar de mediana o gran envergadura (Alonso-Suárez, 2017). Su medición precisa es esencial para estimar el potencial energético de una ubicación geográfica y para maximizar el rendimiento, por ejemplo, de los sistemas solares fotovoltaicos (PV) (Cabrera et al., 2024). Sin embargo, uno de los principales desafíos en el análisis de datos solares es la presencia de huecos (datos faltantes) en las series temporales de GHI, que pueden surgir por fallas en los equipos de medición, interrupciones por mantenimiento o condiciones meteorológicas extremas (Muneer y Fairooz, 2002). Estos huecos no solo afectan la calidad de los análisis realizados, sino que también limitan la capacidad de modelar o predecir con precisión el recurso solar, lo que impacta negativamente en la toma de decisiones para proyectos solares (Cabrera et al., 2024). Como ejemplo de la magnitud del problema, Schwandt et al. (2014) reportaron que la red SRRA de India (51 estaciones) presenta aproximadamente 7 % de datos faltantes, con duraciones que van desde minutos hasta varios días, evidenciando que los huecos reales requieren métodos confiables de imputación incluso sin disponibilidad de datos auxiliares externos.

En el reporte técnico de Sengupta et al. (2021) se destaca que la calidad y continuidad de los datos solares son esenciales para todas las etapas de un proyecto solar, desde la selección inicial del sitio hasta la operación del sistema. Este reporte subraya la importancia de implementar controles de calidad

Esta obra está bajo Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional. https://creativecommons.org/licenses/by-nc-sa/4.0/

rigurosos y métodos confiables para el manejo de datos faltantes, especialmente en regiones con alta variabilidad climática. En este contexto, se han propuesto numerosos métodos para llenar estos huecos (o gap-filling) en los datos meteorológicos. En el caso de la GHI, las técnicas más comunes incluyen métodos estadísticos (como la interpolación lineal o polinómica), el uso de datos satelitales o de reanálisis, y enfoques más recientes basados en Machine Learning (ML) (Cabrera et al., 2024; Iturbide et al., 2023). Los modelos de ML han demostrado ser particularmente efectivos, ya que pueden aprender relaciones complejas entre variables y adaptarse a patrones locales específicos. Entre estos modelos, la Regresión Lineal Simple (SLR), la Regresión Lineal Múltiple (MLR) y las redes neuronales artificiales, como el Perceptrón Multicapa (MLP), han sido ampliamente utilizados en aplicaciones similares (He et al., 2025). Sin embargo, todavía existe una necesidad de estudios que comparen su rendimiento específicamente en el contexto de los datos de GHI.

En este trabajo se propone y evalúa un enfoque basado en ML para la reconstrucción (gap-filling) de datos de GHI registrados con una resolución temporal de 15 minutos en la localidad El Rosal, Salta, Argentina. Para ello, se entrenaron modelos de SLR, MLR y MLP utilizando estimaciones para radia ción solar satelitales del Servicio de Monitoreo de la Atmósfera de Copernicus (CAMS, por sus siglas en inglés) (Inness et al., 2019), obtenidos del sitio https://www.tsv.soda-pro.com/web-services/ radiation/cams-radiation-service, datos de reanálisis ERA5 del Centro Europeo de Predicción Meteorológica a Plazo Medio (ECMWF, por sus siglas en inglés) (Hersbach et al., 2020), obtenidos del si tio https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab= overview y datos de GHI de cielo claro obtenidos con el modelo ARGP_V2, el cual está basado en el modelo McClear ajustado empíricamente para el noroeste argentino (Ledesma et al., 2023). Tanto las estimaciones satelitales como los datos de reanálisis proporcionan información a través de variables adicionales que afectan directamente los patrones de Irradiancia solar.

Tomando como referencia la metodología semi-Monte Carlo propuesta por Demirhan y Renwick (2018), se diseñaron escenarios de huecos sintéticos controlados seleccionando primero el 10 %, 30 % y 50 % de los días del registro; sobre cada día elegido se generaron huecos de 1 y 4 pasos de 15 min con distribución temporal aleatoria uniforme, lo que permite evaluar el desempeño de los modelos de ML bajo patrones realistas de datos faltantes sin depender de los huecos reales de la serie.

Este trabajo busca contribuir en mejorar la calidad de los conjuntos de datos solares en la región, y también ofrecer una metodología replicable para otras ubicaciones con características climáticas similares. Además, resalta el valor de combinar datos de múltiples fuentes, como así también técnicas de aprendizaje automático para abordar los desafíos del manejo de datos faltantes en aplicaciones solares.

METODOLOGÍA

La metodología incluye la generación de huecos sintéticos en series de datos medidos para evaluar el rendimiento de los modelos en diferentes escenarios, con tasas de datos faltantes en hasta un 50 % de los días de la serie de datos y huecos de tamaños variados (aleatorios de mayor frecuencia y continuos de menor frecuencia). Asimismo, las métricas utilizadas para cuantificar la precisión de los modelos son el error de desvío estándar (MBE), el error cuadrático medio (RMSD) y el error medio absoluto (MAE) en términos relativos.

Descripción del set de datos

Los datos utilizados en este trabajo corresponden a la localidad El Rosal (latitud = -24.393°, longitud = -65.768°, a una altitud aproximada de 3355 m.s.n.m), en la Provincia de Salta, al noreste de Argentina, con un clima semiárido frío, correspondiente al clima Bsk en la clasificación climática de Köppen-Geiger (Peel et al., 2007). Se caracteriza por inviernos fríos y veranos templados o cálidos, con precipitaciones escasas. Las mediciones se realizaron utilizando un piranómetro de la marca Kipp & Zonen, modelo CMP3 cuya constante de calibración se obtuvo mediante intercomparación “outdoor”

frente a un piranómetro Kipp & Zonen, modelo CM21 (N° de serie 051481), con trazabilidad al patrón mundial, y se registraron minutalemnte en un datalogger de la marca Campbell Scientific, modelo CR1000.

El set incluye GHI medida (variable de referencia) y estimaciones de CAMS, ERA5 y del modelo de cielo claro ARGP_V2. (Ledesma et al., 2023).

Los datos de ERA5 son horarios por defecto; CAMS tiene resolución nativa de 15 min y las mediciones tiene una frecuencia de registro de 1 min. Para igualar la resolución temporal, ERA5 se interpola 15 min y los registros minutales se integraron al mismo paso. El período analizado va del 3 de enero de 2016 al 30 de mayo de 2017, totalizando 20 886 registros de 15 min que superaron los filtros de calidad de Nollas et al. (2023).

Con el objetivo de reducir la dimensionalidad y evitar el ingreso de variables irrelevantes, se construyó una matriz de correlación entre todas las variables disponibles. Esta matriz permitió identificar aquellas variables con mayor correlación con la variable objetivo, las cuales fueron seleccionadas como variables de entrada para los modelos. En la Fig. 1 se presentan los resultados de manera gráfica, siendo las variables consideradas como posibles entradas para los modelos las que figuran en la Tabla 1.

Tabla 1: Variables propuestas y sus definiciones.

Símbolo / variable	**Definición**
sza	Ángulo cenital solar (°)
GHIargp2	GHI del modelo ARGP2 para cielo claro (W/m^2)
delta	Declinación solar (°)
mak	Masa de aire (Kasten) (Ledesma et al., 2023)
TOA	Radiación solar incidente a tope de la atmósfera (W/m^2)
GHIcams	GHI CAMS para toda condición de cielo
GHIcamscc	GHI CAMS para cielo claro
tco3	Ozono total (Dobson)
tcwv	Vapor de agua (kg/m^2)
AOD BC	AOD de carbono negro a 550 nm
AOD DU	AOD de polvo a 550 nm
AOD SS	AOD de sal marina a 550 nm
AOD OR	AOD de materia orgánica a 550 nm
AOD SU	AOD de sulfato a 550 nm
Cloud coverage	Cobertura nubosa (%)
GHIera	GHI ERA5 para toda condición de cielo (W/m^2)
GHIeracc	GHI ERA5 para cielo claro (W/m^2)

msl	Presión al nivel del mar
ie	Flujo instantáneo de humedad
hcc	Cobertura de nubes altas
d2m	Punto de rocío a 2 m
ghi	GHI medida en tierra (W/m^2)

Se puede observar que la serie de referencia (GHI) tiene una mayor correlación, en magnitud, con las variables: sza, GHIargp2, mak, TOA, GHIcams, GHIcamscc, GHIera, GHIeracc y, por lo tanto, son estas variables las que se utilizaron como datos de entrada.

Si bien algunas variables presentan una correlación negativa, esto refleja una relación inversa entre ellas: cuando una aumenta, la otra disminuye de manera proporcional (por ejemplo, a mayor ángulo cenital menor valor de irradiancia).

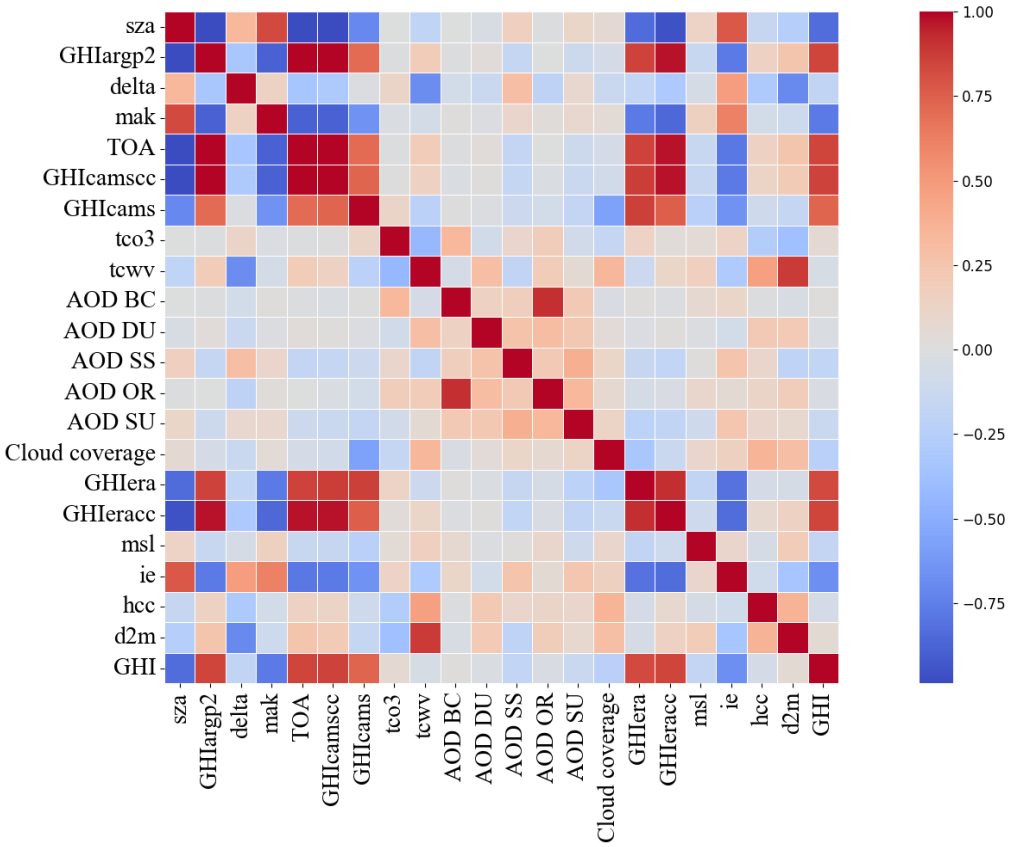


Figura 1: Matriz de correlación entre las variables correspondientes al archivo de datos de trabajo.

Generación de huecos sintéticos

Para evaluar el desempeño de los modelos, se generaron huecos sintéticos en la serie de datos. Este enfoque permite comparar los valores predichos por los modelos con los valores reales, lo que facilita la evaluación de la precisión. Con el objetivo de poder simular datos faltantes reales, se consideraron tres parámetros fundamentales: **porcentaje de días** a los cuales se le generarán los huecos, **cantidad de huecos** en los días seleccionados y **tamaño de los huecos**. Cabe destacar que, tanto la ubicación de

los huecos en el día, como la elección de los días se realizan de manera aleatoria, con la consideración de que no se repita la misma posición y evitar solapamientos.

Se plantean tres escenarios en los que se introducen datos faltantes en el 10 %, 30 % y 50 % del total de días analizados. Para cada uno de estos escenarios, se consideran dos configuraciones: una de corta duración y mayor frecuencia (12 huecos por día, cada uno con 1 dato faltante) y otra de larga duración pero menor frecuencia (3 huecos por día, cada uno con 4 datos faltantes).

En cada escenario los modelos se entrenan con la totalidad de la serie excepto los puntos que fueron simulados como faltantes; de esta forma se busca reproducir una situación operativa en la que solo se dispone de fragmentos medidos.

Modelos de machine learning

Regresión Lineal Simple (SLR).

El modelo SLR se utilizó para estimar los valores en función de una única variable dependiente, en este caso se consideraron cinco opciones: GHIcams, GHIera GHIcamscc, GHIeracc y GHIargp2. La ecuación del modelo es:

$$y_{adaptado} = ax_{estimado} + b \tag{1}$$

donde a y b son los coeficientes del modelo, los cuales se obtienen del entrenamiento considerando todos los datos disponibles, excepto aquellos correspondientes a los huecos generados en cada escenario planteado. Este modelo es simple y rápido de implementar, pero puede no capturar relaciones complejas entre las variables.

Regresión Lineal Múltiple (MLR).

El modelo MLR incluyó múltiples variables independientes: GHIcamscc, TOA, GHIargp2, GHIeracc, GHIera, GHIcams, mak y ie. El procedimiento es análogo al caso anterior, solo que en este caso se utiliza una regresión lineal múltiple. La forma de la función que realiza la adaptación tiene la forma de la Ec. 2.

$$y_{adaptado} = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b \tag{2}$$

Se deben encontrar ahora los coeficientes a_i (con $1 \leq i \leq n$, con n la cantidad de variables consideradas) y b. Para ello se consideran todos los datos disponibles, excepto aquellos correspondientes a los huecos generados en cada escenario planteado.

Perceptrón Multicapa (MLP).

El modelo MLP es una arquitectura de red neuronal artificial que sigue un esquema feed-forward, compuesta por capas densamente conectadas, donde cada neurona aplica una transformación no lineal (función de activación) a una combinación lineal de sus entradas. Para optimizar su rendimiento, se realizó una búsqueda de cuadrícula (grid search) evaluando diferentes configuraciones de arquitectura de hiperparámetros como se presenta en la Tabla 2.

Durante el entrenamiento del MLP se aplicó validación cruzada de 3 folds. En este procedimiento, la serie se dividió en tres subconjuntos de igual tamaño: en cada iteración, dos tercios de los datos se utilizaron para el entrenamiento y el tercio restante para validación, rotando los subconjuntos de modo

que cada partición fue usada una vez como conjunto de validación. Este esquema no constituye una partición fija en conjuntos de entrenamiento y prueba como en problemas de predicción generalizable, sino que permitió evaluar de manera robusta el desempeño de los modelos dentro de la misma serie y aplicar early stopping para reducir el riesgo de sobreajuste. Tanto la validación cruzada como el early stopping fueron implementadas mediante un GridSearchCV (Pedregosa et al., 2012). La validación cruzada permite estimar de manera robusta el desempeño del modelo mediante la partición repetida del conjunto de datos, mientras que el early stopping interrumpe el entrenamiento cuando no se observa mejora en el rendimiento sobre un subconjunto de validación, evitando así un ajuste excesivo a los datos de entrenamiento.

Tabla 2: Configuración del MLP para GridSearchCV

Aspecto	**Configuración**
Arquitecturas	• $1capa: (n), (2n)$ • $2capas: (n, n), (n, 2n)$ • $3capas: (n, n, n), (n, n, 2n)$
Función de activación	ReLU
Regularización (α)	$10^{-4}, 10^{-3}, 10^{-2}$
Validación cruzada	3-fold

Cabe destacar que no se realiza una separación tradicional de los datos entre conjunto de entrenamiento y testeo como se hace habitualmente en problemas de predicción generalizables. Esto se debe a que el objetivo principal no es desarrollar modelos que sean capaces de generalizar a nuevos datos, sino estimar de manera precisa los valores faltantes dentro de un conjunto de datos específico.

Métricas de desempeño

Para evaluar los modelos se utilizaron las siguientes métricas: la desviación media del sesgo (MBD), que indica si el modelo tiende a sobreestimar o subestimar sistemáticamente los valores; el error absoluto medio (MAE), que mide el tamaño promedio del error sin considerar su dirección; y el desvío cuadrático medio (RMSD), que penaliza más los errores grandes y refleja la dispersión general del error.

$$MBE = \frac{1}{n}\sum_{i=1}^{n} y_i - x_i \tag{3}$$

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{5}$$

donde:

x = Valor medido
y = Valor estimado
n = Tamaño de la muestra

Estas métricas pueden expresarse en términos relativos como porcentaje de la media de los valores medidos $(\overline{x})$, denominados aquí rMBE, rRMSD y rMAE respectivamente.

RESULTADOS

Comparación entre la serie de referencia y las series estimadas

Antes de comparar estadísticamente las series, se analiza la distribución de condiciones de cielo mediante el índice de claridad $k_t = GHI/TOA$. La Fig. 2 muestra que el valor más frecuente es $k_t \approx 0{,}75$, indicando que la mayor parte de los registros corresponden a cielo despejado y que los períodos fuertemente nublados son minoritarios en el sitio.

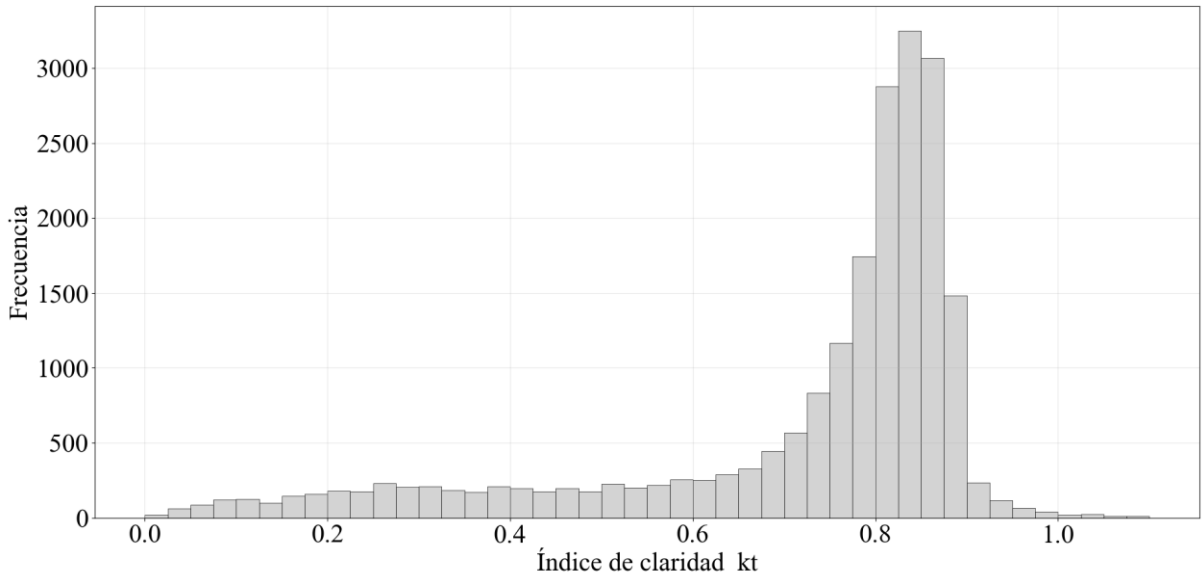


Figura 2: Distribución del índice de claridad (k_t)para la serie de referencia (15 min).

En la Fig. 3 se presentan los diagramas de caja (boxplot) que resumen la distribución de los valores de las distintas series de GHI, incluyendo sus cuartiles, medianas y valores extremos. Esto nos permite caracterizar y comparar las distintas series de irradiancia consideradas.

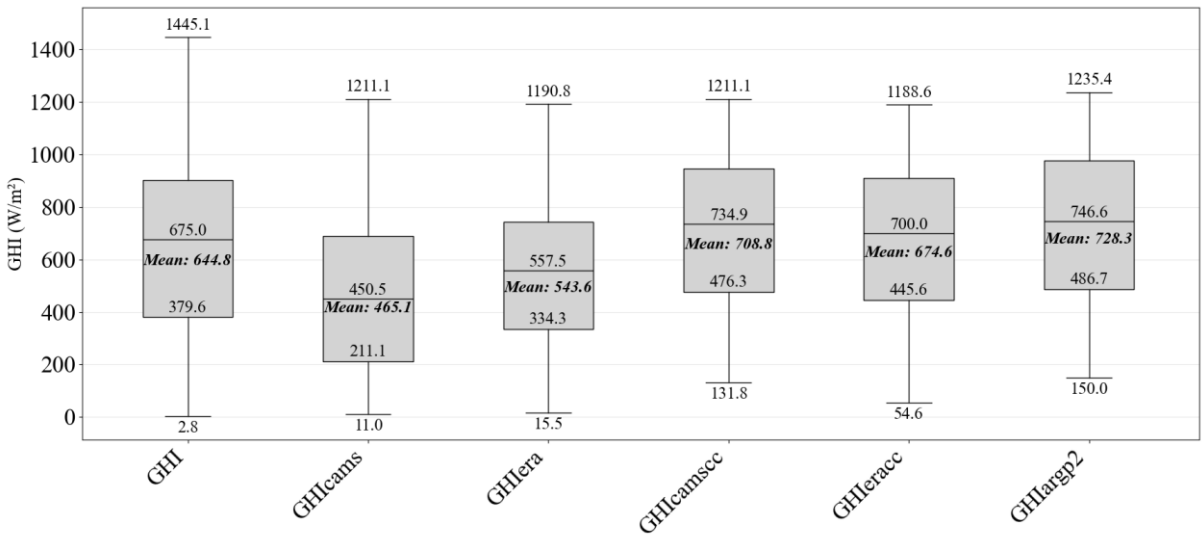


Figura 3: Diagrama de caja (boxplot) de las cinco series de irradiancia global horizontal (GHI) consideradas en el estudio, correspondientes a diferentes fuentes de datos (CAMS, ERA5 y ARGP2). Las series presentan una frecuencia temporal de 15 minutos.

El comportamiento relativo de los productos de GHI modelado frente a las mediciones se resume en la Fig. 4. A simple vista, tanto CAMS como ERA5 subestiman la GHI ($\sim 27\%$ y $\sim 15\%$ respectivamente)

bajo toda condición de cielo, mientras que los modelos de cielo claro, aunque sobreestiman, exhiben desviaciones menores. Esta diferencia confirma la limitación de CAMS para representar la GHI en sitios como el estudiado, señalada por los autores Qu et al. (2017) y Ledesma et al. (2025). Llama la atención que GHIcamscc muestre un mejor desempeño que GHIargp2, aunque este último fue calibrado incluyendo El Rosal entre sus estaciones de ajuste; sin embargo, ese resultado es coherente con lo reportado por los propios autores de ARGP2 (Ledesma et al., 2023).

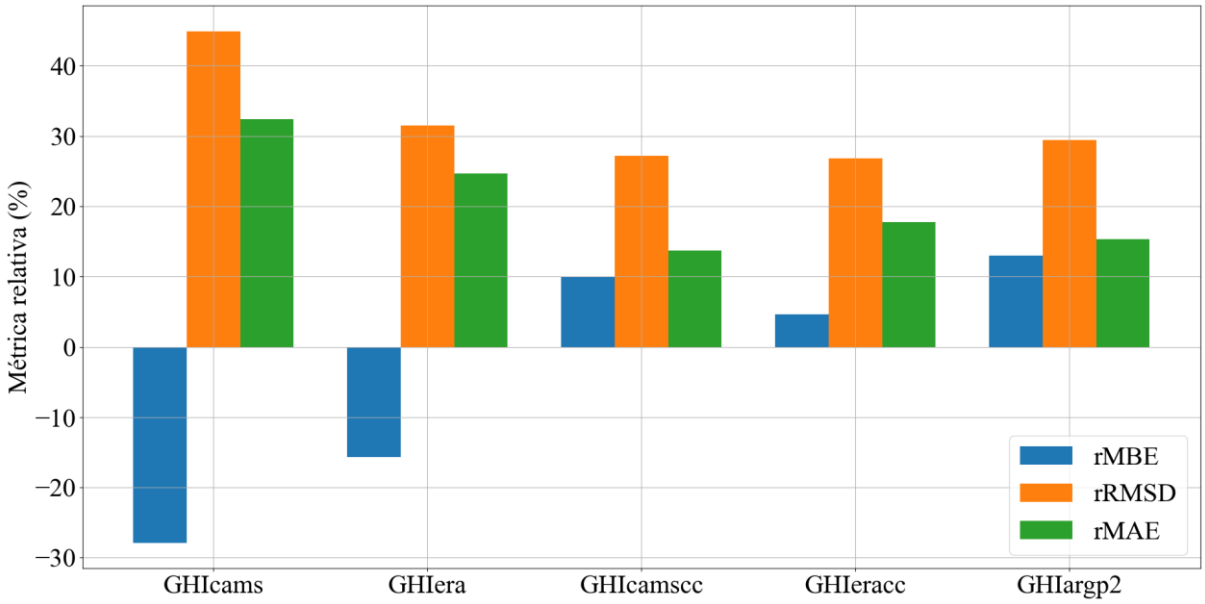


Figura 4: Métricas de desempeño entre los valores observados (GHI) y los valores estimados por CAMS, ERA5 y ARGP2, considerando los valores medios presentados en al Fig.2

En cuanto a las métricas con residuos cuadráticos, GHIeracc presentó un mejor desempeño, seguido por GHIcamscc, GHIargp2, GHIera, y finalmente GHIcams, que registró los mayores errores. Sin embargo, al considerar el error absoluto medio relativo, el orden cambia: el valor más bajo corresponde a GHIcamscc, seguido por GHIargp2, GHIeracc, GHIera y, nuevamente, GHIcams. La predominancia de cielo claro detectada en la distribución de k_t (ver Fig. 2) explica por qué los modelos de cielo claro registran los errores más bajos.

Estas diferencias sugieren que mientras GHIeracc minimiza mejor los errores grandes, GHIcamscc podría estar más cerca de los datos de la serie de referencia, lo que indica un comportamiento más consistente y robusto frente a valores extremos.

Evaluación de los modelos

En las Tablas 3 y 4 se presentan los valores de rMBE, rRMSD y rMAE obtenidos para cada modelo y conjunto de datos en los distintos escenarios de huecos sintéticos, para los casos de mayor frecuencia y corta duración, y para los casos de menor frecuencia y larga duración respectivamente. Los valores relativos están referidos a la media de los valores medidos correspondientes a los huecos generados en cada escenario. Los resultados muestran que el modelo MLP —arquitectura de dos capas ocultas con 9 y 18 neuronas, y regularización L2 con $\alpha = 0{,}01$ (factor de penalización de los pesos grandes para reducir el sobre-ajuste)— obtuvo el mejor desempeño en términos generales, logrando los valores más bajos de error relativo en comparación con los demás enfoques.

En particular, los modelos basados en ML (SLR, MLR y MLP) superan en precisión a las estimaciones de los conjuntos de datos satelitales CAMS y ERA5. Estos resultados resaltan la eficacia de los enfoques de ML en la imputación de datos faltantes en series temporales de GHI, especialmente en escenarios con grandes pérdidas de información.

Tabla 3: Resultados de rMBE, rRMSD y rMAE para diferentes modelos y porcentajes de días con datos faltantes en el caso de mayor frecuencia y corta duración.

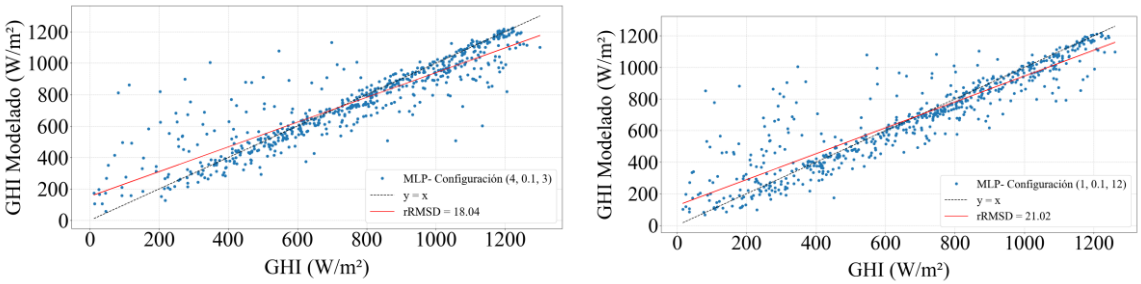
Modelo	**10%**			**30%**			**50%**		
	rMBE	**rRMSD**	**rMAE**	**rMBE**	**rRMSD**	**rMAE**	**rMBE**	**rRMSD**	**rMAE**
CAMS	-25.5	43.4	30.1	-26.2	43.1	30.3	-25.9	42.1	29.6
ERA5	-14.7	31.3	25	-15.8	31	24.7	-16.1	30.4	24.2
SLR (GHIcamscc)	-0.4	25.8	17.2	-0.7	24.4	16.7	-1.7	23.4	16
SLR (GHIcams)	0.6	33.2	26.3	1.5	32.5	25.7	1.2	31.5	25
SLR (GHIeracc)	0.7	26.4	19.1	0.7	25.7	18.6	-1.6	24.9	17.9
SLR (GHIera)	0.7	27.6	21.8	0.6	26.7	20.7	-0.5	25.7	19.8
SLR (GHIargp2)	-0.7	26.6	17.9	-1	25.3	17.4	-2.4	24.2	16.8
MLR	0.6	24.1	15.5	0.2	24.4	15.7	-0.5	21.4	13.9
MLP	0.9	21	12.6	0.1	20.2	12.3	-0.3	19	11.4

Tabla 4: Resultados3 de rMBE, rRMSD y rMAE para diferentes modelos y porcentajes de días con faltantes en el caso de menor frecuencia y larga duración.

Modelo	**10%**			**30%**			**50%**		
	rMBE	**rRMSD**	**rMAE**	**rMBE**	**rRMSD**	**rMAE**	**rMBE**	**rRMSD**	**rMAE**
CAMS	-25.7	41.6	29.4	-26.9	42.4	30.4	-26.5	41.7	29.8
ERA5	-15.8	28.1	22.3	-16.5	29.3	23.5	-16.8	29.2	23.4
SLR (GHIcamscc)	-2.4	21.2	14.5	-1.5	22.3	15.7	-2	22.4	15.6
SLR (GHIcams)	-4.4	31	24.4	-1.7	30.9	24.4	-1.4	30.1	23.7
SLR (GHIeracc)	-2.7	21.3	15.3	-2	23.3	17.2	-2.5	23.4	17
SLR (GHIera)	-2	23.4	18.2	-1.6	24.2	18.8	-1.8	24	18.4
SLR (GHIargp2)	-2.9	21.9	15.3	-1.9	23.1	16.6	-2.5	23.2	16.4
MLR	-1	19.9	13.1	-0.6	20.4	13.8	-0.7	20.4	13.5
MLP	-0.9	18	10.8	-0.4	18.4	11.3	-0.4	18.1	11.1

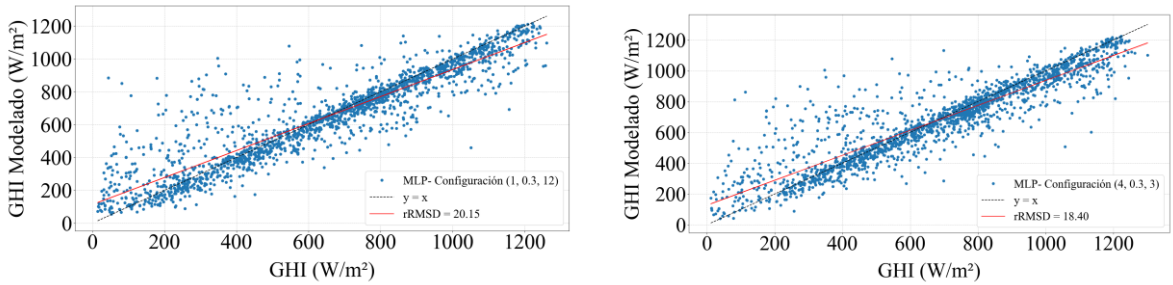
En las Figs. 5, 6 y 7 se presentan los gráficos de dispersión entre la medición terrestre y la estimación

obtenida con el MLP, para las configuraciones correspondientes al 10 %, 30 % y 50 % de días con datos faltantes en la serie respectivamente. A simple vista se observa que, en todos los casos considerados, el modelo presenta una sobreestimación en condiciones de baja irradiancia (ej: cielos nublado, amanecer/atardecer), mientras que en condiciones de alta irradiancia (ej: cielos despejados, mediodía solar) el modelo subestima.



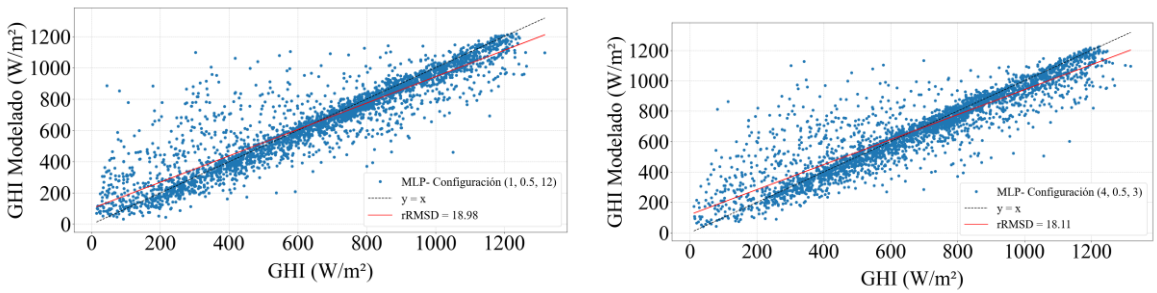
(a) Configuración: 12 huecos de tamaño 1. (b) Configuración: 3 huecos de tamaño 4.

Figura 5: Gráfico de dispersión entre la medida y la estimación de GHI para el caso de un 10 % de días con datos faltantes.



(a) Configuración: 12 huecos de tamaño 1. (b) Configuración: 3 huecos de tamaño 4.

Figura 6: Gráfico de dispersión entre la medida y la estimación de GHI para el caso de un 30 % de días con datos faltantes.



(a) Configuración: 12 huecos de tamaño 1. (b) Configuración: 3 huecos de tamaño 4.

Figura 7: Gráfico de dispersión entre la medida y la estimación de GHI para el caso de un 50 % de días con datos faltantes.

Desempeño de los modelos ante distintos patrones de huecos

Aunque se evaluaron dos patrones extremos de huecos -corta duración y alta frecuencia frente a larga duración y baja frecuencia-, no se observan diferencias significativas en el error atribuibles al tamaño o frecuencia de los huecos. Como muestran las Tablas 3 y 4, las variaciones en rRMSD entre ambos escenarios son menores al 3 % en todos los casos.

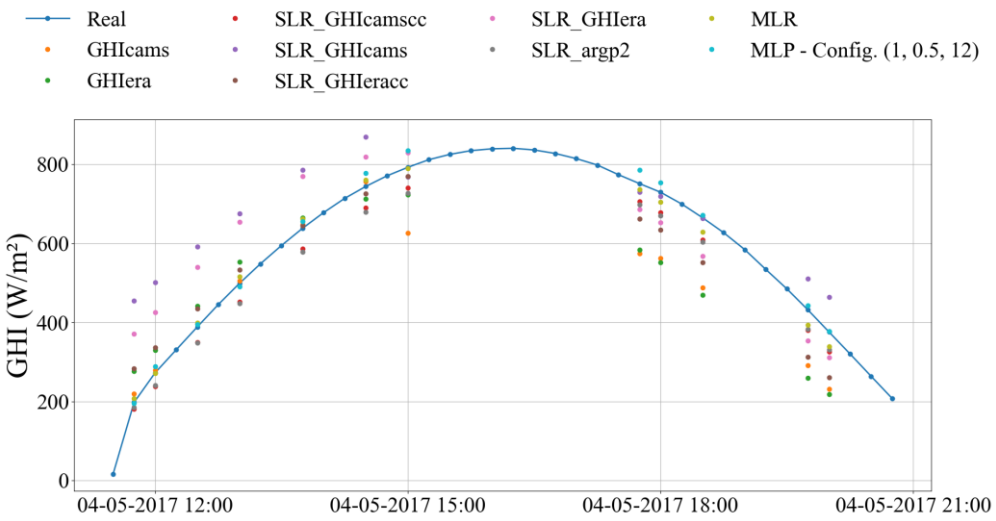
En cambio, las diferencias relevantes provienen del modelo utilizado. El MLP logra los menores errores en ambos escenarios, con rRMD entre 18.0 % y 21.0 %, seguido por la MLR, con valores entre 19.9 % y 24.1 %. Los modelos SLR presentan mayor dispersión, dependiendo de la variable regresora empleada. La GHIcamscc es la que mejores resultados aporta dentro de este grupo, con rRMSD entre 21.2 % y 25.8 %, muy por debajo de otras opciones como GHIcams o GHIera.

Para ilustrar la capacidad de reconstrucción dentro de los huecos, la Fig. 8 (cielo claro) y la Fig. 9 (cielo nublado) presentan un día de ejemplo para las configuraciones extremas:

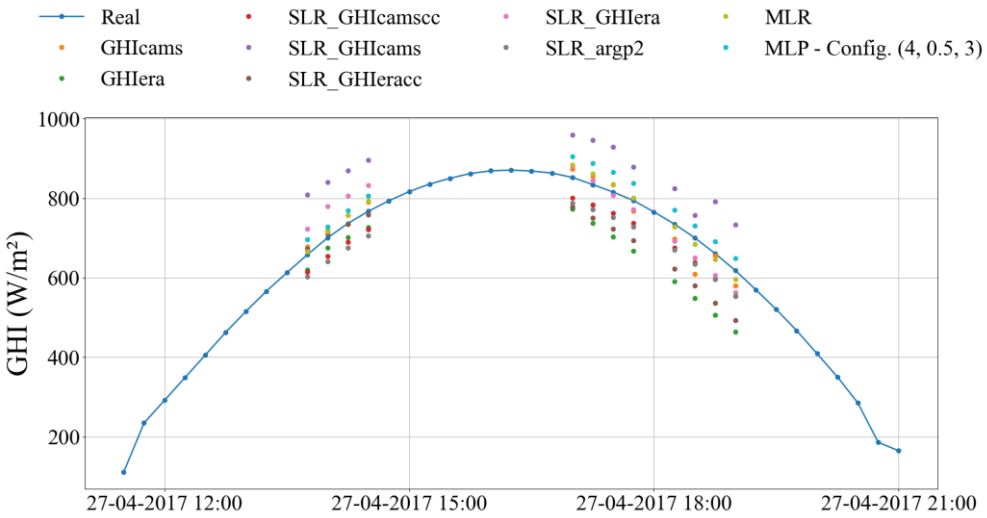
- 12 huecos de tamaño 1 (frecuente y breve)
- 3 huecos de tamaño 4 (poco frecuente y prolongado)

ambas evaluadas con 10 % y 50 % de días con datos faltantes. En cada gráfico se indica la serie medida (en azul) junto con las estimaciones de todos los modelos, permitiendo observar directamente cómo cada método recupera los datos de irradiancia dentro de los intervalos perdidos.

Cabe destacar que para el entrenamiento de los modelos se consideraron todos los datos disponibles en la serie, teniendo en cuenta el escenario de datos faltantes generados en cada caso.



(a) Configuración: 12 huecos de tamaño 1. – 50 % de días con datos faltantes

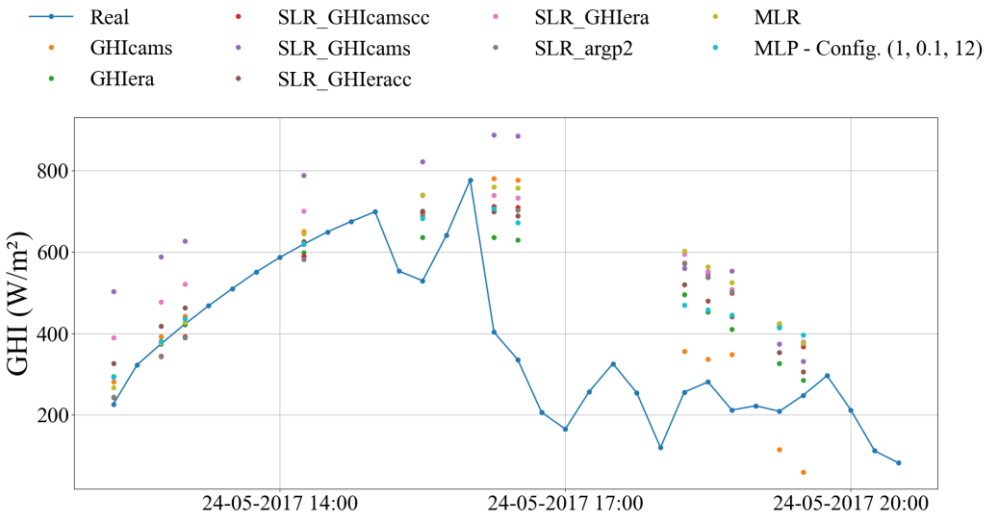


(b) Configuración: 3 huecos de tamaño 4. – 50 % de días con datos faltantes

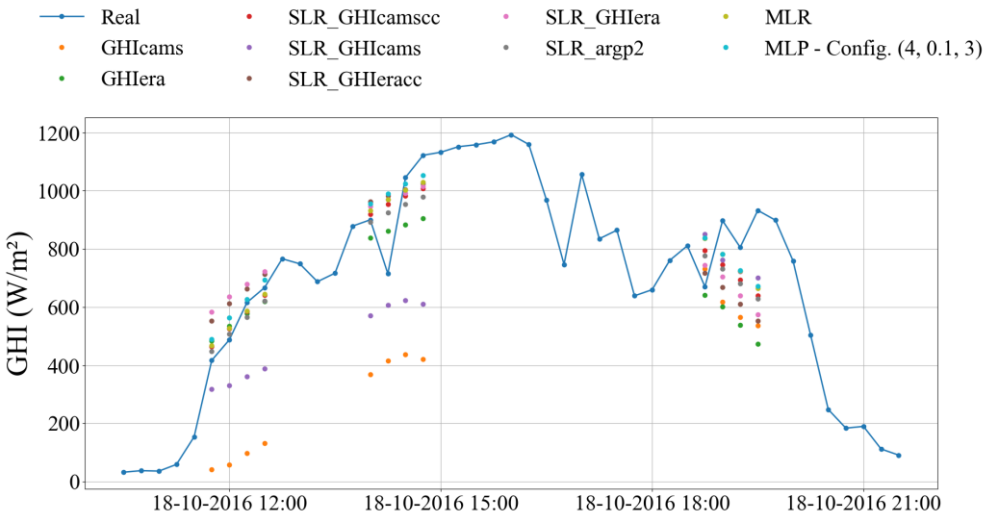
Figura 8: Día de cielo claro ejemplo para 2 configuraciones.

DISCUSIÓN

Los resultados obtenidos demuestran que el modelo MLP es superior en la reconstrucción de series de GHI con huecos sintéticos, superando tanto a los modelos más sencillos que solo consideran un ajuste lineal de los datos (SLR y MLR) como a los conjuntos de datos satelitales y de reanálisis considerados en este trabajo (CAMS y ERA5). Esto se ve evidenciado en las métricas de desempeño (rRMSD, rMBE y rMAE).



(a) Configuración: 12 huecos de tamaño 1. – 10 % de días con datos faltantes



(b) Configuración: 3 huecos de tamaño 4. – 10 % de días con datos faltantes

Figura 9: Día de cielo nublado ejemplo para 2 configuraciones.

El mejor desempeño del modelo MLP puede atribuirse a su capacidad para capturar relaciones no lineales en los datos, lo que es especialmente útil dada la variabilidad inherente de la radiación solar. Sin embargo, es importante destacar que los modelos lineales (SLR y MLR) también mostraron un desempeño aceptable, particularmente en configuraciones con huecos de mayor tamaño y menor frecuencia. El modelo MLR, en particular, demostró ser una alternativa viable al mantener el segundo mejor desempeño en términos de rRMSD (entre 19.9 % y 20.4 %), mientras que el modelo SLR utilizando la GHI de cielo claro de CAMS (GHIcamscc) como variable regresora también mostró resultados prometedores (rRMSD entre 21.2 % y 22.3 %).

Al comparar los resultados de los modelos de ML con las estimaciones proporcionadas por CAMS y ERA5, se observó que ambos conjuntos de datos presentaron un desempeño inferior en términos de rRMSD, rMBE y rMAE. Esto sugiere que, aunque CAMS y ERA5 son herramientas útiles para la estimación de GHI a gran escala, su precisión puede verse comprometida cuando se considera un sitio de superficie reducida.

Es importante señalar que la irradiancia global horizontal (GHI) presenta un marcado ciclo intradiario asociado a la geometría solar, lo cual impacta directamente en la matriz de correlación. Variables como el ángulo cenital solar, la masa de aire o la irradiancia en el tope de la atmósfera aparecen fuertemente correlacionadas con la GHI. Una estrategia habitual para reducir este sesgo es el empleo de índices de claridad (k_t o k), que normalizan la irradiancia respecto a su valor teórico máximo y aíslan en ma yor medida la variabilidad atmosférica. En este trabajo, sin embargo, se optó por trabajar directamente con GHI, dado que la magnitud absoluta de irradiancia es la variable de mayor interés en aplicaciones energéticas y de dimensionamiento de sistemas solares.

Además, se reconoce que algunas de las variables seleccionadas presentan redundancia, en particular las asociadas a la geometría solar (sza, TOA, mak, ie) y las distintas estimaciones de irradiancia (GHIcams, GHIcamscc, GHIera, GHIeracc). No se aplicó un filtrado adicional para eliminar estas redundancias, ya que el objetivo del estudio fue evaluar el desempeño relativo de cada fuente de datos y su utilidad en la reconstrucción de huecos. Aunque la redundancia puede influir en la interpretación de las correlaciones y en la importancia relativa de los predictores, no altera las conclusiones principales, dado que los modelos se entrenaron y evaluaron bajo condiciones consistentes.

A pesar de los resultados prometedores, es importante reconocer algunas limitaciones del estudio. Por

ejemplo, el desempeño de los modelos podría variar en regiones con condiciones climáticas diferentes a la considerada en este trabajo. Asimismo, este trabajo no tiene como objetivo la obtención de un modelo generalizable a nuevas series temporales, sino la reconstrucción de patrones dentro de la misma serie. En este contexto, cierto grado de sobreajuste controlado puede ser aceptable o incluso útil, siempre que se gestionen adecuadamente los riesgos asociados, como se hace mediante el uso de técnicas como early stopping y validación cruzada KFold.

Cabe destacar que los modelos se aplicaron a una serie real previamente filtrada mediante el control de calidad; únicamente la posición y el tamaño de los huecos fueron controlados para simular condiciones operativas. Por tanto, las métricas reportadas deben tomarse como referencia: en la práctica la cantidad y extensión de los huecos son estocásticos y podrían incrementar tanto el sesgo como la dispersión.

CONCLUSIONES

En este trabajo se evaluaron métodos de imputación atemporal de irradiancia solar global (GHI) mediante tres enfoques de machine learning (SLR, MLR y MLP) y bases de datos satelitales y de reanálisis (CAMS, ERA5). El objetivo fue identificar la estrategia más precisa para estimar datos faltantes bajo distintos porcentajes de días (10 %, 30 % y 50 %) con datos faltantes, y dos patrones de huecos sintéticos (corta y larga duración).

Los resultados muestran que el tamaño o la frecuencia de los huecos no influyen significativamente en el error; las diferencias de rRMSD entre ambos patrones fueron siempre menores al 3 %. La jerarquía de desempeño está determinada por el modelo: MLP alcanza los menores errores (rRMSD entre 18 % y 21 %), seguido por MLR (rRMSD entre 20 % y 24 %) y SLR con GHIcamscc (rRMSD entre 21 % y 26 %). Todas las métricas se calcularon respecto a la media local de cada escenario, lo que refuerza el carácter atemporal del análisis.

La capacidad del MLP para reconstruir series, incluso con altos porcentajes de ausencia, resulta clave para la gestión de sistemas fotovoltaicos y la estimación de generación solar. Los modelos lineales (MLR y SLR) siguen siendo útiles cuando se prioriza simplicidad o bajo costo computacional.

Cabe señalar que las series de cielo claro de CAMS y ERA5 reproducen mejor la variabilidad diaria de la GHI medida que sus homólogas para toda condición de cielo, lo cual se justifica por el predominio local de días despejados. En consecuencia, es esperable que la imputación de huecos correspondientes a días no claros presente una precisión menor.

Como trabajo futuro se propone explorar las diferentes configuraciones de los modelos que pueden ser aplicados para la imputación de datos faltantes, como así también su aplicabilidad en otras ubicaciones geográficas y bajo diferentes condiciones atmosféricas.

REFERENCIAS

Alonso-Suárez, R. (2017). *Estimación del recurso solar en Uruguay mediante imágenes satelitales*. Tesis doctoral, Universidad de la República.

Cabrera, S. R., Teixeira-Branco, V., Medeiros, J. V. F. F., y Alonso-Suárez, R. (2024). Accurate estimation of solar pv power plant capacity factors in Uruguay through detailed quality control and satellite gap filling. En *2024 IEEE URUCON*, pp. 1–5.

Demirhan, H. y Renwick, Z. (2018). Missing value imputation for short to mid-term horizontal solar irradiance data. *Applied Energy*, 225:998–1012.

He, M., Luo, Z., Xie, X., Wang, P., Wang, H., y Zapata-Lancaster, G. (2025). Gap filling crowdsourced air temperature data in cities using data-driven approaches. Building and Environment, 271:112593.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., y Thépaut, J.-N. (2020). The era5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730):1999–2049.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., y Suttie, M. (2019). The cams reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, 19(6):3515–3556.

Iturbide, P., Alonso-Suarez, R., y Ronchetti, F. (2023). An analysis of satellite-based machine learning models to estimate global solar irradiance at a horizontal plane. En Naiouf, M., Rucci, E., Chichizola, F., y De Giusti, L., editores, *Cloud Computing, Big Data & Emerging Topics*, pp. 118–128, Cham. Springer Nature Switzerland.

Ledesma, R., Alonso-Suárez, R., Salazar, G., Nollas, F., y Vilela, O. (2025). Evaluation of satellite and reanalysis models for solar irradiance estimation in northwest argentina. *IEEE Latin America Transactions*, 23(8):706–717.

Ledesma, R. D., Salazar, G. A., y de Castro Vilela, O. (2023). Argp-v2 un modelo práctico para la estimación de irradiancia global horizontal en condiciones de cielo claro para sitios de altura. *Avances en Energías Renovables y Medio Ambiente - AVERMA*, 26:283–289.

Muneer, T. y Fairooz, F. (2002). Quality control of solar radiation and sunshine measurements – lessons learnt from processing worldwide databases. *Building Services Engineering Research and Technology*, 23(3):151–166.

Nollas, F. M., Salazar, G. A., y Gueymard, C. A. (2023). Quality control procedure for 1-minute pyranometric measurements of global and shadowband-based diffuse solar irradiance. *Renewable Energy*, 202:40–55.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenho fer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., y Louppe, G. (2012). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12.

Peel, M. C., Finlayson, B. L., y McMahon, T. A. (2007). Updated world map of the köppen-geiger climate classification. *Hydrology and Earth System Sciences*, 11(5):1633–1644.

Qu, Z., Oumbe, A., Blanc, P., Espinar, B., Gesell, G., Gschwind, B., Klüser, L., Lefèvre, M., Saboret, L., Schroedter-Homscheidt, M., y Wald, L. (2017). Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method. *Meteorologische Zeitschrift*, 26(1):33–57.

Schwandt, M., Chhatbar, K., Meyer, R., Fross, K., Mitra, I., Vashistha, R., Giridhar, G., Gomathinaya gam, S., y Kumar, A. (2014). Development and test of gap filling procedures for solar radiation data of the indian srra measurement network. *Energy Procedia*, 57:1100–1109. 2013 ISES Solar World Congress.

Sengupta, M., Habte, A., Wilbert, S., Gueymard, C., y Remund, J. (2021). Best practices handbook for the collection and use of solar resource data for solar energy applications: Third edition. Technical report, National Renewable Energy Lab. (NREL), Golden, CO (United States).

RECONSTRUCTION OF GLOBAL HORIZONTAL IRRADIANCE SERIES WITH SYNTHETIC GAPS USING MACHINE LEARNING MODELS AND SATELLITE DATA. CASE STUDY: EL ROSAL, SALTA

ABSTRACT: Global Horizontal Irradiance (GHI) is a key variable for the design and optimization of solar energy systems, but its records often contain missing data due to instrument failures or maintenance. This work compares three machine-learning-based imputation strategies —simple linear regression (SLR), multiple linear regression (MLR) and multilayer perceptron (MLP)— for filling missing data. Satellite products (CAMS), reanalysis data (ERA5) and estimates from the ARGP2 model are also used to improve prediction accuracy. To evaluate model performance, synthetic gaps were generated semi-randomly in a GHI time series from El Rosal, Salta, Argentina. In all cases MLP achieved the lowest errors (rRMSD entre 18 % y 21 %), followed by MLR (rRMSD entre20 % y 24 %) and SLR (rRMSD entre 21 % y 26 %).

Keywords: Solar irradiance, machine learning, CAMS, ERA5, time series, gap-filling.