UNIVERSIDAD NACIONAL DE SALTA

FACULTAD DE CIENCIAS

MEJORA DE LA REPRESENTATIVIDAD ESPACIAL DE ESTIMACIONES DE RADIACIÓN SOLAR DE BASES DE DATOS SATELITALES, PARA SALTA Y JUJUY, USANDO HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL.

Tesis para optar al grado de Doctor en Ciencias

RUBÉN D. LEDESMA

Director: Dr. Germán A. Salazar.

Salta, Argentina.
—-, 2025

Esta tesis está dedicada a...

Agradecimientos

En este trabajo agradezco a...

Contenido

Índice de figuras

Índice de tablas

Mejora de la representatividad espacial de estimaciones de radiación solar de Bases de Datos Satelitales, para Salta y Jujuy, usando herramientas de Inteligencia Artificial.

Rubén D. Ledesma

Resumen

Escribir aquí el resumen del trabajo de grado.

Capítulo 1

Introducción

La radiación solar, también conocida como energía solar, es la vasta cantidad de energía emitida por el Sol (Wald, 2007, 2021). El Sol, con una temperatura superficial de aproximadamente 5780 K (alrededor de 5500 °C), irradia energía a través de un amplio espectro de longitudes de onda, principalmente entre 200 nm y 4000 nm (Wald, 2021). Esta energía es el resultado de procesos termonucleares (Wald, 2021).

La cantidad promedio de radiación solar recibida justo fuera de la atmósfera terrestre, por unidad de área, es de aproximadamente 1361.7 W/m^2, conocida como la constante solar. Sin embargo, la radiación solar real que llega fuera de la atmósfera (radiación extraterrestre) varía anualmente entre 1412 W/m^2 (principios de julio) y 1322 W/m^2 (cambio de año), una variación del 3.3 % debido a la órbita elíptica de la Tierra alrededor del Sol.

Para describir la radiación, se utilizan dos cantidades principales: Irradiancia: Es la potencia recibida por unidad de área, con unidades en vatios por metro cuadrado (W/m^2) Irradiación: Es la energía recibida por unidad de área, con unidades en julios por metro cuadrado (J/m^2). En aplicaciones de energía solar, el vatio-hora por metro cuadrado (Wh/m^2) es una unidad de uso común para la irradiación, aunque no forma parte del Sistema Internacional (SI). La conversión es 1 $Wh/m^2 = 3600\ J/m^2$.

La radiación solar es fundamental para la vida en la Tierra y para numerosas actividades humanas. Su impacto abarca diversos ámbitos, desde el clima hasta la energía.

En primer lugar, regula el equilibrio energético del planeta: la energía absorbida se transforma en calor y se distribuye a través de la atmósfera y los océanos, lo que origina procesos meteorológicos como la convección, la evaporación, la formación de nubes, los vientos y las precipitaciones.

En el diseño arquitectónico, la radiación solar se aprovecha en estrategias pasivas y en la iluminación natural. Gracias a ello, es posible calcular de manera adecuada el tamaño de ventanas y acristalamientos para optimizar la entrada de luz, mejorar el confort térmico y equilibrar las ganancias y pérdidas de calor.

En la agricultura, resulta determinante para el crecimiento y desarrollo de los cultivos, influyendo en procesos como la maduración de la uva en los viñedos o la gestión del riego en invernaderos.

En el medio ambiente, interviene en reacciones fotoquímicas, como la fotólisis de contaminantes atmosféricos, lo que puede generar sustancias secundarias dañinas para la salud y la vegetación. Asimismo, la radiación ultravioleta acelera la degradación de materiales poliméricos expuestos al sol.

En cuanto a la producción de energía, constituye la base de tecnologías renovables como los paneles fotovoltaicos y los sistemas solares térmicos. Por ello, una estimación precisa y el pronóstico de la radiación son esenciales para diseñar y ubicar adecuadamente las plantas de energía solar.

Finalmente, en la salud humana, la radiación solar influye en el estado de ánimo y en los ritmos biológicos. Los rayos ultravioleta favorecen la síntesis de vitamina D, indispensable para la fijación del calcio en los huesos. No obstante, una exposición excesiva puede provocar efectos negativos como quemaduras, envejecimiento prematuro de la piel e incluso cáncer.

La radiación solar que llega a la superficie terrestre está compuesta por tres contribuciones:

- **Radiación directa**, que proviene del disco solar sin desviaciones.
- **Radiación difusa**, resultado de la dispersión por moléculas, aerosoles y nubes, que alcanza la superficie desde todas las direcciones del cielo.
- **Radiación reflejada**, correspondiente a la fracción devuelta por el suelo o elementos circundantes.

La suma de estas tres componentes constituye la **radiación global**. Para su medición se emplean principalmente dos instrumentos: los *piranómetros*, que registran la radiación global (directa + difusa), y los *pirheliómetros*, que miden la radiación directa en un ángulo sólido muy reducido orientado hacia el sol. Además, los *piranómetros con anillo de sombra* o los *irradiómetros de banda rotatoria* permiten aislar el componente difuso (Duffie and Beckman, 2013).

La precisión en estas mediciones es crucial para la **ingeniería solar** y para aplicaciones energéticas. Conocer la disponibilidad de radiación solar en un lugar permite dimensionar adecuadamente sistemas fotovoltaicos y térmicos, optimizar el diseño de colectores y predecir la producción a lo largo del tiempo. Asimismo, estas mediciones son esenciales en climatología y meteorología para modelar balances de energía en la atmósfera y en la superficie terrestre (Wald, 2018; , IEA-PVPS).

No obstante, la instrumentación terrestre presenta limitaciones: las estaciones de medición suelen ser escasas y su cobertura en espacio y tiempo es reducida, lo que dificulta la construcción de series largas y continuas de datos. Estas restricciones explican el creciente interés en el uso de modelos numéricos y productos satelitales como complemento a las observaciones directas.

En síntesis, medir la radiación solar es fundamental tanto para el desarrollo de tecnologías renovables como para comprender los procesos energéticos que gobiernan el sistema climático.

Ante estas limitaciones, se han desarrollado métodos de estimación alternativos. Los modelos meteorológicos numéricos y de reanálisis permiten simular los procesos radiativos de la atmósfera para estimar la radiación solar. De manera complementaria, las imágenes satelitales de observación de la Tierra, obtenidas mediante sensores multiespectrales, se procesan con algoritmos especializados para calcular la radiación a nivel del suelo, como ocurre con servicios tales como HelioClim. También se emplean técnicas de interpolación espacial que estiman valores en un sitio de interés a partir de mediciones de estaciones cercanas, considerando la variabilidad regional de la radiación. Asimismo, se utilizan modelos empíricos basados en correlaciones estadísticas, que permiten derivar los distintos componentes de la radiación a partir de la global medida o bien estimarla en función de variables meteorológicas como la insolación, la temperatura del aire o la nubosidad. La calidad de estas mediciones y estimaciones es un aspecto crítico, por lo que se aplican procedimientos rigurosos de control y validación de datos, incluyendo inspecciones visuales y métodos automáticos de detección de valores atípicos, con el fin de garantizar la confiabilidad de la información disponible.

1.1. Estudio de la Radiación Solar en Argentina y en el Noroeste Argentino

Las primeras estaciones de medición de la Red Solarimétrica en Argentina datan desde el año 1978, a partir de un proyecto financiado inicialmente por la Organización de los Estados Americanos (O.E.A.) Grossi Gallegos et al. (1999).

A fines de 1997 se publicaron los resultados de las primeras evaluaciones a nivel de superficie de radiación solar global en la República Argentina, donde se proceso toda la información disponible en el país complementadas con datos registrados en países vecinos Grossi Gallegos (1998a,b). En estos trabajos se reporta que las cartas elaboradas responden adecuadamente a los datos disponibles en Argentina, dentro de las condiciones que se impusieron en la metodología, siendo compatibles con el mejor nivel del estado del conocimiento del recurso en esta parte del continente. Desde el punto de vista nacional, deberán transcurrir no menos de cinco años para que puedan registrarse modificaciones de importancia a las isolíneas presentadas, las que no superan la incerteza del 10 %.

En Mayo del 2007 se publica el ATLAS DE ENERGÍA SOLAR DE LA REPÚBLICA ARGENTINA, mismo que se declara de interés cultural y educativo por la Dirección General de Cultura y Educación de la Provincia de Buenos Aires Apoyado por la ASADES (Asociación Argentina de Energías Renovables y Ambiente) Grossi Gallegos and Righini (2007). El objetivo planteado para el presente trabajo fue actualizar la evaluación a nivel de superficie del campo de la radiación solar global en Argentina, procesando para ello toda la información disponible en el país hasta el año 1997, proveniente ya sea de mediciones directas del parámetro (28 estaciones

piranométricas) o de estimaciones obtenidas a partir de información meteorológica terrestre (24 estaciones heliográficas) o satelital, complementada con la de los países vecinos, evaluándose la precisión y validez de los resultados obtenidos.

El estudio de la radiación solar en Argentina, particularmente en la región del Noroeste Argentino (NOA), ha evolucionado significativamente a lo largo de las décadas, impulsado por la necesidad de caracterizar este recurso renovable para diversas aplicaciones energéticas y ambientales. La escasez de mediciones terrestres sistemáticas y de alta calidad en muchas áreas ha llevado al desarrollo y la evaluación de modelos empíricos, físicos y satelitales.

A principios de la década de 2000, se empezaron a desarrollar herramientas computacionales clave para facilitar el cálculo y la estimación de la radiación solar. En 2003, Alejandro L. Hernández presentó GEOSOL, un programa para Windows diseñado para calcular coordenadas solares y estimar la irradiación solar horaria. GEOSOL ofrecía funcionalidades gráficas útiles para visualizar la trayectoria solar en 2D y 3D, y permitía el análisis de obstáculos, lo cual es vital para el diseño de instalaciones solares. El programa incorporaba tres métodos de estimación de irradiación: el de Page y el de Hottel para días claros, y el de Liu-Jordan para días medios mensuales, todos validados con mediciones realizadas en la Universidad Nacional de Salta (UNSa) Hernández (2003).

Estos desarrollos iniciales sentaron las bases para estudios más amplios, como la creación de mapas de radiación solar. En Belmonte et al. (2006) utilizaron GEOSOL y Sistemas de Información Geográfica (SIG) para desarrollar mapas de radiación solar en el Valle de Lerma, Salta. Su metodología combinaba cálculos de radiación con tratamiento estadístico y procesamiento SIG, destacando una alta correlación lineal entre la radiación solar total y la altitud, así como entre los valores de diferentes meses. Esto sugirió que las ecuaciones de regresión eran un método eficaz para el mapeo, especialmente en zonas montañosas. El trabajo también reconoció la ausencia de series de mediciones históricas en el área, lo que justificaba la necesidad de generar modelos digitales para definir este parámetro climático esencial.

Continuando con la caracterización del recurso, en 2007, Germán A. Salazar, Luis A. Hernández, Luis R. Saravia y Graciela G. Romero realizaron un estudio para determinar los coeficientes de la relación de Ångström-Prescott para la ciudad de Salta, utilizando datos recopilados entre abril de 2006 y abril de 2007 . Esta relación empírica vincula la irradiación global con las horas de insolación (heliofanía). Observaron que los coeficientes 'a' y 'b' se ven afectados por factores como la latitud, la altura y el vapor de agua en la atmósfera. Al comparar sus resultados con otros métodos de correlación, encontraron que el método de Rietveld mostró una mejor correlación con los datos medidos que el de Glover y McCulloch. Este estudio resaltó la importancia de la relación Ångström-Prescott para estimar la radiación global en sitios con datos limitados de heliofanía Salazar et al. (2007).

En 2008, Salazar, Hernández, Cadena, Saravia y Romero avanzaron en la caracterización

de la radiación solar global para día claro en sitios de altura en el NOA, analizando datos de irradiancia en Salar El Rincón (3730 m), Huacalera (2680 m) y Salta Capital (1190 m) Salazar et al. (2008a). Propusieron tres modelos basados en una ecuación tipo ASHRAE, que estiman la irradiancia instantánea (G) utilizando solo la altura sobre el nivel del mar (A) y la masa de aire (ma) como variables. Los modelos mostraron una muy buena correlación con los datos medidos, con errores porcentuales RMSE promedio inferiores al 3 %. Se destacó que el índice de claridad representativo (Kt-R) se incrementa con la altura, lo cual es consistente con la menor atenuación atmosférica a mayor altitud. Ese mismo año, Salazar et al. continuaron desarrollando un modelo para estimar irradiancia e irradiación solar para día claro, ahora incorporando datos de Buenos Aires para evaluar el comportamiento en bajas altitudes. La versión "Modelo A"demostró una mejor aproximación a los resultados de GEOSOL en un rango más amplio de alturas.

Paralelamente, el estudio de la radiación difusa también se benefició de nuevas metodologías. En Salazar et al. (2008b) Saravia exploraron el uso del método geoestadístico Kriging para estimar los valores de irradiación difusa en la bóveda celeste. Compararon la radiación difusa registrada en Salta (1200 m) y El Rosal (3350 m), confirmando que Salta presentaba valores más altos debido a una mayor dispersión atmosférica a menor altitud. La metodología implicó la proyección de la superficie en un plano (gnomónica o cilíndrica) y el uso del software SURFER 7.0 para el procesamiento de datos y la visualización de mapas de contorno.

La década de 2010 marcó una profundización en el desarrollo de modelos y una evaluación crítica de las fuentes de datos. En Salazar et al. (2010) presentaron modelos prácticos (Modelos 3 y 4) para estimar la irradiancia horizontal en condiciones de cielo claro, especialmente útiles para sitios de altura en el NOA. Estos modelos utilizaban la altitud para generar un índice de claridad representativo (kt-R-p) y mostraron errores no superiores al 5 % para masas de aire corregidas por presión (AMc$<$2). Se destacó su potencial para estudios de factibilidad en la instalación de plantas de energía solar térmica. Los autores también señalaron la necesidad de una nueva convención para clasificar los días de cielo claro, ya que el criterio previamente utilizado (Kt $>$0.7) clasificaba incorrectamente días parcialmente nublados como claros.

En el mismo año 2010, en Salazar (2010) se aplicó el modelo híbrido de Yang a datos climáticos medios mensuales de diez localidades de Argentina. Este modelo, validado previamente en Japón, busca estimar las componentes directa y difusa de la irradiación global. El estudio encontró una muy buena correlación con los valores medidos de irradiación global horizontal, con un RMSE % de aproximadamente 6 % después de realizar correcciones por exceso en los datos de heliofanía. Aunque el modelo de Yang fue calificado como “altamente confiable” por Gueymard (2003), se señaló la necesidad de continuar investigando para estimar variables relacionadas con las transmitancias de los componentes atmosféricos para el contexto argentino.

La disponibilidad de bases de datos satelitales también fue un foco de análisis. En Laspiur et al. (2013) trazaron mapas medios anuales de energía solar (global, directa, difusa y Tilt) para

las provincias de Salta y Jujuy, utilizando la base de datos satelital SWERA y el método geoestadístico Kriging. Este trabajo buscaba proporcionar una herramienta inicial para el estudio de la distribución del recurso solar en el Norte de Argentina.

Una evaluación más profunda de estas bases de datos fue realizada por Salazar et al. (2013), comparando los datos de irradiación solar global media mensual medidos en Salta Capital (periodos 1968-2007) con las estimaciones de las bases de datos SWERA, SoDa y SSE. El estudio concluyó que los datos de SWERA mostraban la mejor correlación con los valores medidos (RMSE % promedio del 14 %), mientras que SoDa y SSE presentaban errores superiores al 24 %, atribuidos al mayor tamaño de sus celdas satelitales. Este análisis evidenció la importancia de verificar la validez de los datos satelitales para cada región específica.

En 2014, Germán Salazar y Carlos Raichijk llevaron a cabo una evaluación de las condiciones de cielo claro en sitios de altura, desafiando la aplicabilidad del criterio de Iqbal para clasificar la nubosidad en estas ubicaciones Salazar and Raichijk (2014). Encontraron que el criterio de Iqbal a menudo clasificaba incorrectamente días parcialmente nublados como días de cielo claro en sitios a gran altitud. Para ello, utilizaron los índices de claridad (Kt) y de cielo claro (Kc), revelando que el valor más probable de Kc para un día de cielo claro dependía del modelo utilizado para estimar la radiación de cielo claro (ESRA y ARG-P).

El interés se extendió también a la radiación ultravioleta (UV). En Suárez et al. (2021) realizaron un estudio sobre la variabilidad diaria y anual de la radiación solar eritémica (UVER) en Salta, San Carlos y El Rosal (periodo 2012-2013). Los resultados indicaron elevados niveles de riesgo solar en las tres localidades, con un incremento de la UVER con la altura. Por ejemplo, en verano, se midieron valores máximos de Índice UV (IUV) de 17 en El Rosal y 15 en Salta, mientras que en invierno los promedios eran de 6 y 5 respectivamente. El estudio proporcionó una caracterización detallada de la distribución anual y diaria de la UVER, crucial para la salud pública.

En el trabajo Vilela et al. (2015) se exploró la caracterización de la radiación directa, difusa y global en localidades de Brasil (Recife, Botucatu) y Argentina (Salta, Luján). Desarrollaron un software para aplicar filtros físicos y estadísticos a los datos de radiación, calculando correlaciones entre las fracciones difusa (kd) y directa (kn) en función del índice de claridad (kt). A pesar de las diferencias climáticas entre las localidades, las relaciones kd vs. kt y kn vs. kt se mostraron consistentes, lo que sugiere su aplicabilidad generalizada.

En la recta final de la década de 2010, el enfoque en el control de calidad de los datos medidos se hizo más prominente. En Romano Armada et al. (2017) aplicaron protocolos de control de calidad de la red BSRN a datos de radiación solar global medidos en Salta entre 2013 y 2015. Sus hallazgos sugirieron que los datos de referencia históricos para Salta subestimaban el recurso, subrayando la importancia de filtrar los datos para asegurar su validez estadística y

representatividad.

Las aplicaciones de la energía solar también continuaron evolucionando. Hongn et al. (2018) simularon el funcionamiento de una planta solar térmica de gran escala (30 MWe) en San Carlos, Salta, utilizando el modelo analítico FAE. Compararon los resultados con el software de referencia SAM (NREL) y con la planta real SEGS VI en el desierto de Mojave. Descubrieron que, si bien la producción anual en San Carlos sería ligeramente menor que en Mojave, la generación eléctrica resultaría más uniforme a lo largo del año, lo que la convierte en una opción viable para la inyección de energía a la red.

En Sarmiento Barbieri et al. (2019) desarrollaron una herramienta SIG para la provincia de Salta. Validaron los datos satelitales LSA-SAF con modelos empíricos y mediciones de cinco estaciones terrestres regionales (Abra Pampa, El Pongo, La Viña, El Rincón, Cafayate) durante un período de siete años. Este trabajo subrayó la escasez de datos de medición en tierra en el norte de Argentina, y señaló que la información derivada de imágenes satelitales puede contribuir a llenar las brechas existentes.

En Ledesma et al. (2023) se repotaron avances en la estimación de irradiancia solar en Salta y Jujuy mediante imágenes satelitales GOES-16. Ante la persistente falta de estaciones radiométricas con mediciones sistemáticas y prolongadas en Argentina, la estimación por satélite se presenta como una alternativa eficiente para cubrir grandes áreas geográficas. El estudio evaluó dos modelos simples basados en un índice de nubosidad (SUNY y Cano) aplicados a imágenes del canal visible de GOES-16, comparando las estimaciones con valores medidos de irradiancia global horizontal (GHI) en tres sitios. Los modelos locales (Cano et al. y SUNY) mostraron una clara ganancia en desempeño en comparación con el modelo Heliosat-4, confirmando el potencial de las imágenes GOES para esta región.

Uno de los estudios más recientes sobre la radiación solar evalúa el rendimiento de diversos modelos satelitales y de re-análisis (CAMS Heliosat-4, NREL NSRDB, GOES DSR, LSA-SAF MDSSFTD, GOES G-CIM, MERRA-2 y ERA-5) para la estimación de GHI en el Noroeste Argentino Ledesma et al. (2025). Este análisis compara estos modelos con mediciones terrestres de alta calidad (2020-2023) en La Quiaca y Salta. Los resultados preliminares indican que los modelos G-CIM (desarrollado localmente y ajustado con imágenes GOES-16) y NSRDB ofrecen las estimaciones más precisas, incluso en entornos complejos con altitudes extremas o reflectividad variable del terreno. Este estudio reitera la falta de redes operativas de estaciones radiométricas con sensores trazables y control de calidad adecuado en el Noroeste Argentino, enfatizando la importancia de usar datos satelitales validados localmente y la necesidad de ajustes específicos de sitio para los modelos globales de irradiancia.

En el apartado 1.3 se explican en detalle modelos de estimación de radiación solar.

1.2. Redes de Medidas

Una de las redes de observación más reconocidas en la comunidad especializada en energía solar es la **Baseline Surface Radiation Network (BSRN)**. Se trata de una red internacional integrada por estaciones distribuidas estratégicamente en distintos entornos climáticos, cuyo objetivo es obtener mediciones de alta calidad de los flujos radiativos de onda corta y larga en la superficie terrestre, con elevada frecuencia de muestreo. Su propósito central es monitorear los componentes radiativos de fondo menos influidos por la actividad humana, validar y evaluar las estimaciones satelitales de radiación superficial, así como proveer datos de referencia para modelos climáticos y estudios de climatología regional. Gracias a la consistencia y precisión de sus registros, la BSRN contribuye a una mejor comprensión de los procesos climáticos, a la evaluación de modelos de circulación y al apoyo de programas científicos internacionales como el *World Climate Research Programme (WCRP)* y el *Global Energy and Water Cycle Experiment (GEWEX)*.

En la actualidad, la red cuenta con 51 estaciones en operación, 16 cerradas (de manera temporal o definitiva), 9 clasificadas como inactivas y varias con estatus de candidatas. Las estaciones realizan distintos tipos de mediciones radiativas: algunas se limitan a las *mediciones básicas*, mientras que otras incluyen *mediciones adicionales*, observaciones sinópticas, sondeos atmosféricos y registros de ozono. Asimismo, se proyecta la incorporación de nuevas estaciones, algunas de las cuales deberían entrar en funcionamiento durante el año en curso. Los datos generados se almacenan en PANGAEA, un repositorio de acceso abierto especializado en datos georreferenciados del sistema terrestre, donde se encuentran disponibles los conjuntos de mediciones y sus metadatos, junto con la información de los investigadores responsables, siguiendo los lineamientos establecidos para su liberación.

En Latinoamérica, la presencia de estaciones BSRN es escasa y, en particular, Argentina no cuenta actualmente con ninguna. Los proyectos nacionales destinados al sostenimiento de redes radiométricas han sido limitados. Entre ellos, uno de los más destacados fue el proyecto ENARSOL, iniciado en 2012, con el objetivo de coordinar esfuerzos entre el INTA, YPF y el grupo GERSolar de la UNLu. La iniciativa contemplaba la instalación de 40 estaciones distribuidas estratégicamente para la medición del recurso solar Aristegui and Righini (2012); sin embargo, el proyecto fue discontinuado y los registros obtenidos no se distribuyen a través de fuentes oficiales.

El **Servicio Meteorológico Nacional (SMN)** es actualmente una de las pocas instituciones argentinas que mantiene estaciones radiométricas activas. Sus mediciones pueden solicitarse mediante contacto oficial (cim@smn.gob.ar). El SMN opera 10 estaciones distribuidas en diferentes regiones del país.

Otro aporte relevante es el proyecto **SAVER-Net**, que monitorea en casi tiempo real aerosoles, ozono y radiación UV, difundiendo la información desde el CEILAP y la Universidad de Magallanes a las instituciones competentes.

En la misma línea, el programa **SATREPS** constituye una colaboración científica entre Japón y países en desarrollo para enfrentar desafíos globales como el cambio climático, la energía,

las enfermedades y los desastres naturales. Desde 2013, Argentina, Chile y Japón llevan adelante un proyecto trinacional que permite monitorear aerosoles, radiación UV y el agujero de ozono en la región, con 10 estaciones radiométricas operativas en territorio argentino.

Por su parte, el **Grupo de Estudios de la Radiación Solar (GERSolar)** fue creado el 2 de septiembre de 2002 en la División Física del Departamento de Ciencias Básicas de la Universidad Nacional de Luján (UNLu). Su propósito ha sido conformar una pequeña red de estaciones para la medición de radiación solar global en la región de la Pampa Húmeda Argentina, zona de mayor relevancia agrícola del país. Las estaciones se han instalado en distintas instituciones y, gracias a la colaboración de sus operadores, generan integrales horarias y diarias de radiación solar global en superficie. Actualmente, la red mantiene 9 estaciones activas en la región pampeana.

Asimismo, el **Grupo de Evaluación y Estudio del Recurso Solar (GEERS)**, del Instituto de Investigaciones en Energía No Convencional (INENCO), mantiene una pequeña red de medidas ubicada en el Noroeste Argentino. Dicha red ha posibilitado el desarrollo de trabajos destinados a evaluar el recurso solar en sitios de altura.

1.3. Modelos de Estimación

Como se ha comentado anteriormente son escasas la redes de medición de irradiancia solar, lo que limita el desarrollo de proyectos que necesiten cuantificar la disponibilidad de este recurso. Añadido a que en las estaciones donde se llevan registros de medidas de GHI no comprenden periodos extensos, lo que presenta una limitante tanto para la bancabilidad de proyectos de energía solar como para análisis ambientales, o cualquier aplicación que requiera un conocimiento preciso de este recurso.

Como un complemento a las medidas en tierra pueden encontrarse modelos de estimación de GHI. Estos modelos pueden ubicarse en dos grandes grupos. Por un lado los modelos de re-análisis y por otro lado los modelos satelitales.

Modelos de reanálisis

Los modelos de reanálisis son sistemas que integran observaciones meteorológicas históricas con modelos físicos de predicción numérica del tiempo, con el objetivo de generar una descripción continua, coherente y físicamente consistente de la atmósfera y el clima a lo largo del tiempo Thejll and Gleisner (2015). Aunque las observaciones provienen de múltiples fuentes —como estaciones meteorológicas, radiosondas, aeronaves y satélites—, su cobertura espacial y temporal es incompleta y su calidad varía, por lo que una simple interpolación matemática no resulta suficiente.

El reanálisis supera esta limitación mediante el uso de modelos físicos que asimilan las observaciones, simulan la evolución atmosférica y ajustan iterativamente las condiciones iniciales para

reducir al mínimo las discrepancias con los datos reales. Esto permite estimar con coherencia física las condiciones en lugares y momentos donde no existen mediciones directas. Estos modelos trabajan con datos históricos —algunos con más de un siglo de antigüedad— y se actualizan conforme se digitalizan nuevos registros o se perfeccionan las representaciones de los procesos físicos. Debido a su alta demanda computacional y elevado coste, solo un número limitado de proyectos de reanálisis se encuentran actualmente en operación.

En el caso particular de la radiación solar, los modelos de reanálisis combinan observaciones históricas con modelos numéricos del clima para estimar de forma coherente y completa variables atmosféricas que incluyen la radiación global, directa y difusa en la superficie terrestre. Este enfoque resulta especialmente útil cuando los datos disponibles son escasos, incompletos o no homogéneos, y se aplica en ámbitos como la investigación climática, la planificación energética y los estudios agrícolas.

Entre los modelos de reanálisis más utilizados en el estudio de la radiación solar destacan MERRA-2 y ERA5. El Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) proporciona datos desde 1980 y fue desarrollado para sustituir al conjunto original MERRA, incorporando mejoras en la asimilación de datos que permiten integrar observaciones modernas de radiancia hiperespectral y microondas, así como mediciones de ocultación de radio GPS. También asimila perfiles de ozono obtenidos por la NASA desde 2004 e incluye avances en el modelo GEOS y el sistema de asimilación GSI. Con una resolución espacial cercana a 50 km, MERRA-2 es el primer reanálisis global de largo plazo que asimila observaciones satelitales de aerosoles y modela sus interacciones con otros procesos del sistema climático, además de representar las capas de hielo en Groenlandia y la Antártida.

Por su parte, ERA5 constituye la quinta generación de reanálisis atmosférico global desarrollado por el European Centre for Medium-Range Weather Forecasts (ECMWF) y cubre el periodo desde enero de 1940 hasta la actualidad. Producido por el Copernicus Climate Change Service (C3S), ERA5 ofrece estimaciones horarias de una amplia gama de variables atmosféricas, terrestres y oceánicas, con cobertura global y resolución espacial de 31 km. La atmósfera se representa mediante 137 niveles verticales, desde la superficie hasta unos 80 km de altitud, e incluye estimaciones de incertidumbre para todas las variables, aunque a resoluciones espaciales y temporales más bajas.

Modelos satelitales

La estimación de la irradiancia solar por satélite es un campo de estudio crucial que busca calcular la radiación solar que llega a la superficie terrestre utilizando datos obtenidos desde sensores remotos Hay (1993); Alonso Suárez et al. (2012). Esta metodología resulta esencial debido a la alta variabilidad espacial de la radiación solar, influenciada principalmente por la presencia y características de las nubes Laguarda et al. (2022). Las redes de monitoreo en superficie suelen carecer de la densidad necesaria para capturar esta complejidad, mientras que los satélites ofrecen amplia cobertura geográfica, alta resolución espacial y un muestreo temporal frecuente —a menudo horario o subhorario—, ventajas que ningún otro sistema de observación

puede igualar.

Desde el trabajo pionero de Fritz et al. (1964), que correlacionó el albedo terrestre medido por satélite con la irradiancia solar en superficie Fritz et al. (1964), se han desarrollado numerosos modelos para estimar la radiación solar a partir de observaciones satelitales. En general, estos modelos se agrupan en dos grandes enfoques: métodos cualitativos y subjetivos, y técnicas objetivas que pueden ser estadísticas o basadas en fundamentos físicos.

Los métodos subjetivos requieren la interpretación manual de imágenes satelitales —a menudo impresas y de baja resolución— para estimar la cobertura nubosa y aplicar relaciones estadísticas que determinen la transmitancia atmosférica. En cambio, los métodos objetivos incluyen varias subcategorías. Los modelos empíricos o estadísticos se apoyan en relaciones funcionales derivadas de mediciones simultáneas de radiación solar y datos satelitales en un mismo lugar. Debido a su carácter empírico, su capacidad de extrapolación es limitada Laguarda (2025), aunque presentan la ventaja de ser simples y eficientes computacionalmente. Dentro de ellos, los denominados "estadísticos puros" seleccionan variables independientes —como nivel de brillo, ángulo cenital solar, agua precipitable o cobertura nubosa estimada— únicamente por su capacidad de explicar la variabilidad de la radiación solar, como es el caso del modelo de Tarpley Tarpley (1979).

Por su parte, los modelos basados físicamente con componentes empíricas parten del balance radiativo Tierra–atmósfera e incorporan coeficientes ajustados mediante observaciones, como los propuestos por Hanson Hanson (1976) y Ellis Ellis et al. (1978). Los modelos teóricos simulan de manera explícita los intercambios radiativos en el sistema Tierra–atmósfera, evitando la calibración empírica, aunque requieren datos ambientales auxiliares dependientes del tiempo y la localización. Entre ellos se distinguen los modelos de banda ancha, que utilizan el balance global de radiación solar, con el trabajo pionero de Gautier et al. Gautier et al. (1980), y los modelos espectrales, que resuelven la ecuación de transferencia radiativa en una atmósfera absorbente y dispersora.

Finalmente, los modelos híbridos o semiempíricos combinan fundamentos físicos con parametrizaciones empíricas de bajo número de coeficientes ajustables. Ejemplos destacados son Heliosat-4 Qu et al. (2017) y SUNY Perez et al. (2002), en los que la irradiancia bajo cielo arbitrario se estima multiplicando la irradiancia de cielo despejado (proveniente de un modelo físico) por un factor de nubosidad derivado de índices satelitales.

La complejidad de estos métodos varía en función de los datos de entrada requeridos, que pueden ir desde imágenes impresas de baja resolución hasta información digital de alta definición. También difieren en la necesidad de datos atmosféricos adicionales —como agua precipitable, aerosoles u ozono—, que pueden provenir tanto de los propios satélites como de otras fuentes. La precisión de las estimaciones puede verse afectada por errores de navegación, limitaciones temporales de muestreo o imprecisiones en el cálculo del flujo radiativo. La calidad de los datos para modelos de cielo despejado es un factor crítico, y la calibración con mediciones terrestres de alta calidad resulta indispensable para optimizar el desempeño. En particular, los modelos empíricos no deben extrapolarse a otras regiones o periodos sin una validación local, ya que el ajuste de sus parámetros es fundamental para reducir el sesgo y mejorar la exactitud.

Puede verse que existe una amplia cantidad de alternativas referidas a la estimación de la irradiancia solar. Sin embargo, debe notarse que estos modelos no sustituyen de ninguna manera a las mediciones en tierra y son más bien un complemento a las mismas o una opción a considerar cuando no se disponen de dichas medidas.

Capítulo 2

Adaptación al Sitio

Existen diferentes enunciados a los que se hace referencia mediante el término Adaptación al Sitio (SA por sus siglas en ingles).

«*todos los métodos estadísticos desarrollados para reducir la incertidumbre en el recurso solar local que buscan mejorar los datos de irradiancia derivados de satélites (disminuyendo sus errores aleatorios y, sobre todo, su sesgo) utilizando características de observaciones terrestres correspondientes durante períodos de tiempo superpuestos*»(Polo et al., 2016).

«*los procedimientos para corregir errores sistemáticos en un período prolongado de datos modelados en cuadrícula, utilizando un período corto de mediciones terrestres como referencia objetiva*»(Yang and Gueymard, 2021b).

«*la aplicación de un método de corrección a productos DSR en cuadrícula mediante el uso de mediciones del sitio*» (Ye et al., 2025).

Aunque cada definición tiene un matiz distinto, todas coinciden en lo esencial: ajustar los datos de irradiancia solar obtenidos por satélite o modelos para que representen con mayor precisión las condiciones reales de un sitio específico.

Una interpretación de estas definiciones nos permite expresar que este proceso consiste en aplicar correcciones estadísticas a los datos de radiación solar obtenidos de satélites o modelos numéricos (como los re-análisis), con el fin de hacerlos más representativos del sitio específico donde se quiere aplicar —por ejemplo, una futura planta solar.

Dado que los datos modelados tienen errores, especialmente errores sistemáticos (sesgos) y errores aleatorios, esta técnica usa mediciones reales de radiación tomadas en el sitio (aunque sea durante un período corto) como referencia para calibrar los datos. El objetivo es reducir la incertidumbre y mejorar la precisión en las estimaciones de largo plazo.

La Figura 2.1 ilustra una comparación entre tres conjuntos de datos de irradiancia solar a

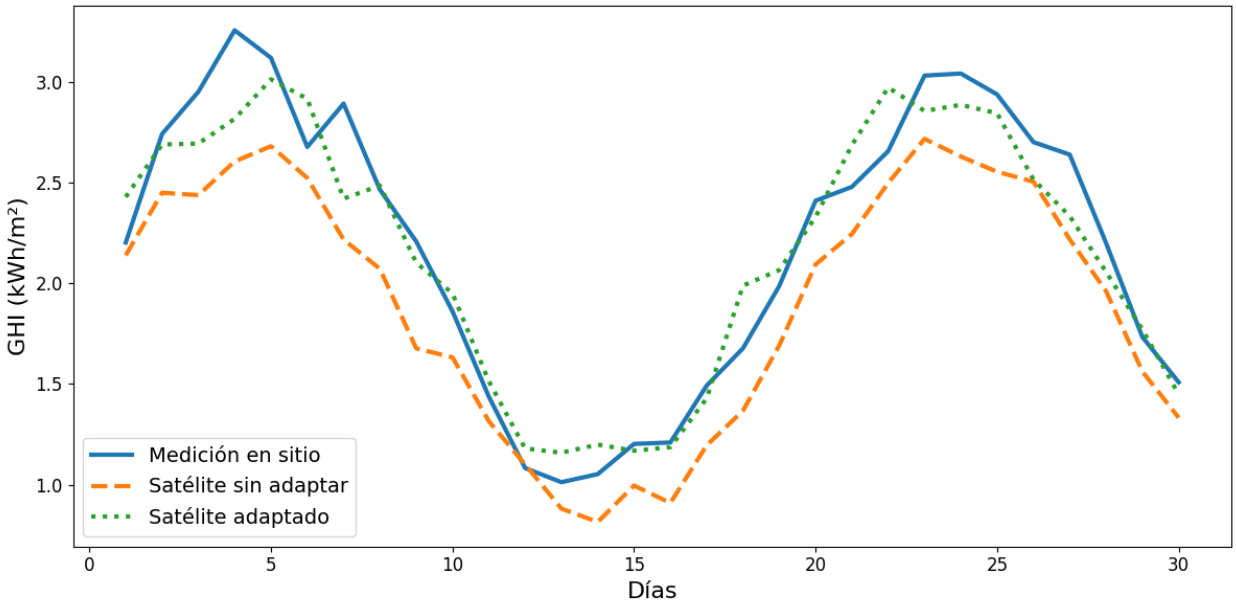


Figura 2.1: Comparación de irradiancia solar diaria entre datos de satélite sin adaptar, datos de satélite adaptados y mediciones locales en el sitio de referencia. Se observa cómo la adaptación al sitio corrige el sesgo sistemático y mejora la representatividad de los datos satelitales.

lo largo de un mes:

Medición en sitio (línea sólida): representa la referencia real tomada con instrumentos locales.

Satélite sin adaptar (línea discontinua): se observa un sesgo, ya que la serie está sistemáticamente desplazada respecto a la referencia. Además, los picos y valles no coinciden con exactitud, reflejando errores aleatorios.

Satélite adaptado (línea punteada): tras aplicar el procedimiento de adaptación al sitio, la serie corregida se ajusta mucho mejor a la referencia. El sesgo desaparece en gran medida y la forma de la curva sigue más de cerca la variabilidad de las mediciones locales.

Puede verse cómo la adaptación al sitio reduce tanto el error sistemático como la dispersión, logrando que los datos derivados de satélite sean más representativos de las condiciones reales de irradiancia en el lugar de interés.

En la Figura 2.2 puede verse el resultado que busca obtenerse al aplicar una adaptación al sitio. En color naranja se muestra la serie de medidas en tierra y en color azul la serie modelada, ambos casos con estilo de linea lleno. Debe tenerse en cuenta que el fondo cambiante de color (blanco y salmón) está realizado así a propósito.

Vamos a presentar una pequeña analogía que ayude a explicar de manera más cotidiana la idea general del preceso de SA.

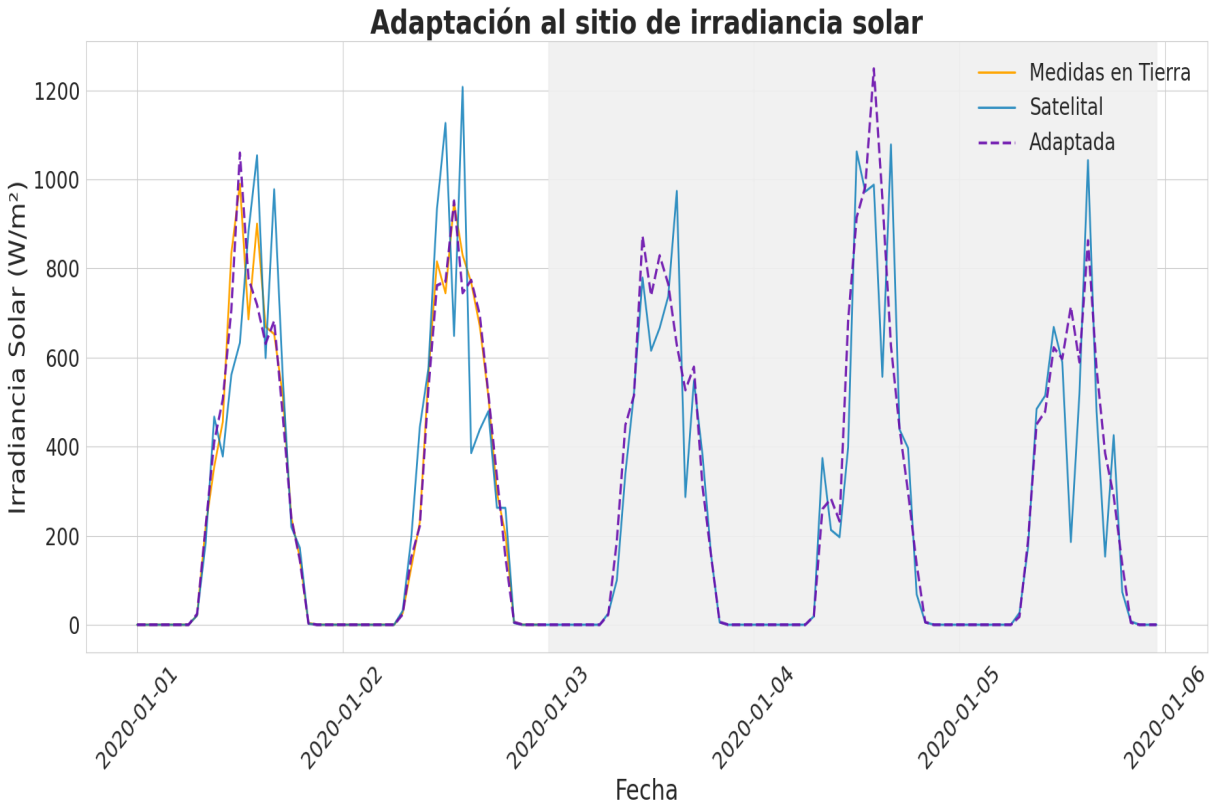


Figura 2.2: Adaptación al Sitio sobre una serie genérica

Imagina que tienes un **termómetro económico** en tu casa (equivalente a los datos de un *modelo*), pero al lado colocas uno profesional del hospital (equivalente a la *medición en sitio*).

Al comparar, notas que tu termómetro casero siempre marca $-2\,^{\circ}\mathrm{C}$ respecto al valor real.

Si corriges ese error sumando esos 2 grados, tu termómetro económico comienza a ser **útil y confiable** para el uso diario.

$\Rightarrow$ Eso es, en esencia, la **adaptación al sitio**: reducir el error de un modelo a partir de una referencia local.

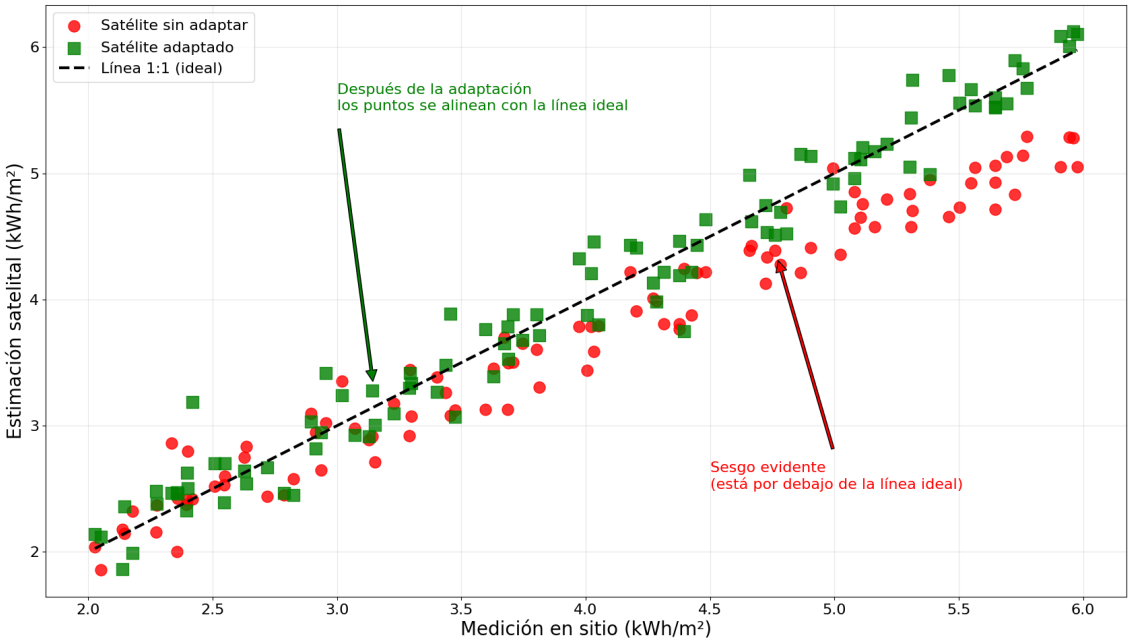


Figura 2.3: Diagrama de dispersión entre mediciones en sitio y estimaciones satelitales. Antes de la adaptación (puntos rojos), los datos muestran un sesgo claro al situarse por debajo de la línea 1:1. Después de la adaptación (puntos verdes), las estimaciones se alinean mucho mejor con la referencia, reduciendo el error sistemático.

La Figura 2.3 muestra de manera evidente el efecto de la adaptación al sitio:

Antes de la corrección (rojo): los puntos están sistemáticamente alejados de la línea 1:1, lo que indica que el satélite subestima los valores reales de irradiancia. Existe un sesgo claro.

Después de la corrección (verde): los puntos se agrupan alrededor de la línea 1:1, lo que significa que las estimaciones satelitales ahora representan con mayor fidelidad las mediciones en sitio.

2.1. Historia

Uno de los principales aportes sobre el tema fue realizado en el trabajo de Polo et al. (2016) en el marco de la Tarea 46 del Programa de Calefacción y Refrigeración Solar de la Agencia Internacional de la Energía Evaluación y pronóstico de recursos solares. En este trabajo se indica que la idea general de este proceso de corrección o calibración de datos modelados es similar a lo que se ha desarrollado en la industria eólica en el pasado (Potter et al., 2008).

Según Polo y sus colaboraores AS en un término actualmente utilizado en proyectos de energía solar para referirse a la mejora que puede lograrse en la irradiancia solar derivada de satélite y los datos del modelo cuando se utilizan mediciones terrestres locales a corto plazo para corregir errores sistemáticos y sesgos en el conjunto de datos original. A partir de esta idea general se han agrupado a los diversos métodos para SA en cinco categorías en las cuales se comprenden las principales estrategias para mejorar la precisión y reducir la incertidumbre en las estimaciones de radiación solar derivadas de satélites mediante el uso de mediciones locales de corta duración.

1. Métodos basados en modelos físicos (Physically based methods)
   Ajustan los datos de entrada atmosféricos (como la turbidez por aerosoles o el vapor de agua) para que los resultados coincidan mejor con las observaciones en tierra. Ejemplos: uso del modelo REST2 y corrección del AOD (Aerosol Optical Depth)

2. Métodos estadísticos (Statistical methods)
   Ajustan los datos del modelo para eliminar errores sistemáticos (sesgo) y mejorar la concordancia con datos medidos localmente.

   - Eliminación de sesgo mediante adaptación lineal
   - Métodos no lineales (transformación de características, polinomios, etc.)
   - MOS (Model Output Statistics)
   - MCP (Measure–Correlate–Predict)
   - Adaptación regional (usando estaciones cercanas)
   - Adaptación usando funciones de distribución acumulativa (CDF)

3. Adaptación de parámetros de entrada del modelo satelital
   Se modifican directamente los parámetros de entrada del modelo (como el índice de claridad o datos atmosféricos) en lugar de los resultados de irradiancia.

4. Técnicas MCP aplicadas a datos satelitales y de reanálisis
   Uso de métodos de correlación-predicción (común en energía eólica) para extender los datos de corto plazo a largo plazo con apoyo en modelos de reanálisis.

5. Combinación de datos satelitales con modelos meteorológicos numéricos (NWP)
   Mejora de la precisión mediante regresiones no paramétricas (como modelos aditivos generalizados - GAM) que combinan datos satelitales con modelos meteorológicos.

Además de presentar la clasificación anterior, se han indicado algunas directrizces que deben ser tomadas en cuenta en un proceso de adaptación.

A partir del año 2020 se comenzaron a publicar diversos trabajos sobre Adaptación al Sitio (SA). En (Polo et al., 2020) se presentan los resultados de adaptaciones realizadas en diez estaciones a escala horaria, considerando ocho modelos satelitales y dos de re-análisis. En (Fernández-Peruchena et al., 2020) se evaluaron los resultados de la adaptación al sitio utilizando regresiones lineales múltiples (MLR) y se combinaron estos resultados con un ajuste mediante mapeo de cuantiles (QM). Este enfoque de ensamble de modelos mostró una reducción promedio del 1,7 % en el sesgo relativo y del 3,3 % en la desviación cuadrática media relativa a escala horaria. El estudio también sugiere que un año de datos es suficiente para entrenar los modelos de ajuste.

En el trabajo de (Babar et al., 2020) se evaluó el desempeño de la SA en los valores medios diarios del modelo satelital CLARA-A2 y del modelo de re-análisis ERA5. Para ello, se utilizó un modelo de Random Forest para la regresión, incorporando como predictores los valores de GHI modelados por CLARA-A2 y ERA5, sus respectivos índices de cielo despejado y el ángulo cenital solar medio. Los resultados demostraron una reducción de la desviación cuadrática media de 17,9 W/m^2 a 16,2 W/m^2 y una corrección completa del sesgo inicial de -1,5 W/m^2. La motivación para adoptar un modelo de regresión basado en aprendizaje automático surgió de la hipótesis de que este tipo de algoritmo podría aplicar funciones de regresión distintas a cada subconjunto del espacio de datos predictivos. Este trabajo es uno de los primeros en utilizar modelos de aprendizaje automático como Random Forest para determinar una función de ajuste que calibra las mediciones en tierra y puede aplicarse a series de largo plazo.

A partir del trabajo de Babar y colaboradores, se han publicado diversos estudios en los que se exploran modelos de aprendizaje automático como alternativa para determinar funciones de ajuste sobre series modeladas, ya sean satelitales o de re-análisis.

En (Narvaez et al., 2021) se evaluó la adaptación al sitio en series horarias, explorando el uso de redes neuronales y los modelos Random Forest y AdaBoost. El rendimiento de estos modelos se evaluó en función de las métricas obtenidas en comparación con los ajustes realizados mediante QM y MLR, donde Random Forest logró una mejora aproximada del 38 % con respecto a QM y MLR.

En (Yang and Gueymard, 2021a) se presenta una alternativa denominada Adaptación al Sitio Probabilística, que aprovecha la disponibilidad de múltiples modelos de estimación simultáneamente. Este enfoque combina regresiones de cuantiles sobre múltiples modelos a la vez para mejorar la precisión de la adaptación al sitio.

En el trabajo de (Salamalikis et al., 2022), se aplicó SA al modelo Heliosat-4 con una resolución horaria mediante el entrenamiento de diferentes modelos de regresión para condiciones de cielo despejado y nublado, considerando segmentos de datos basados en rangos de 15° del ángulo cenital solar. En este estudió se reportó que el sesgo se redujo en un 50 % y la desviación

cuadrática media disminuyó en un 3,8 % en comparación con las estimaciones sin SA. Los mejores resultados de este trabajo se obtuvieron utilizando modelos de regresión basados en árboles de decisión, en particular Random Forest. Además, una característica a destacar sobre este trabajo es que los autores han buscado eliminar la estacionalidad de la serie definiendo Δ_{GHI} como la diferencia entre la GHI medida y la GHI modelada. Esta variable se utiliza entonces como objetivo para los modelos de regresión, y la GHI adaptada se calcula como la suma del GHI modelada y Δ_{GHI}. Puede considerarse que los autores en este trabajo han buscado quitar la «*estacionalidad*» de la serie. Esta práctica es ampliamente conocida y estudiada en el área del análisis de series temporales, es recomendada y analizada en estudios como (Thornton, 2013; Claveria et al., 2015).

En (Zainali et al., 2024) se evaluó un proceso de SA en tres sitios del norte de Europa utilizando diversos ajustes de mapeo de cuantiles y modelos de aprendizaje automático para ajustar los datos del modelo STRANG. Los hallazgos indican que los modelos de aprendizaje automático generalmente obtienen un rendimiento superior al de los métodos estadísticos, logrando una mejora de hasta el 9,2 % en la reducción del sesgo.

2.2. Características

2.3. Ventajas y Desventajas

2.4. Modelos de Aprendizaje Automático

Aunque Arthur Samuel no fue el primero en publicar un artículo que empleara el término .ªprendizaje automático", se le atribuye la creación y definición de este concepto como el campo especializado que conocemos hoy. En su trabajo 'Algunos estudios sobre aprendizaje automático usando el juego de damas' Samuel (1959), presentó el aprendizaje automático como una rama de la informática que permite a las computadoras mejorar su desempeño sin necesidad de ser programadas de forma explícita.

Si bien la definición original de Samuel no lo menciona directamente, un aspecto fundamental del aprendizaje automático es el autoaprendizaje. Este concepto implica el uso de modelos estadísticos para identificar patrones y optimizar el rendimiento con base en datos e información empírica, sin requerir instrucciones de programación explícitas Theobald (2024).

Samuel no infirió que las máquinas puedan tomar decisiones sin programación previa. Al contrario, el aprendizaje automático depende en gran medida de la entrada de código. En cambio, observó que las máquinas pueden realizar una tarea específica utilizando datos de entrada en lugar de depender de un comando de entrada directo.

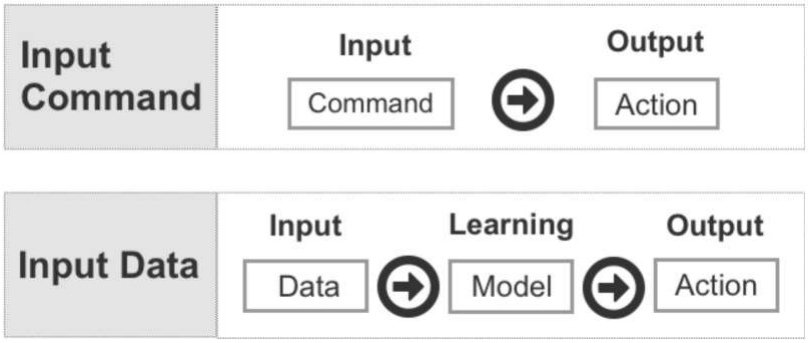


Figura 2.4: Comparación entre el comando de entrada y los datos de entrada

2.4.1. Modelo de red neuronal

No existe una definición única de lo que es una red neuronal artificial, pero distintas fuentes proponen descripciones complementarias. Algunas de las más comunes son:

- Un modelo computacional, paralelo, compuesto por unidades procesadoras adaptativas altamente interconectadas.
- Un sistema de procesamiento de la información que aplica principios inspirados en la organización del cerebro humano.
- Un modelo matemático diseñado para emular, de manera simplificada, el funcionamiento del cerebro.
- Un sistema de información con características de funcionamiento similares a las redes neuronales biológicas.
- Una red adaptativa que combina técnicas de procesamiento paralelo de la información.
- Una extensión de los métodos clásicos estadísticos, especialmente útil en el reconocimiento de patrones.

En todas estas definiciones se aprecia un componente de *simulación biológica*. Las redes neuronales artificiales se inspiran en el cerebro humano en el sentido de que el procesamiento de la información se distribuye entre elementos básicos llamados *neuronas*. Estas están interconectadas mediante *pesos sinápticos*, que se ajustan a lo largo del tiempo durante un proceso denominado *aprendizaje*. En términos simples, aprender consiste en modificar la intensidad de las conexiones entre neuronas para resolver una tarea determinada.

Neurona artificial

Los componentes básicos de una neurona artificial son:

1. Un conjunto de conexiones ponderadas (pesos sinápticos).
2. Un sesgo, que actúa como umbral de activación.
3. Un sumador, que agrega las entradas multiplicadas por sus pesos correspondientes.
4. Una función de activación no lineal, que permite ampliar la capacidad de representación del modelo.

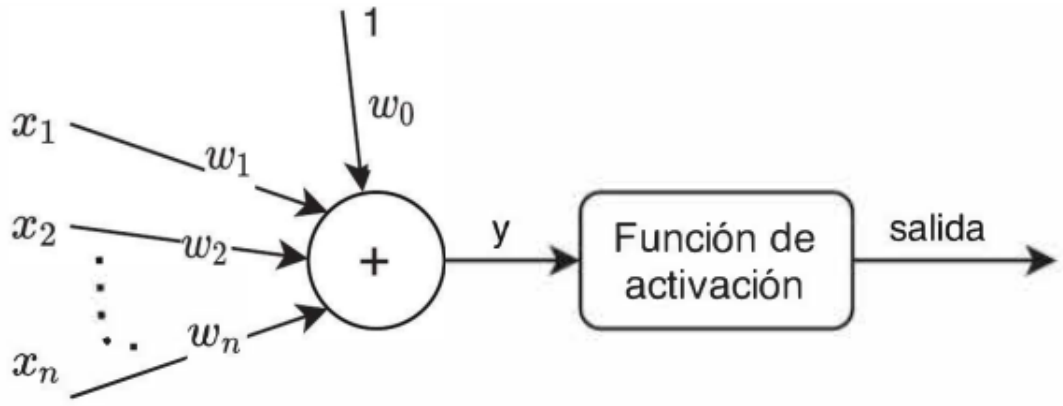


Figura 2.5: Esquema de una neurona artificial

El funcionamiento matemático de una neurona artificial puede expresarse como:

$$salida = f(y) = f\left(\sum_{k=0}^{n} w_k x_k\right) \tag{2.1}$$

donde x_i son las entradas, w_i los pesos sinápticos y $x_0 = 1$ corresponde al sesgo con coeficiente w_0. La función f representa la activación de la neurona.

Funciones de activación

Las funciones de activación más comunes se resumen en la Tabla 2.1. Estas permiten introducir no linealidad en el modelo, lo cual es fundamental para que la red pueda aproximar funciones complejas.

Función	Expresión	Comentario
Signo	$f(x) = \begin{cases} 1, & \text{si } x \geq 0 \\ -1, & \text{si } x < 0 \end{cases}$	Usada en los primeros modelos de neuronas artificiales.
Sigmoide	$f(x) = \frac{1}{1+e^{-x}}$	Transición suave entre 0 y 1, útil en clasificación binaria.
Tangente hiperbólica	$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	Similar a la sigmoide, pero con valores entre -1 y 1.
ReLU	$f(x) = \text{máx}(0, x)$	Una de las más usadas actualmente; evita problemas de gradiente.
Softmax	$f(x_i) = \frac{e^{x_i}}{\sum_k e^{x_k}}$	Utilizada en la salida de modelos de clasificación multiclase.

Tabla 2.1: Funciones de activación más comunes en redes neuronales artificiales.

Perceptrón y sus limitaciones

El perceptrón simple, basado en esta estructura, funciona como un clasificador binario que sólo puede resolver problemas linealmente separables. El procedimiento de aprendizaje consiste en:

1. Inicializar aleatoriamente los pesos w_k.
2. Establecer el parámetro de aprendizaje α.
3. Calcular la salida: $salida = signo(\sum_{k=0}^{n} w_k x_k)$.
4. Calcular el error: $error = salida_{deseada} - salida$.
5. Actualizar los pesos: $w_k = w_k + \alpha \cdot error \cdot x_k$.
6. Repetir el proceso.

Aunque este modelo es simple y útil, su capacidad de representación es limitada. Para superar estas restricciones se introducen las redes **multicapa**.

Perceptrón Multicapa (MLP)

El *Perceptrón Multicapa* (MLP, por sus siglas en inglés) organiza neuronas en diferentes capas: una capa de entrada, una o varias capas ocultas y una capa de salida (Fig. 2.6). Gracias a esta estructura, el MLP puede aproximar funciones no lineales y resolver problemas más complejos.

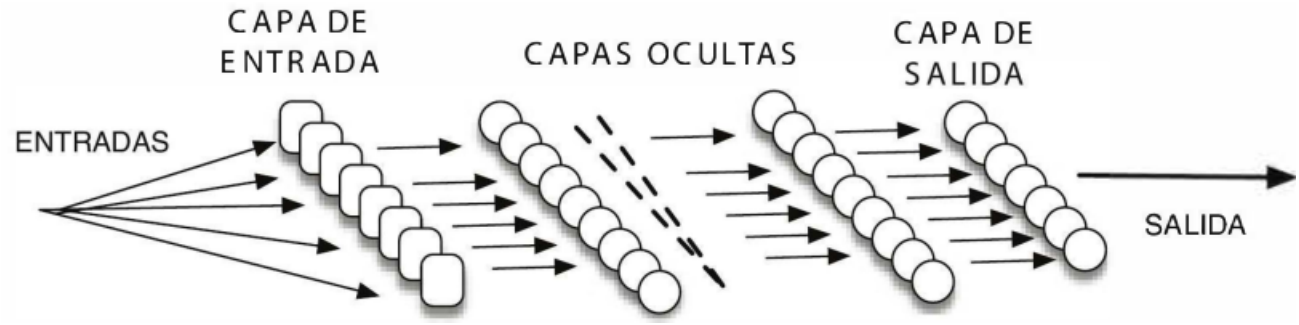


Figura 2.6: Esquema de una red neuronal multicapa (MLP).

Un MLP con una capa oculta se puede definir como:

$$\hat{y} = f^{(2)}\Big(W^{(2)} \cdot f^{(1)}(W^{(1)}x + b^{(1)}) + b^{(2)}\Big), \tag{2.2}$$

donde x es el vector de entrada, $W^{(1)}, W^{(2)}$ son matrices de pesos, $b^{(1)}, b^{(2)}$ los sesgos y $f^{(1)}, f^{(2)}$ las funciones de activación de cada capa Ramchoun et al. (2016).

MLP en la estimación de irradiancia solar

En el contexto de la estimación de la irradiancia global horizontal (GHI), el MLP puede verse como un *equipo de expertos neuronales*. Cada neurona aporta una "opinión parcial" a partir de las variables meteorológicas (nubosidad, humedad, ángulo solar, aerosoles), mientras que las capas ocultas integran estas opiniones para refinar la predicción de manera no lineal Toro Bayona and Lizarazo Salcedo (2012).

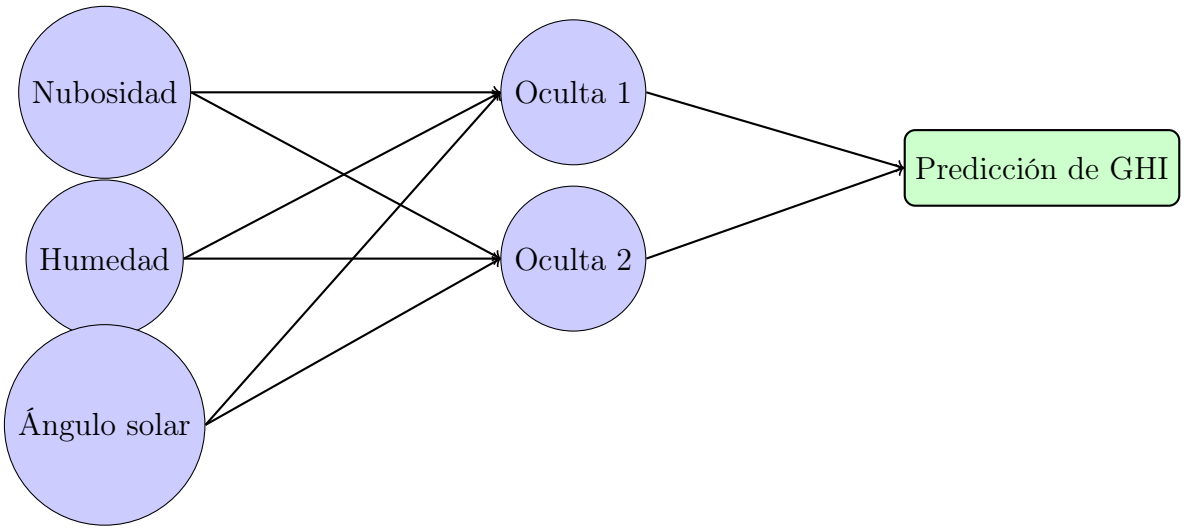
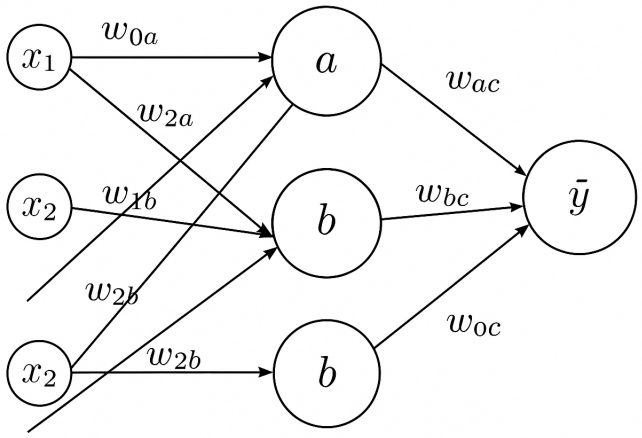


Figura 2.7: Analogía del MLP como un conjunto de expertos neuronales que refinan la predicción de la GHI.

El MLP ofrece una estructura potente y flexible capaz de modelar relaciones no lineales complejas, esto permite considerarlo como una herramienta adecuada para la estimación de irradiancia solar.

Ejemplo práctico de entrenamiento en un MLP

Para ilustrar el funcionamiento del *Perceptrón Multicapa* (MLP), consideremos un ejemplo sencillo con dos entradas, una capa oculta con dos neuronas y una salida (Figura 2.8).



Ejemplo de un MLP con dos entradas, dos neuronas ocultas y una salida

Figura 2.8: Ejemplo de un MLP con dos entradas, dos neuronas ocultas y una salida.

El cálculo de las salidas se realiza en dos fases: propagación hacia adelante (*forward pass*) y retropropagación del error (*backpropagation*).

1. Propagación hacia adelante Las entradas x_1 y x_2 llegan a las dos neuronas ocultas a y b:

$$y_a = w_{0a} + w_{1a}x_1 + w_{2a}x_2, \quad O_a = f(y_a) \tag{2.3}$$

$$y_b = w_{0b} + w_{1b}x_1 + w_{2b}x_2, \quad O_b = f(y_b) \tag{2.4}$$

Las salidas O_a y O_b alimentan a la neurona de salida c:

$$y_c = w_{0c} + w_{ac}O_a + w_{bc}O_b, \quad \hat{y} = f(y_c) \tag{2.5}$$

donde $f(\cdot)$ es una función de activación diferenciable (sigmoide, tangente hiperbólica, ReLU, etc.).

2. Cálculo del error Dado un valor deseado y, el error cuadrático medio (ECM) para un ejemplo es:

$$E = \frac{1}{2}(y - \hat{y})^2 \tag{2.6}$$

3. Retropropagación El objetivo es calcular cómo varía el error respecto a cada peso y ajustar estos valores en la dirección del gradiente descendente.

- Para la salida c:

$$\delta_c = (y - \hat{y}) f'(y_c) \tag{2.7}$$

- Para las neuronas ocultas a y b:

$$\delta_a = f'(y_a) \cdot (w_{ac}\delta_c), \quad \delta_b = f'(y_b) \cdot (w_{bc}\delta_c) \tag{2.8}$$

4. Actualización de pesos Finalmente, los pesos se actualizan usando una tasa de aprendizaje η:

$$w_{ij} \leftarrow w_{ij} + \eta \cdot \delta_j \cdot x_i \tag{2.9}$$

donde x_i es la entrada a la neurona j. Este procedimiento se repite para todos los ejemplos del conjunto de entrenamiento hasta que el error sea lo suficientemente pequeño.

5. Resumen del ciclo de aprendizaje

1. Inicializar pesos y sesgos aleatoriamente.
2. Realizar la propagación hacia adelante para obtener la salida $\hat{y}$.
3. Calcular el error E comparando con el valor real y.
4. Retropropagar el error para obtener los deltas δ de cada neurona.
5. Actualizar los pesos usando la regla de gradiente descendente.
6. Repetir el proceso hasta la convergencia.

Este procedimiento es la base del entrenamiento de redes neuronales modernas. A pesar de su simplicidad, este esquema permite que los MLP aproximen relaciones altamente no lineales, siendo especialmente útiles para la estimación de la irradiancia solar, donde influyen múltiples variables meteorológicas de manera simultánea.

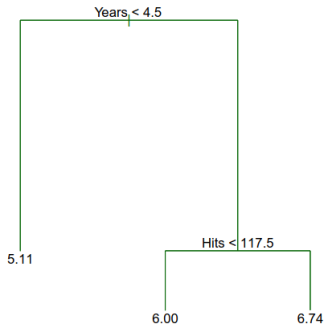


Figura 2.9: Ejemplo Árbol de Regresión. Para los datos de Hitters, se construye un árbol de regresión para predecir el logaritmo del salario de un jugador de béisbol, en función del número de años que ha jugado en las grandes ligas y el número de hits que realizó en el año anterior.

2.4.2. Arboles de Decisión

Los árboles de decisión son modelos no paramétricos (es decir que no se no toman suposiciones previas sobre la forma de distribución de los datos) que se utilizan principalmente para la resolución de problemas de clasificación o regresión. También son conocidos como árboles de clasificación y regresión (CART, classification and regression trees). Este tipo de modelo fue propuesto por Leo Breiman en el libro Breiman et al. (1984)

Los árboles de decisión se basan en una serie de reglas de decisión para dividir el espacio de características predictoras en un número menor de regiones disjuntas en cada una de las cuales los valores de la variable respuesta son similares.

Un árbol de decisión parte del conjunto de datos de entrenamiento, correspondiente a un nodo raíz, y lo va dividiendo recursivamente en subconjuntos de datos homogéneos, dando lugar a nuevos nodos. La manera de formar los subgrupos es mediante la formulación de preguntas con respuesta binaria (si la variable respuesta es 'jugar al tenis' se formula la pregunta ¿Sí o No juega al tenis?; si es 'pesa más o menos de 75 kg.' la pregunta es ¿el peso es $<=$75 o $>$75?).

Los árboles de decisión se pueden clasificar en función del tipo de variable respuesta, si la variable de respuesta y es cuantitativa el árbol es de regresión, si y es cuantitativa el árbol es de clasificación.

El proceso general de construcción de un árbol de regresión puede describirse en dos pasos:

1. Dividir el espacio de predictores: es decir, el conjunto de valores posibles de $X_1, X_2, ..., X_p$ se divide en J regiones distintas y no superpuestas, $R_1, R_2, .., R_J$.

2. Hacer una predicción para cada región: para cada observación que cae en una región R_j, la predicción será simplemente la media de los valores de respuesta de las observaciones de entrenamiento dentro de R_j.

2.4.3. Random Forest

El algoritmo **Random Forest** (RF) es un modelo de aprendizaje supervisado de tipo *ensemble* que se utiliza para resolver problemas de **clasificación** y **regresión** Louppe (2015); Salman et al. (2024). Este método, conceptualizado por Leo Breiman, es una extensión del concepto de *bagging* y se ha establecido como una técnica muy precisa y robusta en la minería de datos Cutler et al. (2011).

El fundamento de RF es la construcción de un conjunto de múltiples árboles de decisión, donde cada árbol se entrena sobre un subconjunto de datos generado de forma aleatoria Salman et al. (2024). La aleatoriedad se introduce en dos niveles principales para garantizar que los árboles sean diversos y que el modelo no se sobreajuste:

1. **Muestreo con reemplazo (Bootstrapping)**: Se generan múltiples submuestras del conjunto de datos de entrenamiento original, con la posibilidad de que una misma observación aparezca varias veces en la misma submuestra. Cada submuestra se usa para entrenar un árbol de decisión individual Salman et al. (2024).

2. **Selección aleatoria de variables**: En cada nodo de un árbol, el algoritmo selecciona un subconjunto aleatorio de las variables predictoras disponibles. La mejor división del nodo se determina a partir de este subconjunto reducido Cutler et al. (2011); Salman et al. (2024). Esta técnica reduce la correlación entre los árboles y mejora la capacidad de generalización del modelo Cutler et al. (2011).

La división óptima en cada nodo del árbol se basa en la minimización de una función de costo, que varía según el tipo de problema.

Para problemas de **clasificación**, se utilizan medidas de impureza como el **índice Gini** o la **Entropía**. El Índice Gini mide la probabilidad de que una muestra elegida al azar de un nodo sea clasificada erróneamente, y se calcula de la siguiente manera:

$$I_G(p) = 1 - \sum_{i=1}^{c} p_i^2$$

donde c es el número de clases y p_i es la proporción de muestras de la clase i en el nodo.

Para problemas de **regresión**, el criterio de división se basa en la minimización de la varianza o el **Error Cuadrático Medio (MSE)** de las predicciones en los nodos hijos:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

donde n es el número de muestras, y_i es el valor real y $\hat{y}_i$ es el valor predicho.

Para realizar una predicción final, el algoritmo agrega los resultados de todos los árboles del *bosque* Salman et al. (2024). En problemas de **clasificación**, la predicción se basa en el **voto**

mayoritario de los árboles. Para problemas de **regresión**, la predicción final es el promedio de los resultados de cada árbol individual:

$$\hat{y}_{RF}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} h_b(\mathbf{x})$$

donde B es el número total de árboles en el bosque y $h_b(\mathbf{x})$ es la predicción del b-ésimo árbol.

Random Forest ofrece varias ventajas significativas para el análisis de datos Cutler et al. (2011):

- **Resistencia al sobreajuste (*overfitting*)**: El algoritmo es intrínsecamente resistente al sobreajuste, ya que el conjunto de árboles se entrena en subconjuntos aleatorios de datos Salman et al. (2024).
- **Manejo de variables**: Es capaz de manejar eficientemente un gran número de variables, incluso cuando hay valores perdidos, sin necesidad de imputación previa Cutler et al. (2011).
- **Estimación de la importancia de las variables**: El algoritmo proporciona una medida integrada que indica la contribución relativa de cada variable al modelo, lo que facilita la interpretación de los resultados Cutler et al. (2011).

2.4.4. XGBoost

El algoritmo *Extreme Gradient Boosting* (XGBoost) es una implementación optimizada del método de *Gradient Boosted Decision Trees* (GBDT) Chen and Guestrin (2016b). Su funcionamiento consiste en combinar múltiples árboles de decisión simples para construir un modelo robusto, de manera semejante a cómo un equipo de especialistas aporta su conocimiento colectivo para tomar mejores decisiones. XGBoost se ha consolidado como estado del arte en aprendizaje automático debido a su alto desempeño en tareas de clasificación y regresión en diversos dominios Espinosa-Zúñiga (2020).

Para entender el algoritmo y su aplicación en la estimación de la GHI, podemos usar una analogía. Supongamos que deseamos predecir la GHI en un lugar específico.

Un único árbol de decisión actúa como un *experto solitario*, que utiliza reglas simples para emitir un juicio, por ejemplo: ès la nubosidad es alta, entonces la irradiancia será baja". Aunque este razonamiento es útil, tiene limitaciones: la irradiancia también depende de la altura solar, aerosoles, humedad relativa o estacionalidad. Por ello, un solo árbol puede fallar al no capturar la complejidad completa del fenómeno.

XGBoost supera esta limitación mediante un *comité de expertos*, es decir, un conjunto de árboles que se construyen secuencialmente para aprender de los errores de sus predecesores. Cada árbol nuevo se centra en corregir las predicciones incorrectas de los anteriores, especializándose

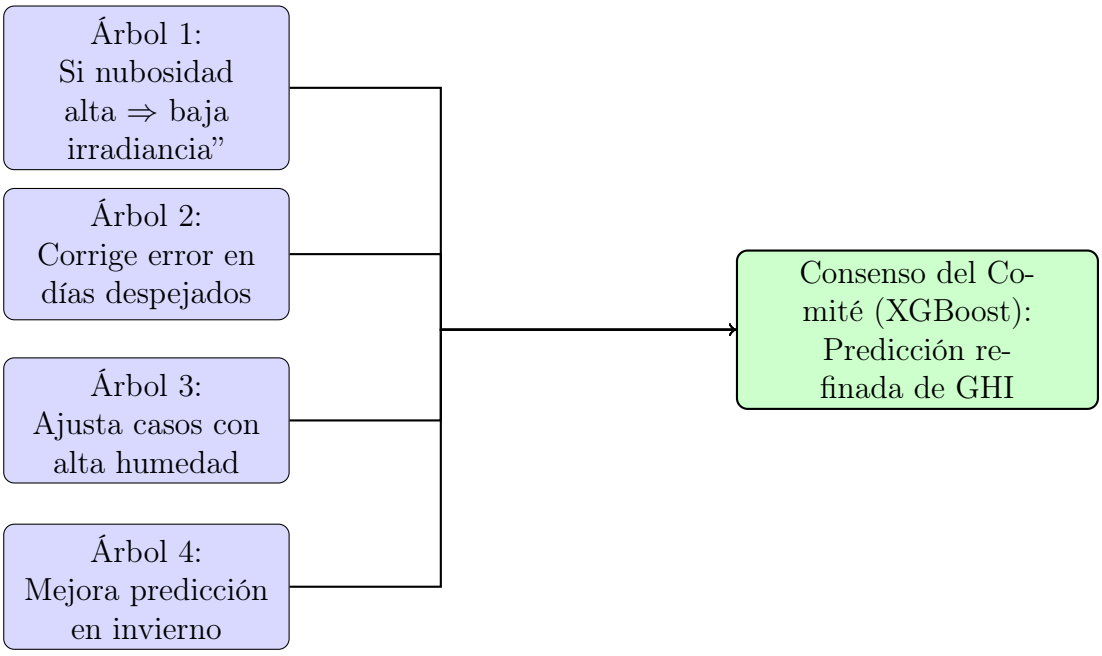


Figura 2.10: Analogía del comité de expertos en XGBoost: cada árbol corrige los errores de sus predecesores y contribuye a una predicción final más precisa de la GHI.

en las regiones donde estos fallaron. Así, el modelo colectivo refina sus predicciones de manera progresiva. Esta dinámica se ilustra en la Figura 2.10.

Podemos compararlo con un proceso de deliberación científica: el primer investigador propone un modelo básico; otro detecta que no captura los días parcialmente nublados y lo corrige; un tercero ajusta los errores en días despejados con baja humedad. Con el tiempo, el grupo obtiene un modelo más completo que cualquiera de los expertos individuales.

Esta idea refleja la esencia del *gradient boosting*: aprendizaje aditivo y secuencial, donde cada árbol se incorpora para reducir la pérdida residual del conjunto previo Chen and Guestrin (2016b). La fuerza del modelo final no está en la exactitud de cada árbol individual (clasificador débil), sino en la sinergia de todos ellos, formando un predictor colectivo altamente robusto. En el caso de la estimación de GHI, el ensamble de árboles de XGBoost permite capturar relaciones complejas entre variables meteorológicas y solares, mientras controla el sobreajuste mediante regularización y técnicas adicionales como *shrinkage* y submuestreo Developers (2018).

En resumen, si un árbol es un experto solitario con visión parcial, XGBoost funciona como un *panel de expertos* que deliberan y corrigen mutuamente sus errores para alcanzar predicciones más precisas.

Formalmente, el modelo se define como:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}, \tag{2.10}$$

donde cada f_k es un árbol de regresión ($CART$) y $\mathcal{F}$ es el espacio de todos los árboles posibles. La función objetivo combina el error de predicción y la complejidad del modelo:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2.11}$$

con regularización definida como:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{2.12}$$

donde T es el número de hojas, w los pesos de cada hoja, γ penaliza la complejidad del árbol y λ regula la magnitud de los pesos Chen and Guestrin (2016b).

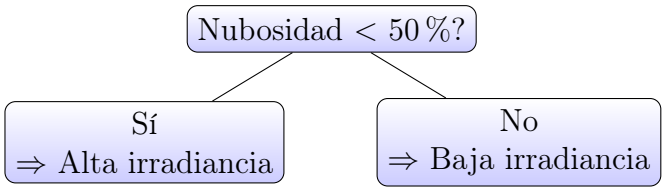


Figura 2.11: Ejemplo de un árbol de decisión simple. XGBoost combina cientos de estos árboles débiles para construir un modelo poderoso.

XGBoost entrena los árboles de manera *aditiva*, es decir, cada iteración añade un árbol que corrige los errores acumulados. Para ello utiliza una expansión de segundo orden de la función objetivo:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \tag{2.13}$$

donde g_i y h_i son el gradiente y la segunda derivada de la función de pérdida en la predicción previa. Esta formulación otorga estabilidad y precisión al proceso de entrenamiento Chen and Guestrin (2016b).

El proceso iterativo se representa en la Figura 2.12.

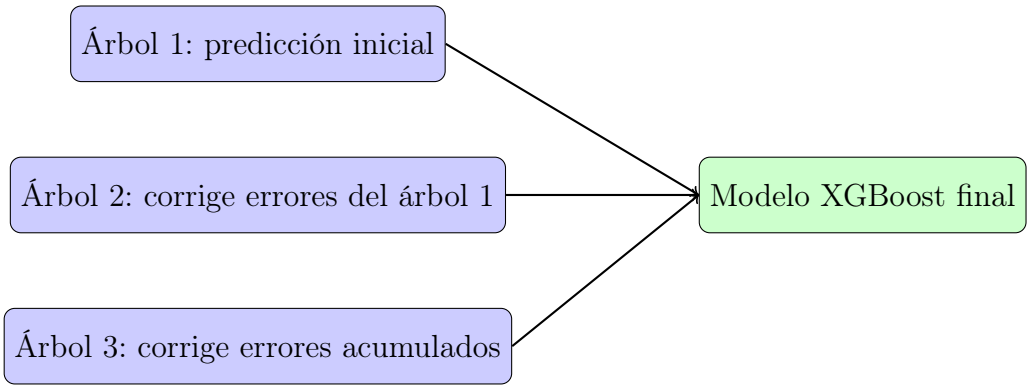


Figura 2.12: Proceso iterativo: cada nuevo árbol corrige los errores del modelo acumulado.

XGBoost también incluye estrategias adicionales para mejorar la generalización Developers (2018):

- **Shrinkage (η):** funciona como una 'moderación" en las decisiones del comité, reduciendo el impacto de cada nuevo árbol.
- **Submuestreo:** cada árbol se entrena con una muestra parcial de datos y características, como consultar a un subgrupo de expertos para ganar diversidad en las opiniones.
- **Regularización adicional:** actúa como una *disciplina* que limita la complejidad de cada experto, evitando que se vuelva demasiado específico.

En este estudio se utilizó la librería `xgboost` Developers (2018), con soporte para paralelización y GPU, lo que permitió un entrenamiento eficiente incluso con grandes volúmenes de datos. Los principales hiperparámetros (`max_depth`, `eta`, `n_rounds`, λ, γ) se ajustaron mediante *grid search*.

De acuerdo a las especificaciones teóricas presentadas, consideramos que el uso de XGBoost es especialmente adecuado para el análisis de irradiancia solar porque:

1. Captura relaciones no lineales entre variables atmosféricas y solares.
2. Reduce el sobreajuste mediante mecanismos internos de regularización.
3. Permite escalabilidad y eficiencia en grandes bases de datos.
4. Suele superar a otros ensambles como Random Forest Espinosa-Zúñiga (2020).

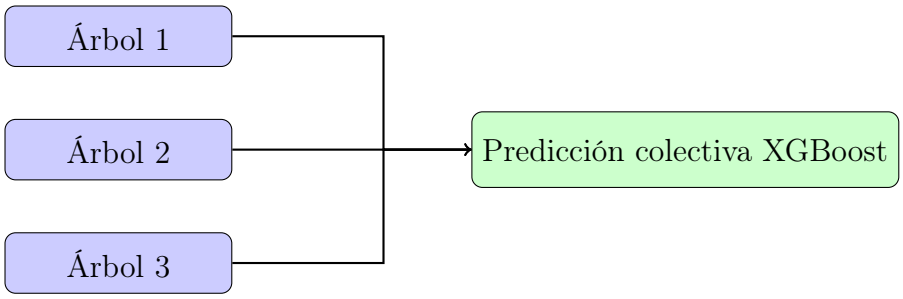


Figura 2.13: Esquema conceptual: múltiples árboles (*expertos*) corrigen iterativamente sus errores para formar un modelo robusto.

La Figura 2.13 resumen la idea para una predicción generica utilizando XGBoost como modelo regresor.

En resumen, mientras que un árbol de decisión actúa como un experto solitario y XGBoost combina múltiples árboles secuencialmente, un MLP funciona como un sistema de neuronas interconectadas que colectivamente aprenden patrones complejos en los datos, logrando predicciones precisas de irradiancia solar.

Capítulo 3

Adaptación al Sitio en NOA

En este capítulo se presentan los resultados obtenidos a partir de las distintas evaluaciones realizadas a la AS.

3.1. Medidas en tierra

Los sitios analizados en esta tesis se resumen en la Tabla 3.1, donde se incluyen sus códigos de identificación, coordenadas geográficas, altitud sobre el nivel del mar, periodos de medición, clasificación climática según Köppen–Geiger Peel et al. (2007) y el tipo de piranómetro utilizado. Estas estaciones están ubicadas en el noroeste de Argentina como puede verse en la Figura 3.1 y abarcan diversas condiciones climáticas y geográficas, que van desde tierras bajas subtropicales hasta altiplanos andinos de gran altitud.

La estación Sa se encuentra en el campus experimental del INENCO, en la Universidad Nacional de Salta. Está situada en un entorno urbano preandino dentro del Valle de Lerma, una zona donde es común la formación de nubes debido a la cercanía de la cordillera. Su clima es subtropical de montaña, con inviernos secos y veranos frescos (Cwb). Las mediciones se realizaron entre 2009 y 2020 utilizando un piranómetro Eppley PSP.

La estación Lq, localizada en La Quiaca, presenta un clima estepario frío semiárido (BSk) típico de regiones andinas, y destaca por registrar uno de los mayores niveles de horas de sol anuales en Argentina. Contó con un piranómetro Kipp & Zonen CMP11 y recopiló datos entre 2021 y 2023.

La estación Yu, ubicada en una zona de clima subtropical húmedo (Cwa), registró datos entre 2017 y 2018 con un sensor CMP11. La estación Sca, también con clima Cwb, operó entre 2012 y 2013 con un piranómetro CMP3. Finalmente, la estación Ero, situada en una región de gran altitud con clima BSk, recopiló datos entre 2016 y 2018, igualmente con un CMP3.

Todas las estaciones utilizan piranómetros que cumplen con los estándares de la norma ISO 9060:2018 Clase A o B para la medición de la irradiancia global horizontal (GHI). Los datos se registraron a intervalos de un minuto, y cada valor corresponde al promedio de seis muestras

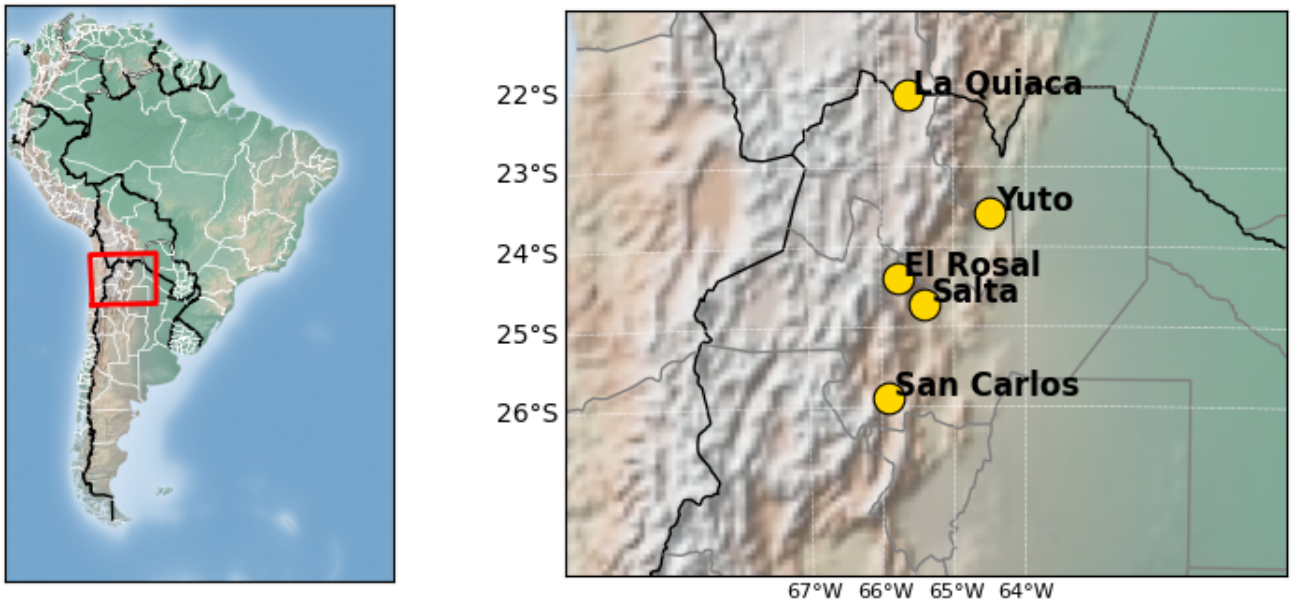


Figura 3.1: Ubicación de la estaciones de medida

tomadas cada 10 segundos.

3.1.1. Control de Calidad en las Medidas

Las medidas fueron sometidas a un control de calidad (QC) siguiendo un procedimiento simplificado basado en Nollas et al. (2023), con una etapa preliminar de filtrado por inspección visual según las recomendaciones de Abal et al. (2020). Dado que este estudio se basa únicamente en mediciones de irradiancia global horizontal (GHI) y no incluye componentes difusas, se aplicó una versión reducida del procedimiento original.

ID	Provincia	Localidad	Latitud	Longitud	Altitud (m.s.n.m)	Clima
Yu	Jujuy	Yuto	-23.58	-64.5	401	**Cwa**
Sa	Salta	Salta	-24.72	-65.4	1233	**Cwb**
Sca	Salta	San Carlos	-25.8951	-65.925	1624	**Cwb**
Er	Salta	El Rosal	-24.39278	-65.76806	3355	**Bsk**
Lq	Jujuy	La Quiaca	-24.39278	-65.76806	3355	**Bsk**

Tabla 3.1: Estaciones de medidas utilizadas en este trabajo

La Tabla 3.2 resume los filtros utilizados, donde E es la constante solar, S el factor de corrección de la distancia Tierra–Sol, θ_z el ángulo cenital solar y kt el índice de claridad, definido como la razón entre la GHI y la irradiancia teórica en el tope de la atmósfera sobre un plano horizontal.

Tabla 3.2: Filtros de control de calidad aplicados a las mediciones.

Filtro	**Descripción**
F1	$GHI < 1{,}5\ E\ S\ (\cos(\theta_z))^{1{,}2} + 100\,\mathrm{W/m^2}$
F2	$GHI > (6{,}5331 - 0{,}065502\ \theta_z + 1{,}8312\text{E-}4\ \theta_z^2)/(1 + 0{,}01113\ \theta_z)$
F3	$kt < 1{,}4\ \&\ (90 - \theta_z) < 10^\circ$

Los filtros aplicados pueden describirse de la siguiente manera:

- **F1**: Rechaza valores que superan un límite físicamente razonable en función de la posición solar.
- **F2**: Descarta mediciones utilizando un umbral empírico dependiente del ángulo cenital.
- **F3**: Elimina valores del índice de claridad superiores a 1.4 cuando el Sol se encuentra a menos de 10° sobre el horizonte.

El porcentaje de datos diurnos retenidos varió entre estaciones. En particular, el 73 % de los registros de Yu cumplieron los criterios establecidos, frente al 82 % en Sa, 72 % en Sca, aproximadamente 84 % en Ero y 69 % en Lq.

3.1.2. Métricas de desempeño

Los indicadores de desempeño más comunes en el campo de la evaluación del recurso solar han sido abordados por Zhang et al. (2015); estos incluyen el Error Medio de Sesgo (MBE), el Error Medio Absoluto (MAE) y el Error Cuadrático Medio (RMSE). Las tres métricas se definen de la siguiente manera:

$$\mathrm{MBE} = \frac{\sum_{i=1}^{n}(y_i - x_i)}{n}, \tag{3.1}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}, \tag{3.2}$$

Tabla 3.3: Métricas de desempeño (MBE, MAE, RMSE) para cada modelo y conjunto de datos satelitales en los cinco sitios. Los valores están normalizados y expresados como porcentajes relativos al promedio de GHI en cada sitio: 396.8 W/m^2 (Yu), 397 W/m^2 (Sa), 557.1 W/m^2 (Sca), 690.6 W/m^2 (Ero) y 673.7 W/m^2 (Lq).

	YU			SA			SCA			ERO			LQ		
Modelo	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE
Resolución Temporal: 15 minutos															
CAMS	0.5	18.4	28.4	3.5	23.9	33.2	2.7	23	30.4	-23.7	27.8	41.2	-7.3	16.2	25.3
LSA-SAF	10.7	19	28.8	17.3	26.9	38.8	11.5	22.2	30.6	-8.1	16.5	26.8	3.7	12.3	22.3
Resolución Temporal: horaria															
CAMS	0.5	16	24.1	3.6	20.5	28.8	2.9	21	27.3	-23.7	26.8	39.5	-6.1	14.6	22
LSA-SAF	10.7	16.9	24.9	17.3	24.8	35	11.6	20.5	27.1	-8.1	15	24.3	4.6	10.9	18.7
ERA-5	-4.2	45.4	61.9	8.5	26.8	37.5	7.4	21.2	28.9	-13.7	19.1	25.3	-1.7	12	19.3
MERRA-2	26.9	35	51.9	42.1	47	63.6	12.7	21.9	29.3	-3.1	13.1	20.5	1.0	13.4	21.1

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - x_i\Big)^2}, \tag{3.3}$$

donde x y y son los valores medidos y estimados, respectivamente, y n es el tamaño de la muestra. El MBE mide el sesgo sistemático que un modelo puede introducir en una evaluación a largo plazo, mientras que el MAE y el RMSE miden la dispersión del error utilizando normas absolutas y cuadráticas, respectivamente. Debido a su mayor sensibilidad a los valores atípicos, el RMSE se utiliza frecuentemente en esta área. Ambas métricas de dispersión se reportan aquí por completitud. Los tres indicadores se presentan en términos relativos como un porcentaje del promedio de los valores medidos, denominados aquí como MBE (%), MAE (%) y RMSE (%).

3.2. Desempeño de los modelos de GHI en el NOA

Previo a la presentación del desempeño de los distintos procesos de adaptación al sitio se evaluaron los modelos de estimación de GHI disponibles en la región. Siendo este uno de los aportes que se pretende en este trabajo. A continuación se muestran las métricas de desempeño de los modelos calculadas sobre el conjunto de datos de de cada sitio.

3.2.1. Análisis de las estimaciones 15-minutales

La Figura 3.2 muestra la evaluación comparativa de los modelos CAMS y LSA-SAF en los sitios de estudio. En el caso del MBE, se observa que LSA-SAF presenta sesgos positivos en la mayoría de los sitios, indicando una tendencia a la sobreestimación sistemática de las variables

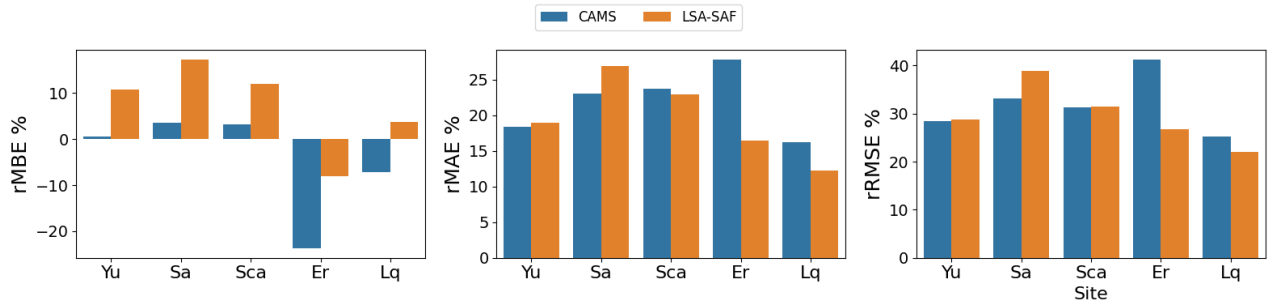


Figura 3.2: Comparación del desempeño de los modelos CAMS y LSA-SAF en cinco sitios de estudio (Yu, Sa, Sca, Er y Lq) mediante las tres métricas estadísticas: Mean Bias Error (MBE), Mean Absolute Error (MAE) y Root Mean Square Error (RMSE) expresadas en términos relativos a escala 15-minutal.

simuladas, mientras que CAMS exhibe valores más cercanos a cero o incluso negativos, lo que refleja un comportamiento más balanceado, aunque con subestimaciones marcadas en sitios como Er y Lq.

En relación al MAE, ambos modelos presentan magnitudes relativamente similares, aunque LSA-SAF tiende a mostrar errores absolutos ligeramente mayores en sitios como Yu y Sa, mientras que CAMS presenta valores más altos en Sca y Er. Esto sugiere que ninguno de los dos modelos logra una reducción clara y consistente del error en todos los sitios.

Por último, en la métrica RMSE, que penaliza los errores grandes, se mantiene un patrón semejante: LSA-SAF suele exhibir errores algo superiores a los de CAMS en Yu y Sa, mientras que en Er y Lq la diferencia favorece al modelo satelital. En general, los resultados muestran que el desempeño relativo de los modelos depende fuertemente del sitio, sin que exista un claro ganador en todos los casos.

La Figura 3.3 muestra la variación estacional del rRMSE (%) para los productos CAMS y LSA-SAF con resolución de 15 minutos en todos los sitios de estudio. Se observa una clara tendencia estacional en el comportamiento del error: en general, el rRMSE tiende a aumentar durante el verano, lo que indica que la precisión de las estimaciones disminuye en este período en la mayoría de los sitios. La excepción es Yuto, donde el error permanece relativamente estable a lo largo de las estaciones. Este patrón sugiere que durante el verano hay una mayor presencia de nubosidad, lo cual podría estar afectando la capacidad de los modelos satelitales para estimar correctamente la radiación solar, mientras que en otras estaciones la cobertura nubosa es menor, permitiendo estimaciones más precisas.

En el caso de RRMSD vs Kt (Figura 3.4), se observa una marcada disminución del error a medida que aumenta la claridad atmosférica. Para condiciones de cielo más nuboso (Kt $<$0.3), los valores de RRMSD son elevados en todos los sitios, superando en algunos casos los 200 %, especialmente en la estimación proveniente de LSASAF. Sin embargo, conforme Kt aumenta

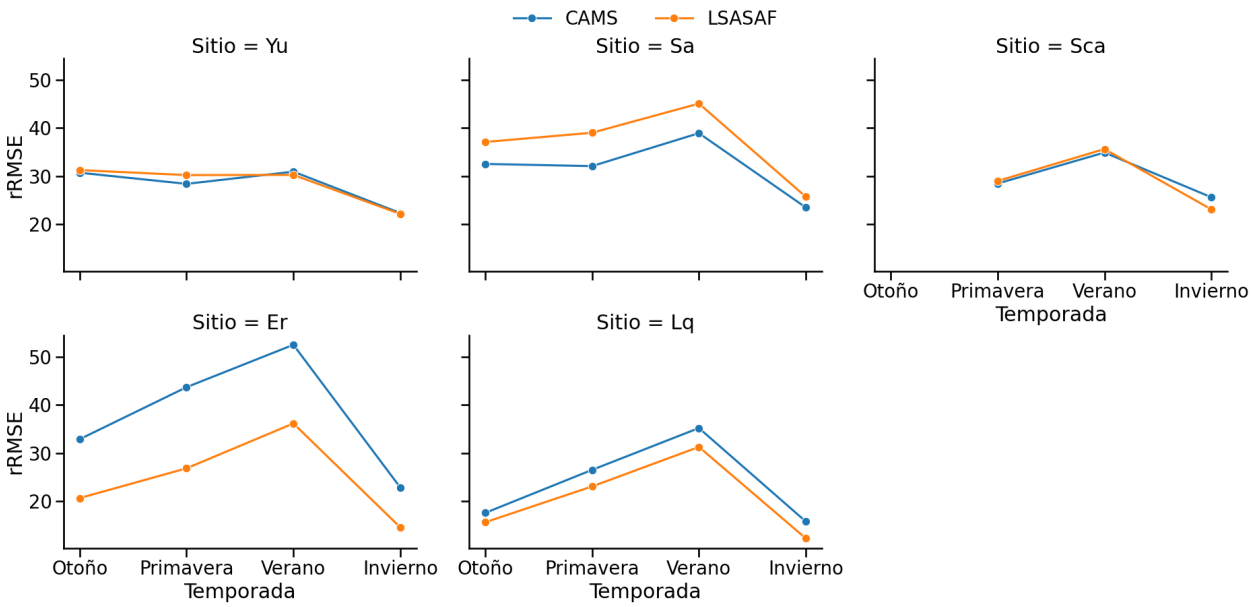


Figura 3.3: Figura 3.3. Variación estacional del rRMSE (%) para los productos CAMS y LSA-SAF con resolución de 15 minutos en todos los sitios de estudio. Se evidencia un aumento del error durante el verano en la mayoría de los sitios, mientras que en Yuto el rRMSE se mantiene estable, sugiriendo que la presencia de nubosidad estacional afecta de manera diferenciada la precisión de las estimaciones de radiación solar.

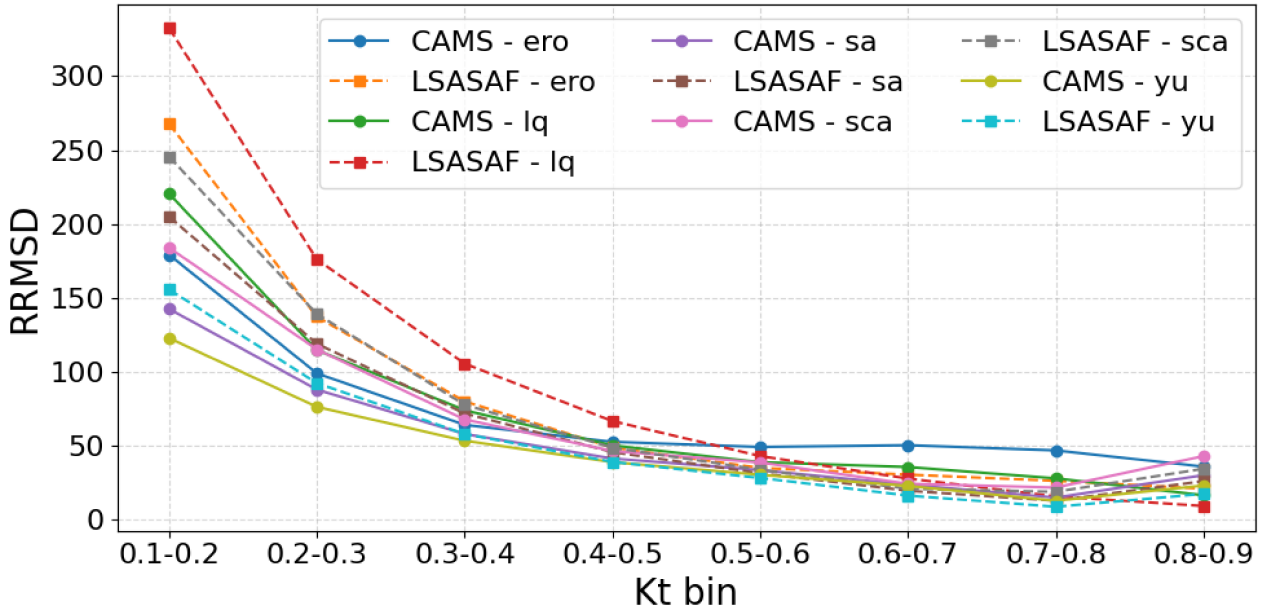


Figura 3.4: Variación del error cuadrático medio relativo (RRMSD) en función del indice de claridad (kt) para los cinco sitios analizados (YU, SA, SCA, ERO y LQ). Resultados para CAMS (líneas continuas) y LSASAF (líneas punteadas).

(>0.5), el error desciende rápidamente y tiende a estabilizarse por debajo del 50 %, alcanzando valores mínimos en condiciones de cielo despejado (Kt >0.7). Esta tendencia se mantiene consistente en ambos productos (CAMS y LSASAF), aunque LSASAF presenta errores relativamente mayores en las condiciones más turbias.

Por otro lado, la relación RRMSD vs SZA (Figura 3.5) evidencia un comportamiento en forma de "U" invertida: los menores errores se concentran en ángulos intermedios (30°–50°), mientras que hacia ángulos bajos (<20°) y altos (>70°) el error aumenta de manera significativa. Esta tendencia se observa en los cinco sitios, con un incremento más pronunciado en ERO y SA en las condiciones de SZA más extremas. En general, CAMS muestra un mejor desempeño que LSASAF en la mayoría de los intervalos, aunque las diferencias se reducen en los ángulos medios.

En conjunto, estos resultados indican que la precisión de ambos productos depende fuertemente de las condiciones atmosféricas (Kt) y de la geometría solar (SZA). En cielos despejados y ángulos intermedios, los errores se reducen notablemente, mientras que en condiciones nubosas y en situaciones de baja o alta elevación solar los modelos presentan las mayores limitaciones.

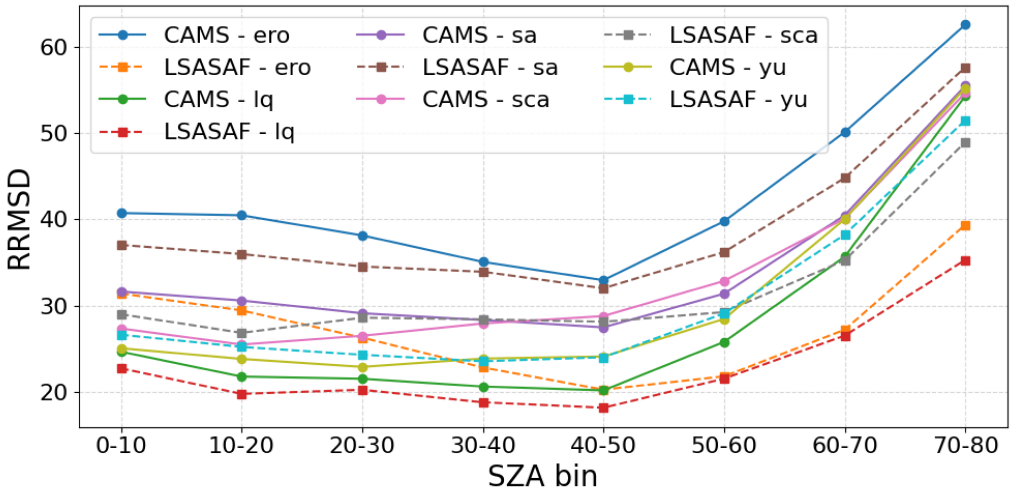


Figura 3.5: Variación del error cuadrático medio relativo (RRMSD) en función del ángulo cenital solar (SZA) para los cinco sitios analizados (YU, SA, SCA, ERO y LQ). Resultados para CAMS (líneas continuas) y LSASAF (líneas punteadas).

3.2.2. Análisis de las estimaciones horarias

3.2.3. División del conjunto de datos

De acuerdo a lo indicado en la sección 2.4 el conjunto de datos debe ser segmentado con fin de evaluar el rendimiento de los modelos de aprendizaje automático utilizados en la regresión del proceso de adaptación al sitio y controlar el sobre-entrenamiento de los modelos. En el contexto de AS en los trabajos Polo et al. (2016, 2020) se recomienda tomar al menos un año de medidas para la calibración de los modelos. Siguiendo estas ideas el conjunto de medidas fue dividido en subconjunto de entrenamiento, validación y prueba.

ID	Entrenamiento	Validación	Prueba
YU	2017	2017	2018
SA	2009	2009	2010 - 2024
SCA	2013	2013	2012-2014
ERO	2013	2013	2014 - 2024
LQ	2020	2020	2021 - 2024

Tabla 3.4: División del conjunto de datos en entrenamiento, validación y prueba

Tabla 3.5: Métricas de desempeño (MBE, MAE, RMSE) para cada modelo y conjunto de datos satelitales en los cinco sitios en el **conjunto de pruebas**. Los valores están normalizados y expresados como porcentajes relativos al promedio de GHI en cada sitio: 396.8 W/m^2 (Yu), 397 W/m^2 (Sa), 557.1 W/m^2 (Sca), 690.6 W/m^2 (Ero) y 673.7 W/m^2 (Lq).

	YU			SA			SCA			ERO			LQ		
Modelo	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE
Resolución Temporal: 15 minutos															
CAMS	-0.2	17.4	27.2	3.9	23.4	33.7	2.6	21.8	29.8	-23.6	27.6	40.9	-6.9	16.5	25.8
LSA-SAF	7.6	16.5	25.5	17.9	27.4	39.4	13.1	22.2	30.9	-7.7	16.2	26.5	4.0	12.6	22.6
Resolución Temporal: horaria															
CAMS	-0.2	15.3	23.5	4.0	20.9	29.3	3.0	19.7	26.1	-23.6	26.6	39.3	-4.8	14.5	21.6
LSA-SAF	7.6	14.5	21.9	17.9	25.3	35.7	13.3	20.3	27.1	-7.7	14.8	24.0	4.8	10.9	18.2
ERA-5	-4.0	43.5	60.2	9.4	27.1	37.7	1.9	20.2	29.7	-14.0	19.3	25.6	-0.8	11.3	18.4
MERRA-2	25.0	33.6	51.2	43.4	48.3	65.0	10.9	21.7	30.3	-3.8	13.1	20.4	1.4	13.4	20.8

3.3. Adaptación al sitio con una variable descriptiva

En los trabajos citados en las secciones anteriores sobre la evaluación del proceso de Adaptación al Sitio (AS), se ha prestado escasa atención a las razones subyacentes por las cuales ciertos modelos de ML superan a otros en dicho proceso. Considerando que algunos modelos de ML presentan una naturaleza inherentemente más compleja que otros, podría pensarse que modelos más complejos ofrecerían un mejor rendimiento en comparación con aquellos de menor complejidad. En este contexto, el término complejidad hace referencia tanto a la complejidad computacional —es decir, la cantidad de operaciones necesarias para ejecutar el modelo— como a la complejidad conceptual asociada al nivel de conocimiento requerido para comprender su funcionamiento.

En esta sección, se presenta una implementación del proceso de AS basada en uno de los enfoques clásicos, el cual consiste en adaptar una serie temporal modelada (proveniente de datos satelitales o de reanálisis) utilizando como referencia una serie de mediciones in situ.

Dado que el modelo de Regresión Lineal Simple (RLS) puede considerarse el menos complejo entre los modelos empleados en este estudio, la primera evaluación del proceso de SA en la región de interés se realiza utilizando dicho modelo. El objetivo es comparar el desempeño de la RLS con el de otros modelos de regresión más complejos, tales como Perceptrón Multicapa (MLP) y XGBoost, utilizando las mismas métricas de evaluación.

Para ambos modelos, MLP y XGBoost, determinar la configuración óptima de hiperparámetros fue esencial para lograr un desempeño consistente a través de las particiones de validación cruzada. Esto se llevó a cabo mediante una búsqueda exhaustiva en rejilla implementada con

la función **GridSearchCV** de la biblioteca **Scikit-learn** en Python Pedregosa et al. (2012). Los rangos específicos de hiperparámetros evaluados para los modelos MLP y XGB se describen en la Tabla 3.6. La selección de hiperparámetros óptimos es un paso crítico en el aprendizaje automático, ya que influyen directamente en la complejidad del modelo, su capacidad de generalización y su rendimiento predictivo global Goodfellow et al. (2016).

Tabla 3.6: Espacio cartesiano de hiperparámetros para las técnicas de aprendizaje supervisado.

Hiperparámetro	**Inferior**	**Superior**	**Paso**	**Función de transformación**
MLP				
Capas ocultas	1	3	1	-
Nodos ocultos	1	4	1	2^x
Fracción de dropout	0	0.3	0.1	-
Tasa de aprendizaje	-3	-1	1	10^x
XGBoost				
Booster	gbtree			
Estimadores	1	50	10	-
Profundidad máxima	2	5	1	2^x
Tasa de aprendizaje	-3	-1	1	10^x

En cuanto al preprocesamiento de los datos de entrada, la normalización o el escalado de características es una práctica común en aprendizaje automático para mejorar la convergencia y la estabilidad, particularmente en modelos sensibles a la magnitud de las variables de entrada. Sin embargo, en este estudio **no se aplicó normalización**, ya que no se consideró necesaria para los modelos y datos utilizados.

En el caso del modelo **SLR**, el escalado de la variable independiente no es necesario porque el modelo es inherentemente invariante al escalado: multiplicar la entrada por un factor constante provoca un cambio inversamente proporcional en el coeficiente de la pendiente, dejando las predicciones sin alterar.

De manera similar, **XGB**, cuando se implementa con árboles de decisión como aprendices base, **no requiere escalado de características**, ya que las divisiones en los nodos dependen del **orden relativo** de los valores y no de sus magnitudes absolutas Soria Olivas et al. (2022); Chen and Guestrin (2016a).

En el caso del modelo **MLP**, el escalado puede ser importante cuando las variables de entrada difieren significativamente en su rango o unidades. Sin embargo, en el presente estudio

Tabla 3.7: Métricas de desempeño (MBE, MAE, RMSE) para cada modelo adaptado en los cinco sitios en el **conjunto de pruebas**

	YU			SA			SCA			ERO			LQ		
Modelo	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE	MBE	MAE	RMSE
Resolución Temporal: 15 minutos															
CAMS SLR	-0.9	17.1	26.4	3.8	21.3	31.5	-1.5	18.4	26.0	2.5	24.8	31.5	2.1	15.9	23.5
CAMS MLP	-4.4	17.7	26.2	4.8	21.4	31.6	7.7	17.8	27.7	0.5	24.1	31.2	9.1	19.4	25.6
CAMS XGB	-1.3	17.1	26.0	3.8	21.4	31.4	-1.8	18.9	26.2	2.5	23.9	30.9	2.2	15.9	23.5
LSASAF SLR	-5.5	18.4	25.0	4.4	23.9	34.6	1.0	18.0	26.7	2.7	17.1	25.4	2.2	12.9	22.3
LSASAF MLP	-7.1	18.6	25.0	4.5	23.6	34.5	5.9	18.4	26.6	2.5	17.0	25.3	-0.4	13.6	22.1
LSASAF XGB	-6.0	18.2	24.9	4.3	23.9	34.6	0.5	18.6	27.0	2.4	17.2	25.1	2.3	13.1	22.3
Resolución Temporal: horaria															
CAMS SLR	-1.3	14.7	22.7	3.5	18.5	27.0	-1.7	15.9	22.0	2.1	23.4	29.6	2.6	13.8	19.4
LSASAF SLR	-5.5	16.0	21.2	4.3	21.2	30.5	1.1	15.6	22.7	2.5	15.7	22.8	0.3	10.6	17.4
ERA-5 SLR	1.2	44.0	55.1	6.4	26.5	36.9	-6.4	20.1	29.1	-0.6	15.4	21.4	2.1	11.6	18.4
MERRA-2 SLR	-3.6	34.7	44.1	7.1	35.2	46.0	-2.2	20.3	27.7	0.7	12.9	20.1	0.3	13.1	20.4
CAMS MLP	-4.9	16.2	23.0	4.3	18.6	27.2	-4.0	16.6	22.4	3.8	23.5	29.5	-0.5	13.0	19.2
LSASAF MLP	-5.4	15.4	20.9	2.5	21.4	30.2	8.3	16.1	24.1	9.6	19.4	24.6	-3.8	12.1	17.8
ERA-5 MLP	0.2	43.7	55.1	9.6	27.1	37.6	-1.8	19.0	28.5	-7.5	16.7	23.1	0.1	11.3	18.3
MERRA-2 MLP	0.7	33.7	43.9	0.8	36.3	45.4	-0.3	19.5	27.5	11.7	17.5	23.4	-2.8	13.4	20.6
CAMS XGB	-1.6	14.8	22.2	3.4	18.9	27.1	-2.4	16.6	22.4	2.1	22.8	29.2	2.7	13.9	19.6
LSASAF XGB	-6.1	16.0	21.3	4.1	21.4	30.6	0.0	16.9	23.4	2.2	15.9	22.6	0.2	10.9	17.4
ERA-5 XGB	1.4	44.6	55.4	6.3	27.3	37.4	-7.3	20.7	29.3	-0.5	15.2	21.2	2.0	11.8	18.6
MERRA-2 XGB	-4.9	35.3	44.5	6.9	35.6	46.1	-2.8	20.8	28.2	0.8	13.3	20.3	0.2	13.5	20.6

se utilizó **una sola variable de entrada** (GHI derivado de satélite), expresada en las **mismas unidades y rango** que la variable objetivo (GHI medido en superficie). Por lo tanto, no se consideró necesaria ninguna normalización adicional.

Los resultados obtenidos son expresados en la Tabla 3.3.

La evaluación comparativa de los modelos SLR,MLP y XGB, utilizando como variables de entrada los productos satelitales *CAMS* y *LSA-SAF*, se realizó en cinco sitios con características climáticas y geográficas diversas. El desempeño se expresó en términos de error medio de sesgo (MBE), error absoluto medio (MAE) y raíz del error cuadrático medio (RMSE), todos normalizados respecto al promedio de la irradiancia global horizontal (GHI) en cada estación (Tablas 3.3 y 3.2.3).

En términos generales, los errores normalizados se mantuvieron dentro de un rango moderado en todas las combinaciones de modelos y conjuntos de datos, sin observarse diferencias sustanciales entre los enfoques lineales y los no lineales. El modelo RLS mostró un desempeño competitivo en ambos conjuntos de datos, con métricas de error cercanas a las obtenidas por los modelos más complejos.

En el caso de **CAMS**, aunque el MLP logró reducir ligeramente el RMSE en algunos sitios, esto se produjo a costa de sesgos más pronunciados en el MBE, lo cual evidencia una compensación entre reducción de varianza e incremento del error sistemático. El modelo XGB, por su parte, presentó un comportamiento muy similar al de RLS, con diferencias marginales.

Por otro lado, al emplear **LSA-SAF** se observó una ligera mejora respecto a CAMS, particularmente en estaciones de mayor altitud (ERO y LQ), donde tanto XGB como MLP alcanzaron menores valores de MAE y RMSE. Esto sugiere que la mayor resolución temporal o la representación más detallada de nubes en LSA-SAF aportan información adicional útil para el proceso de adaptación local.

No obstante, el incremento de la complejidad del modelo no se tradujo en ganancias sustanciales de desempeño. Estos resultados indican que, dadas las condiciones actuales de calidad de los datos satelitales, los modelos simples como RLS son capaces de capturar gran parte de la relación entre los insumos derivados de satélite y las mediciones de GHI en superficie, ofreciendo además mayor robustez frente al ruido y a las limitaciones en los datos de entrenamiento.

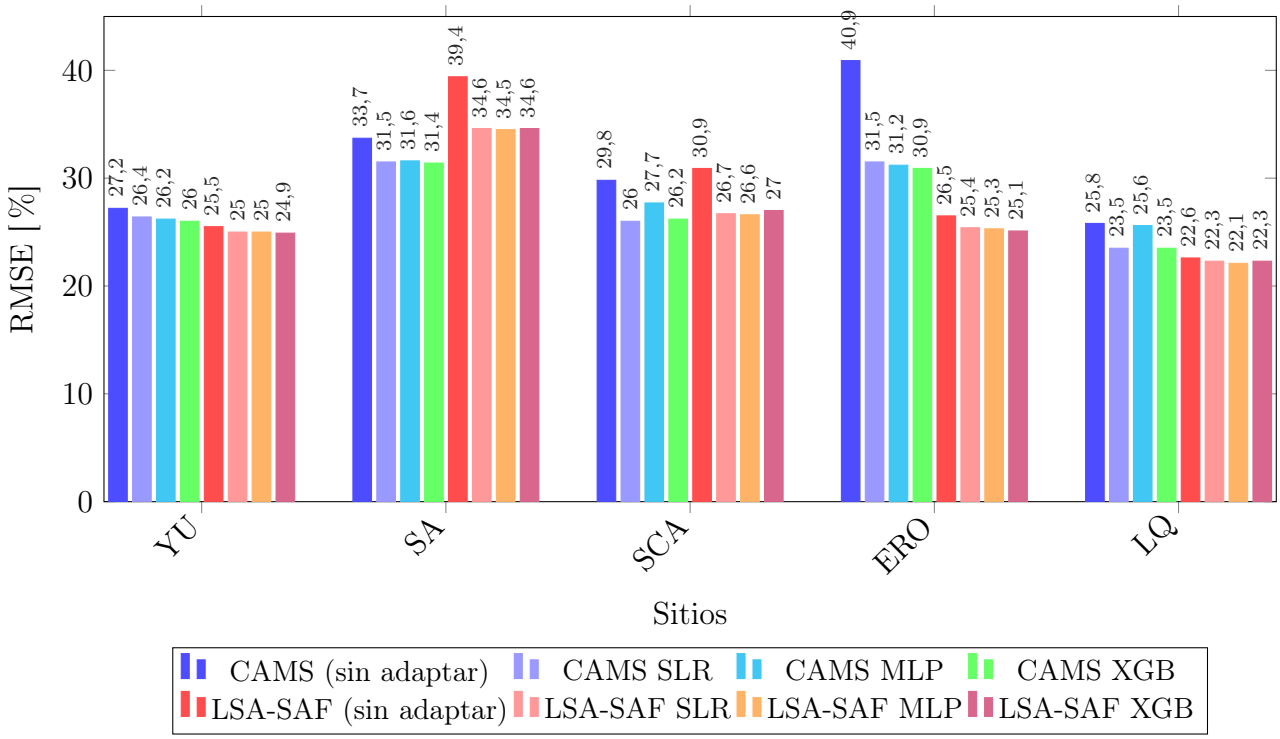


Figura 3.6: RMSE en resolución de 15 minutos para cada modelo y sitio, comparando modelos sin adaptación y adaptados.

La Figura 3.6 muestra el comportamiento del RMSE (%) en cada sitio para los diferentes modelos de regresión utilizados. Se observa que todas las propuestas de adaptación logran reducir el RMSE en cada sitio. Aunque en algunos casos la mejora puede ser modesta, como en YU y

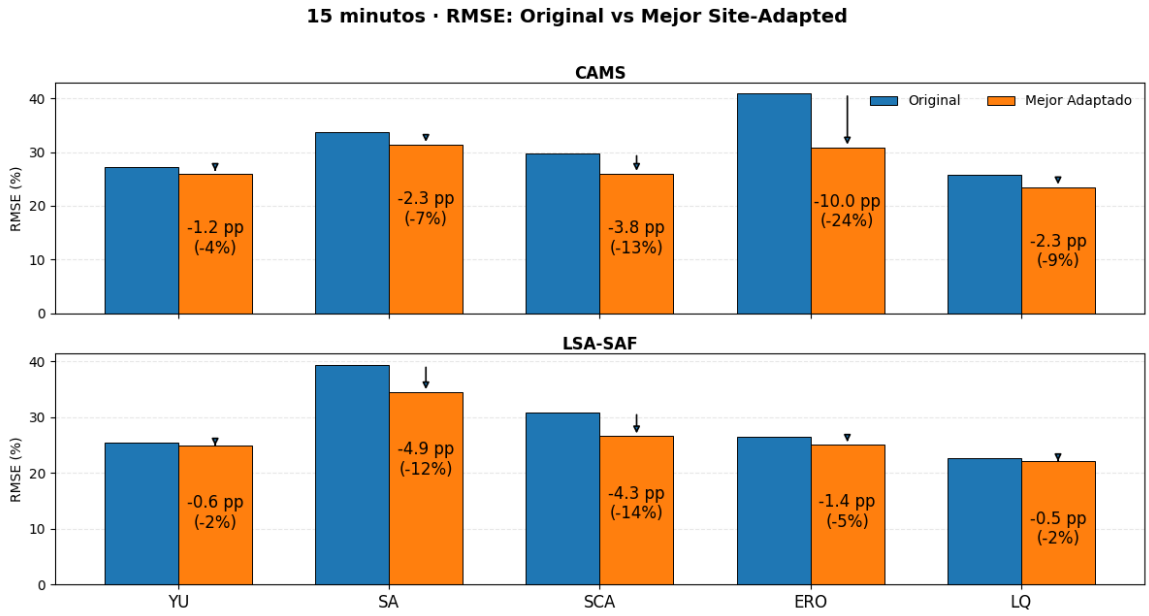


Figura 3.7: Variación del error cuadrático medio relativo (RRMSD) en función del ángulo cenital solar (SZA) para los cinco sitios analizados (YU, SA, SCA, ERO y LQ). Resultados para CAMS (líneas continuas) y LSASAF (líneas punteadas).

LQ, queda evidenciado que un simple ajuste específico puede mejorar el desempeño del modelo para una ubicación determinada.

Además, los resultados de la adaptación dependen del modelo de estimación empleado. Cada modelo impone un límite sobre la precisión de la serie resultante, y optimizar su salida no garantiza necesariamente un desempeño superior frente a otro modelo que, de forma natural, ya proporciona una estimación más adecuada para un sitio específico.

Los hallazgos de este primer estudio indican que, en el contexto de la adaptación del sitio para GHI utilizando productos derivados de satélite como CAMS y LSA-SAF, los modelos de aprendizaje automático evaluados SLR, XGB y MLP mostraron un rendimiento comparable en términos de métricas de error estándar, con diferencias mínimas entre ellos. A pesar de su simplicidad, SLR demostró una notable capacidad para capturar la relación entre los datos derivados de satélite y las mediciones terrestres, logrando niveles de error relativo similares a los obtenidos por los modelos XGB y MLP, más complejos. Estos resultados sugieren que la calidad de los datos de entrada impone un límite superior a las mejoras de precisión alcanzables que ofrecen los modelos complejos. Los productos satelitales empleados presentan incertidumbres inherentes y carecen de la resolución necesaria para considerar fenómenos localizados de alta frecuencia, como los efectos microclimáticos o la variabilidad inducida por el terreno. En consecuencia, aumentar la complejidad del modelo no mejora sustancialmente el rendimiento predictivo en estas condiciones de datos. En este sentido, la capacidad predictiva de los modelos complejos

se vuelve redundante cuando la relación subyacente entre las variables es predominantemente lineal o solo ligeramente no lineal, lo que explica el excelente rendimiento del SLR. Además, el análisis reveló que los modelos más flexibles, como el MLP, son propensos al sobreajuste, lo que genera un comportamiento inconsistente en diferentes métricas de rendimiento. Esto subraya la importancia de equilibrar la complejidad del modelo con la calidad de los datos para evitar comprometer la capacidad de generalización del modelo. En resumen, dada la calidad actual de los datos y las características del sitio, modelos simples como el SLR representan una solución robusta, interpretable y computacionalmente eficiente para la estimación del GHI adaptada al sitio. Es más probable lograr mejoras significativas en la precisión de la estimación incorporando variables meteorológicas adicionales, aplicando técnicas de descomposición temporal o desarrollando enfoques híbridos que combinen modelos físicos con correcciones estadísticas, en lugar de simplemente aumentar la complejidad de los algoritmos de aprendizaje automático.

3.4. Adaptación al sitio con múltiples variables descriptivas

Los modelos de aprendizaje automático se emplean cada vez más para identificar patrones y extraer información significativa de grandes volúmenes de datos. No obstante, la efectividad de estos modelos depende en gran medida de la calidad de las características utilizadas durante el entrenamiento. La **selección de características**, un paso crucial en el preprocesamiento de datos, consiste en identificar las variables más relevantes y eliminar aquellas que sean redundantes o irrelevantes Liu et al. (2023); Huang et al. (2024). Este proceso no solo incrementa la interpretabilidad del modelo, sino que también mejora su eficiencia computacional y su capacidad predictiva. Un exceso de características o la inclusión de variables irrelevantes puede causar **sobreajuste**, donde el modelo presenta un buen desempeño con los datos de entrenamiento pero falla al generalizar a datos no vistos Che et al. (2024b). La selección de características ayuda a mitigar este problema al reducir la dimensionalidad, fomentando la creación de modelos más robustos y generalizables. Además, contribuye a disminuir los costos computacionales y el tiempo requerido para entrenar el modelo, consolidándose como una herramienta esencial tanto para investigadores como para profesionales Cheng (2024).

Las técnicas de selección de características se agrupan en tres categorías principales: **métodos de filtro, métodos envolventes y métodos embebidos**, cada uno con su propia metodología, ventajas y limitaciones:

Métodos de filtro. Los métodos de filtro evalúan la relevancia de cada característica mediante criterios estadísticos como correlación, información mutua o varianza. Son computacionalmente eficientes y no dependen de un algoritmo de aprendizaje específico. Sin embargo, pueden no captar las interacciones entre variables. Ejemplos comunes incluyen la correlación de Pearson, las pruebas chi-cuadrado y la ganancia de información.

Métodos envolventes. Estos métodos consisten en entrenar y evaluar un modelo de aprendizaje automático múltiples veces para identificar el subconjunto óptimo de características. Técnicas como la selección hacia adelante, la eliminación hacia atrás y la eliminación recursiva de características (RFE) son ejemplos típicos Che et al. (2024a); Liu et al. (2024). Aunque suelen proporcionar mayor precisión, son costosos en términos computacionales y pueden no escalar eficientemente con conjuntos de datos grandes.

Métodos embebidos. Los métodos embebidos incorporan la selección de características directamente en el proceso de entrenamiento del modelo. Ejemplos incluyen técnicas de regularización como LASSO (regularización L1) y modelos basados en árboles de decisión. Este enfoque logra un equilibrio entre eficiencia y rendimiento, convirtiéndolo en una opción ampliamente utilizada en distintas aplicaciones.

Comprender estas técnicas permite a los profesionales seleccionar el método más adecuado según las características del conjunto de datos, el dominio del problema y las limitaciones computacionales.

En esta sección se documentan los resultados obtenidos al seleccionar tres métodos complementarios para la selección de variables con el objetivo de identificar los mejores regresores que contribuyan a la mejora de la estimación de la GHI en el contexto de la adaptación al sitio: RFE, LASSO y Stepwise.

Eliminación recursiva de características (RFE). RFE permite evaluar de manera iterativa la importancia de cada variable en el modelo, eliminando progresivamente las menos relevantes. Este enfoque es particularmente útil en el análisis de variables de geometría solar, donde pueden existir correlaciones complejas entre diferentes parámetros. Al utilizar RFE, se asegura que las variables seleccionadas aporten información significativa al modelo y reduzcan el riesgo de sobreajuste.

LASSO (Least Absolute Shrinkage and Selection Operator). LASSO integra la selección de variables en el propio proceso de entrenamiento mediante regularización L1. Este método penaliza los coeficientes de variables menos relevantes, promoviendo modelos más simples y generalizables. La utilización de LASSO en nuestro estudio permite manejar de manera eficiente la multicolinealidad entre parámetros solares y resaltar únicamente los regresores que contribuyen de manera significativa a la predicción de la GHI.

Stepwise (selección hacia adelante y hacia atrás). Los métodos Stepwise combinan criterios estadísticos de inclusión y exclusión de variables, permitiendo construir un modelo óptimo de manera secuencial. Este enfoque es especialmente adecuado cuando se busca un balance entre interpretabilidad y rendimiento predictivo. En el contexto de variables solares,

Stepwise ayuda a identificar combinaciones de regresores que optimizan la estimación de la GHI sin introducir redundancias innecesarias.

La elección de los tres métodos de selección de variables se justifica por su complementariedad: RFE se centra en la importancia iterativa de cada predictor, LASSO introduce regularización para reducir la complejidad y Stepwise optimiza la construcción del modelo desde un enfoque estadístico. La combinación de estas técnicas permite una selección robusta de variables, maximizando la eficiencia del modelo y mejorando la precisión de la estimación de la irradiancia global horizontal (GHI) en el posprocesamiento.

En esta investigación se trabajó con un conjunto de variables de carácter astronómico, atmosférico, satelital y meteorológico, definidas de la siguiente manera:

- `N`: día del año.
- `delta`: declinación solar.
- `Fn`: factor de corrección orbital.
- `w`: ángulo horario.
- `SZA`: ángulo cenital solar.
- `alphaS`: altura solar.
- `E0`: ecuación del tiempo.
- `TOA`: irradiancia en el tope de la atmósfera.
- `GHIargp2`: estimación de la GHI en condiciones de cielo claro según el modelo Argpv2.
- `mr`: masa de aire óptica relativa.
- `kt`: índice de claridad.
- `tm`: temperatura del aire a 2 m (ERA5).
- `uw`: componente zonal del viento (ERA5).
- `vw`: componente meridional del viento (ERA5).
- `cams`: estimación satelital de la GHI a partir del modelo CAMS.
- `lsasaf`: estimación satelital de la GHI a partir del modelo LSA-SAF.

Estas magnitudes fueron seleccionadas bajo la premisa de que son relativamente fáciles de calcular o de obtener a partir de modelos satelitales y de reanálisis, y representan parámetros clave de la geometría solar, la atmósfera y la dinámica meteorológica. La combinación de estas fuentes garantiza la aplicabilidad práctica de los modelos, al tiempo que permite capturar la complejidad del recurso solar.

El análisis comparativo de los métodos de selección de variables (RFE, LASSO y Stepwise) aplicado a distintos sitios (YU, SA, SCA, ERO y LQ) permitió evaluar el impacto de los predictores en el desempeño de los modelos de estimación de la GHI.

En términos generales, se observó que el método *Stepwise* ofreció el mejor rendimiento en la mayoría de los casos, presentando los valores más bajos de *rrmsd*. Por ejemplo, en ERO y LQ se obtuvieron errores relativos cercanos al 21 %, mientras que en YU y SCA se situaron en torno al 25–27 %. Esto confirma que la selección progresiva de predictores, incorporando únicamente aquellos que generan mejoras significativas en el ajuste, resulta más eficiente que estrategias exhaustivas (RFE) o con regularización estricta (LASSO).

Respecto a la comparación entre las fuentes satelitales, los resultados muestran que no existe un modelo universalmente superior, sino que la conveniencia de utilizar CAMS o LSA-SAF depende del sitio específico. En YU y ERO los modelos con LSA-SAF alcanzaron menor error, mientras que en SCA y LQ los mejores resultados se obtuvieron con CAMS. En SA, ambos conjuntos presentaron un desempeño inferior, con errores significativamente más altos (*rrmsd* entre 31 % y 34 %), lo que sugiere la influencia de condiciones locales complejas o limitaciones en la representatividad de los productos satelitales. En este sitio, sin embargo, las variables meteorológicas derivadas de ERA5 (temperatura y viento) fueron seleccionadas con mayor frecuencia, lo que sugiere que aportan información adicional relevante.

En cuanto a la composición de los conjuntos de predictores, se constató que la variable satelital correspondiente (`cams` o `lsasaf`) fue incluida de manera consistente en todos los modelos, confirmando su papel central. Asimismo, variables geométricas como el ángulo cenital solar (`SZA`), el ángulo horario (`w`), el día del año (`N`) y el índice de claridad (`kt`) aparecieron recurrentemente en las configuraciones con mejor desempeño. Otros predictores astronómicos y atmosféricos, como `TOA`, `GHIargp2`, la temperatura (`tm`) y las componentes del viento (`uw`, `vw`), fueron seleccionados en varios sitios, reforzando la importancia de integrar tanto la geometría solar como las condiciones meteorológicas en el ajuste. Variables como `Fn`, `delta`, `alphaS`, `E0` y `mr` aparecieron con menor frecuencia, sugiriendo que su aporte es más dependiente de condiciones locales particulares.

En conjunto, los resultados permiten concluir que:

- El método Stepwise constituye la estrategia más recomendable para la selección de variables, al balancear complejidad y precisión.

- La fuente satelital óptima depende del sitio de estudio: LSA-SAF fue más favorable en YU y ERO, mientras que CAMS resultó superior en SCA y LQ.
- Las variables astronómicas y atmosféricas son complementarias a los predictores satelitales, y su integración mejora de forma significativa la capacidad explicativa de los modelos.
- La incorporación de variables meteorológicas de ERA5 (`tm`, `uw`, `vw`) aporta beneficios adicionales, especialmente en sitios donde los modelos satelitales muestran limitaciones.
- Existen diferencias notables entre sitios: ERO y LQ son los más favorables ($rrmsd \approx 21\,\%$), YU y SCA presentan un desempeño intermedio (25–27 %), y SA muestra los peores resultados (>31 %).

Finalmente, el análisis de recurrencia de predictores revela un núcleo de variables esenciales que debería formar parte de cualquier configuración robusta:

$$\{N,\ w,\ TOA,\ kt,\ SZA,\ GHIargp2,\ \text{cams/lsasaf},\ tm,\ uw,\ vw\}$$

Este conjunto concentra los predictores más consistentes y con mayor fundamentación física, garantizando un buen compromiso entre precisión y aplicabilidad transversal. Según el sitio, se podrían añadir variables secundarias (`Fn`, `delta`, `alphaS`, `E0`, `mr`), cuyo aporte es más dependiente de condiciones locales específicas.

Es importante remarcar que los métodos de selección utilizados en este análisis (RFE, LASSO y Stepwise) no ordenan explícitamente las variables por importancia, como lo haría, por ejemplo, un algoritmo de tipo Random Forest que entrega un ranking de importancia. Lo que sí podemos inferir es la frecuencia y consistencia con la que cada predictor es seleccionado a lo largo de los sitios y métodos. En este sentido, variables como `N` (día del año), `w` (ángulo horario), `SZA` (ángulo cenital solar) y `kt` (índice de claridad) aparecen de forma reiterada y estable, confirmando su relevancia central. En contraste, predictores como la temperatura del aire (`tm`) o las componentes de viento (`uw`, `vw`) tienden a ser seleccionados en casos específicos, sugiriendo un aporte complementario más que universal.

Seguramente sería útil contar algún ranking de importancia que busque establecer cuál es el orden de importancia con el que se deberían escoger las variables regresoras. Como se mencionó en el párrafo anterior, el algoritmo RF permite definir este ranking. Si bien este modelo no ha mostrado ser superior en comparación al modelo XGBoost en el contexto de la adaptación al sitio, según lo reportado en Salamalikis et al. (2022), conocer el orden de importancia de las variables regresoras para nuestros sitios de estudio puede ser relevante al momento de optimizar costo computacional y tiempo en el entrenamiento de los modelos, en trabajos que busquen replicar los procesos de adaptacion al sitio en la región.

En este sentido hemos tomado RF como modelo de regresión para realizar la adaptación en cada uno de los sitios, utilizando una búsqueda en rejilla según se especifica en la Tabla T. Como resultado se ha generado un ranking de importancia de las variables regresoras que se muestra en la

3.5. Adaptación al sitio tomando consideraciones de una serie temporal

Una «serie temporal» es un conjunto de observaciones de una o más variables recolectadas y ordenadas cronológicamente. El orden temporal es esencial para su análisis e interpretación Olivas et al. (2022).

El análisis de series temporales es aplicable en múltiples disciplinas. En Economía, permite estudiar tendencias de precios y demanda; en Marketing, ayuda a comprender la evolución de ventas. En Medicina, las bioseñales como ECG, EEG o EOG son ejemplos de series temporales. También se aplican en la gestión hospitalaria, por ejemplo, para analizar la afluencia de pacientes a urgencias o la demanda de especialistas. Otro campo fundamental es la meteorología, de especial interés en este trabajo, donde se utilizan para caracterizar, clasificar y predecir variables como la GHI.

La «estacionalidad» de una serie temporal es una característica por la cual los datos experimentan cambios regulares y predecibles con una frecuencia constante. Esta frecuencia puede ser, por ejemplo, diaria, semanal o mensual. Cualquier fluctuación predecible de frecuencia constante que aparezca en una serie temporal se dice que es estacional.

En esta sección se ha evaluado el impacto en el proceso de AS la tener en cuenta la estacionalidad de la serie GHI. Se han comparado dos enfoques distintos respecto al enfoque tradicional.

3.6. Adaptación al sitio usando celdas satelitáles adyacentes al sitio de interés

Las distintas evaluaciones realizadas tanto en este trabajo como en los estudios previamente referenciados sobre adaptación al sitio (AS) se han basado exclusivamente en el uso de datos medidos en una única estación meteorológica. Estos datos se combinan con una o más variables modeladas —provenientes de una celda satelital o de un modelo de reanálisis— que contienen geográficamente a dicha estación. En otras palabras, la información utilizada para entrenar y validar los modelos de corrección se limita a la celda específica en la cual se encuentra el punto de medición, sin considerar el contexto espacial más amplio que podría aportar información adicional relevante.

Todos los enfoques previos de SA se han aplicado exclusivamente a mediciones terrestres tomadas en una única estación asociada a una única celda de la cuadrícula satelital. Como resultado, cualquier técnica de aprendizaje automático implementada bajo este marco inevitablemente encuentra una limitación de aprendizaje, no debido al sobreajuste, sino a que la

información disponible de los productos derivados de satélite es inherentemente finita. Para abordar esta restricción, proponemos un enfoque alternativo que amplía el alcance espacial de los datos de entrada mediante la incorporación de valores de GHI modelados de celdas satelitales adyacentes. Esta metodología se basa en la hipótesis de que la variabilidad de la irradiancia solar en una ubicación determinada no es aislada, sino que forma parte de una estructura espacialmente coherente gobernada por una dinámica atmosférica más amplia. Al integrar las estimaciones de irradiancia de celdas vecinas como predictores adicionales, el modelo accede a información espacial más rica, lo que mejora su capacidad para aprender patrones complejos y generalizar eficazmente.

Esta estrategia, basada en información espacial, aprovecha correlaciones bien documentadas en los campos de irradiancia solar, a menudo impulsadas por el movimiento de las nubes, el transporte de aerosoles y los sistemas meteorológicos a escala sinóptica (**?**). A diferencia de los métodos tradicionales de SA, que se basan únicamente en variables atmosféricas locales dentro de una sola celda, el marco propuesto introduce un novedoso uso de la irradiancia obtenida por satélite de las celdas de la cuadrícula circundante como características de entrada para los modelos de aprendizaje automático. Ningún trabajo previo ha explorado explícitamente esta dirección. Al integrar el contexto espacial directamente en el proceso de aprendizaje, este enfoque busca superar la saturación inherente del rendimiento observada en las técnicas clásicas de SA cuando se limitan al modelado aislado de una sola celda.

Capítulo 4

Conclusiones

Se escriben las conclusiones del trasasdabajo.

Bibliografía

Gonzalo Abal, Rodrigo Alonso-Suárez, and Agustín Laguarda. *Radiación Solar: Notas del curso Fundamentos del Recurso Solar*. Laboratorio de Energía Solar, Uruguay, versión 4.0 edition, junio 2020. URL http://les.edu.uy/.

R. Alonso Suárez, G. Abal, R. Siri, and P. Musé. Brightness-dependent tarpley model for global solar radiation estimation using goes satellite images: Application to uruguay. *Solar Energy*, 86(11):3205–3215, 2012. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2012.08.012. URL https://www.sciencedirect.com/science/article/pii/S0038092X12003040.

R. Aristegui and R. Righini. Discusión sobre el proceso de selección de sitios apropiados para la ubicación de estaciones de una futura red solarimétrica nacional. *GERSolar-INEDES, Departamento de Ciencias Básicas, Universidad Nacional de Luján*, 2012. Recibido: 31/07/12; Aceptado: 02/10/12, Tel. 02323-440241, e-mail: gersolar@yahoo.com.ar.

Bilal Babar, Luigi Tommaso Luppino, Tobias Boström, and Stian Normann Anfinsen. Random forest regression for improved mapping of solar irradiance at high latitudes. *Solar Energy*, 198:81–92, 2020. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2020.01.034. URL https://www.sciencedirect.com/science/article/pii/S0038092X20300426.

Silvina Belmonte, Virgilio Núñez, Judith Franco, and José Viramonte. Mapas de radiación solar para el valle de lerma (salta – argentina). *Avances en Energías Renovables y Medio Ambiente*, 10:1149–1156, 2006. ISSN 0329-5184.

L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984. ISBN 9780412048418. URL https://books.google.com.ar/books?id=JwQx-WOmSyQC.

Chang Che, Chen Li, and Zengyi Huang. The integration of generative artificial intelligence and computer vision in industrial robotic arms. *International Journal of Computer Science and Information Technology*, 2(3):1–9, May 2024a. doi: 10.62051/ijcsit.v2n3.01. URL https://wepub.org/index.php/IJCSIT/article/view/1820.

Chang Che, Qunwei Lin, Xinyu Zhao, Jiaxin Huang, and Liqiang Yu. Enhancing multimodal understanding with clip-based image-to-text transformation, 2024b. URL https://arxiv.org/abs/2401.06167.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016a. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016b. doi: 10.1145/2939672.2939785.

Xueyi Cheng. A comprehensive study of feature selection techniques in machine learning models. *Insights in Computer, Signals and Systems*, 1:65–78, 11 2024. doi: 10.70088/xpf2b276.

Oscar Claveria, Enric Monte, and Salvador Torra. Effects of removing the trend and the seasonal component on the forecasting performance of artificial neural network techniques. IREA Working Papers 201503, University of Barcelona, Research Institute of Applied Economics, January 2015. URL https://ideas.repec.org/p/ira/wpaper/201503.html.

Adele Cutler, D. Richard Cutler, and John R. Stevens. Random forests. In *Machine Learning*, pages 157–175. Springer, 2011.

XGBoost Developers. *XGBoost Documentation, Release 0.80*, 2018. URL https://xgboost.readthedocs.io/en/release_0.80/.

John A. Duffie and William A. Beckman. *Solar Engineering of Thermal Processes*. Wiley, Hoboken, NJ, 4th edition, 2013.

J. S. Ellis, T. H. Vonder Haar, S. Levitus, and A. H. Oort. The annual variation in the global heat balance of the earth. *Journal of Geophysical Research: Oceans*, 83(C4):1958–1962, 1978. doi: https://doi.org/10.1029/JC083iC04p01958. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC083iC04p01958.

Javier Jesús Espinosa-Zúñiga. Aplicación de algoritmos random forest y xgboost en una base de solicitudes de tarjetas de crédito. *Ingeniería Investigación y Tecnología*, 21(3):1–16, 2020. doi: 10.22201/fi.25940732e.2020.21.3.022.

Carlos M. Fernández-Peruchena, Jesús Polo, Luis Martín, and Luis Mazorra. Site-adaptation of modeled solar radiation data: The siteadapt procedure. *Remote Sensing*, 12(13), 2020. ISSN 2072-4292. doi: 10.3390/rs12132127. URL https://www.mdpi.com/2072-4292/12/13/2127.

S. Fritz, P. Krishna Rao, and M. Weinstein. Satellite measurements of reflected solar energy and the energy received at the ground. *Journal of Atmospheric Sciences*, 21(2):141 – 151, 1964. doi: 10.1175/1520-0469(1964)021<0141:SMORSE>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/21/2/1520-0469_1964_021_0141_smorse_2_0_co_2.xml.

Catherine Gautier, Georges Diak, and Serge Masse. A simple physical model to estimate incident solar radiation at the surface from goes satellite data. *Journal of Applied Meteorology and Climatology*, 19(8):1005 – 1012, 1980. doi: 10.1175/1520-0450(1980)019<1005:ASPMTE>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/19/8/1520-0450_1980_019_1005_aspmte_2_0_co_2.xml.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org. Book in preparation for MIT Press.

Hugo Grossi Gallegos. Distribucion de la radiacion solar global en argentina. i. analisis de la informacion. *Energías Renovables y Medio Ambiente*, 4:13–17, 01 1998a.

Hugo Grossi Gallegos. Distribución de la radiación solar global en la república argentina. ii. cartas de radiación. *Energías Renovables y Medio Ambiente*, 5:33–42, 01 1998b.

Hugo Grossi Gallegos and R. Righini. *Atlas de Energía Solar de la República Argentina*. 05 2007. ISBN 978-987-9285-36-7.

Hugo Grossi Gallegos, Alejandro Roberti, and Graciela Renzini. Evaluacion de las bases de datos de radiacion global disponibles en la republica argentina. *Avances en Energías Renovables y Medio Ambiente*, 3:11.13–11.17, 10 1999.

Chris Gueymard. Direct solar transmittance and irradiance predictions with broadband models. part i: Detailed theoretical performance assessment. *Solar Energy*, 74:355–379, 05 2003. doi: 10.1016/S0038-092X(03)00195-6.

Kirby J. Hanson. A new estimate of solar irradiance at the earth's surface on zonal and global scales. *Journal of Geophysical Research (1896-1977)*, 81(24):4435–4443, 1976. doi: https://doi.org/10.1029/JC081i024p04435. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC081i024p04435.

John E. Hay. Satellite based estimates of solar irradiance at the earth's surface—i. modelling approaches. *Renewable Energy*, 3(4):381–393, 1993. ISSN 0960-1481. doi: https://doi.org/10.1016/0960-1481(93)90105-P. URL https://www.sciencedirect.com/science/article/pii/096014819390105P. Solar radiation, environment and climate change.

Alejandro Hernández. Geosol: Una herramienta computacional para el cÁlculo de coordenadas solares y la estimaciÓn de irradiaciÓn solar horaria. *Avances en Energías Renovables y Medio Ambiente*, 7:19 – 24, 10 2003.

M. Hongn, R. Lozano, and G. Salazar. Simulación del funcionamiento de una planta solar térmica de gran escala situada en la localidad de san carlos, salta. *Avances en Energías Renovables y Medio Ambiente*, 22:02.119–02.129, 2018.

Zengyi Huang, Haotian Zheng, Chen Li, and Chang Che. Application of machine learning-based k-means clustering for financial fraud detection. *Academic Journal of Science and Technology*, 10(1):33–39, Mar. 2024. doi: 10.54097/74414c90. URL https://drpress.org/ojs/index.php/ajst/article/view/19142.

International Energy Agency Photovoltaic Power Systems Programme (IEA-PVPS). Worldwide benchmark of modelled solar radiation data. Technical report, IEA PVPS Task 16, Paris, France, 2023.

A. Laguarda. *Modelado de la irradiancia solar sobre la superficie terrestre: Modelos físicos e híbridos utilizando información satelital sobre la Pampa Húmeda*. Tesis de doctorado, Universidad de la República, Uruguay, 2025.

A. Laguarda, P. Iturbide, X. Orsi, M. J. Denegri, S. Luza, B. L. Burgos, V. Stern, and R. Alonso-Suárez. Validación de modelos satelitales heliosat-4 y cim-esra para la estimación de irradiancia solar en la pampa húmeda. *Energías Renovables y Medio Ambiente*, 48:1–9, may 2022. URL https://portalderevistas.unsa.edu.ar/index.php/erma/article/view/2877.

R. Laspiur, G. A. Salazar, J. Zerpa, and M. Watkins. 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. *Avances en Energías Renovables y Medio Ambiente*, 17:08.47–08.52, 2013.

R. Ledesma, G. Salazar, and O. Vilela. Avances en la estimación de irradiancia solar en las provincias de salta y jujuy mediante imágenes satelitales goes-16, 2023. Manuscrito no publicado.

Rubén Ledesma, Rodrigo Alonso-Suárez, Germán Salazar, Fernando Nollas, and Olga Vilela. Evaluation of satellite and reanalysis models for solar irradiance estimation in northwest argentina. *IEEE Latin America Transactions*, 23(8):706–717, 2025. doi: 10.1109/TLA.2025.11072498.

Bo Liu, Liqiang Yu, Chang Che, Qunwei Lin, Hao Hu, and Xinyu Zhao. Integration and performance analysis of artificial intelligence and computer vision based on deep learning algorithms, 2023. URL https://arxiv.org/abs/2312.12872.

Houze Liu, Chongqing Wang, Xiaoan Zhan, Haotian Zheng, and Chang Che. Enhancing 3d object detection by using neural network with self-adaptive thresholding, 2024. URL https://arxiv.org/abs/2405.07479.

Gilles Louppe. *Understanding Random Forests from theory to practice*. PhD thesis, Université de Liège, 2015.

Gabriel Narvaez, Luis Felipe Giraldo, Michael Bressan, and Andres Pantoja. Machine learning for site-adaptation and solar radiation forecasting. *Renewable Energy*, 167:333–342, 2021. ISSN 0960-1481. doi: https://doi.org/10.1016/j.renene.2020.11.089. URL https://www.sciencedirect.com/science/article/pii/S0960148120318395.

Fernando M. Nollas, German A. Salazar, and Christian A. Gueymard. Quality control procedure for 1-minute pyranometric measurements of global and shadowband-based diffuse solar irradiance. *Renewable Energy*, 202:40–55, 2023. ISSN 0960-1481. doi: https://doi.org/10.1016/j.renene.2022.11.056. URL https://www.sciencedirect.com/science/article/pii/S0960148122016962.

Emilio Salia Olivas, Pablo Rodríguez Belenguer, Quique García Vidal, Fran Vaquer Estalr1ich, Juan Vicent Camisón, and Jorge Vila Tomás. *Inteligencia artificial: Casos prácticos con aprendizaje profundo*. Ediciones de la U, 2022. ISBN 9789587924411. URL https://books.google.com.ar/books?id=XHugEAAAQBAJ.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay, and Gilles Louppe. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 01 2012.

M. C. Peel, B. L. Finlayson, and T. A. McMahon. Updated world map of the köppen-geiger climate classification. *Hydrology and Earth System Sciences*, 11(5):1633–1644, 2007. doi: 10.5194/hess-11-1633-2007. URL https://hess.copernicus.org/articles/11/1633/2007/.

Richard Perez, Pierre Ineichen, Kathy Moore, Marek Kmiecik, Cyril Chain, Ray George, and Frank Vignola. A new operational model for satellite-derived irradiances: description and validation. *Solar Energy*, 73(5):307–317, 2002. ISSN 0038-092X. doi: https://doi.org/10.1016/S0038-092X(02)00122-6. URL https://www.sciencedirect.com/science/article/pii/S0038092X02001226.

J. Polo, S. Wilbert, J.A. Ruiz-Arias, R. Meyer, C. Gueymard, M. Súri, L. Martín, T. Mieslinger, P. Blanc, I. Grant, J. Boland, P. Ineichen, J. Remund, R. Escobar, A. Troccoli, M. Sengupta, K.P. Nielsen, D. Renne, N. Geuder, and T. Cebecauer. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. *Solar Energy*, 132:25–37, 2016. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2016.03.001. URL https://www.sciencedirect.com/science/article/pii/S0038092X16001754.

Jesus Polo, Carlos Fernández-Peruchena, Vasileios Salamalikis, Luis Mazorra-Aguiar, Mathieu Turpin, Luis Martín-Pomares, Andreas Kazantzidis, Philippe Blanc, and Jan Remund. Benchmarking on improvement and site-adaptation techniques for modeled solar radiation datasets. *Solar Energy*, 201:469–479, 2020. ISSN 0038-092X. doi: https://doi.org/

10.1016/j.solener.2020.03.040. URL https://www.sciencedirect.com/science/article/pii/S0038092X20302784.

Cameron W. Potter, Debra Lew, Jim McCaa, Sam Cheng, Scott Eichelberger, and Eric Grimit. Creating the dataset for the western wind and solar integration study (u.s.a.). *Wind Engineering*, 32(4):325–338, 2008. doi: 10.1260/0309-524X.32.4.325. URL https://doi.org/10.1260/0309-524X.32.4.325.

Zhipeng Qu, Armel Oumbe, Philippe Blanc, Bella Espinar, Gerhard Gesell, Benoît Gschwind, Lars Klüser, Mireille Lefèvre, Laurent Saboret, Marion Schroedter-Homscheidt, and Lucien Wald. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The heliosat?4 method. *Meteorologische Zeitschrift*, 26(1):33–57, 02 2017. doi: 10.1127/metz/2016/0781. URL http://dx.doi.org/10.1127/metz/2016/0781.

Hassan Ramchoun, Youssef Ghanou, Mohammed Amine, Mohammed A. Janati Idrissi, and Hamid Tairi. Multilayer perceptron: Architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(1):26–30, 2016. doi: 10.9781/ijimai.2016.415.

Martín Romano Armada, Hugo César Suligoy, Carlos Fernández, and Germán Ariel Salazar. Aplicación de protocolos de control de calidad de datos de radiación solar medidos en salta (argentina). *Avances en Energías Renovables y Medio Ambiente*, 21:61–66, 2017. ISSN 2314-1433, 2796-8111. URL http://portalderevistas.unsa.edu.ar/index.php/averma/article/view/1292. Peer-reviewed.

Vasileios Salamalikis, Panayiotis Tzoumanikas, Athanassios A. Argiriou, and Andreas Kazantzidis. Site adaptation of global horizontal irradiance from the copernicus atmospheric monitoring service for radiation using supervised machine learning techniques. *Renewable Energy*, 195:92–106, 2022. ISSN 0960-1481. doi: https://doi.org/10.1016/j.renene.2022.06.043. URL https://www.sciencedirect.com/science/article/pii/S0960148122008758.

G. Salazar and C. Raichijk. Evaluation of clear-sky conditions in high altitude sites. *Renewable Energy*, 64:197–202, 2014. doi: 10.1016/j.renene.2013.11.003. URL https://doi.org/10.1016/j.renene.2013.11.003.

Germán A. Salazar. Aplicación del modelo híbrido de yang a datos climáticos medios mensuales de 10 localidades de argentina. *Energías Renovables y Medio Ambiente*, 25:15–21, 2010. ISSN 0328-932X.

Germán A. Salazar, Alejandro L. Hernández, Luis R. Saravia, and Guillermo G. Romero. Determinación de los coeficientes de la relación de ångström–prescott para la ciudad de salta (argentina) a partir de datos tomados durante un año. *Avances en Energías Renovables y Medio Ambiente*, 11:–, 2007. ISSN 0329-5184.

Germán A. Salazar, Alejandro L. Hernández, Carlos Cadena, Luis R. Saravia, and Guillermo G. Romero. Caracterización de valores de radiación solar global para día claro en sitios de altura en el noroeste de la república argentina. *Avances en Energías Renovables y Medio Ambiente*, 12:11.33–11.42, 2008a. ISSN 0329-5184.

Germán A. Salazar, Alejandro L. Hernández, and Luis R. Saravia. Estudio de la radiación solar difusa en la bóveda celeste, utilizando kriging como método estimador. In *II Congresso Brasileiro de Energia Solar e III Conferência Regional Latino-Americana da ISES*, Florianópolis, Brasil, 2008b.

Germán A. Salazar, Alejandro L. Hernández, and Luis R. Saravia. Practical models to estimate horizontal irradiance in clear sky conditions: Preliminary results. *Renewable Energy*, 2010. doi: 10.1016/j.renene.2010.01.033.

Germán Salazar, Alejandro Hernández, Ricardo Echazú, Luis Saravia, and Graciela Romero. Comparison between measured mean monthly solar insolation data and estimates from swera database for salta city (northwestern argentina). *Electronic Journal of Energy and Environment*, 3:1, 11 2013. doi: 10.7770/ejee-V0N0-art531.

Hasan Ahmed Salman, Ali Kalakech, and Amani Steiti. Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024:69–79, 2024. doi: 10.58496/BJML/2024/007.

A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. doi: 10.1147/rd.33.0210.

Nilsa Maria Sarmiento Barbieri, Silvina Belmonte, Pablo Roberto Dellicompagni, Ada Judith Franco, Karina Natalia Escalante, and et al. A solar irradiation gis as decision support tool for the province of salta, argentina. *Renewable Energy*, 132:68–80, March 2019.

Emilio Soria Olivas, Pablo Rodríguez Belenguer, Quique García Vidal, Fran Vaquer Estalrich, Juan Vicent Camisón, and Jorge Vila Tomás. *Inteligencia artificial: Casos prácticos con aprendizaje profundo*. Ediciones de la U, Bogotá, Colombia, 1 edition, 2022. ISBN 978-958-792-440-4.

H. Suárez, J. Castillo, G. Salazar, D. Acosta, C. Cadena, M. J. Marin, P. Utrillas, and L. J. A. Martínez. Variabilidad diaria y anual de radiación solar eritémica en tres regiones de la provincia de salta. *Avances en Energías Renovables y Medio Ambiente - AVERMA*, 18:53–61, oct. 2021. URL https://portalderevistas.unsa.edu.ar/index.php/averma/article/view/2015.

J. D. Tarpley. Estimating incident solar radiation at the surface from geostationary satellite data. *Journal of Applied Meteorology and Climatology*, 18(9):1172 – 1181, 1979. doi: 10.1175/1520-0450(1979)018<1172:EISRAT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/18/9/1520-0450_1979_018_1172_eisrat_2_0_co_2.xml.

Peter Thejll and Hans Gleisner. *Reanalysis data*. 03 2015. ISBN 2759817334.

O. Theobald. *Machine Learning for Absolute Beginners: A Plain English Introduction (Third Edition)*. Repro India Limited, 2024. ISBN 9789362056863. URL https://books.google.com.ar/books?id=wUmv0AEACAAJ.

Michael A. Thornton. Removing seasonality under a changing regime: Filtering new car sales. *Computational Statistics & Data Analysis*, 58:4–14, 2013. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2011.06.021. URL https://www.sciencedirect.com/science/article/pii/S0167947311002386. The Third Special Issue on Statistical Signal Extraction and Filtering.

Guillermo Antonio Toro Bayona and Iván Alberto Lizarazo Salcedo. Evaluación de las redes neuronales artificiales perceptron multicapa y fuzzy-artmap en la clasificación de imágenes satelitales. *Ingeniería*, 17(1):61–72, 2012. URL http://www.redalyc.org/articulo.oa?id=498850174008.

O. C. Vilela, M. H. O. Pedrosa Filho, J. F. Escobedo, A. Dal Pai, G. Salazar, C. Raichjk, R. Righini, H. Grossi, and N. Fraidenraich. Qualificação y caracterización da radiação direta, difusa e global em diferentes localidades do brasil e argentina. In *Acta de la XXXVIII Reunión de Trabajo de la Asociación Argentina de Energías Renovables y Medio Ambiente*, volume 3, pages 11.123–11.130, 2015.

Lucien Wald. *Solar radiation energy (fundamentals)*. 01 2007.

Lucien Wald. Basics in solar radiation at earth surface. Technical report, MINES ParisTech, Paris, France, 2018.

Lucien Wald. *Fundamentals of Solar Radiation*. 05 2021. ISBN 9781003155454. doi: 10.1201/9781003155454.

Dazhi Yang and Christian A. Gueymard. Probabilistic post-processing of gridded atmospheric variables and its application to site adaptation of shortwave solar radiation. *Solar Energy*, 225: 427–443, 2021a. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2021.05.050. URL https://www.sciencedirect.com/science/article/pii/S0038092X21004205.

Dazhi Yang and Christian A. Gueymard. Probabilistic post-processing of gridded atmospheric variables and its application to site adaptation of shortwave solar radiation. *Solar Energy*, 225: 427–443, 2021b. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2021.05.050. URL https://www.sciencedirect.com/science/article/pii/S0038092X21004205.

Lianlian Ye, Mengqi Liu, Disong Fu, Hao Wu, Hongrong Shi, and Chunlin Huang. Probabilistic site adaptation for high-accuracy solar radiation datasets in the western sichuan plateau. *Remote Sensing*, 17(10), 2025. ISSN 2072-4292. doi: 10.3390/rs17101720. URL https://www.mdpi.com/2072-4292/17/10/1720.

Sebastian Zainali, Dazhi Yang, Tomas Landelius, and Pietro Elia Campana. Site adaptation with machine learning for a northern europe gridded global solar irradiance product. *Energy and AI*, 15:100331, 2024. ISSN 2666-5468. doi: https://doi.org/10.1016/j.egyai.2023.100331. URL https://www.sciencedirect.com/science/article/pii/S2666546823001039.

Jie Zhang, Anthony Florita, Bri-Mathias Hodge, Siyuan Lu, Hendrik F. Hamann, Venkat Banunarayanan, and Anna M. Brockway. A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, 111:157–175, 2015. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2014.10.016. URL https://www.sciencedirect.com/science/article/pii/S0038092X14005027. Available: https://doi.org/10.1016/j.solener.2014.10.016.

Anexo 1

Se escribe el anexo correspondiente.