ASADES

Avances en Energías Renovables y Medio Ambiente
Vol. 28, pp. 308-320, 2024, ISSN 2796-8111
ASADES 2024

ANÁLISIS DE DESEMPEÑO DE DIFERENTES TÉCNICAS DE APRENDIZAJE AUTOMÁTICO EN UNA ADAPTACIÓN AL SITIO DE IRRADIANCIA SOLAR GLOBAL PARA SALTA (ARGENTINA).

Germán Salazar[1,3], Rubén Darío Ledesma[1,3], Constanza López Ruiz[1], Olga de Castro Vilela[2]
[1]Grupo de Estudio y Evaluación de la Radiación Solar (GEERS) - INENCO - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
[2]Centro de Energías Renovaveis (CER) – Universidade Federal de Pernambuco - Brasil
[3]Facultad de Ciencias Exactas – Universidad Nacional de Salta - Argentina
e-mail: german.salazar@conicet.gov.ar

RESUMEN: En este trabajo se analizan diferentes métricas comparando valores medidos de Irradiancia Solar Global contra valores estimados usando diferentes técnicas de Inteligencia Artificial, particularmente Aprendizaje Automático (Machine Learning), con el objetivo de realizar una Site Adaptation a través de la Base de Datos Satelitales CAMS-Rad para la ciudad de Salta (1200 metros sobre el nivel del mar). Se analizaron los datos con integraciones temporales de 5 minutos y 15 minutos. Se usaron dos periodos diferentes de datos medidos: un año (2014) o dos años (2014-2015) para entrenar y validar. Las técnicas de Machine Learning usadas fueron Regresión Lineal Simple y Múltiple, Quantile Mapping, Multilayer Perceptron, XGBoost y regresión Clusterwise. Todas las métricas indican una mejora sobre los estimados por CAMS-Rad, destacándose las regresiones Clusterwise.

Palabras clave: irradiancia solar global, aprendizaje automático, adaptación al sitio, Salta.

INTRODUCCIÓN

El conocimiento de las características de distribución espaciotemporal del recurso solar resulta de interés para poder aprovecharlo de manera óptima, ya sea por medio de sistemas fotovoltaicos o de concentración (Vignola et al, 2012; Huld et al, 2012). Se sabe que el Noroeste argentino (NOA) es una de las regiones del planeta con una irradiación superior a la media mundial, pero dicha distribución no es homogénea en la región debido a su orografía (Solargis.com). Los mapas actuales de distribución de la radiación solar en el NOA no provienen de redes de medición, sino de estimaciones de Bases de Datos Satelitales (BDS) (Laspiur et al, 2013): esto es porque no existen redes radiométricas en la región que midan de manera constante y con protocolos de mantenimiento y recalibración. Se han realizado estudios que muestran que las estimaciones con modelos satelitales tienen diferencias respecto de los valores reales (Sengupta et al, 2018), lo que implica que usar valores de irradiancia solar estimada usando imágenes satelitales inducirá un error en las posteriores estimaciones que se puedan llegar a hacer usando esa información, por ejemplo, generación de energía (Viana et al, 2011).

Es posible adaptar los valores de irradiancia solar de una BDS a partir de los valores medidos si se dispone de una serie de mediciones, en un sitio especifico, durante un periodo temporal simultaneo dentro de la extensión de la BDS (Polo et al, 2016, 2020). La idea es encontrar una función que *adapte*, es decir, mejore los estimados satelitales acercándolos a los valores efectivamente medidos en tierra. Es importante mencionar que no se habla de "corrección" sino de "adaptación" ya que corregir implica que los valores finales sean iguales a los medidos. Este método se denomina habitualmente Adaptación al Sitio (AS) o Site Adaptation, en inglés. Así, encontrada dicha función, se podría aplicar la misma a todos los valores de la BDS *fuera del periodo* de simultaneidad, adaptando años de datos satelitales a partir de una muestra de datos medidos en superficie.

Se han estudiado varias metodologías y modelos para encontrar esa función siendo el Machine Learning (ML) el más utilizado (Chu et al, 2024). En este trabajo analizamos los resultados obtenidos al utilizar

técnica de ML para adaptar los valores de la BDS CAMS-Rad (SoDa-pro.com) para la ciudad de Salta (Argentina). La eficacia de cada técnica se muestra analizando las métricas de comparar los valores irradiancia solar medidos en Salta, durante el periodo 2013-2015, contra valores modelados usando técnicas de ML, siendo las mismas Regresión Lineal Simple (RLS) y Múltiple (RLM), Quantile Mapping (QM), Multilayer Perceptron (MLP), XGBoost (XGB) y regresión Clusterwise (RCws). Resulta destacable mencionar que todos los modelos utilizados en este trabajo mejoran los estimados por CAMS-Rad.

MATERIALES UTILIZADOS

Sitio de Estudio e *Instrumentos utilizados*

La ciudad de Salta es la capital de la provincia de Salta (Argentina). Las mediciones de Irradiancia Solar Global (GHI por sus siglas en inglés) se realizaron en el Campus de la Universidad Nacional de Salta La posición geográfica de la estación radiométrica del Grupo de Estudio y Evaluación de la Radiación Solar (GEERS) es latitud 24.7288° Sur, longitud 65.4095° Oeste. La altitud es de 1233 metros sobre el nivel del mar. La clasificación del clima en la ciudad de Salta es Cwb según Köppen-Geiger (Peel y Fynlayson, 2007).

El radiómetro utilizado para adquirir los datos analizados fue uno marca Eppley modelo PSP. Dado que es un piranómetro antiguo, el mismo se recalibra cada 6 meses contra un piranómetro marca Kipp & Zonen modelo CM21, cuya contante de calibración esta referenciada a instrumentos contrastados en Davos (Suiza) por la WRC.

El radiómetro PSP está conectado a un datalogger marca Campbell Scientifics modelo CR1000. La frecuencia de medición fue programada a 5 segundos, guardándose el promedio cada 1 minuto.

Base de datos satelital usada

La BDS utilizado en este trabajo es el Servicio Atmosférico Copernicus (CAMS), que integra modelos atmosféricos de aerosoles de última generación con datos de observación de la Tierra para proporcionar servicios de información que cubren la calidad del aire europeo, la composición atmosférica global, el clima, la energía solar y radiación Ultravioleta (SoDa-pro.com). El Servicio de Radiación Solar CAMS (CAMS-Rad) proporciona una parametrización rápida de la transferencia radiativa en la atmósfera. Vincula parámetros del cielo despejado, como los aerosoles, el vapor de agua y el ozono, con información sobre las nubes obtenida por satélite. Dado que la información sobre nubes de alta resolución se infiere directamente de las observaciones por satélite, las series temporales de irradiación para condiciones nubosas sólo están disponibles actualmente para el campo de visión del satélite Meteosat de Segunda Generación (MSG), que abarca aproximadamente Europa, África, el Océano Atlántico y Oriente Medio (+66° a -66° tanto en latitud como en longitud). El sitio estudiado se sitúa prácticamente en el borde oeste de la imagen del satélite, con una longitud de 65.4095°O para Salta. El modelo utilizado para la estimación de la irradiancia es Heliosat-4 (SoDa-pro.com)

Características de los datos medidos:

Los valores de GHI usados se registraron desde el 1/1/2013 al 31/6/2015, con frecuencia de 1 minuto. Se aplicaron a los datos una serie de filtros de Control de Calidad (Ecuaciones 1) diseñados para ser aplicados solo a valores de GHI (Nollas et al, 2022):

$$GHI < 1.5\ S\ Esc\ (\cos SZA)^{1.2} + 100\ W/m^2 \quad (1.1)$$

$$GHI > (6.5331 - 0.065502\ SZA + 1.8312 \times 10^{-4}\ SZA^2) / (1 + 0.01113\ SZA) \quad (1.2)$$

$$K_t < 1.4 \quad (1.3)$$

donde S es el factor de corrección de la distancia Tierra-Sol, Esc es la Irradiancia Solar Total (1361 W/m^2), SZA es el ángulo cenital solar y K_t es el índice de claridad, igual a GHI/I_{ext}. Del 100% de datos

diurnos utilizados, 95.45% pararon el primer filtro, 95.44% pasaron el filtro 2 y 3. En este trabajo solo se analizaron datos para SZA>80°, lo que representó el 82.86% de los datos.

Se usaron dos tipos de integración temporal para comparar los valores de GHI medidos y modelados: 5 minutos y 15 minutos. La integración de 5 minutos se hizo para explorar las altas frecuencias en las técnicas de ML usadas. La integración de 15 minutos se hizo para establecer una base temporal inferior a la de las imágenes de satélite en su formato nativo. Dado que este trabajo pretende comparar las prestaciones de diferentes ML técnicos para realizar AS, esta frecuencia parece óptima para esta tarea.

TÉCNICAS DE MACHINE LEARNING USADAS

El ML aborda el problema de desarrollar algoritmos capaces de mejorarse a sí mismos con la experiencia (Mitchell, 1997). Los algoritmos de ML se dividen en dos categorías principales: Aprendizaje Supervisado (ApSu), en el que el algoritmo crea una función que relaciona las entradas con los resultados esperados, y Aprendizaje No Supervisado (ApNSu), en el que el algoritmo modela un conjunto de entradas sin disponer de ejemplos etiquetados. Los algoritmos de ApSu se dividen en *Clasificación* y *Regresión*. La *Clasificación* es una técnica utilizada para predecir la pertenencia a un grupo de instancias de datos (Faouzi y Colliot, 2024). Por otro lado, el ApNSu incluye algoritmos de *Agrupación* (clustering) y *Reducción Dimensional*, del tipo *Análisis de Componentes Principales* (ACP). En el estudio del AS, así como en otros procesos de ML relacionado a la radiación solar, se han descripto y utilizado muchas técnicas (Narvaez et al, 2021; Datha et al, 2022). En este trabajo se han utilizado técnicas de ML para hacer correlaciones, es decir, encontrar una función que relacione la variable regresora (o variables regresoras) con la variable objetivo, minimizando el error. Las más simples son las regresiones lineales. La que sigue en simpleza es el Mapa de Cuantiles (Quantile Mapping en inglés), la que utiliza una función de Distribución Acumulativa para corregir tendencias estadísticas en la distribución de los valores de irradiancia solar entre las series medidas y estimadas.

El perceptron multicapa (multilayer perceptrón, en inglés, MLP) es quizás la estructura más utilizada para el aprendizaje automático, por estar directamente relacionada con el concepto de Redes Neuronales Artificiales: un MLP es una Red Neurona Artificial. Otra técnica usada habitualmente en el análisis de series de irradiancia solar es el XGBoost (Li et al, 2022; Obiora et al, 2021; Zhang et al , 2023; Huang et al, 2025) que es un algoritmo de la familia de los Árboles de Decisión . Por último, se utiliza una técnica que no es habitual de usar en el análisis de series temporales de radiación solar, como lo es el Clusterwise: dicha técnica se encuentra bajo análisis de varios grupos de investigación de la región (Rodrigues et al, 2024; Ledesma et al, 2024)

Regresión Lineal Simple (RLS): esta función es la más simple posible ya que solo tiene un solo regresor.

En la ecuación 2 se muestra la forma de la función.

$$\mathrm{GHI_{adap}} = \mathrm{a.GHI_{BDS}} + \mathrm{b} \tag{2}$$

donde los coeficientes a y b se calculan a través de $\mathrm{GHI_{med}} = \mathrm{a.GHI_{BDS}} + \mathrm{b}$.

Regresión Lineal Múltiple (RLM): es una función lineal que utiliza más de un regresor. En la ecuación 3 se describe su forma general.

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_j x_{\mathrm{i,j}} + \mathrm{e}_i \tag{3}$$

En resumen, se trata de ajustar los datos a un modelo bajo los siguientes supuestos:

- Los residuos e_i son normales de media 0 y varianza común desconocida σ; además, estos residuos son independientes.
- El número de variables explicativas (j) es inferior al número de observaciones (i); esta hipótesis se conoce como rango completo.
- No existen relaciones lineales exactas entre las variables explicativas.

El estimador del vector paramétrico **β** es

$$\beta=(X^T X)^{-1} X^T y \tag{4}$$

donde las matrices ***y*** e ***X*** son:

$$y=\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_{i-1} \\ y_i \end{pmatrix} \qquad \boldsymbol{X}=\begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,j} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,j} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{i,1} & x_{i,2} & \dots & x_{i,j} \end{pmatrix}$$

Quantile Mapping (QM): El mapeo de cuantil (QM) es una técnica sencilla utilizada en modelado climático y meteorología para corregir la distribución de un parámetro modelado comparándolo con la distribución empírica de las observaciones (Panofsky and Brier, 1968). La metodología consiste en transformar los datos al dominio de probabilidad (cuantiles) e invertir la transformación utilizando la función de distribución acumulativa (CDF) como operador.

$$y_c = \mathrm{CDF}_0^{-1}[\mathrm{CDF}_m(x_m)] \tag{5}$$

XGBoost (XGB): El Extreme Gradient Boosting (Chen y Guestrin, 2016) es una técnica avanzada de Machine Learning que ha ganado gran popularidad debido a su eficiencia y precisión. En el contexto de la regresión, XGBoost se utiliza para predecir valores continuos, como por ejemplo temperaturas. Utiliza un enfoque iterativo basado en árboles de decisión y un método de optimización llamado boosting para mejorar continuamente la precisión de sus predicciones. El objetivo principal de la regresión con XGBoost es minimizar el error de predicción, construyendo un modelo fuerte a partir de la combinación de múltiples modelos débiles (árboles de decisión) de manera iterativa.

Para entender cómo funciona XGBoost, es esencial conocer algunos conceptos básicos: los árboles de decisión, el boosting y la función objetivo. Los árboles de decisión dividen los datos en segmentos basados en decisiones sobre características específicas. El boosting es una técnica de ensemble learning donde varios modelos se entrenan secuencialmente, cada uno intentando corregir los errores de los anteriores. La función objetivo de XGBoost combina una función de pérdida, que mide el error del modelo, con un término de regularización, que penaliza la complejidad del modelo. El proceso de XGBoost incluye varias etapas: inicialización con una predicción simple, construcción de árboles para predecir errores residuales, ajuste de predicciones, construcción de árboles sucesivos y optimización mediante regularización.

Multilayer Perceptron (MLP): El perceptrón multicapa (MLP)(Rumelhart et at, 1986) es una clase de red neuronal artificial (RNA) de tipo feedforward, es decir que la información fluye en una sola dirección: desde las entradas hacia las salidas a través de una serie de capas intermedias llamadas capas ocultas. Generalmente, cuando se hace referencia a una RNA, se trata de un MLP. Su estructura básica consta de tres capas de nodos: una capa de entrada, una o más capas ocultas y una capa de salida. Los nodos de las capas oculta y de salida son neuronas artificiales que utilizan una función de activación no lineal. El MLP utiliza una técnica de aprendizaje supervisado para el proceso de entrenamiento. El algoritmo de aprendizaje más utilizado en los MLP es la retro propagación combinada con un algoritmo de optimización. Este algoritmo permite al MLP minimizar el error en sus salidas en comparación con los resultados esperados, normalmente medidos por una función, a lo largo de múltiples iteraciones de entrenamiento o épocas.

Regresión Clusterwise (RCws): Se habla de regresión por conglomerados o regresión Clusterwise (Hastie et al, 2009) cuando se concatena un proceso de clustreing seguido de un proceso de regresión aplicado a cada grupo, con el objetivo de obtener mejores predicciones al reconocer que diferentes grupos dentro de los datos pueden tener relaciones distintas entre las variables explicativas y la variable de respuesta. En este caso se realizó primero un proceso de agrupación (clustering) y luego para la regresión se usó MLR. Este método ha mostrado ser de interés para el proceso de AS (Miranda et al,

2024) y ha generado una investigación (Ledesma et al, 2024). El método mirandaMLR se inicia realizando un proceso de clustering para asignar una de cinco clases a cada instante medido en el conjunto de entrenamiento, para lo que utiliza tres parámetros, VI (índice de variablidad), Kc (indice de cielo claro) y KDE (estimación de la densidad del kernel), este último es una forma no paramétrica para estimar la función de densidad de probabilidad de una variable aleatoria. En este caso el clustering se realiza mediante el algoritmo Kmeans, y utiliza centroides iniciales definidos por el autor (Miranda). En cambio, en el método randomMLR se realiza también un proceso de clustering sobre el conjunto de entrenamiento, pero utilizando dos parámetros, VI y Kc, aplicando el algoritmo Kmeans con una inicialización de centroides aleatorios. Es necesario destacar que para calcular KDE es necesario determinar un ancho de banda óptimo, aunque en la propuesta de mirandaMLR no es especifica el valor escogido, en este trabajo se utilizó 'la regla de oro' de (Silverman, 1986) para determinar el valor óptimo, indicando un valor de 0,1.

Una vez que los datos del conjunto fueron etiquetados con una clase, a partir del paso anterior, en el modelo MirandaMLR se realiza una clasificación, utilizando el algoritmo Random Forest, con el objetivo de poder inferir una clase (etiqueta) a partir de las variables regresoras modeladas, nótese que para definir estas clases se utilizaron variables derivadas la medida en el paso anterior (indice de variabilidad VI, índice de cielo claro Kc, Kernel Density Estimate KDE) (Miranda et al, 2024). Esta clasificación es aplicada y evaluada sobre el conjunto de validación. En el método RandomMLR esta clasificación se realiza utilizando el mismo algoritmo y las mismas variables, pero en este caso el algoritmo de clasificación se entrena en el conjunto de entrenamiento, pero se evalúa en el conjunto de prueba, esto implica una estrategia más rigurosa en cuanto al análisis en la obtención de un modelo óptimo para la clasificación, teniendo en cuenta en cómo se han definido los conjuntos de entrenamiento, validación y prueba.

Como paso final, en el método MirandaMLR se entrena un modelo de regresión lineal múltiple por cada una de las clases obtenidas en el conjunto de validación, luego se evalúa el resultado obtenido en el conjunto de prueba. En cambio, en el método RandomMLR cada modelo de regresión es entrenado a partir de los datos del conjunto de entrenamiento, utilizando el conjunto de validación nada más que para el evaluar el rendimiento de los distintos hiper-parámetros de los modelos, luego los modelos son evaluados en el conjunto de prueba.

Metodología de Entrenamiento, Validación y Prueba

Los procesos de ML requieren de un entrenamiento, que es donde "aprenden" a corregir la función que describirá la relación entre la serie de datos medidos y los estimados. Ese entrenamiento en realidad se compone de dos procesos: entrenamiento y validación, el que se aplica dentro de un mismo periodo temporal, separándose cierta cantidad de datos para entrenar y el restante para validar. Durante el entrenamiento, se trata de determinar la forma de la función/ecuación que correlaciona los valores medidos y estimados (emparejados) dentro del periodo de simultaneidad. La veracidad de esa función (entendiendo por veracidad al error que genera entre los valores medidos y los que estima) se prueba contra los datos de validación. Por lo general, se destinan 80% de los datos para entrenar y 20% para validar. Es recomendable que la asignación de esos datos (como dato de entrenamiento o dato de validación) sea aleatoria, para evitar la inducción de sesgos estacionales en las evaluaciones.

Sin embargo, esto no evita lo que se denomina overfitting, es decir, la sobre especialización de una función de correlación sobre una serie de datos. Resulta sorprendente que el desempeño de las técnicas de ML se mida generalmente con las métricas hechas sobre los datos de entrenamiento, de validación o de la concatenación de estos (Fig.1-b). Estas métricas mostrarán valores que indicarán una alta correlación entre los valores medidos y los adaptados, pero la real medida de esa correlación solo podrá ser percibida al medirse fuera del periodo de simultaneidad, al ser aplicada sobre los valores que no han sido usados para entrenar y validar (Fig.1-a). Este procedimiento se denomina de prueba (o Test) y las métricas sobre estos datos indicaran si el overfitting del periodo de simultaneidad afecta la implementación de la función sobre la extensión completa de los datos estimados por la BDS. Esta información será presentada para cada técnica utilizada en este artículo. Se usaron los datos de un año (2014) y de dos años (2014 y 2015) para entrenar y validar. Para el caso de un año/dos años, el 80% de

los datos de ese periodo se usaron para entrenar y el restante 20% de ese mismo periodo se usaron para validar. El periodo de prueba o test fue realizado con los datos del año 2013.

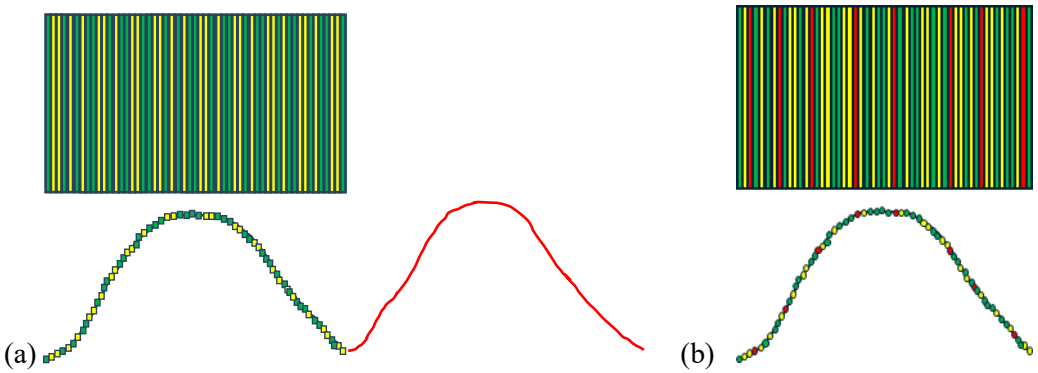

Figura 1. Comparación del procedimiento de Entrenamiento, Validación y Prueba realizado de dos maneras distintas. En (a) se usan los valores del periodo de simultaneidad para entrenar (puntos verdes) y validar (puntos amarillos), pero la prueba se hace fuera de ese periodo (puntos rojos). En (b) se usa el periodo de simultaneidad para realizar las tres operaciones.

RESULTADOS

A continuación, se muestran los resultados de la aplicación de los diferentes modelos a los datos de GHI medidos en Salta, en frecuencia de 5 minutos y de 15 minutos. Las métricas (Tabla 1 y 2) utilizadas para comparar los resultados fueron el error cuadrático medio (RMSE), el error de desvío medio (MBE) y el error absoluto de desvío medio (MABE)

$$\mathrm{RMSE}=\sqrt{\sum_{i=1}^{n}\left(\mathrm{GHI}_{\mathrm{med,i}}-\mathrm{GHI}_{\mathrm{mod,i}}\right)^{2}/n} \tag{6}$$

$$\mathrm{MBE}=\sum_{i=1}^{n}\left(\mathrm{GHI}_{\mathrm{med,i}}-\mathrm{GHI}_{\mathrm{mod,i}}\right)/n \tag{7}$$

$$\mathrm{MABE}=\sum_{i=1}^{n}\left|\mathrm{GHI}_{\mathrm{med,i}}-\mathrm{GHI}_{\mathrm{mod,i}}\right|/n \tag{8}$$

donde el subíndice *med* indica medido y el subíndice *mod* indica modelado. Para generar los valores porcentuales (rRMSE, rMBE y rMABE) se deben dividir por el promedio de los valores medidos.

Tabla 1. Métricas de la comparación de los valores medidos vs los adaptados por las distintas técnicas usadas, para una frecuencia de datos de 5 minutos. Entrenamiento y Validación se realizan sobre el mismo periodo (1 año = 2014, 2 años = 2014+2015), solo que el primero toma aleatoriamente el 80% de los datos mezclados y el segundo el 20% restante. El testeo o Prueba se realiza sobre todos los datos de 2013.

	5min					
	CAMS					
	Entrenamiento	Validación	Testeo	Entrenamiento	Validación	Testeo
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	37.9	38.3	37.5	37.9	37.2	37.5
rMBE(%)	5	3.64	5.5	4.9	3.3	5.5
rMABE(%)	23.8	23.9	22.9	23.8	23.9	22.9
	SLR					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	35.4	36	35.4	35.4	35.2	35.4
rMBE(%)	1.7	-1	0.11	-2.1	-0.9	0.2
rMABE(%)	23.5	23.5	23.2	23.5	23.3	23.2
	MLR					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	34.4	35.3	35.4	34.5	35.4	35.5
rMBE(%)	4.9	-0.8	-1.5	-4.2	-3.5	-1.45
rMABE(%)	22.7	22.9	23.7	22.7	23.6	23.7
	QM					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	36.7	37.3	36.9	36.7	36.6	36.9
rMBE(%)	0	1.38	-2	0	2.7	-2.1
rMABE(%)	22.4	22.6	22.8	22.4	23.1	22.8
	MLP					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	33.7	34.5	35	34.5	35.3	35.8
rMBE(%)	-0.9	-1.7	-3.4	-3	-3	-3.5
rMABE(%)	22.3	22.3	23.8	24.5	23.3	24.5
	XGBoost					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	33.5	33.5	35.6	34.9	34.9	35.5
rMBE(%)	-0.9	-0.9	-2.4	-4	-4	-2.1
rMABE(%)	21.7	21.7	23.8	23.6	23.6	23.7
	Miranda MLR					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	24.67	26.6	38.2	24.8	25	38.2
rMBE(%)	-0.1	-1.2	0	-0.2	-1	0
rMABE(%)	14.8	15.7	21.8	14.9	15	21.8
	Random MLR					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	34.9	34.1	34.1	34.1	34.2	34.9
rMBE(%)	1.5	1.7	1.7	1.6	-0.5	1.5
rMABE(%)	22.6	22	22	22	22.3	22.6

El calcular los valores de RMSE, MBE y MABE para evaluar los procedimientos de Entrenamiento, Validación y Prueba (o Test) resulta de utilidad ya que permite apreciar las performances de los modelos en estas etapas.

Lo primero que se debe observar es que los valores de RMSE entre Entrenamiento y Validación deben ser muy similares: esto es porque los procedimientos utilizan el valor de la validación para detener el proceso de entrenamiento, pero como se realizan en dos sets diferentes de datos, no son exactamente iguales (Figura 2 y 3).

La métrica del periodo de Prueba es quizás la más importante porque la misma ha sido realizada en un periodo fuera del Entrenamiento y Validación. Es decir, probar fuera del periodo de entrenamiento es la prueba definitiva del correcto funcionamiento de un modelo. El modelo que se elige es precisamente el que presenta la mejor performance en los datos de Testeo o Prueba.

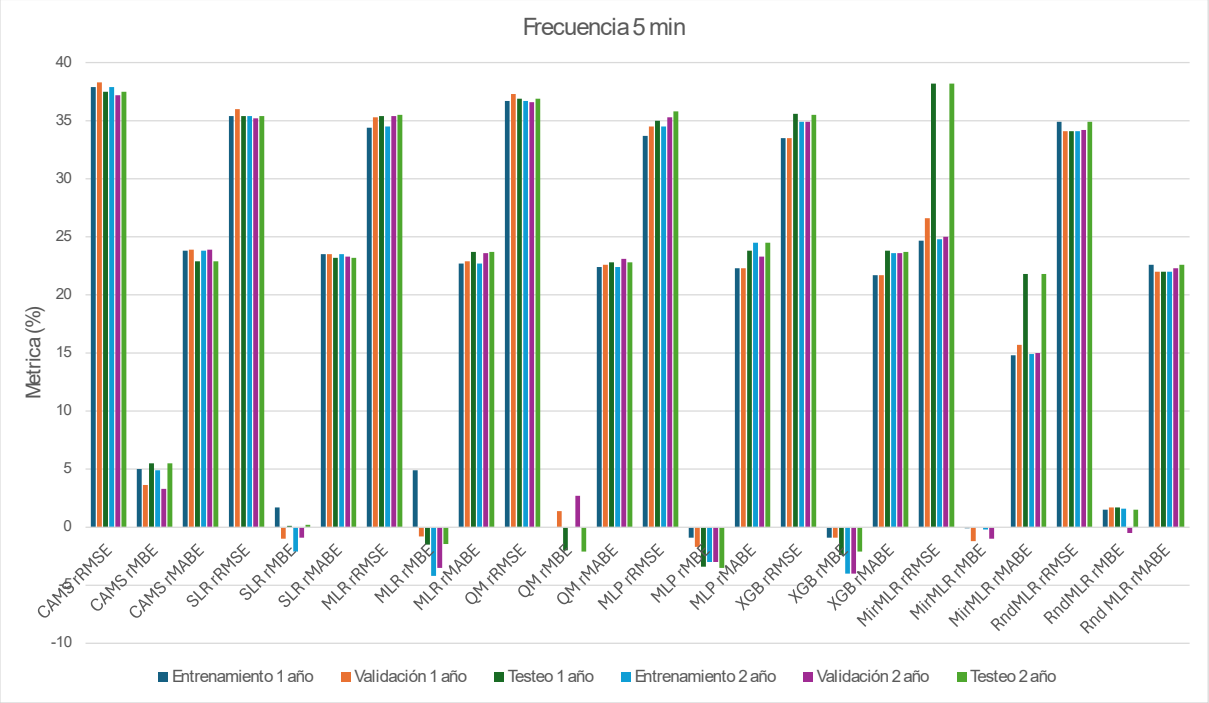

Figura 2. Valores de rRMSE, rMBE y rMABE para la frecuencia de datos de 5 minutos para Salta, de acuerdo con la Tabla 1

Un caso interesante es el que se observa para el modelo Miranda MLR. Para los periodos de Entrenamiento y Validación este modelo da métricas (para RMSE) que son menores a los de los demás modelos: para el caso de 5 minutos el rRMSE de entrenamiento y validación es de 24.67% y 26.6% respectivamente. Estos valores son aproximadamente 10% más bajos que los otros modelos para los mismos procedimientos. Sin embargo, para la Prueba, el rRMSE (38.2%) es superior al de los demás modelos. Esto es un claro indicio de que el modelo entrenado y validado en 2014 (o 2014-2015) no funciona tan bien en 2013. Esto no ocurre para Random MLR, el que mantiene sus métricas similares en todas sus etapas de entrenamiento, validación y prueba.

Las demás técnicas ofrecen resultados similares, siendo el mejor el mostrado por Random MLR. Puede decirse que el peor de los métodos es el de QM. Quizás la mejor cualidad de esta última técnica mencionada es que corrige el bias de los datos entre los valores medidos y los adaptados, lo que se aprecia en su rMBE de entrenamiento. Pero esto solo ocurre en ese procedimiento: en la Validación y la Prueba, esta característica no se aprecia más.

Tabla 2. Métricas de la comparación de los valores medidos vs los adaptados por las distintas técnicas usadas, para una frecuencia de datos de 15 minutos. Entrenamiento y Validación se realizan sobre el mismo periodo (1 año = 2014, 2 años = 2014+2015), solo que el primera se toma de manera aleatoria el 80% de los datos mezclados y el segundo el 20% restante de esos datos. El testeo o Prueba se realiza sobre todos los datos de 2013

	15 min					
	CAMS					
	1 año	1 año	1 año	2 año	2 año	2 año
	Entrenamiento	Validación	Testeo	Entrenamiento	Validación	Testeo
rRMSE(%)	37.49	38.1	35.7	36.2	35.6	35.7
rMBE(%)	5.4	7	5.4	3.8	3.4	5.4
rMABE(%)	23.66	23.8	21.9	23.5	23.3	21.9
	SLR					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	34.6	34.8	35.4	33.9	33.3	35.3
rMBE(%)	-4.1	1.6	-0.4	9.4	-0.5	0.5
rMABE(%)	23.2	23.3	23.4	22.5	22.1	23
	MLR					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	33.4	33.8	33.8	33.5	33.1	33.2
rMBE(%)	1	1.2	-2.6	3.2	-0.5	1.2
rMABE(%)	22.4	22.6	23.1	22.1	21.8	21.6
	MQ					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	35.7	36.3	34.6	34.9	34.4	34.8
rMBE(%)	0	2.9	-2.6	0	-1.4	-5.2
rMABE(%)	21.9	22	21.7	21.7	21.9	22.7
	MLP					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	33.2	33.8	33.7	33.5	33	32.9
rMBE(%)	1	2.2	-1.6	-2.3	-2.7	-1.4
rMABE(%)	21.9	22.3	22.8	22.3	21.9	21.8
	XGBoost					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	30.9	32.4	34.4	32.2	31.5	33.2
rMBE(%)	0	1.39	-4.5	0	-0.5	1
rMABE(%)	20.8	21.9	23.9	21.2	20.6	21.7
	Rodrigues MLR					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	21.5	22.1	36.5	21.6	21.1	36.5
rMBE(%)	0.5	1.6	-0.05	-0.2	0	0
rMABE(%)	13.2	13.4	21.1	13	13	21.1
	Random MLR					
	1 año	1 año	1 año	2 año	2 año	2 año
rRMSE(%)	33.1	33.5	33.1	32.8	32.3	32.9
rMBE(%)	2.1	3.9	2.1	0.1	-0.4	1.5
rMABE(%)	21.4	21.6	21.4	21.5	21.1	21.5

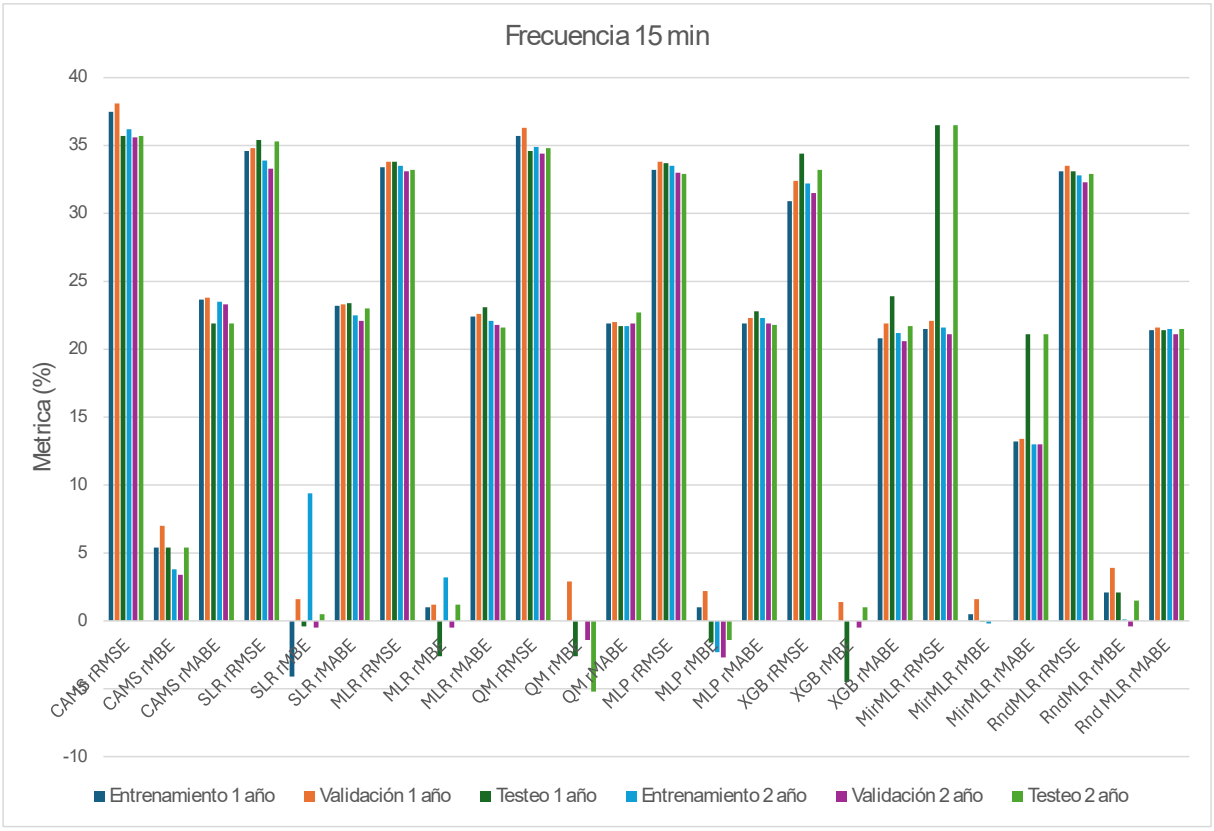

Figura 3. Valores de rRMSE, rMBE y rMABE para la frecuencia de datos de 15 minutos para Salta, de acuerdo con la Tabla 2.

DISCUSIÓN

En este análisis se han detectado varias situaciones para tener en cuenta, las que enumeraremos a continuación:

i) **Overfitting:** el overfitting es un fenómeno que ocurra al entrenar un modelo con un cierto set especifico de datos. Ocurre que el modelo/función resultante se ha especializado en resolver la relación entre los regresores y los valores medidos de ese set de datos, siendo altamente probable que al aplicar ese modelo/función a otro set de datos diferente al set con el que se entrenó, los resultados tengan un error mayor al calculado dentro del periodo de entrenamiento. Además, es común que como salida del entrenamiento se brinden muchos posibles modelos/funciones; y es común que se elija como modelo/función final al que ofrece el menor error. Sin embargo, debe tenerse muy claro que esto no asegura el correcto desempeño de este fuera del periodo de simultaneidad. En este trabajo las configuraciones asociadas a los hiperparámetros de los modelos XGBoost y MLP fueron determinadas a partir de una optimización realizada mediante la técnica de búsqueda en rejilla (grid search), donde se buscaron los modelos que brinden el menor RMSE en el conjunto de pruebas, esto fue realizado mediante la librería scikit-learn de Python.

Puede notarse que todos los modelos mostrados no indican overfitting, salvo por MirandaMLR, que en su Entrenamiento y Validación muestra métricas muy inferiores a los demás modelos. Sin embargo, en el Prueba las métricas se disparan superando el desempeño de CAMS. Eso no sucede en Random MLR, indicando que sus estimaciones fuera de la región de simultaneidad son mejores, es decir, con un error menor y con mayor estabilidad.

ii) **Calidad de la BDS y de los valores medidos**: durante el proceso de aprendizaje el sistema busca correlacionar valores simultáneos de la base de datos satelital con los datos medios. Si la base de datos satelital tiene período del día o rango de ángulo cenital que presentan

errores y no correlacionan con los valores medidos (más aún si estos errores son aleatorios) el proceso de aprendizaje generará un sistema que se "equivocará" mucho. Lo mismo aplica para los valores medidos lo que deben ser de calidad y no tener errores. Este es un problema presente para CAMS-Rad para ángulos cenitales altos (> 60°) cerca del atardecer: este problema puede deberse al ángulo del satélite respecto de la posición de la ciudad de Salta. Sin embargo, este problema no es sistemático.

CONCLUSIONES

En este trabajo se analizan los resultados de realizar un proceso de Adaptación al Sitio usando datos medidos de Irradiancia Solar Global en la ciudad de Salta (Argentina) contra valores estimados por la Bases de Datos Satelital CAMS-Rad. Se han utilizado diferentes técnicas de Aprendizaje Automático como Regresión Lineal Simple y Múltiple, Quantile Mapping, Multilayer Perceptron, XGBoost y regresión Clusterwise. Se usaron dos integraciones temporales de 5 y 15 minutos para los datos de Entrenamiento, Validación y Prueba. Los valores de entrenamiento y Validación abarcaron un año (2014) y dos años (2014-2015), mientras que los datos de Prueba se realizaron usando un año (2013). Los resultados muestran que todas las técnicas mejoran los estimados por CAMS-Rad. Solo el modelo Miranda MLR muestra efectos de overfitting en su etapa de Prueba. El modelo Random MLR es el que ha mostrado los mejores resultados en la etapa de Prueba, indicando una mejora de entre 2.8% y 3.4% respecto de los estimados por CAMS. Los demás modelos analizados tienen mejoras de entre 2.5% a 0.6%, sin manifestar overfitting. Esto es importante para que los modelos entrenados puedan ser usados fuera de la región de simultaneidad. Solo QM ha mostrado realizar una corrección de bias de los datos analizados, con un rMBE igual a 0. Estos resultados indican que es posible realizar una Adaptación al Sitio de manera eficiente para los datos de irradiancia solar en la ciudad de Salta. Estos procedimientos serán replicados en otros sitios de la provincia de Salta, en la búsqueda de la confección de un Mapa de Radiación Solar con altísimo detalle en sus valores reales de ocurrencia.

FUENTES DE FINANCIAMIENTO

Este trabajo fue financiado con el Proyecto "A" Nº2751/0 del Consejo de Investigaciones de la Universidad Nacional de Salta (Argentina).

REFERENCIAS

Vignola F, Grover C, Lemon N, McMahan A. Building a bankable solar radiation dataset. Sol Energy 2012; 86:2218–29. https://doi.org/10.1016/j. solener.2012.05.013.

Huld T, Müller R, Gambardella A. A new solar radiation database for estimating PV performance in Europe and Africa. Sol Energy 2012;86(6):1803–15. https://doi. org/10.1016/j.solener.2012.03.006

Laspiur MR, Salazar GA, Zerpa J, Watkins M. TRAZADO DE MAPAS MEDIOS ANUALES DE ENERGÍA SOLAR GLOBAL, DIRECTA, DIFUSA Y TILT, USANDO LA BASE DE DATOS DE SWERA. CASO DE ESTUDIO: PROVINCIAS DE SALTA Y JUJUY. Acta de la XXXVI Reunión de Trabajo de la Asociación Argentina de Energías Renovables y Medio Ambiente Vol. 1, pp. 08.157-08.162, 2013. ISBN 978-987-29873-0-5

Sengupta M, Xie Y, Lopez A, Habte A, Maclaurin G, Shelby J. The national solar radiation data base (NSRDB). Renew Sustain Energy Rev 2018; 89:51–60. https:// doi.org/10.1016/j.rser.2018.03.003.

Viana TS, Rüther R, Martins FR, Pereira EB. Assessing the potential of concentrating solar photovoltaic generation in Brazil with satellite-derived direct normal irradiation. Sol Energy 2011; 85:486–95. https://doi.org/10.1016/j. solener.2010.12.015.

Polo J, Wilbert S, Ruiz-Arias JA, Meyer R, Gueymard C, Súri M, et al. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. Sol Energy 2016; 132:25–37. https://doi.org/10.1016/j. solener.2016.03.001

Polo, J.; Fernández-Peruncheña, C.; Salamalikis, V.; Mazorra-Aguiar, L.; Turpin, M.; Martín-Pomares, L.; Kazantzidis A.; Blanc, P.; Remund, J.; Benchmarking on improvement and site-adaptation techniques for modeled solar radiation datasets, Solar Energy, Volume 201, 2020, Pages 469-479, ISSN 0038-092X.

Chu Y, Wang Y, Yang D, Chen S, Li M. A review of distributed solar forecasting with remote sensing and deep learning. Renewable and Sustainable Energy Reviews 198,2024, 114391, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2024.114391

https://www.soda-pro.com/web-services/radiation/cams-radiation-service

https://www.soda-pro.com/help/cams-services/introduction

Peel MC, Finlayson BLMT. Updated world map of the Koppen-Geiger € climate classification. Hydrol Earth Syst Sci 2007:1633–44

Nollas FM, Salazar GA, Gueymard CA, Quality control procedure for 1-minute pyranometric measurements of global and shadowband-based diffuse solar irradiance, Renewable Energy (2022), doi: https://doi.org/10.1016/j.renene.2022.11.056.

Mitchell TM. Machine Learning (1997) ISBN: 0070428077

Faouzi J, Colliot O. Classic machine learning algorithms. Machine Learning for Brain Disorders, Springer, In press. hal-03830094v1

Narvalez, G.; Giraldo, L. F.; Bressan, M. Pantoja, A.; Machine learning for site-adaptation and solar radiation forecasting, Renewable Energy, Volume 167, 2021, Pages 333-342, ISSN ISSN 0960-1481, https://doi.org/10.1016/j.renene.2020.11.089.

Dhata, E.F.; Kim, C.K.; Kim, H.-G.; Kim, B.; Oh, M. Site-Adaptation for Correcting Satellite-Derived Solar Irradiance: Performance Comparison between Various Regressive and Distribution Mapping Techniques for Application in Daejeon, South Korea. Energies 2022, 15, 9010. https://doi.org/10.3390/en15239010

Miranda DR, Vilela OC, Salazar GA, Alonso-Suárez R, Costa ACA, Costa RSS, Ing Ren T. A Novel Methodology for Site Adaptation of Solar Radiation Using Supervised and Non-Supervised Procedures. 2024. Pre-print:https://ssrn.com/abstract=4804417, http://dx.doi.org/10.2139/ssrn.4804417

Ledesma, Ruben and Salazar, Germán Ariel and Vilela, Olga de Castro and Lopez Ruíz, Constanza, Evaluation of Data Splitting Strategies for Improving Satellite-Derived Solar Irradiance Time Series. 2024. Available at SSRN: https://ssrn.com/abstract=4944326 or http://dx.doi.org/10.2139/ssrn.4944326

Huang, Liexing and Kang, Junfeng and Wan, Mengxue and Fang, Lei and Zhang, Chunyan and Zeng, Zhaoliang, Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events, Frontiers in Earth Science9, 2021, DOI 10.3389/feart.2021.596860 ISSN 2296-6463.

Xianglong Li, Longfei Ma, Ping Chen, Hui Xu, Qijing Xing, Jiahui Yan, Siyue Lu, Haohao Fan, Lei Yang, Yongqiang Cheng, Probabilistic solar irradiance forecasting based on XGBoost, Energy Reports,

Volume 8, Supplement 5, 2022, Pages 1087-1095, ISSN 2352-4847, https://doi.org/10.1016/j.egyr.2022.02.251.

Chibuzor N Obiora, Ahmed Ali, Ali N Hasan. Implementing Extreme Gradient Boosting (XGBoost) Algorithm in Predicting Solar Irradiance. 2021 IEEE PES/IAS PowerAfrica. IEEE

Chunxiao Zhang, Yingbo Zhang, Jihong Pu, Zhengguang Liu, Zhanwei Wang, Lin Wang, An hourly solar radiation prediction model using eXtreme gradient boosting algorithm with the effect of fog-haze, Energy and Built Environment, Volume 6, Issue 1, 2025, Pages 18-26, ISSN 2666-1233

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Panofsky, H.A. and Brier, G.W. (1968) Some Applications of Statistics to Meteorology. Earth and Mineral Sciences Continuing Education, College of Earth and Mineral Sciences.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC London

PERFORMANCE ANALYSIS OF DIFFERENT MACHINE LEARNING TECHNIQUES IN A GLOBAL SOLAR IRRADIANCE SITE ADAPTATION FOR SALTA (ARGENTINA)

ABSTRACT: In this work, different metrics are analyzed comparing measured values of Global Solar Irradiance against estimated values using different Artificial Intelligence techniques, particularly Machine Learning, to perform a Site Adaptation through the CAMS-Rad Satellite Database for the city of Salta (1200 meters above sea level). The data were analyzed with temporal integrations of 5 minutes and 15 minutes. Two periods of measured data were used: one year (2014) or two years (2014-2015) for training and validation. Machine Learning techniques were Simple and Multiple Linear Regression, Quantile Mapping, Multilayer Perceptron, XGBoost and Clusterwise regression. All metrics indicate an improvement over CAMS-Rad estimates, with Clusterwise regressions standing out.

Keywords: global solar irradiance, machine learning, site adaptation, Salta.