UNIVERSIDAD DE CONCEPCIÓN - CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

Modelación y predicción de la potencia generada por plantas fotovoltaicas

por
Andrés Ignacio Salgado Hernández

Profesor guía
Daniel Gerónimo Sbárbaro Hofer

Concepción, septiembre de 2024

Tesis presentada a la

ESCUELA DE GRADUADOS
DE LA UNIVERSIDAD DE CONCEPCIÓN

para optar al grado de
MAGÍSTER EN INGENIERÍA ELÉCTRICA

Modelación y predicción de la potencia generada por plantas fotovoltaicas

Andrés Ignacio Salgado Hernández

Una Tesis del

Departamento de Ingeniería Eléctrica

Presentada en Cumplimiento Parcial de los Requerimientos del Grado de Magíster en Ciencias de la Ingeniería con mención en Ingeniería Eléctrica de la Escuela de Graduados de la Universidad de Concepción, Chile

Septiembre 2024

Resumen

Modelación y predicción de la potencia generada por plantas fotovoltaicas

Andrés Ignacio Salgado Hernández
Universidad de Concepción, 2024

La integración a gran escala de fuentes de energía renovables como la solar fotovoltaica en la red eléctrica se ve dificultada por la incertidumbre asociada a su generación, especialmente por su efecto en la confiabilidad y operación económica del sistema. Para mitigar esta incertidumbre, una de las estrategias más prometedoras es la creación de pronósticos de generación, en los cuales incluso una leve mejora puede causar un alto impacto económico, dada la escala de la red. Los enfoques basados en Machine Learning se han vuelto cada vez más populares para este propósito, sobre todo en los horizontes de predicción intradía y día-adelante. Sin embargo, existe un creciente reconocimiento del valor de incorporar conocimiento basado en principios físicos para la mejora de estos modelos, el cual puede provenir de la modelación de la planta o del recurso solar. Además, es importante realizar estudios con datos propios de la zona debido a que las condiciones climáticas y geográficas locales inciden en la optimización de modelos de Machine Learning

El estudio se llevó a cabo mediante el desarrollo de modelos no-lineales autorregresivos con entradas exógenas (NARX), elegidos por su simplicidad, para la predicción de series de tiempo utilizando redes neuronales. Los modelos fueron sometidos a un proceso de selección de características en el cual se incorporó información física de modelos simples, y fueron ajustados y probados con datos de la zona centro-sur de Chile, donde la producción de energía está fuertemente ligada a las condiciones climáticas y hacen falta estudios sobre su predicción.

Los resultados obtenidos muestran que la inclusión de variables de modelos físicos logra una mejora modesta de la precisión de los pronósticos en distintos horizontes de tiempo. En particular, se disminuyó la raíz del error cuadrático medio entre un 0.2% y un 0.4% de la potencia nominal con respecto a los modelos que no incluyen estas variables. Por último, se concluyó que las variables adicionales más importantes fueron los ángulos de posición solar y las temperaturas calculadas de los paneles, en el proceso de selección de características.

Agradecimientos

Quisiera expresar mis más sinceros agradecimientos a las personas y entidades que hicieron posible el desarrollo de esta tesis:

A la Universidad de Concepción, por el otorgamiento de la beca de articulación pregrado – postgrado, que cubrió los costos de arancel de mis primeros años de magíster.

Al proyecto FONDAP "Solar Energy Research Center (SERC-Chile)" Pycto.1523A0006, por otorgarme financiamiento para el desarrollo de mi tesis, cubriendo los costos de arancel de los años restantes.

A mi profesor guía, el Dr. Daniel Sbárbaro, por su interminable paciencia, comprensión, optimismo y buena disposición al momento de prestar ayuda para salir adelante y completar esta tesis.

A la Fundación Energía Comunitaria por proveer la Plataforma Integrada para la Capacitación en Energía Solar Fotovoltaica (PIACE) de la cual se descargó la base de datos utilizada en el estudio.

A mi familia, en especial a mi padre por incentivarme a aprender inglés, muy necesario para facilitar el desarrollo de mis estudios, y a mi tío por prestarme un computador en el que ejecutar las simulaciones cuando el mío se quemó.

Al Dr. Alok Kanojia por sus aportes en el canal de YouTube “HealthyGamerGG”, que provee una gran fuente de conocimientos sobre salud mental, psiquiatría y muchos consejos sobre cómo enfrentar tareas difíciles como una tesis.

A mis amigos César, Gabriel, Marcos, J.R. y Nhân, por su constante apoyo y compañía.

Finalmente, a Enna Alouette por ser una inspiración para seguir adelante y alcanzar mis metas.

Tabla de contenidos

Nomenclatura

ACF	Autocorrelation Function
AI	Artificial Intelligence
AIC	Akaike Information Criterion
ANN	Artificial Neural Network.
ARIMA	Autoregressive Integrated Moving Average.
ARMA	Autoregressive Moving Average.
BP	Backpropagation.
BR	Bayesian Regularization.
CNN	Convolutional Neural Network.
CSRM	Clear Sky Radiation Model
DCNN	Deep Convolutional Neural Network.
DL	Deep Learning
DRNN	Deep Recurrent Neural Network
FFNN	Feed-forward Neural Network.
GAN	Generative Adversarial Networks.
GHI	Global Horizontal Irradiance.
GRU	Gated Recurrent Unit.
k-NN	k-Nearest Neighbors.
LMBP	Levenberg-Marquardt Backpropagation.
LSTM	Long Short-Term Memory.
MAE	Mean Average Error.
MAPE	Mean Absolute Percentage Error.
MARS	Multivariate Adaptive Regression Splines.
ML	Machine Learning.
MLP	Multilayer Perceptron.
MSE	Mean Square Error.
NAR	Nonlinear Autoregressive.
NARX	Nonlinear Autoregressive with exogenous input.
nMAPE	Normalized Mean Absolute Percentage Error.
nRMSE	Normalized Root Mean Square Error.
NWP	Numerical Weather Prediction
PACF	Partial Autocorrelation Function
PXCF	Partial Cross-correlation Function
PHANN	Physical Hybrid ANN
PV	Photovoltaic
PVPF	Photovoltaic Power Forecasting
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network.
SVM	Support Vector Machine.
XCF	Cross-correlation Function.
WT	Wavelet Transform.

Lista de tablas

Lista de figuras

1 Introducción

1.1 Introducción general

El cambio climático es uno de los problemas más importantes y urgentes a nivel mundial. Este implica cambios a largo plazo en las temperaturas y patrones meteorológicos, lo que conlleva diversas consecuencias catastróficas como sequías, escasez de agua, aumento del nivel del mar e inundaciones [1]. En los últimos años, se ha reforzado la evidencia que vincula estos cambios con la influencia humana y se ha observado un aumento de fenómenos meteorológicos extremos como olas de calor, ciclones tropicales, tormentas y fuertes nevadas [2, 3]. La causa principal del cambio climático es el uso de combustibles fósiles como el carbón, el petróleo y el gas, los cuales liberan gases de efecto invernadero en la atmósfera que atrapan el calor del sol y contribuyen al aumento de las temperaturas [1].

La mayoría de las emisiones de gases de efecto invernadero provienen del sector energético a través del uso de combustibles fósiles para generar electricidad [4]. A pesar de que la electricidad es importante para el desarrollo económico, la dependencia de los combustibles fósiles para su generación trae consigo consecuencias para el medioambiente [3]. Para reducir las emisiones y protegerlo, se requiere de grandes cambios como la reducción del uso de combustibles fósiles, y el despliegue de fuentes de energías renovables como la eólica y solar [5], así como aumentos de la eficiencia energética y su conservación [2], metas que se alinean con el Objetivo de Desarrollo Sostenible 7 (ODS 7) de la Organización de las Naciones Unidas (ONU) [6].

La energía solar fotovoltaica (PV) ha surgido como la fuente de energía renovable más prometedora para aplicaciones residenciales, comerciales e industriales [7]. Esto es debido a que es abundante, asequible y fácilmente escalable, lo que la hace apropiada tanto para sistemas domésticos como de gran escala [8].Sin embargo, su integración a gran escala presenta desafíos relacionados principalmente con su incertidumbre, la cual es causada por variaciones según la estación del año, la hora del día, temperaturas, condiciones geográficas y meteorológicas características de la ubicación de la planta, como la presencia de nubes [3, 9], o su cercanía a zonas áridas o costeras. [10]

La incertidumbre en la generación amenaza la confiabilidad y estabilidad de los sistemas eléctricos, causando problemas de balance de potencia, compensación de potencia reactiva y respuesta en frecuencia de la red [3, 5, 7] además de una gran cantidad de problemas económicos y

complicaciones para la operación del sistema eléctrico y el despacho [3, 9, 11], dificultando incluso el equilibrio del mercado eléctrico y el manejo de licitaciones [8].

Una forma prometedora de solucionar estos problemas es mediante la creación de pronósticos de generación de energía fotovoltaica (Photovoltaic Power Forecasting [PVPF]) [7], la cual ha demostrado ser una solución simple y económica [10]. Las predicciones son necesarias en varios horizontes de tiempo y resultan útiles para la planificación de la operación del sistema eléctrico y la gestión del almacenamiento de energía. Incluso una leve mejora en los pronósticos conlleva un impacto económico considerable. Según [9], los costos asociados a la intermitencia de las energías solar y eólica se encuentran de forma global entre 1 y 12€ por MWh, con variaciones en cada estudio. Por otro lado, en [11] se menciona que para un 9% de penetración solar, una mejora del 25% en los pronósticos día-adelante permitiría ahorrar 0.33 USD$ por MWh, equivalente a 3.92 millones de USD$ al año en el sistema de potencia ISO-NE.

1.2 Hipótesis de trabajo

Es posible encontrar una metodología que combine modelos de Machine Learning e información física, de modo que mejore los pronósticos de generación de energía fotovoltaica en comparación con los enfoques convencionalmente utilizados en la literatura, utilizando datos de la zona centro-sur de Chile.

1.3 Objetivos

1.3.1 Objetivo general

Realizar una comparación del desempeño de diversos modelos, principalmente redes neuronales artificiales, para la predicción de generación de energía de una planta PV ubicada en la zona centro-sur de Chile y proponer un enfoque que mejore las predicciones mediante la incorporación de aspectos físicos.

1.3.2 Objetivos específicos

1) Recrear metodologías con redes neuronales haciendo uso de técnicas vistas en la literatura, incluyendo: amplificación del conjunto de datos, agrupamiento de datos (clustering), uso de clasificación del clima y uso de ventana móvil.

2) Desarrollo de al menos un modelo alternativo para su evaluación respecto a las metodologías ya estudiadas.

3) Evaluar el desempeño de los modelos en el conjunto de datos de prueba para su validación cruzada y comparación.

1.4 Limitaciones y alcances

El proyecto se enfoca en investigar estrategias que es posible poner en práctica con la base de datos disponible. Esta no cuenta con lo necesario para implementar algunas de las técnicas más sofisticadas como el uso de imágenes satelitales o cámaras del cielo. Además, los datos están limitados a mediciones en ubicación de la planta, descartando el desarrollo de modelos NWP.

El enfoque se limita a redes neuronales poco profundas, específicamente de una sola capa oculta, con un mayor interés en la simplicidad, robustez y poder de generalización de los modelos.

Los horizontes de tiempo considerados se limitan a los intervalos intradía y día-adelante, específicamente 1-6 horas adelante, 24 horas adelante, y "día-adelante" en el caso de predecir un día completo a la vez en lugar de hacerlo hora por hora.

El muestreo original de los datos es cada 1 minuto, pero se agrupa en promedios cada 1 hora para reducir el costo computacional. La predicción con una mayor frecuencia de muestreo es posible, pero se deja en caso de trabajo futuro.

1.5 Temario y metodología

El resto del informe se organiza de la siguiente manera: En el capítulo 2 se presenta la discusión bibliográfica, analizando aspectos relevantes y presentando el aporte de la investigación. En el capítulo 3 se explican las metodologías utilizadas para el desarrollo de la tesis, presentando el marco teórico de los modelos y decisiones sobre los métodos utilizados. También se presentan modelos de referencia provenientes de la literatura. En el capítulo 4 se resumen los modelos utilizados para la obtención de resultados, presentando las ecuaciones que los describen y las configuraciones que fueron seleccionadas para cada uno. En el capítulo 5 se presentan los resultados obtenidos y se comparan los distintos enfoques. Por último, la sección 6 presenta las conclusiones finales de la investigación.

2 Estado del arte

2.1 Introducción

El problema de PVPF y su optimización ha sido el tema estudio de cientos de investigaciones, conformando una extensa literatura con tendencias claras que se identifican en artículos de revisión [3, 5, 7, 8]. Sin embargo, todavía existen posibilidades de mejora que se discuten a continuación.

2.2 Tipos de modelo

Un primer aspecto importante es el tipo de modelo. Entre los métodos más prometedores destacan los que utilizan Inteligencia Artificial (AI) o modelos de Machine Learning (ML) como las redes neuronales artificiales (ANN), máquinas de vectores de soporte (SVM), k-NN (k-Nearest Neighbors), modelos híbridos y ensambles, entre otros [3, 7, 8, 12]. Los modelos de Deep Learning (DL) reciben gran atención, algunos ejemplos son el análisis de imágenes con DCNN (Deep Convolutional Neural Network), o la predicción de series de tiempo con redes LSTM (Long Short-Term Memory) o GRU (Gated Recurrent Units) [5].

Algunas de las desventajas de los modelos de ML provienen de ignorar las relaciones físicas entre variables [12], además, estos presentan problemas como la falta de interpretabilidad y explicabilidad de sus resultados, lo que ha llevado a destacar modelos más simples como MARS (Multivariate Adaptive Regression Splines) con un desempeño similar al de modelos de DL y una mayor interpretabilidad [8]. Además, la optimización de modelos basados en ANN y DL resulta compleja y su aleatoriedad inherente dificulta su implementación, haciéndolos menos atractivos para fines prácticos [8, 13].

Otras estrategias prometedoras involucran el uso de imágenes de satélite o imágenes del cielo: las de satélite son útiles para predicciones en una gran área, sin embargo, sufren de problemas de resolución temporal o espacial. Por su parte, las imágenes del cielo sirven para ubicaciones específicas y su popularidad está en aumento [8]. En un estudio señalan que el uso de cámaras del cielo resultó útil para la predicción 10 minutos adelante en días con alta variabilidad de irradiancia, pero no necesariamente en los días más soleados o nublados. [14] Para la presente investigación no se cuenta con este tipo de instrumento, pero vale la pena señalarlo como una posible mejora.

2.3 Horizontes de tiempo

Para la selección de horizonte de tiempo, se ha señalado que es importante tener predicciones en múltiples horizontes, existiendo múltiples razones para predecir en cada uno, y distintas estrategias óptimas [3, 7-9, 15]. Ahmed et al. [3] destacan la necesidad de predicciones en los intervalos intradía y un día adelante, indicando que en último es más común el uso de modelos con algún tipo de predicción meteorológica numérica (NWP). Notton et al. [9] hacen énfasis en el impacto económico de las predicciones en varios horizontes. En estudios como [10] se analizan los errores de predicción en múltiples horizontes a la vez. Alcañiz et al [8] mencionan que un gran número de estudios se enfocan en la predicción un día adelante, y en menor medida en la predicción a muy corto plazo o "nowcasting" (hasta 15 minutos adelante).

Diagne et al. [16] explican que la selección de modelos depende del horizonte de predicción y la base de datos, recomendando el uso de NWP para horizontes superiores a 6 horas adelante, y métodos como ARIMA o ANN para horizontes más cortos. También señalan que en el orden de los minutos es más difícil superar a un modelo de persistencia. En [15], los autores mencionan que las limitaciones de NWP han llevado a un mayor enfoque en los modelos estadísticos y técnicas de IA para la predicción a corto plazo. Sobri et al. [7] mencionan que el uso de ANN es más apropiado para los horizontes de tiempo cortos. Por otro lado, estudios como [17, 18] ofrecen ejemplos de predicción día-adelante con ANN.

2.4 Comparación de modelos y ubicación geográfica

La comparación apropiada de modelos es una parte esencial de la investigación, sin embargo, esta presenta dificultades debido a que su desempeño depende de la base de datos y su calidad, lo cual hace prácticamente imposible realizar una comparación justa sin una base de datos en común. Aparte, comparar indicadores de error entre estudios diferentes suele no ser conclusivo [8, 12]. En consecuencia, para comparar un enfoque nuevo con los que existen en la literatura es necesario replicarlos y probarlos en la misma base de datos. Esto presenta un desafío adicional debido a que, a pesar del gran número de estudios disponible, su replicabilidad es muy baja, llevando a comparaciones inapropiadas u erróneas [19].

Por su parte, la ubicación de la planta y las condiciones meteorológicas locales pueden alterar de manera considerable el proceso de optimización de los modelos, haciendo importante su

evaluación en datos locales [8]. Sin embargo, varios artículos de revisión muestran que existen muy pocos estudios en Sudamérica, y aun menos en Chile [7, 8, 20].

2.5 Integración de información física en modelos de ML

Aunque entre los métodos más prometedores destacan los métodos de Machine Learning (ML), estos raramente incorporan las propiedades físicas de los sistemas PV u otras características bien conocidas como la periodicidad solar en los modelos [10], ignorando las relaciones físicas entre variables. Además, su confiabilidad depende de la base de datos en la que se les entrena [12]. Pombo et al. [19] proponen amplificar el conjunto de datos inicial con un modelo de planta PV, y hacer una selección de características basada en el conocimiento a priori de la relación física entre las variables. Otra estrategia propuesta en [10] es la inclusión de información sobre la periodicidad solar diaria y estacional, la cuál es altamente predecible y ha sido incluida en muy pocas publicaciones. Para esto se pueden incorporar los ángulos de azimut y elevación solar para la ubicación de la planta en la base de datos.

2.6 Técnicas comunes para mejorar los pronósticos

Entre las técnicas utilizadas para mejorar las predicciones se incluyen: la clasificación del clima [21, 22], la selección de variables de entrada (Feature Selection) y el preprocesamiento de datos. Para la clasificación del clima se ha sugerido clasificar los días en 4 categorías: soleado, nublado, parcial y lluvia [10]. Hossain y Mahmood [22] utilizaron un algoritmo de k-means para agrupar valores de irradiancia, y un pronóstico meteorológico sintético para mejorar las predicciones. En selección de variables, un detalle que amerita investigación se ve en el estudio de estudio de Husein y Chung [23], dónde muestran que se puede omitir el uso de la irradiancia solar, la cual es una de las variables más comúnmente utilizadas. Otro aspecto que podría tenerse en cuenta es la periodicidad de variables aparte de la irradiancia, como por ejemplo, la temperatura [13]. Para el preprocesamiento de datos existen diversas técnicas. Las más comúnmente utilizadas son la normalización de datos, y la aplicación de Wavelet Transform (WT) [5, 8], sin embargo, el uso de esta última presenta dificultades que raramente se discuten de manera explícita en la literatura, como los efectos de borde, los que resulta necesario abordar para la implementación en tiempo real.

2.7 Evaluación del error

Respecto a la evaluación de los errores, la forma más común es mediante el uso de indicadores de error como RMSE, MAE, y versiones normalizadas de estos. En artículos como [3, 7, 8] se pueden encontrar compendios de indicadores de error con explicaciones sobre los indicadores comunes, así como también de otras métricas más complejas menos utilizadas. Por otra parte, además del uso de indicadores, resulta conveniente hacer análisis y observaciones directamente en los datos a fin de identificar características y posibilidades de mejora para los modelos que no se revelan en los indicadores más simples.

2.8 Aporte

En conclusión, existe una gran cantidad de aspectos relevantes para realizar predicciones: el tipo de modelo, el horizonte de tiempo, la granularidad de los datos, la factibilidad de su aplicación práctica, las técnicas de procesamiento de datos, las condiciones climáticas de la zona geográfica y su efecto en la utilidad de distintas variables, así como el análisis del error y las técnicas de validación de modelos. Sin embargo, a pesar del amplio estudio en estas áreas, todavía existen aspectos que se pueden mejorar.

Uno de los aportes de este estudio es debido a su relevancia geográfica. Existen muy pocos estudios en Chile sobre predicción de generación de energía fotovoltaica, y es de especial interés su estudio con datos locales en la zona centro-sur del país, donde la generación depende en gran manera de las condiciones meteorológicas. Otros aspectos tienen que ver con la transparencia y replicabilidad. Se busca realizar un aporte mediante la descripción detallada de los pasos utilizados para la obtención de resultados. Por último, se abarcará la inclusión de información física en los modelos mediante la extensión de la base de datos utilizando modelos de posición solar, modelos de cielo despejado, y un modelo de la planta PV.

3 Metodologías y antecedentes

3.1 Selección de horizonte de tiempo y tiempo de muestreo

3.1.1 Horizonte de tiempo

En el presente estudio, la naturaleza de la base de datos limita las técnicas de predicción que se pueden utilizar, siendo estos apropiados para el desarrollo de modelos basados en series de tiempo, pero descartando modelos basados en NWP. El tipo de modelo seleccionado en este caso son modelos en base a ANN, lo cual es determinante para acotar el horizonte de predicción.

En [7] se menciona que los modelos basados en ANN suelen ser apropiados para horizontes de tiempo cortos. A pesar de esto, en [17, 18] se ejemplifica su uso en predicción día-adelante. En [3] se menciona que las predicciones día-adelante se benefician del uso de NWP [3] y en [16] recomiendan usar NWP para horizontes superiores a 6 horas adelante.

Teniendo en cuenta lo anterior, se considera que no vale la pena enfocarse en horizontes más allá de 24 horas adelante, siendo más atractivo el rango entre 1 y 6 horas adelante. Sin embargo, como se cuenta con algunos ejemplos relativamente simples de replicar de la literatura para la comparación de predicciones 24-horas adelante, se considera este horizonte también.

Una justificación para investigar el desempeño en el horizonte intradía es que las predicciones en este horizonte suelen ser más precisas que en el horizonte día-adelante, esto debido a la variación diaria de las condiciones meteorológicas, que tiene un impacto en el mercado eléctrico intradía y hace conveniente contar con correcciones de última hora [8].

De este modo, se seleccionan 7 horizontes de tiempo, de 1 a 6, y 24 horas adelante.

3.1.2 Tiempo de muestreo

La selección de la granularidad de los datos se vio motivada por varias razones. En [16] mencionan que las predicciones día-adelante suelen realizarse con datos con granularidad horaria. Algunos estudios que se toman como referencia para el desarrollo de modelos, también utilizan un muestreo horario [17, 18, 22]. En [15] indican que un tema de interés es la predicción de rampas de potencia, para lo cual es conveniente tener muestras en periodos de hasta 1 hora [15]. Por último, resulta conveniente utilizar un muestreo con baja frecuencia para reducir el número de datos y con

ello costo computacional de los algoritmos. Esto llevó a la elección de un muestreo con frecuencia horaria, agrupando los datos mediante el promedio de sus valores.

3.2 Preparación de la base de datos

3.2.1 Introducción y origen de la base de datos

La base de datos utilizada corresponde a mediciones in situ de variables meteorológicas y eléctricas en la planta PV experimental de 2560W ubicada en el techo de la Facultad de Ingeniería de la Universidad de Concepción. Está compuesta por 8 variables medidas en un lapso de 4 años, desde marzo de 2019 a marzo de 2023, con muestreos cada un minuto.

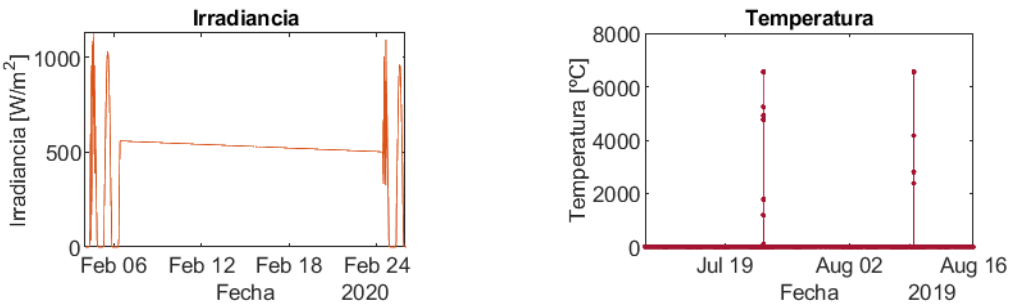
Para la medición de variables meteorológicas, la planta cuenta con un sensor Lufft WS502-UMB. El sensor cuenta con un piranómetro con un rango espectral de 300-1100nm, que se utiliza para medir la irradiancia global horizontal (GHI). El resto de las variables registradas son la temperatura ambiente, la presión atmosférica, la velocidad del viento y la humedad. Además, se tiene como variables la potencia de salida de la planta y los voltajes AC y DC del inversor.

Las secciones siguientes describen el preprocesamiento que se aplicó para prepararla para su uso en el entrenamiento de modelos de ML, incluyendo el proceso de limpieza de datos y la expansión con información física.

3.2.2 Limpieza y agrupación de datos

El primer paso consiste en la preparación y limpieza de la base de datos. Los datos fueron descargados desde el sitio web de la planta y sometidos a un proceso de inspección. Posteriormente fueron reordenados y corregidos, se aplicó un proceso de eliminación de duplicados y se agrupó las variables en una tabla. Del total de datos, aproximadamente un 11.4% de las marcas de tiempo presentó algún tipo de problema causado por diversos motivos: datos perdidos, mediciones erróneas, o marcas de tiempo no registradas. Ejemplos de datos erróneos se muestran en la Fig. 3.1.

Las mediciones incorrectas fueron eliminadas y reemplazados con valores NaN (Not-a-Number) mediante revisión manual y algoritmos ad-hoc. Las marcas de tiempo faltantes fueron incluidas y llenadas con valores NaN. Cuando fuera posible, los datos perdidos fueron reemplazados con la media entre datos adyacentes, sin embargo, en el caso de que los datos perdidos fueran continuos por más de una hora, o extendiéndose por varios días o semanas, simplemente se les excluyó de los análisis y procesos de entrenamiento.



(a) **(b)**

Fig. 3.1 Ejemplos de errores en los datos.
(a) mediciones de irradiancia no válidas; **(b)** valores absurdos de temperatura ambiente.

El total de muestreos por variable resultantes es de 2,106,360, de los cuales 1,866,649 son válidos. Tras la limpieza, se realizó un pequeño recorte en los extremos de la base de datos para uniformar, y se aplicó un submuestreo agrupando los datos según su promedio horario. De este modo, se redujo el número de marcas de tiempo a 35,088, de las cuales 31,134 son válidas.

3.2.3 Separación en subconjuntos para desarrollo de modelos de ML

La base de datos fue dividida en subconjuntos como se recomienda en [24]. En este caso, se hizo una división en cuatro subconjuntos que fueron designados como “entrenamiento”, “validación”, “ajuste” y “prueba”. El subconjunto de entrenamiento se utilizó para ajustar el modelo a los datos. Con este se calcularon los parámetros de las redes utilizando algoritmos de retropropagación. El subconjunto de validación se utiliza para la estimación del error de generalización, siendo útil para prevenir el problema de sobreajuste, evaluar el ajuste de hiperparámetros, y la ejecución de algoritmos de detención anticipada en el entrenamiento mediante validación cruzada. El subconjunto de ajuste también se utiliza para validación cruzada, siendo reservado para la selección de características y modelos, previniendo el sobreajuste al conjunto de validación. Por último, el subconjunto de prueba se reserva para la evaluación del error de generalización final y la comparación entre modelos, por lo que no juega un rol en los procesos de entrenamiento y selección.

La separación de los subconjuntos se realizó en bloques, designando el último año disponible de los datos para el subconjunto de prueba. El penúltimo año se dividió entre los subconjuntos de validación y ajuste en intervalos de 6 semanas, de modo que cada subconjunto recibiera un número similar de muestras por estación del año. El resto de los datos se asignaron al conjunto de entrenamiento. Esta distribución se ilustra en la Fig. 3.2, donde también se muestra la división según estación del año.

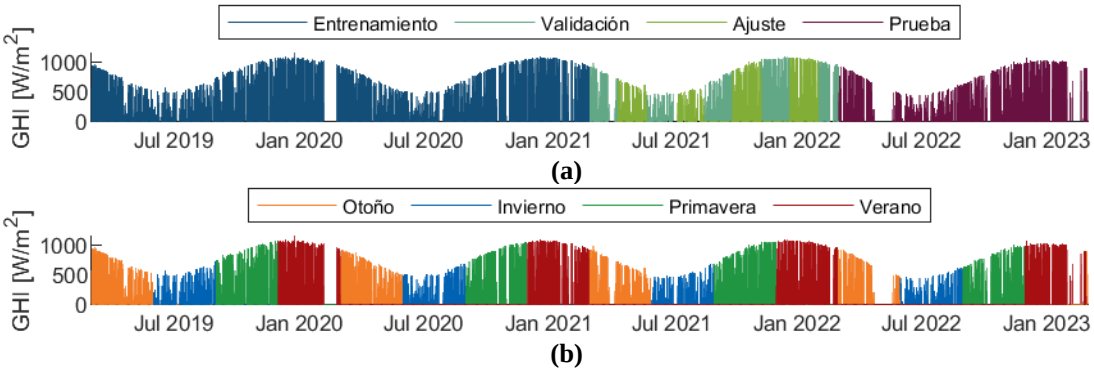


Fig. 3.2 Gráfico de GHI ilustrando divisiones del conjunto de datos.
(a) división según subconjunto; **(b)** división según estación del año.

Se seleccionó la separación en bloques para asegurarse que todos los retardos en los modelos provengan de un mismo subconjunto. Esto tiene el objetivo de evitar un problema de sobreajuste que se había observado previamente al distribuir los datos de forma aleatoria, lo que provocaba una mejora artificial de los indicadores de error en los conjuntos reservados para validación cruzada.

3.2.4 Expansión de la base de datos con información física

Para ampliar la base de datos con información física se emplearon metodologías propuestas en [10, 19], así como modelos e instrucciones disponibles en el sitio web de la PVPMC [25] y en el Toolbox PVLIB para MATLAB [26], el cual contiene una suite de funciones para simular sistemas PV.

Se utilizaron tres modelos principales para la expansión. El primero es un modelo de posición solar, el cual resulta útil para incorporar información de la periodicidad solar mediante la inclusión de los ángulos de azimut y zénit, los cuales se pueden predecir con gran exactitud [10]. El segundo es un modelo genérico de planta PV descrito en [19]. Este modelo utiliza como variable la irradiancia en el plano inclinado E_{POA}, la cual se obtiene a partir de la GHI con ecuaciones descritas más adelante. El tercero es un modelo de cielo despejado que predice la irradiancia solar en base al ángulo de zénit, calculado en función de la ubicación, fecha y hora.

Para obtener la posición solar se utilizó el modelo "Sandia's Ephemeris Model" a través de la función "pvl_ephemeris" disponible en PVLIB. Esta función calcula los ángulos de azimut (Θ_A) y elevación, así como el tiempo solar en función de la ubicación, hora y fecha. El ángulo de zenit (Θ_Z) se obtuvo como el complemento del ángulo de elevación, y fue utilizado en conjunto al día del año y las mediciones de GHI para calcular la irradiancia directa normal (DNI) mediante el uso del modelo DISC, implementado a través de la función "pvl_disc", con la que además se obtiene el índice de claridad (k_t) [26]. Luego, la irradiancia difusa horizontal (DHI) se calculó con (3.1) [25].

$$DHI = GHI - DNI \cdot \cos(\Theta_Z) \tag{3.1}$$

Para calcular la irradiancia en el plano inclinado E_{POA} se utilizó la suma de tres componentes como se muestra en las ecuaciones (3.2) y (3.3) [25]. Donde la irradiancia de haz (E_b) se calcula con el ángulo de incidencia (AOI) obtenido de la función "pvl_getaoi". La irradiancia reflejada en tierra (E_g) se calcula con la función "pvl_grounddiffuse" y requiere un valor de albedo que se estimó en ~0.2 según lo sugerido en [25]. La irradiancia difusa (E_d) se calcula con el modelo de Klucher de 1979, implementado en la función "pvl_klucher1979" [26].

$$E_{POA} = E_b + E_g + E_d \tag{3.2}$$

$$E_b = DNI \cdot \cos(AOI) \tag{3.3}$$

Habiendo obtenido la irradiancia E_{POA}, se implementa el modelo de la planta PV como se explica en [19]. Las temperaturas de módulo (T_m) y celda (T_c) se calculan con ayuda de la función "pvl_sapmcelltemp", que implementa las ecuaciones (3.4) y (3.5), donde a, b y ΔT son parámetros especificados en [27], T_a es la temperatura ambiente, W_s la velocidad del viento y E_{STC} es la irradiancia estándar 1000 W/m^2. Para un módulo y montura tipo "glass/cell/polymer" y "open rack" se utilizaron los valores $a = -3.56, b = -0.0750, \Delta T = 3$.

$$T_m = T_a + E_{POA} \exp(a + b\, W_s) \tag{3.4}$$

$$T_c = T_m + \frac{E_{POA}}{E_{STC}} \Delta T \tag{3.5}$$

La potencia DC del array se obtiene con la ec. (3.6), donde $P_{mp,STC}, T_{STC}, \gamma_{mp}, N_S, N_P$ son la potencia y temperatura peak del panel bajo condiciones de prueba estándar, el coeficiente de temperatura normalizado de máxima potencia, y el número de paneles conectados en serie y paralelo, respectivamente [19]. Por último, se utiliza la eficiencia del inversor para obtener la potencia de salida estimada de la planta (P_{AC-mod}) con la ec. (3.7), dónde η_{max} es la eficiencia máxima del inversor. Este valor de P_{AC-mod} podría ser usado para identificar periodos de restricción de potencia o mediciones imprecisas [19].

$$P_{DC} = N_S N_P \frac{E_{POA}}{E_{STC}} P_{mp,STC} \left[1 + \gamma_{mp}(T_c - T_{STC})\right] \tag{3.6}$$

$$P_{AC-mod} = \begin{cases} \eta_{max}\, P_{DC}, & si\ \eta_{max}\, P_{DC} \leq P_{ACmax} \\ P_{ACmax}\ , & si\ \eta_{max}\, P_{DC} > P_{ACmax} \end{cases} \tag{3.7}$$

Por último, se utilizó el modelo de cielo despejado de Haurwitz mediante la función "pvl_clearsky_haurwitz" para calcular la GHI de cielo despejado a partir del ángulo de zenit Θ_Z [26]. Esta GHI fue utilizada para estimar la irradiancia en el plano inclinado de cielo despejado

($E_{POA-clearsky}$) mediante un proceso análogo al empleado para calcular E_{POA}. Estas variables fueron agregadas a la base de datos.

De esta manera, se amplió la base de datos mediante la inclusión de características que incluyen dependencia física, tomando en cuenta las relaciones entre las variables meteorológicas y la producción de energía, lo que tiene el potencial de simplificar procesos de entrenamiento, y aumentar la robustez y confiabilidad de los modelos [19].

Un par de observaciones importantes que resultan de la expansión de la base de datos tienen que ver con las relaciones entre las nuevas variables y la potencia medida (P_{AC}). En la Fig. 3.3 se muestra una comparación entre la P_{AC} y las irradiancias GHI y E_{POA}. La segunda, que fue calculada teniendo en cuenta la posición solar, presenta una mayor correlación con la potencia, por lo que se esperaría que esta fuera una mejor predictora que la GHI medida. De manera similar, en la Fig. 3.4 se muestra una comparación entre la P_{AC} y la temperatura ambiente, de módulo y de celda.

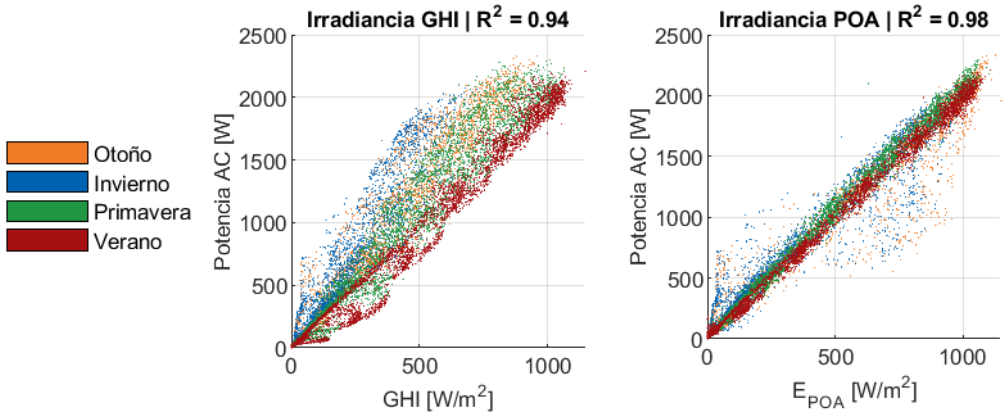


Fig. 3.3 Gráficos de dispersión entre potencia y variables de irradiancia por estación.

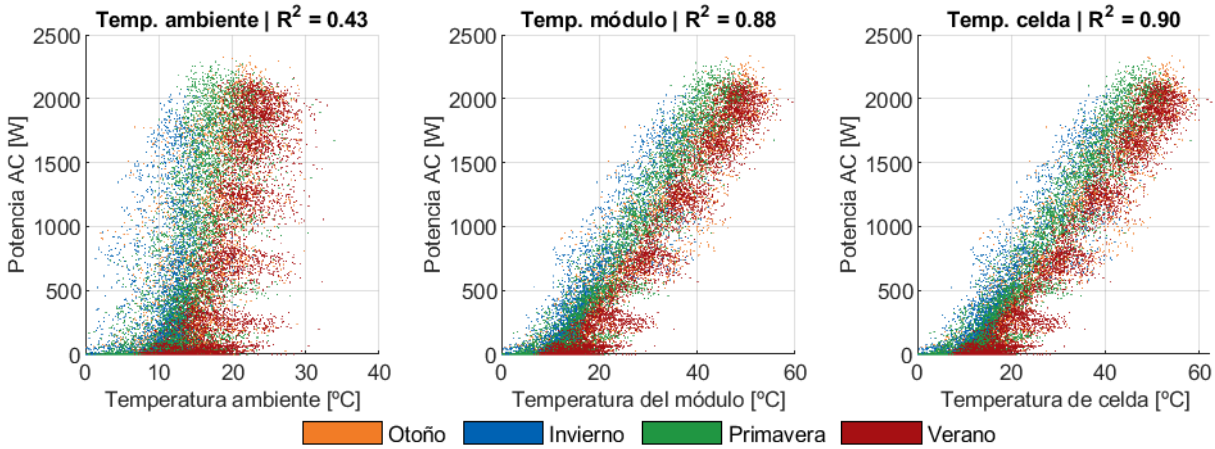


Fig. 3.4 Gráficos de dispersión entre potencia y variables de temperatura por estación.

Por otro lado, la Fig. 3.5 muestra una comparación entre la potencia AC medida por el sensor P_{AC}, y la calculada por el modelo P_{AC-mod}, dónde se observan discrepancias durante los meses de invierno y otoño. Una examinación de los datos reveló que, en su mayoría, estos son casos aislados que se pueden explicar por diferencias entre las sombras que cubren sensor y paneles. Por otro lado, se observa una desviación entre los valores calculados y los medidos, que revela una eficiencia real menor que la considerada en el modelo, pero esto no se atribuye a una causa específica.

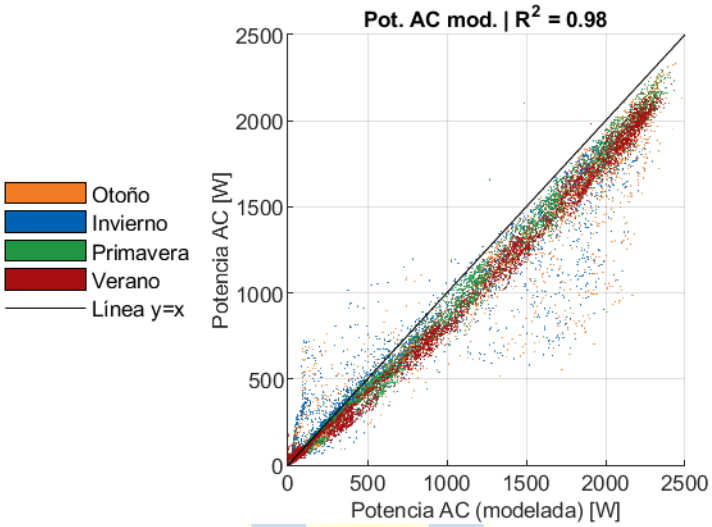


Fig. 3.5 Gráfico de dispersión entre la potencia real y la potencia modelada.

3.2.5 Separación de datos diurnos y nocturnos

Como el objetivo es evaluar la capacidad de los modelos de predecir la potencia que genera la planta PV, resulta de poco interés su desempeño durante las horas de la noche, cuando la predicción resulta trivial. Por esto, se hizo una clasificación de los datos en diurnos y nocturnos utilizando un criterio en base a los valores de irradiancia y potencia.

El criterio utilizado es el siguiente: un valor se considera diurno si cumple al menos una de dos condiciones, que la GHI sea superior a 1.147 [W/m^2], equivalente al 0.1% del valor máximo (1147 [W/m^2]), o que la potencia AC sea superior a 1.28 [W], equivalente al 0.05% del valor nominal (2560 [W]). Si no cumple con ninguna condición, se le considera un valor nocturno. Los límites utilizados para el criterio fueron elegidos en base a una revisión manual de los datos.

La distinción entre valores diurnos y nocturnos fue utilizada para el análisis de los datos, el cálculo de indicadores de error, y para el entrenamiento de modelos, de modo que sólo se consideraran valores de potencia diurnos. Para implementar esto en MATLAB, se le pasó un vector de pesos a la función “train” del mismo tamaño que el vector de salidas, compuesto por valores iguales a 0 para los valores nocturnos y 1 para los valores diurnos.

3.2.6 Resumen de preparación de datos

Tras finalizar la preparación de la base de datos, el conjunto, originalmente compuesto por 8 variables y una marca de tiempo, fue expandido hasta un conjunto con 24 variables que se muestran en la Tabla 3.1. Las primeras 8 variables pertenecen al conjunto original.

Tabla 3.1 Lista de variables del conjunto de datos expandido

1	Potencia AC	9	Azimut solar	17	Temperatura de módulo
2	Irradiancia (GHI)	10	Elevación solar	18	Temperatura de celda
3	Temperatura ambiente	11	Elevación solar aparente	19	Potencia DC (calculada)
4	Presión atmosférica	12	Tiempo solar	20	Potencia AC (calculada)
5	Humedad relativa	13	Zénit solar	21	GHI de cielo despejado
6	Velocidad del viento	14	Zénit solar aparente	22	Irradiancia POA de cielo despejado
7	Voltaje DC	15	Irradiancia POA	23	Hora del día
8	Voltaje AC	16	Índice de claridad	24	Día del año

Algunas de las variables del conjunto son redundantes. y fueron excluidas del espacio de búsqueda de los modelos. Específicamente, se excluyeron los ángulos de elevación, elevación aparente, y zénit aparente, por ser redundantes con el ángulo de zénit solar.

En la Fig. 3.6 se muestran gráficos de algunas de las variables en las que se pueden observar patrones estacionales a simple vista, información que resulta relevante para el desarrollo de los modelos. Además, se muestra la división en subconjuntos mediante colores.

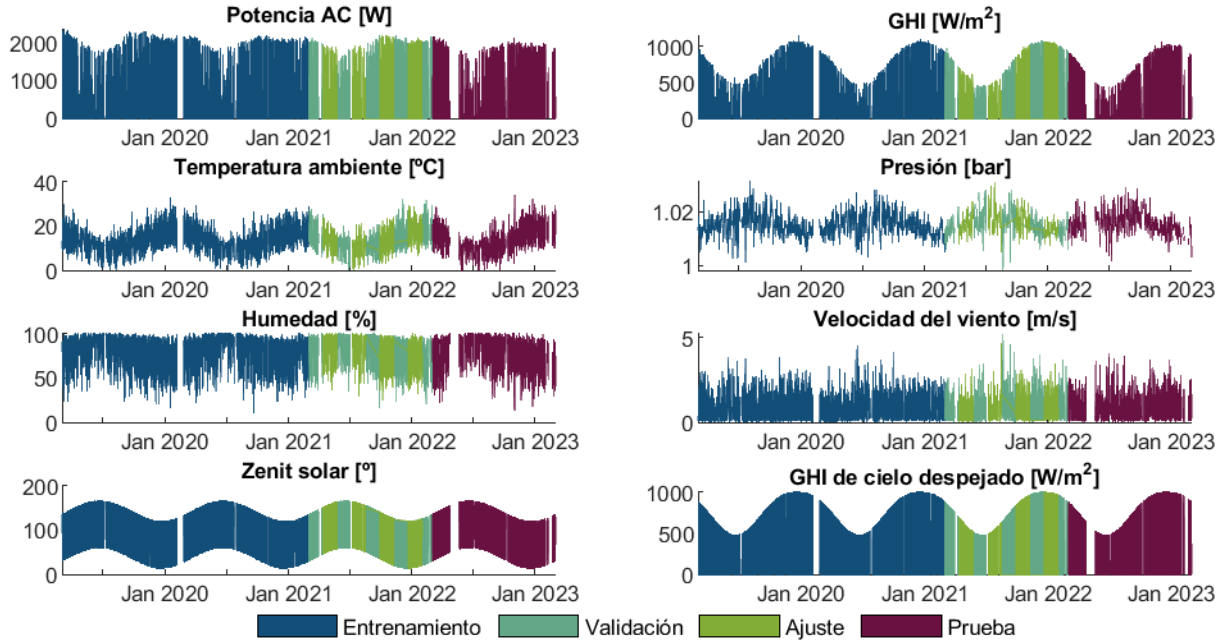


Fig. 3.6 Variables de la base de datos separadas en bloques según subconjunto.

Por último, en la Fig. 3.7 se muestran histogramas de la base de datos según la estación del año, lo que permite ver variaciones estacionales en la distribución de cada variable. Los histogramas fueron creados considerando sólo valores diurnos.

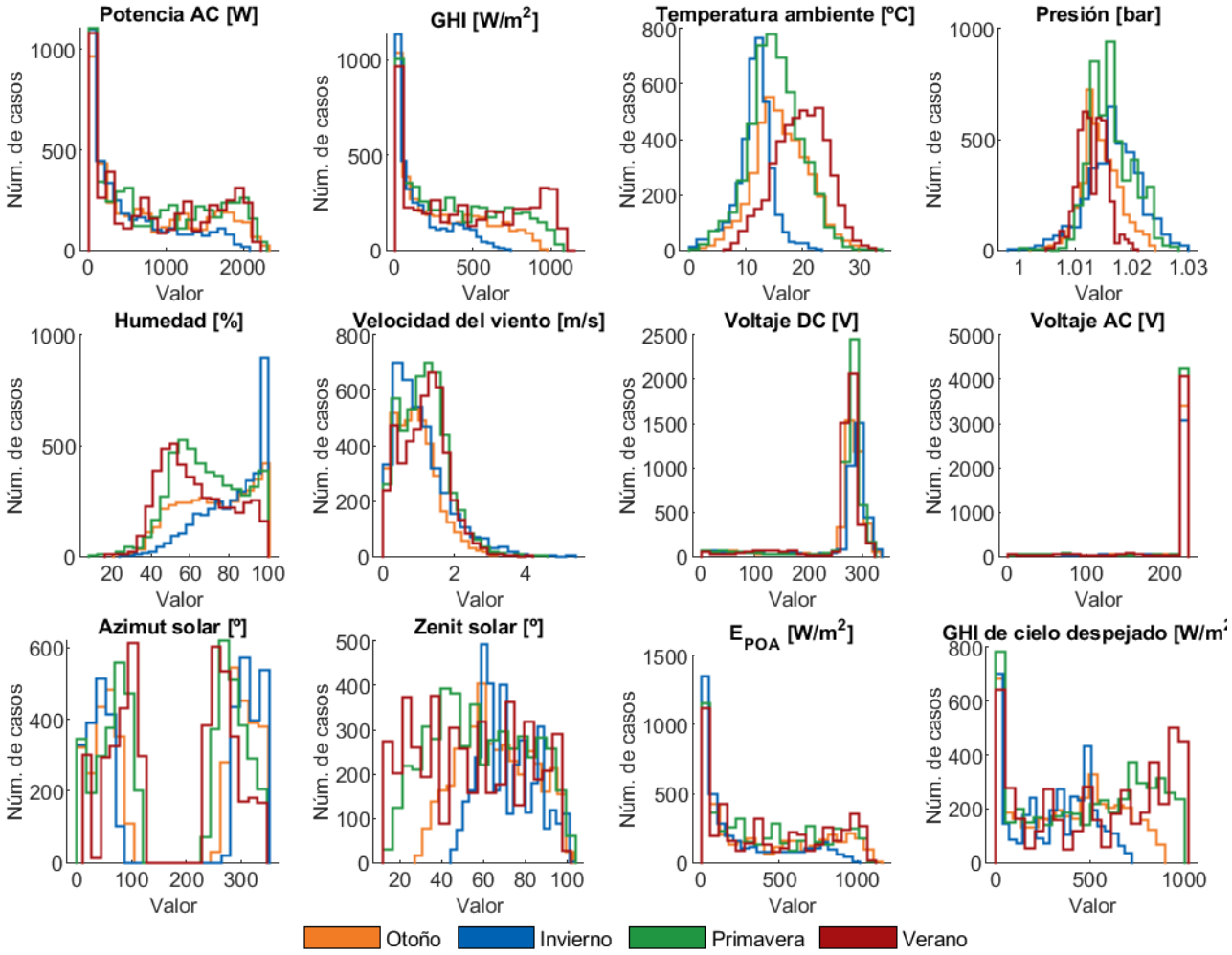


Fig. 3.7 Histogramas de variables de la base de datos separados según la estación del año.

3.3 Análisis estadístico de la base de datos

Para estimar a priori la importancia de las diferentes variables de interés en el desarrollo de modelos de predicción, se realizó un análisis estadístico de la base de datos ocupando herramientas lineales y no-lineales. En este caso la variable principal es la potencia generada por la planta fotovoltaica (P_{AC}).

3.3.1 Función de autocorrelación y autocorrelación parcial

Un primer análisis consiste evaluar la función de autocorrelación (ACF) y la función de autocorrelación parcial (PACF) de la potencia. La primera muestra la relación lineal que existe entre

la serie y sus retardos, mientras que la PACF elimina la influencia de los retardos intermedios con un modelo lineal, siendo útil para descartar retardos con información redundante. Estas funciones se implementaron en base a las ecuaciones disponibles en [28], haciendo uso del coeficiente de correlación de Pearson, calculado con (3.8).

$$\rho_{X,Y} = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^{N}(x_n - \bar{x})^2 \sum_{n=1}^{N}(y_n - \bar{y})^2}} \tag{3.8}$$

Como existe un mayor interés en los valores diurnos de la potencia, se calcularon las correlaciones excluyendo los datos asociados a valores nocturnos con un algoritmo ad-hoc. Los resultados de esto se muestran en la Fig. 3.8, dónde se identificaron como algunos de los retardos más importantes y_{t-1}, y_{t-2}, y_{t-21}, y_{t-19} e y_{t-17}, en orden de importancia, para el caso de potencia diurna. Los resultados concuerdan con la noción de que es conveniente incluir retardos de hasta al menos 24 horas. Posteriormente, se calcularon los casos separando la base de datos por estación, lo que evidenció diferencias estacionales. Esto se muestra en la Fig. 3.9.

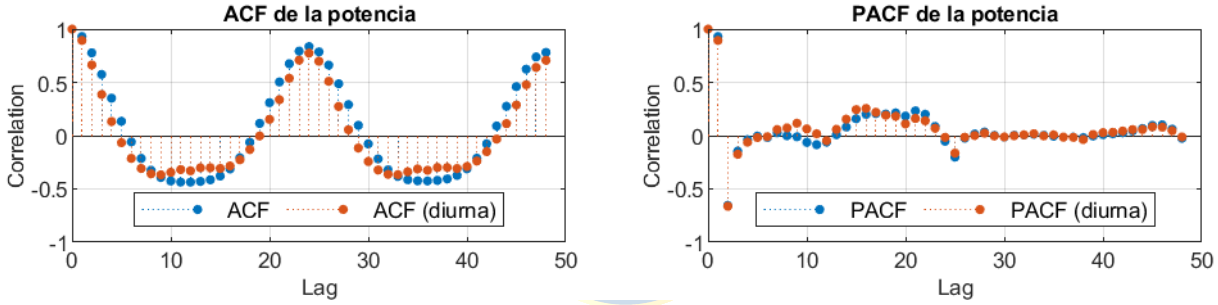


Fig. 3.8 Autocorrelaciones lineales de la potencia con y sin datos nocturnos.

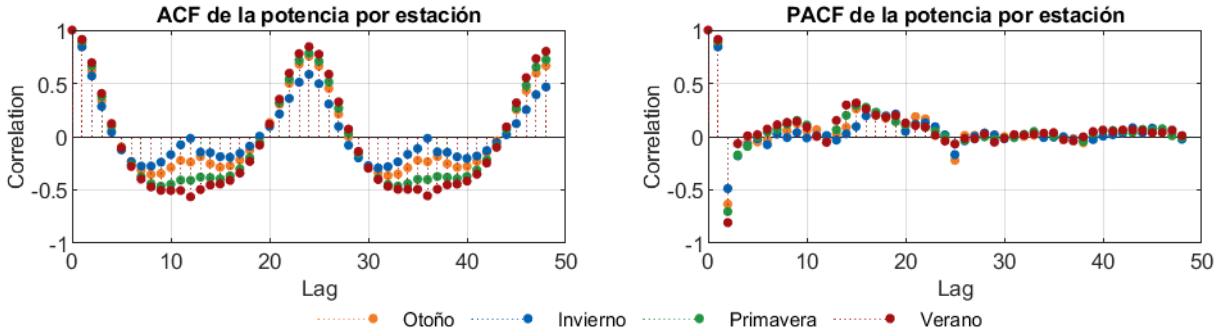


Fig. 3.9 Autocorrelaciones lineales de la potencia según estación del año. Sólo datos diurnos.

3.3.2 Función de correlación cruzada y correlación cruzada parcial

Para analizar la correlación entre la potencia y el resto de las variables se utilizó una función de correlación cruzada (XCF) y una función de correlación cruzada parcial (PXCF), ambas basadas en las ecuaciones mostradas en [28]. En este caso la PXCF elimina la influencia de los retardos intermedios de la variable exógena.

Se seleccionó una variable de ejemplo para ilustrar el cálculo de la XCF y la PXCF. En este caso se utilizó la presión atmosférica, debido a que es una variable poco utilizada [8] y que normalmente se descartaría dado que su correlación con la potencia es baja. Un valor de correlación típico para determinar si una variable es útil es ~0.7 o superior [19]. A pesar de esto, la presión atmosférica mostró mejoras en el desempeño de algunos modelos día-adelante en experimentos preliminares, por lo que es de interés visualizar sus relaciones.

Las XCF y PXCF se muestran en la Fig. 3.10, considerando los casos con y sin valores de potencia nocturnos. Algunos de los retardos más importantes serían x_{t-5}, x_{t-2}, x_{t-4}, x_{t-23} y x_{t-6}, en orden de importancia. La correlación es considerablemente baja. Por otro lado, las XCF y PXCF por estación se muestran en la Fig. 3.11, considerando sólo valores de potencia diurnos. En este caso se observa una mayor correlación en los meses de invierno, lo que podría ser útil para explicar por qué la presión puede ayudar en algunas predicciones.

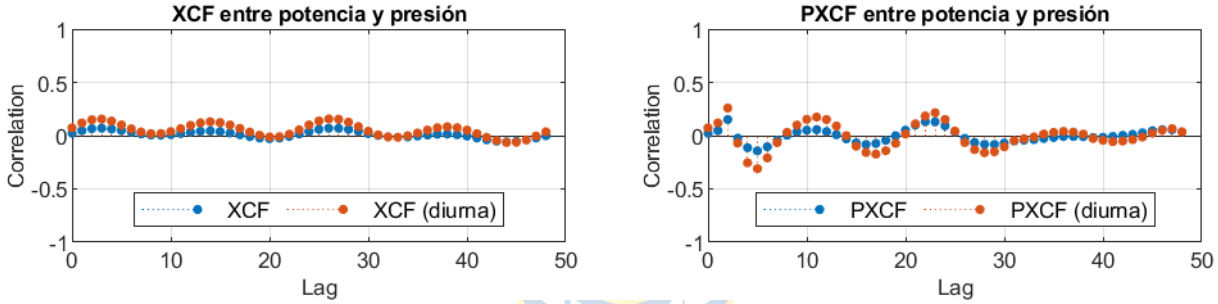


Fig. 3.10 XCF y PXCF entre potencia y presión atmosférica con y sin datos nocturnos.

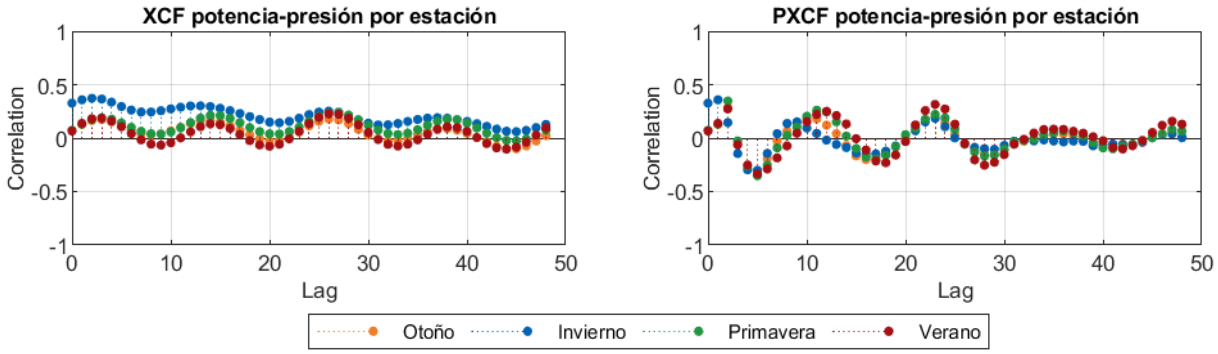


Fig. 3.11 XCF y PXCF entre potencia y presión atmosférica por estación.

El resto de las correlaciones cruzadas se ilustran mediante mapas de calor en la Fig. 3.12, en las que se utilizaron valores absolutos de la XCF y PXCF. A pesar de que existen bastantes variables con alta correlación, estas contienen información redundante, por lo que más adelante se realiza un proceso de selección de características basado en el desempeño.

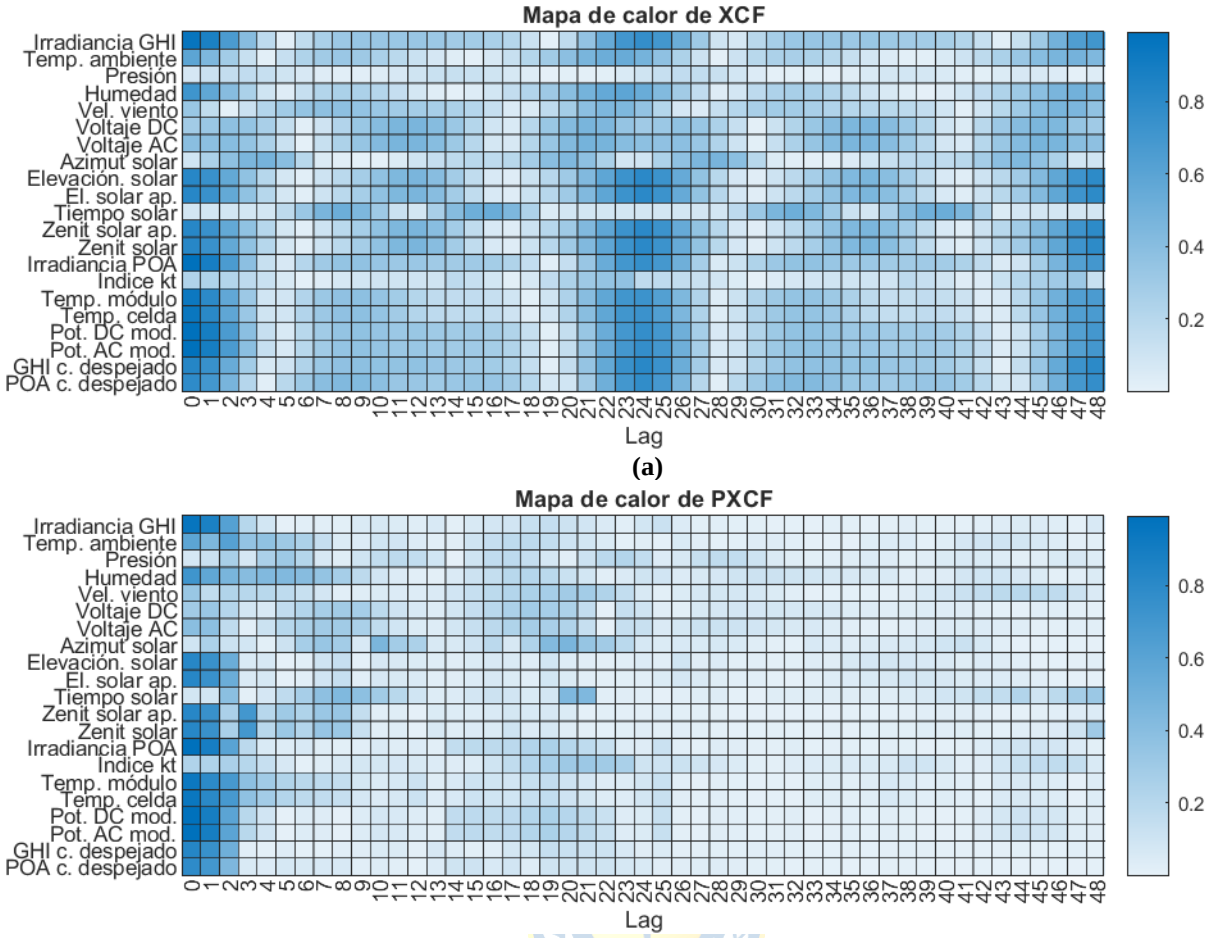


Fig. 3.12 Mapas de calor de las correlaciones entre la potencia y el resto de las variables
(a) valor absoluto de XCF; **(b)** valor absoluto de PXCF.

3.3.3 Correlación de distancia

La correlación de distancia es una medida no-lineal de la dependencia entre dos vectores aleatorios. Su teoría y metodología de cálculo se explican en [29]. Su implementación en MATLAB se realizó con un algoritmo publicado en MATLAB File Exchange [30]. Debido a sus requisitos de memoria, se calculó utilizando una muestra aleatoria de la base de datos, tomando aproximadamente la mitad de los valores diurnos. En total se utilizaron 8738 datos de los 35,088. Los resultados obtenidos se resumen en el mapa de calor de la Fig. 3.13. Comparado con la XCF, la correlación de distancia sugiere una mayor relevancia de variables como el tiempo solar, el ángulo de azimut solar y el índice de claridad k_t.

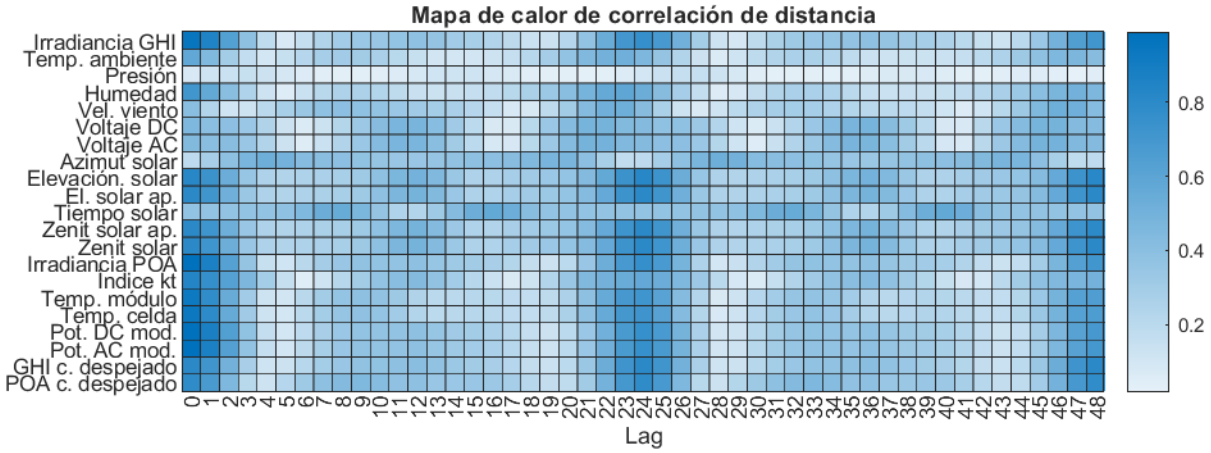


Fig. 3.13 Mapa de calor de la correlación de distancia entre la potencia y el resto de las variables.

3.3.4 Resumen del análisis previo

De la evaluación de las correlaciones se intuye que los retardos más importantes suelen ser los más inmediatos y algunos entre 16 y 24 horas. Por otra parte, una selección óptima puede variar según la estación del año, por lo que se prefiere no elegir retardos específicos, y en su lugar, dejar que se determinen su ponderación durante el entrenamiento de los modelos.

Existen variables con alta correlación, pero con información redundante. Esta puede tener efectos perjudiciales como empeorar los resultados, oscurecer la influencia de variables aparentemente menos relevantes [28], aumentar la complejidad de los modelos innecesariamente y reducir su capacidad de generalización [31]. Para una selección de características apropiada es necesario evaluar la utilidad real de incluir una variable. Esto se hizo mediante el entrenamiento de los modelos y posterior cálculo de indicadores de error como el Criterio de Información de Akaike (AIC), el cual penaliza el aumento de la complejidad que conlleva incluir variables y retardos adicionales [28]. Una alternativa es evaluar el cambio que causa agregar una variable en el indicador RMSE (root-mean-square error). Se puede graficar su mejora en función del número de entradas como se ejemplifica en [31].

Para eliminar variables redundantes también se puede utilizar el entendimiento físico de los datos. Ejemplos incluyen eliminar la temperatura del módulo en favor de la temperatura de celda [19] o eliminar el ángulo de elevación por ser redundante con el de zénit.

3.4 Redes neuronales artificiales (ANN)

3.4.1 Estructura de modelos NAR y NARX

Dos estructuras simples de modelo para predicción de series de tiempo son los modelos no-lineal autorregresivo (NAR) y no-lineal autorregresivo con entrada exógena (NARX). Estos modelos hacen uso de valores pasados de la salida y entradas mediante líneas de retardos (tapped delay lines o TDL) y pueden ser implementados con una red neuronal sencilla, entrenada con algoritmos de retropropagación. En este caso se utilizaron redes de una sola capa oculta con función de activación tangente hiperbólica, lo que las hace aproximadores de funciones universales. [24].

La Fig. 3.14 muestra un esquema general de una red NARX, la cual es descrita por las ecuaciones (3.9) y (3.10). Esta red se puede utilizar para predecir una serie de tiempo un paso adelante como se muestra en la ec. (3.11), donde la salida de la red aproxima la potencia AC. En este caso se omite el retardo 0 de las entradas, dado que estos valores no son conocidos de antemano en tiempo real. Por último, se modificó el modelo especificando un horizonte h que indica el número de horas adelante que se quiere predecir como se muestra en la ec. (3.12). En este caso se utilizaron valores de 1 a 6, y 24 horas adelante. En el caso de omitir el vector de entradas exógenas (**x**), la red aproxima un modelo NAR.

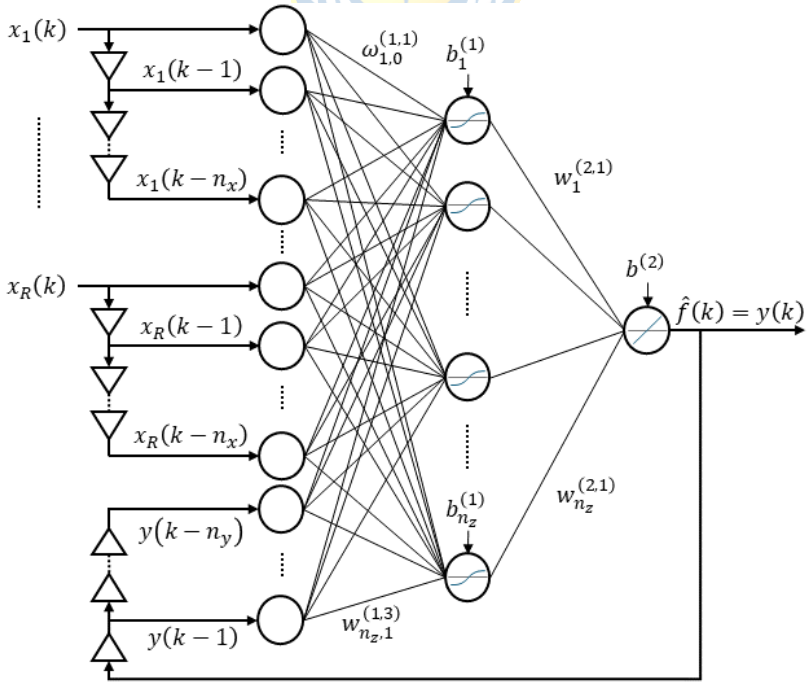


Fig. 3.14 Esquema general de una red neuronal NARX.

$$\hat{f}(k) = b^{(2)} + \sum_{l=1}^{n_z} w_l^{(2,1)} z_l^{(1)}(k) \tag{3.9}$$

$$z_l^{(1)}(k) = \tanh\left(b_l^{(1)} + \sum_{i=1}^{n_y} w_{l,i}^{(1,3)} y(k-i) + \sum_{j=1}^{R}\sum_{i=0}^{n_x} \omega_{l,i}^{(1,j)} x_j(k-i)\right) \tag{3.10}$$

Donde:

Los superíndices indican el número de la capa.

$\hat{f}(k)$: salida de la red NARX en el instante k.

$b^{(2)}$: valor de bias de la capa de salida

n_z : número de nodos en la capa oculta

$w_l^{(2,1)}$: peso que conecta el l-ésimo nodo de la capa oculta a la salida

$z_l^{(1)}(k)$: salida del l-ésimo nodo de la capa oculta en el instante k

$b_l^{(1)}$: bias del l-ésimo nodo de la capa oculta

n_y : número de retardos de la salida

n_x : número de retardos de la entrada

$w_{l,i}^{(1,3)}$: peso que une el i-ésimo retardo de la salida con el l-ésimo nodo de la capa oculta.

$\omega_{l,i}^{(1,j)}$: peso que une el i-ésimo retardo de la j-ésima entrada con la capa oculta

y : valores de la salida que se quiere aproximar.

$x_1(k), \dots, x_R(k)$ conforman un vector de R variables exógenas en el instante k.

$$y_k = \hat{f}\left(y_{k-1}, \dots, y_{k-n_y}, \mathbf{x}_{k-1}, \dots, \mathbf{x}_{k-n_x}\right) + \varepsilon_k \tag{3.11}$$

$$y_{k+h} = \hat{f}\left(y_k, \dots, y_{k-(n_y-1)}, \mathbf{x}_k, \dots, \mathbf{x}_{k-(n_x-1)}\right) + \varepsilon_{k+h} \tag{3.12}$$

Donde:

y_k : es la salida del modelo en el instante k.

$\mathbf{x}_k$: es el vector de entradas exógenas del modelo en el instante k.

h : es el horizonte de tiempo en horas.

n_y : es el número de retardos de la salida.

n_x : es el número de retardos del vector de entradas.

ε_k : es un término de error.

3.4.2 Inicialización de parámetros

La inicialización de parámetros se realiza de forma aleatoria con el método de Nguyen-Widrow [24, 32]. Este proceso está automatizado en Matlab y se realiza en toda capa oculta cuya función de activación sea acotada [33].

Un factor muy importante que se debe considerar es que la inicialización de parámetros puede afectar el desempeño de las ANN, lo que complica la comparación entre modelos. Para asegurarse de que las diferencias en desempeño se deben a la configuración de las redes y no a la aleatoriedad, se deben entrenar varias redes neuronales para cada configuración de modelo, como se recomienda en [31]. En este estudio todos los modelos se comparan considerando promedios de indicadores, calculados a partir de los resultados de un número no menor a 9 redes individuales por modelo.

3.4.3 Retropropagación Levenberg-Marquardt

El ajuste de parámetros de las redes se realizó mediante retropropagación Levenberg-Marquardt (LMPB). Esta es una técnica de basada en el algoritmo del mismo nombre, el cual es una variación del método de Newton para minimizar sumas de cuadrados de funciones no-lineales, lo que la hace útil para entrenar ANN que utilizan el error cuadrático medio como índice de desempeño [24, 34].

3.4.4 Problemas de infrajuste y sobreajuste

Dos problemas relacionados con la selección inadecuada del número de parámetros de las ANN son el infrajuste (underfitting) y el sobreajuste (overfitting). El primero tiene como consecuencia que la red sea incapaz de aprender las relaciones entre entradas y salidas, mientras que el segundo consiste en una mejora del desempeño en datos de entrenamiento a costa de una pérdida de generalización, empeorando el desempeño en datos nuevos. Por otro lado, el problema de sobreajuste no debiera ocurrir si el número de datos es mucho mayor que el número de parámetros de la red [24].

Para enfrentar estos problemas, la selección del número de parámetros se realizó mediante el entrenamiento de redes con varias configuraciones de hiperparámetros y un proceso de selección basado en validación cruzada.

3.4.5 Entrenamiento con criterio de detención anticipada

Se utilizó un criterio de detención anticipada para prevenir el problema de sobreajuste y a la vez reducir el tiempo de entrenamiento. Este criterio consiste en obtener un estimado del error de generalización en cada iteración del entrenamiento, mediante validación cruzada, utilizando el subconjunto de validación. El entrenamiento se detiene una vez que el error de generalización no mejora por varias iteraciones seguidas. La idea tras la detención anticipada es que en cada iteración la red utiliza más de sus parámetros disponibles, aumentando su complejidad. Si se detiene el entrenamiento antes de llegar al mínimo global es menos probable que se sobreajuste [24].

3.5 Análisis del error

3.5.1 Criterio de Información de Akaike

El Criterio de Información de Akaike (AIC) resulta útil para la selección de modelos ya que penaliza la complejidad de estos. Se implementó con la (3.13) vista en [28] y se utilizó principalmente para la optimización de hiperparámetros y selección de características, eligiendo las que minimicen su valor.

$$AIC = N \ln\left(\frac{1}{N}\sum_{n=1}^{N}(y_n - t_n)^2\right) + 2p \tag{3.13}$$

Donde:

N : es el número de muestras.

y_n : es un valor de la salida del modelo.

t_n : es un valor objetivo, un valor real medido.

p : es el número de parámetros del modelo

En el caso de comparar modelos con distinto número de retardos se produce una diferencia menor en el número de muestras del error disponibles que dificulta la comparación. Esto se resolvió asumiendo que se puede rellenar los valores de error faltantes con su valor promedio, para cada modelo según corresponda.

3.5.2 Indicadores de error clásicos

Algunos de los indicadores de error típicos que se utilizan en una variedad de estudios son la raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE), y coeficiente de determinación R^2, siendo el más popular el RMSE [3, 7, 8]. Estos se definen como se muestra en las

ecuaciones de la (3.14) a la (3.18), entre las que se incluyen versiones normalizadas con la potencia nominal de la planta P_{nom} (2560 [W]).

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - t_n)^2} \tag{3.14}$$

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|y_n - t_n| \tag{3.15}$$

$$R^2 = \left(\frac{\sum_{n=1}^{N}(t_n - \bar{t})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^{N}(t_n - \bar{t})^2 \sum_{n=1}^{N}(y_n - \bar{y})^2}}\right)^2 \tag{3.16}$$

$$nRMSE\% = \frac{RMSE}{P_{nom}} \cdot 100\% \tag{3.17}$$

$$nMAE\% = \frac{MAE}{P_{nom}} \cdot 100\% \tag{3.18}$$

Donde:

N : es el número de muestras.

y_n : es un valor de la salida del modelo.

t_n : es un valor objetivo, un valor real medido.

$\bar{y}$: es el valor medio de la salida.

$\bar{t}$: es el valor medio de los objetivos.

P_{nom} : es la potencia nominal de la planta (2560 [W]).

3.5.3 Regresión entre salidas y objetivos

Una manera útil de evaluar los modelos es mediante la regresión lineal entre las salidas de la red y sus valores objetivo, lo cual se hace con la ec. (3.19). Esta regresión se compara con el caso ideal en que $y_q = t_q$ y se calcula el coeficiente de determinación entre salidas y objetivos R^2 con la ec. (3.16). Si este es cercano a 1 indica un mejor desempeño. [24]

$$y_q = mt_q + c + \varepsilon_q \tag{3.19}$$

Donde:

m : es la pendiente de la recta

c : es el offset

t_q : es un valor objetivo

y_q : es el correspondiente valor de salida de la red

ε_q : es el error residual de la regresión

3.5.4 Autocorrelación del error y correlación error-entrada

Que los errores de predicción estén correlacionados consigo mismos, o con las entradas del modelo en el tiempo, es un indicio de que el error podría ser predicho y, en consecuencia, que el modelo se puede mejorar [24]. Para evaluar esto se calculó la función de autocorrelación (ACF) entre el error y sus retardos, y la función de correlación cruzada (XCF) entre el error y los retardos de las entradas.

Si los errores no están correlacionados, los valores de la ACF debiesen ser cercanos a 0 para todos los retardos excepto 0. De ser así, se puede decir que la señal es ruido blanco, y de no serlo, una forma de mejorar la predicción puede ser aumentar el número de retardos de la red. Por otro lado, los valores de la XCF deben ser cercanos a 0 para todo retardo. En una red NARX, la correlación entre entradas y errores puede indicar que se debe aumentar el número de retardos en tanto entradas como realimentaciones [24]. Se consideró un intervalo de confianza de aproximadamente 95%, igual a $\pm 1.96/\sqrt{N}$.

3.6 Metodología de ajuste de hiperparámetros y selección de características

Como la inicialización de parámetros es un proceso aleatorio y afecta el desempeño, se entrenaron al menos 9 redes neuronales para cada configuración de hiperparámetros. En particular, se entrenaron múltiples redes para cada combinación de tamaño de capa oculta, número de retardos de la salida, selección de variables de entradas y número de retardos de la entrada. Luego se utilizaron las medianas de los indicadores de error RMSE y AIC para la comparación entre modelos. Se prefirió el uso de la mediana debido a la presencia de valores atípicos en algunos resultados.

En el caso de los modelos NARX, el ajuste de hiperparámetros y la selección de características se realizó siguiendo un procedimiento paso a paso, que se repitió para cada uno de los siete horizontes de tiempo, explicado a continuación:

El primer paso fue la selección de hiperparámetros en un modelo sin entradas exógenas. En este caso sólo se consideraron redes NAR de una sola capa oculta. El número de nodos de la capa

oculta y el número de retardos de la variable de salida fueron seleccionados mediante una búsqueda en rejilla (grid search), evaluando los resultados en el subconjunto de validación.

El segundo paso, tras haber elegido un modelo NAR como base, consiste en realizar un proceso de selección de características y elegir una primera variable de entrada. Una limitación que se encontró fue que la red NARX disponible en el Deep Learning Toolbox de MATLAB procesa todas las entradas exógenas como un solo vector de entradas [33], lo que hace necesario que cada entrada adicional tenga los mismos retardos que la anterior. El método seleccionado para elegir la primera entrada exógena y el número de retardos fue una segunda búsqueda en rejilla.

El resto de los pasos consiste en añadir el resto de las variables exógenas a los modelos. Estas fueron agregadas una por una siguiendo un procedimiento similar a [31] hasta alcanzar un criterio de detención. En cada paso, las redes fueron entrenadas con una nueva variable y se evaluó su desempeño calculando el AIC en el subconjunto de validación. El proceso concluye cuando agregar una nueva variable deja de reducir la mediana del AIC. Llegado este punto, se reevalúan todos los modelos mediante validación cruzada una última vez usando el subconjunto de ajuste, y se seleccionan las configuraciones con mejor desempeño según la mediana del AIC.

Finalmente, es importante tener en mente que las mejores variables pueden variar según el horizonte de tiempo, el tiempo de muestreo de los datos, el periodo del año y la selección de hiperparámetros de la red. El método de selección de variables empleado no abarca todas las posibilidades en favor de un menor esfuerzo computacional y reducir el tiempo empleado en optimización.

3.7 Estrategias basadas en la literatura para comparación

3.7.1 Modelo de persistencia para benchmarking

Se implementó un modelo de persistencia clásico cuya función es servir como base de comparación para el resto de los modelos. Este modelo es descrito por la ecuación (3.20), donde el valor predicho en el instante $\hat{y}_{k+h}$ es igual al valor presente y_k [20].

$$\hat{y}_{k+h} = y_k \tag{3.20}$$

3.7.2 Modelo día-adelante con clustering de irradiancia

Se puede replicar uno de los modelos comparados por Nespoli et al. [17] utilizando el conjunto de datos local. Este modelo predice las 24 horas del día siguiente a la vez mediante el uso de 3 redes

neuronales MLP y un proceso de clasificación de los días en “soleados” y “nublados” en función de la irradiancia promedio diaria.

Un día se considera “soleado” si la irradiancia en el plano inclinado promedio es mayor o igual a 150 [W/m^2], y en el caso contrario se considera como día “nublado”. La distribución de los días en los subconjuntos se muestra en la Tabla 3.2, y por estación del año en la Tabla 3.3, donde se observa que la mayoría de los días asignados como nublados corresponden a los meses de invierno.

Tabla 3.2 Distribución de los días válidos por subconjunto.

Subconjunto	Entrenamiento	Validación	Ajuste	Prueba
Días soleados	485	122	115	185
Días nublados	182	55	41	95
Total	667	177	156	280

Tabla 3.3 Distribución de los días válidos por estación.

Estación	Otoño	Invierno	Primavera	Verano
Días soleados	229	123	281	274
Días nublados	88	201	60	24
Total	317	324	341	298

La primera red se utiliza para predecir el tipo de día siguiente mediante un pronóstico de la irradiancia, la cual se promedia. Esta red aproxima la función f_G de la ec. (3.21), mientras que las otras dos se utilizan para predecir la potencia, cada una especializada en un tipo de día. Estas aproximan la función f_P de la (3.22) en sus días correspondientes. Un esquema del modelo se muestra en la Fig. 3.15.

$$\{G_1(d+1), G_2(d+1), \dots, G_{24}(d+1)\} = f_G(G_m(d), T_m(d), d) \quad (3.21)$$

$$\{P_1(d+1), P_2(d+1), \dots, P_{24}(d+1)\} = f_P\big(G_m(d), T_m(d), \mathrm{P_m}(d)\big) \quad (3.22)$$

Donde:

$G_1, G_2, \dots, G_{24}$: corresponden a la irradiancia en el plano inclinado en cada hora del día.

$P_1, P_2, \dots, P_{24}$: corresponden a la potencia en cada hora del día.

G_m : es la irradiancia en el plano inclinado promedio diaria.

T_m : es la temperatura promedio diaria.

P_m : es la potencia promedio diaria.

d : es un día arbitrario.

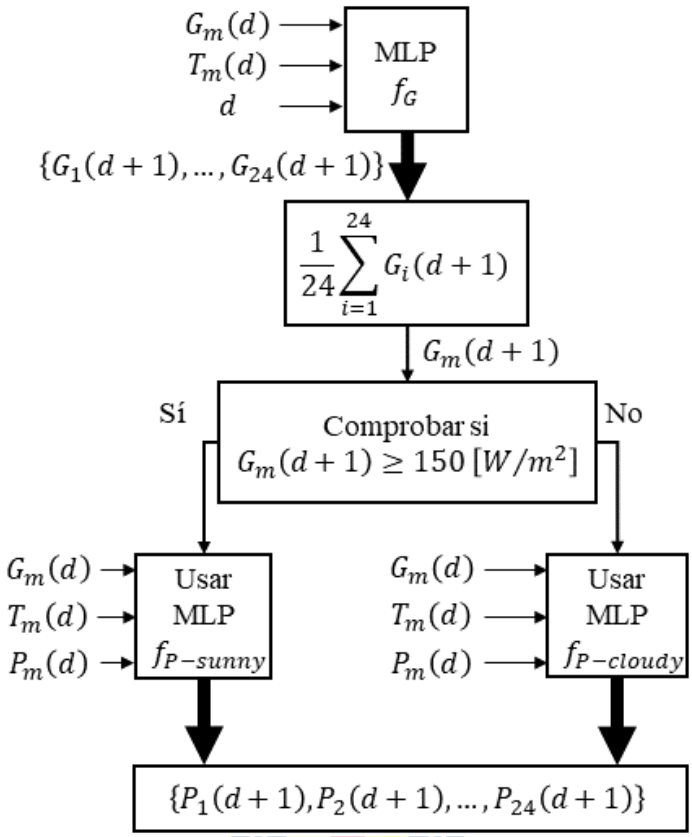


Fig. 3.15 Diagrama de bloques de la metodología de la referencia

El número de nodos de cada red se seleccionó mediante la minimización del indicador RMSE en el subconjunto de validación. Se entrenaron 20 redes por configuración y se buscó en un rango de 2 a 20 nodos. Las configuraciones seleccionadas para cada red se resumen en la Tabla 3.4. La razón por la que se prefirió el valor de RMSE fue que, dado que la red tiene un número considerable de salidas, cada nodo adicional incrementó el número de parámetros de un modo que el AIC penaliza fuertemente, promoviendo la selección de modelos con muy pocos parámetros y un desempeño más bajo de lo deseado. Este comportamiento se muestra en los gráficos de cajas y bigotes obtenidos durante la selección, en la Fig. 3.16.

Tabla 3.4 Número de nodos seleccionados para cada red del modelo

Red	Nodos
Red 1: Irradiancia	5
Red 2: Potencia (soleado)	5
Red 3: Potencia (nublado)	7

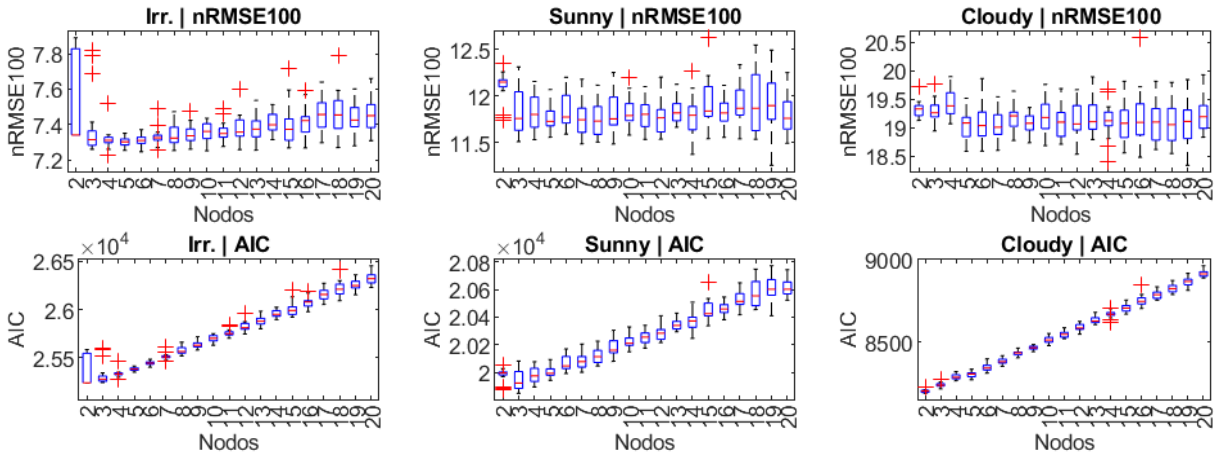


Fig. 3.16 Indicadores obtenidos durante el proceso de selección del número de nodos.

3.7.3 Modelo para predicción 24 horas adelante con retardos cada 24 horas

Alomari et al. [18] entrenaron modelos para crear pronósticos 24 horas adelante utilizando datos de los días anteriores a la misma hora del pronóstico. En su trabajo consideraron 3 variables de entrada: el número de horas desde el inicio del año, la irradiancia, la temperatura y combinaciones de estas. Para comparar su desempeño se realizó una adaptación de su trabajo en la base de datos local, entrenando redes con las combinaciones de entradas correspondientes y realizando una búsqueda para seleccionar el número de nodos en la capa oculta. Las redes entrenadas fueron probadas en el subconjunto de validación y se seleccionó la configuración que minimizó el AIC.

La configuración que obtuvo el mejor desempeño en nuestro conjunto de datos fueron ANN con 4 nodos en la capa oculta, que toma como entradas la hora desde el inicio del año y 5 retardos de GHI. Estas redes son descritas por la ec. (3.23), donde t es el número de horas, G es la GHI e $\hat{y}_k$ es el valor pronosticado de potencia.

$$\hat{y}_k = \hat{f}(t_k, G_{k-24}, G_{k-48}, G_{k-72}, G_{k-96}, G_{k-120}) \tag{3.23}$$

Adicionalmente, su estrategia de tomar datos en intervalos de 24 horas sirvió de base para el proceso de selección de retardos para los modelos NARX en el mismo horizonte.

4 Modelos con redes neuronales

4.1 Modelos NAR

El primer enfoque del estudio consistió en la selección de 7 modelos NAR con ANN de una sola capa oculta, que sirvieran como base para el desarrollo de modelos NARX en cada horizonte de tiempo. Para los horizontes de 1 a 6 horas, estos modelos son descritos por la ec. (4.1). Para el horizonte de 24 horas se utilizó la ec. (4.2). En estas ecuaciones y_k es la salida en el instante k, h es el horizonte de tiempo, n_y es el número de retardos de la salida y ε_k es un término de error residual.

$$y_{k+h} = \hat{f}\left(y_k, \dots, y_{k-(n_y-1)}\right) + \varepsilon_{k+h} \tag{4.1}$$

$$y_{k+24} = \hat{f}\left(y_k, y_{k-24}, \dots, y_{k-24(n_y-1)}\right) + \varepsilon_{k+24} \tag{4.2}$$

Para seleccionar los hiperparámetros se realizó una búsqueda en rejilla con un rango de retardos de salida entre 2 y 48, y un número de nodos en la capa oculta entre 2 y 20, ambos en pasos de a 2. En el caso de 24-horas, se utilizó un espacio reducido de 2 a 10 nodos, y entre 1 y 7 retardos de la salida. Las configuraciones que obtuvieron el mejor desempeño se listan en la Tabla 4.1.

Tabla 4.1 Configuración de modelo NAR seleccionada para cada horizonte de tiempo.

Horizonte (h)	1	2	3	4	5	6	24
N.º de retardos de salida (n_y)	28	30	22	28	22	22	5 (uno por día)
Nodos en la capa oculta	4	4	6	6	6	4	3

En la Fig. 4.1 se muestran mapas de calor que ilustran el proceso de búsqueda en rejilla, mostrando el valor de los indicadores nRMSE% y AIC obtenidos en el subconjunto de validación, para el horizonte 1-hora adelante. Se encontró que incluir retardos en la red tiene un mayor efecto que agregar nodos.

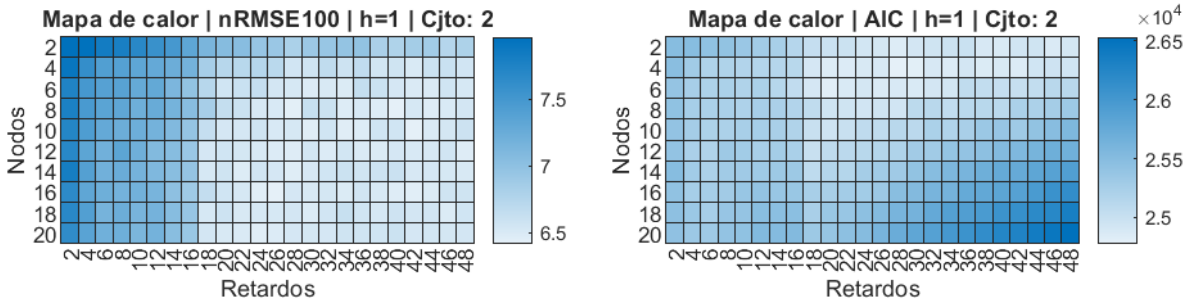


Fig. 4.1 Mapas de calor de la búsqueda en rejilla para el modelo NAR.

4.2 Modelos NARX utilizando el conjunto de variables original

El segundo enfoque del estudio fue el desarrollo de modelos NARX utilizando las configuraciones de los modelos NAR como base. Para los horizontes de 1 a 6 horas, estos modelos son descritos por la ec. (4.3). Para el horizonte de 24 horas se utilizó la ec. (4.4). En estas ecuaciones $\mathbf{x}_k$ es el vector de entradas en el instante k, y n_x es el número de retardos de las entradas.

$$y_{k+h} = \hat{f}\left(y_k, \dots, y_{k-(n_y-1)}, \mathbf{x}_k, \dots, \mathbf{x}_{k-(n_x-1)}\right) + \varepsilon_{k+h} \tag{4.3}$$

$$y_{k+24} = \hat{f}\left(y_k, y_{k-24}, \dots, y_{k-24(n_y-1)}, \mathbf{x}_k, \mathbf{x}_{k-24}, \dots, \mathbf{x}_{k-24(n_x-1)}\right) + \varepsilon_{k+24} \tag{4.4}$$

Dada la restricción de que todas las entradas comparten los mismos retardos, se realizó una búsqueda en rejilla simplificada para la primera variable, considerando un rango retardos de la entrada entre 8 y 48, en pasos de a 8, para cada variable. En el caso de 24-horas, se utilizó un rango de 2 a 7 retardos de las entradas separados en intervalos de 24 horas. En la Fig. 4.2 se muestran mapas de calor que ilustran este proceso de búsqueda, mostrando resultados para el horizonte de 1-hora en el subconjunto de validación.

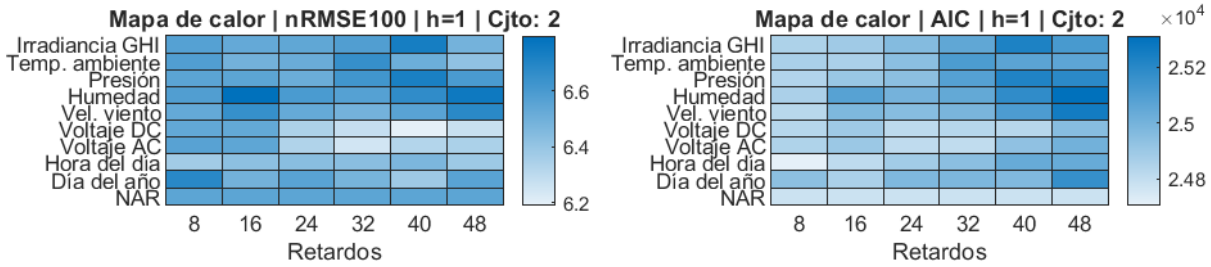


Fig. 4.2 Ejemplo de búsqueda en rejilla para añadir una única variable. Conjunto original.

Luego de la selección de variables, las configuraciones que obtuvieron el mejor desempeño final en el subconjunto de ajuste se listan en la Tabla 4.2. En el caso del primer horizonte, ninguna red NARX mejoró los resultados con respecto al modelo NAR, y las variables más frecuentemente seleccionadas fueron la temperatura ambiente y la hora del día.

Tabla 4.2 Configuración de modelos NARX en el conjunto de variables original.

Horizonte	Nodos	n_y	Entradas exógenas (en orden de selección)	n_x
1	4	28	Ninguna	N/A
2	4	30	Voltaje DC	32
3	6	22	Vel. del viento, hora del día, t.ª ambiente y GHI	8
4	6	28	Temperatura ambiente y hora del día	8
5	6	22	Temperatura ambiente y hora del día	8
6	4	22	Temperatura ambiente y hora del día	8
24	3	5 (días)	Presión y humedad	3 (días)

4.3 Modelos NARX utilizando el conjunto de variables expandido

Se repitió el proceso de entrenamiento y búsqueda de redes NARX utilizando variables del conjunto de datos expandido y las configuraciones que obtuvieron el mejor desempeño final en el subconjunto de ajuste se listan en la Tabla 4.3.

Tabla 4.3 Configuración de modelos NARX en el conjunto de variables expandido.

Horizonte	Nodos	n_y	Entradas exógenas (en orden de selección)	n_x
1	4	28	Zénit y temperatura de módulo	8
2	4	30	Zénit, temperatura ambiente y voltaje DC	8
3	6	22	Zénit y temperatura de celda	8
4	6	28	Zénit y temperatura de celda	8
5	6	22	Temperatura. ambiente y azimut	8
6	4	22	Zénit y temperatura de módulo	8
24	3	5 (días)	Presión, zénit, potencia DC modelada y humedad	3 (días)

En este caso las variables más frecuentemente seleccionadas fueron el ángulo de zénit solar y una de tres variables de temperatura. La selección del zénit como variable exógena en casi todos los modelos sugiere que agregarla al conjunto de datos es una mejora importante.

En la Fig. 4.3 se muestran ejemplos mediante mapas de calor, obtenidos del proceso de búsqueda de la primera variable exógena, dónde se destacan las variables de azimut, zénit y tiempo solar, para el horizonte de 1 hora, evaluando en el subconjunto de validación.

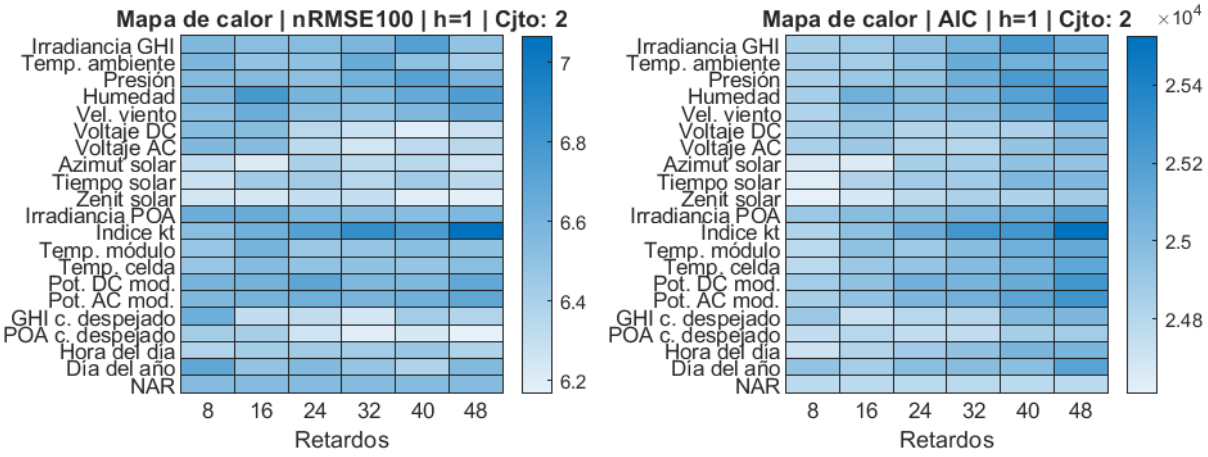


Fig. 4.3 Ejemplo de búsqueda en rejilla para añadir una única variable. Conjunto expandido.

El caso de 24 horas se muestra en la Fig. 4.4, en el que la variable con mayor impacto por sí sola fue la presión atmosférica, lo que llama la atención dada su baja correlación lineal con la potencia como se señaló en el análisis de la sección [3.3.2].

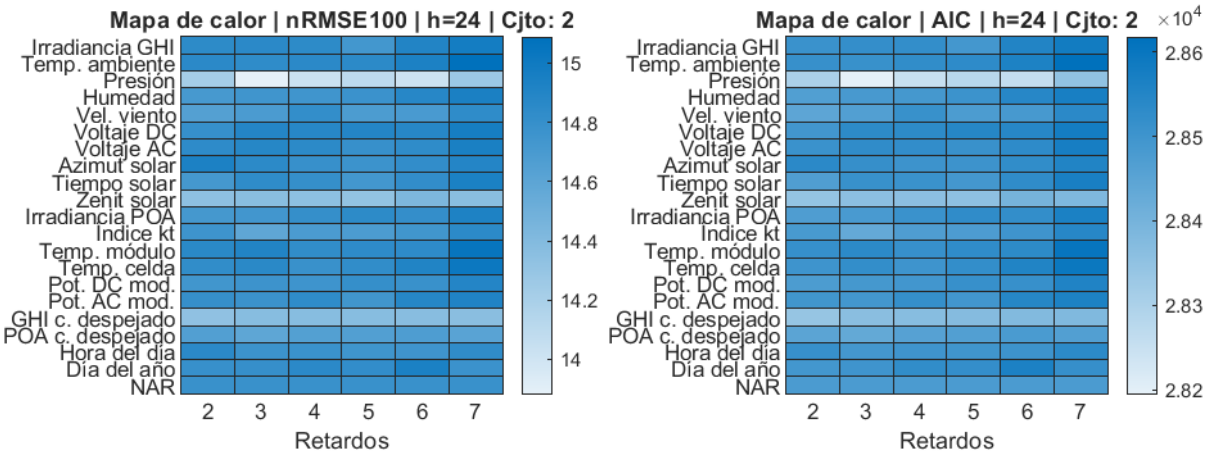


Fig. 4.4 Ejemplo de búsqueda en rejilla para añadir una única variable. Horizonte de 24-horas.

4.4 Modelos NARX utilizando sólo variables predichas

Para mejorar el desempeño de los modelos NARX se intentó aprovechar que algunas de las variables exógenas pueden ser predichas de antemano hasta un horizonte arbitrario. Se probó realizar una selección de características en un espacio de búsqueda reducido solamente a las variables: azimut, zénit, tiempo solar, irradiancias de cielo despejado (GHI y POA), hora del día y día del año. Las cuales fueron adelantadas en el tiempo (time-shifted) en el conjunto de datos.

La ecuación que describe los modelos para los horizontes de 1 a 6 horas es la ec. (4.5), y para el caso de 24 horas se usa la ec. (4.6). En estas, se utilizó el vector $\mathbf{s}_k$ para indicar que la entrada exógena corresponde a una entrada del subconjunto de variables predecibles. De este modo, un valor de n_x igual a 1 representa que sólo se utilizó el retardo 0. Por último, las configuraciones que obtuvieron el mejor desempeño se resumen en la Tabla 4.4. Para mayor claridad se indicaron explícitamente los retardos de entrada respecto del horizonte objetivo.

$$y_{k+h} = \hat{f}\left(y_k, \dots, y_{k-(n_y-1)}, \mathbf{s}_{k+h}, \dots, \mathbf{s}_{k+h-(n_x-1)}\right) + \varepsilon_{k+h} \tag{4.5}$$

$$y_{k+24} = \hat{f}\left(y_k, y_{k-24}, \dots, y_{k-24(n_y-1)}, \mathbf{s}_{k+24}, \mathbf{s}_k, \dots, \mathbf{s}_{k-24(n_x-2)}\right) + \varepsilon_{k+24} \tag{4.6}$$

Tabla 4.4 Configuración de modelos NARX que usan sólo variables predichas.

Horizonte	Nodos	n_y	Entradas exógenas (en orden de selección)	n_x	Retardos efectivos de la entrada
1	4	28	Irr. POA de cielo despejado, zénit y hora del día	2	[0,1]
2	4	30	GHI (cielo despejado) e Irr. POA (cielo despejado)	3	[0,1,2]
3	6	22	GHI (cielo despejado)	4	[0,1,2,3]
4	6	28	Tiempo solar	1	[0]
5	6	22	Tiempo solar y zénit	4	[0,1,2,3]
6	4	22	Irr. POA (cielo despejado) y tiempo solar	2	[0,1]
24	3	5 (días)	GHI (cielo despejado)	1	[0,24,48,72,96]

4.5 Modelos NARX combinando enfoques previos

Otro intento de mejorar los modelos NARX anteriores fue adelantar las variables fácilmente predecibles y reemplazarlas en el conjunto de datos expandido para realizar la selección de características con variables desplazadas y no-desplazadas. En este caso, los retardos de entrada se conservaron iguales a los de los modelos NARX entrenados en el conjunto expandido. El modelo equivalente es descrito por la ec. (4.7) para los horizontes 1-6, y por la ec. (4.8) para el caso de 24 horas. En estas, el vector $\mathbf{s}_k$ representa las entradas desplazables, y $\mathbf{x}_k$ las no-desplazables. Las configuraciones que obtuvieron el mejor desempeño se resumen en la Tabla 4.5.

$$y_{k+h} = \hat{f}\left(y_k, \dots, y_{k-(n_y-1)}, \mathbf{x}_k, \dots, \mathbf{x}_{k-(n_x-1)}, \mathbf{s}_{k+h}, \dots, \mathbf{s}_{k+h-(n_x-1)}\right) + \varepsilon_{k+h} \quad (4.7)$$

$$y_{k+24} = \hat{f}\left(y_k, \dots, y_{k-24(n_y-1)}, \mathbf{x}_k, \dots, \mathbf{x}_{k-24(n_x-1)}, \mathbf{s}_{k+24}, \dots, \mathbf{s}_{k-24(n_x-2)}\right) + \varepsilon_{k+24} \quad (4.8)$$

Tabla 4.5 Configuración de modelos NARX con combinación de variables desplazadas.

Horizonte	Nodos	n_y	Entradas exógenas ("s" para entradas adelantadas)	n_x
1	4	28	GHI de c. despejado (s) y velocidad del viento (x)	8
2	4	30	Irr. POA de c. despejado (s), t.ª ambiente (x) y zénit (s)	8
3	6	22	Zénit (s), t.ª ambiente (x) e irr. POA de c. despejado (s)	8
4	6	28	T.ª ambiente (x) e irr. POA de c. despejado (s)	8
5	6	22	T.ª de celda (x) e irr. POA de c. despejado (s)	8
6	4	22	Tiempo solar (s) y t.ª ambiente (x)	8
24	3	5 (días)	Presión (x), GHI de c. despejado (s), día del año (s), irr. POA (x) y t.ª de celda (x)	3 (días)

4.6 Inclusión de variable adicional en el modelo día-adelante con clustering de irradiancia

De manera exploratoria, se intentó mejorar el desempeño del modelo de Nespoli et al. [17] agregándole una variable adicional. Estas se resumen por red en la Tabla 4.6.

Tabla 4.6 Configuración seleccionada para cada red del modelo

Red	Nodos	Entrada adicional
Red 1: Irradiancia	5	Presión atmosférica media
Red 2: Potencia (soleado)	5	Irradiancia POA de cielo despejado media
Red 3: Potencia (nublado)	7	Presión atmosférica media

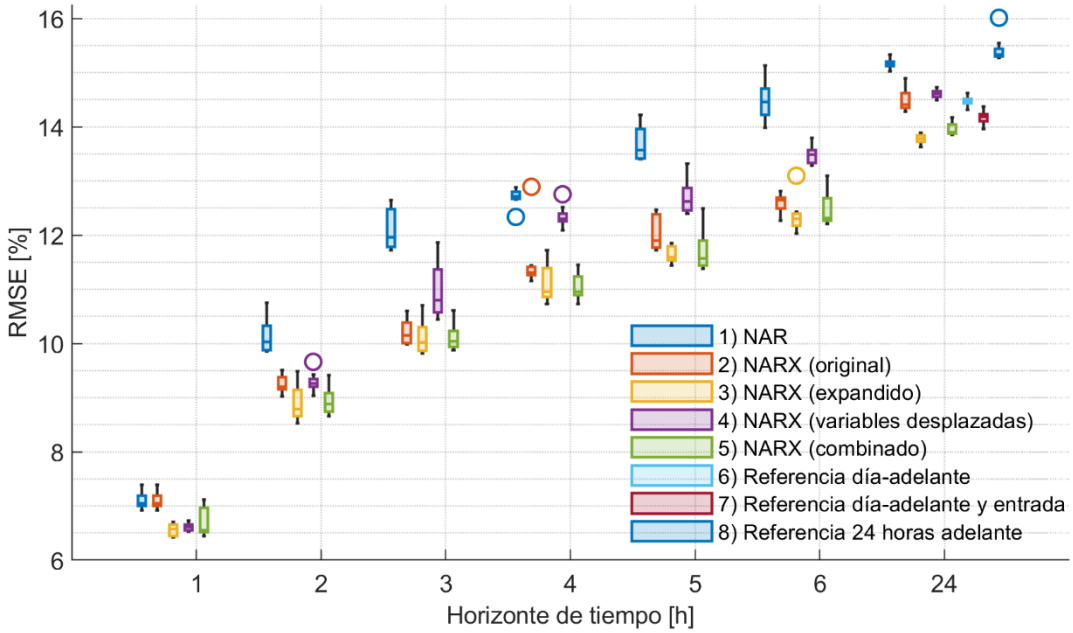
5 Resultados y discusión

5.1 Introducción

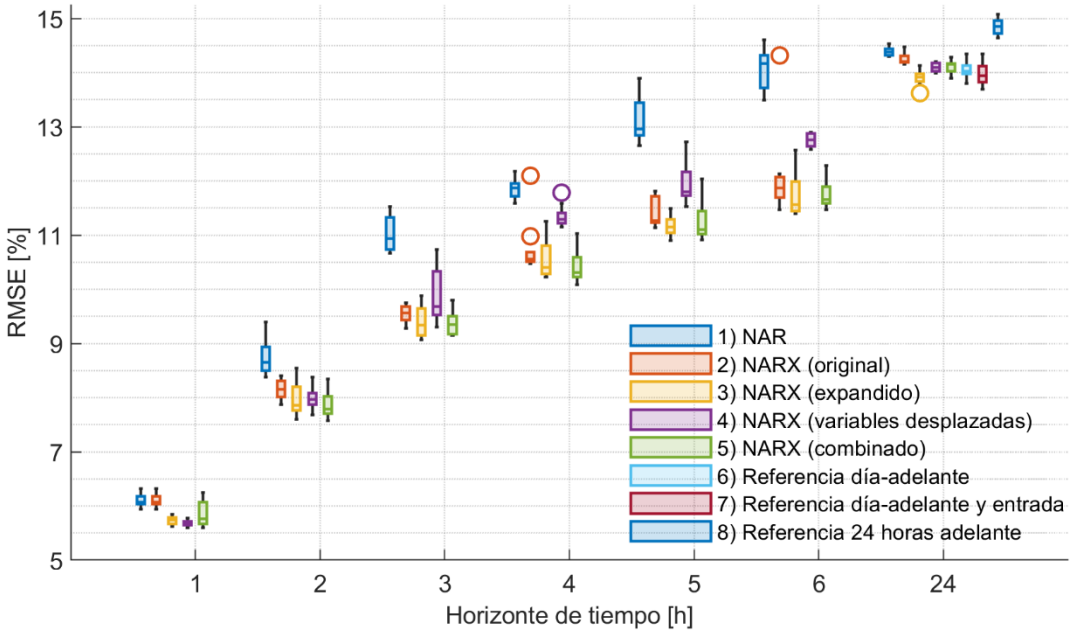
Esta sección presenta los resultados experimentales obtenidos en los distintos horizontes de tiempo. Inicialmente se presenta una comparación de métricas de error de todos los modelos, seguido de un análisis de error para dos variantes relevantes. Por último, se comenta el efecto de incluir variables de los modelos de física en el proceso de selección de características. Para la presentación de los resultados, los subconjuntos de entrenamiento, validación y ajuste son agrupados y denominados subconjuntos de desarrollo.

5.2 Error general por horizonte

Para la comparación general se usó el indicador RMSE. Este fue calculado para cada red y agrupado según tipo de modelo y horizonte de tiempo. Los resultados se presentan por medio de un gráfico de cajas y bigotes en la Fig. 5.1. Se destaca que los modelos que utilizan el conjunto de datos expandido obtienen mejores resultados y con una menor dispersión que los que usan el conjunto original. Por otro lado, los intentos de adelantar variables no tuvieron un impacto considerable y empeoraron los resultados en algunos casos. En el caso de 24-horas, el desempeño de los modelos NARX entrenados fue comparable al de los modelos de la literatura. De manera complementaria a los gráficos, se muestran las medianas del RMSE en el conjunto de prueba en la Tabla 5.1.



(a)



(b)

Fig. 5.1. Comparación general del error según horizonte de tiempo.
(a) subconjuntos de desarrollo; **(b)** subconjunto de prueba

Tabla 5.1 Mediana del RMSE porcentual de los modelos en el subconjunto de prueba.

Modelo / Horizonte	1	2	3	4	5	6	24/día
Persistencia	11,5	20,7	28,6	34,8	39,1	41,7	17,1
NAR	6,1	8,7	10,9	11,9	13,0	14,2	14,4
NARX (original)	6,1	8,2	9,6	10,6	11,3	11,9	14,2
NARX (expandido)	5,7	7,9	9,3	10,4	11,1	11,6	13,9
NARX (variables desplazadas)	5,7	8,0	9,7	11,3	11,8	12,8	14,1
NARX (combinado)	5,8	7,8	9,4	10,3	11,1	11,7	14,0
Referencia día-adelante	-	-	-	-	-	-	14,0
Referencia día-adelante y entrada	-	-	-	-	-	-	13,9
Referencia 24 horas adelante	-	-	-	-	-	-	14,9

5.3 Análisis de error

La comparación más relevante del estudio se realiza entre dos variantes de modelos NARX, la primera entrenada con el conjunto de variables original y las configuraciones de la Tabla 4.2, y la segunda entrenada con el conjunto expandido y las configuraciones de la Tabla 4.3. En esta sección se presentan análisis del error para los casos de predicción 1-hora adelante. Para esto se seleccionó una red representativa, comparando su RMSE con el de la mediana de su grupo en el subconjunto de ajuste y eligiendo la que tuviera el valor más cercano. El proceso de análisis consiste en regresiones entre objetivos y salidas, y correlaciones del error sugeridas en [24].

Primero se muestran los análisis para el caso del modelo que utiliza las variables originales. La Fig. 5.2 muestra los gráficos de regresión e histogramas del error. En este caso el coeficiente R

corresponde al coeficiente de correlación de Pearson. Al comparar el desempeño entre los conjuntos de desarrollo y de prueba, se ve que este es similar, indicando que la capacidad de generalización del modelo es adecuada. Los histogramas muestran que no hay una desviación considerable de la distribución del error. La Fig. 5.3 muestra la autocorrelación del error, para la cual la mayoría de los retardos tienen una correlación cercana a cero, lo que sugiere que existe poca posibilidad de mejora de los resultados, y que su ajuste es apropiado. En este caso no se muestra una correlación cruzada con las entradas debido a que en este horizonte el mejor modelo fue el modelo NAR.

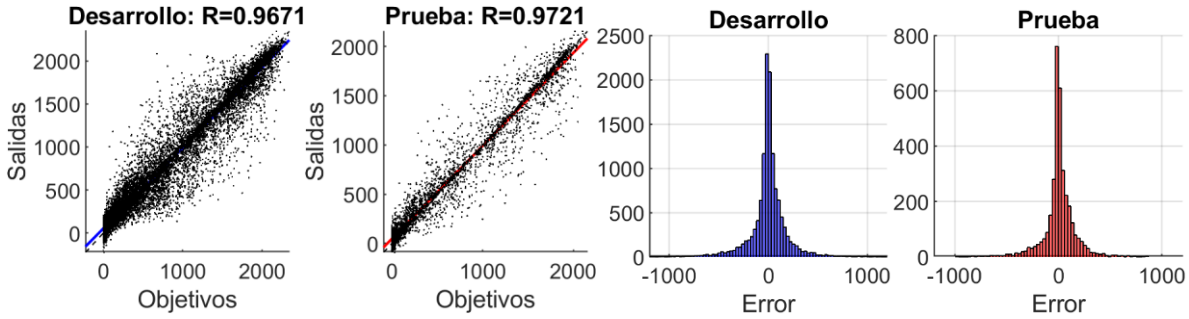


Fig. 5.2 Regresiones e histogramas del modelo NARX en el conjunto de variables original.

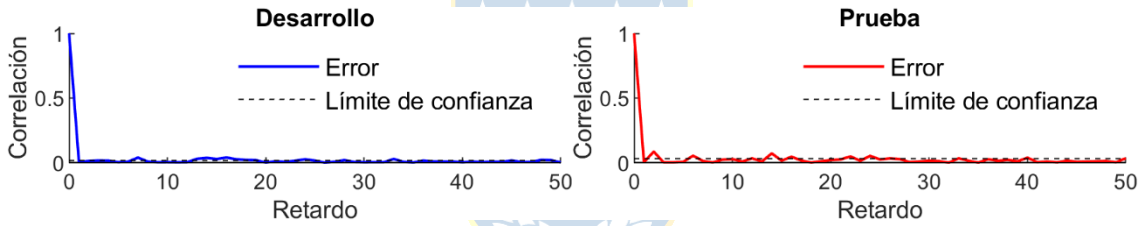


Fig. 5.3 Autocorrelación del error del modelo en el conjunto de datos original.

Luego se muestran los análisis para el caso del modelo que utiliza el conjunto de datos expandido. La Fig. 5.4 muestra los gráficos de regresión e histogramas del error. En este caso la capacidad de generalización del modelo también se ve adecuada, y al compararlos con los del modelo anterior no se observa una diferencia clara. La Fig. 5.5 muestra la autocorrelación de los errores, la cual muestra una pequeña mejora en el conjunto de prueba. La Fig. 5.6 muestra las correlaciones entrada-salida para las dos variables exógenas. En este caso el retardo 0 de la temperatura de módulo presenta una correlación un poco más elevada debido a que este retardo no se incluyó en los modelos. Para el resto de los retardos, la mayoría son cercanos a cero. En el conjunto de prueba, la correlación aumenta de manera leve, lo que sugiere que continuar entrenando la red causará un sobreajuste.

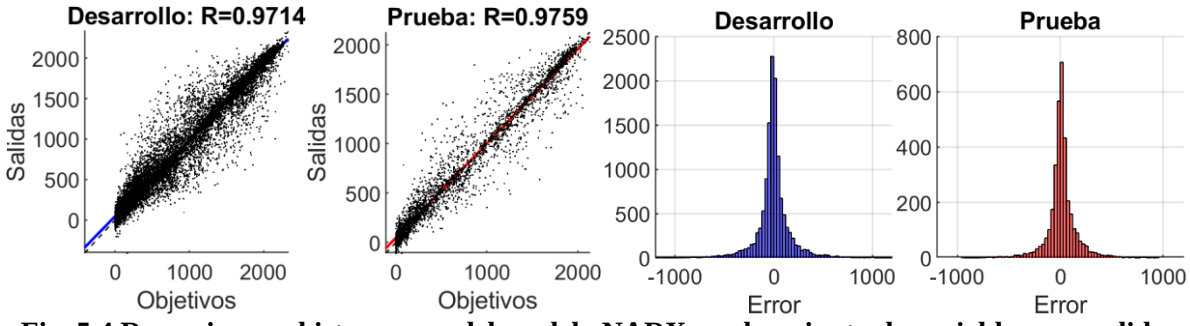


Fig. 5.4 Regresiones e histogramas del modelo NARX en el conjunto de variables expandido.

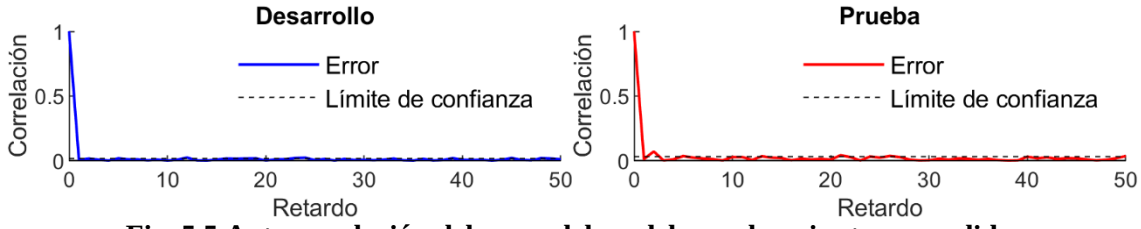


Fig. 5.5 Autocorrelación del error del modelo en el conjunto expandido.

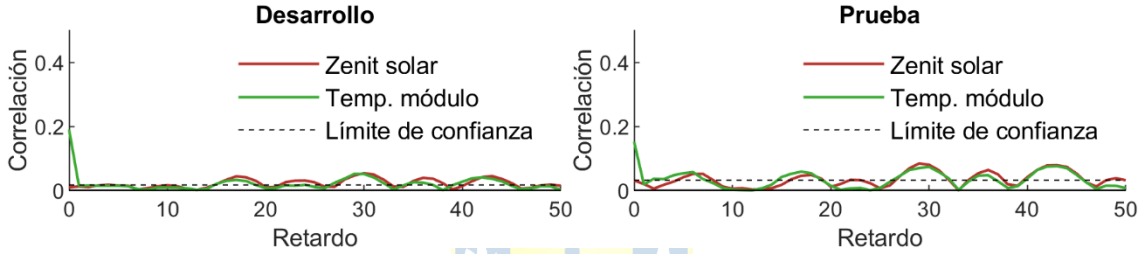


Fig. 5.6 Correlación error-entrada del modelo en el conjunto expandido.

5.4 Efecto de incluir variables de modelos de física

Tomando como referencia la Tabla 5.1 y comparando los modelos NARX del conjunto de datos original con los del conjunto de datos expandido se puede observar que incluir variables de modelos de física en el proceso de selección de características contribuye a reducir el error de los modelos. En términos de RMSE, el error disminuye en un 0.2% a un 0.4% de la potencia nominal de la planta.

De la selección de modelos se determinó que las variables más influyentes fueron el ángulo de zénit solar y las temperaturas de módulo o de celda. Como ejemplo, en la Fig. 5.7 se muestra el impacto que tiene agregar una única variable exógena al modelo NAR en el subconjunto de prueba en dos horizontes de tiempo. Una observación es que, si bien las variables del modelo de cielo despejado tienen un impacto, estas fueron raramente seleccionadas y dependen del ángulo de zénit calculado.

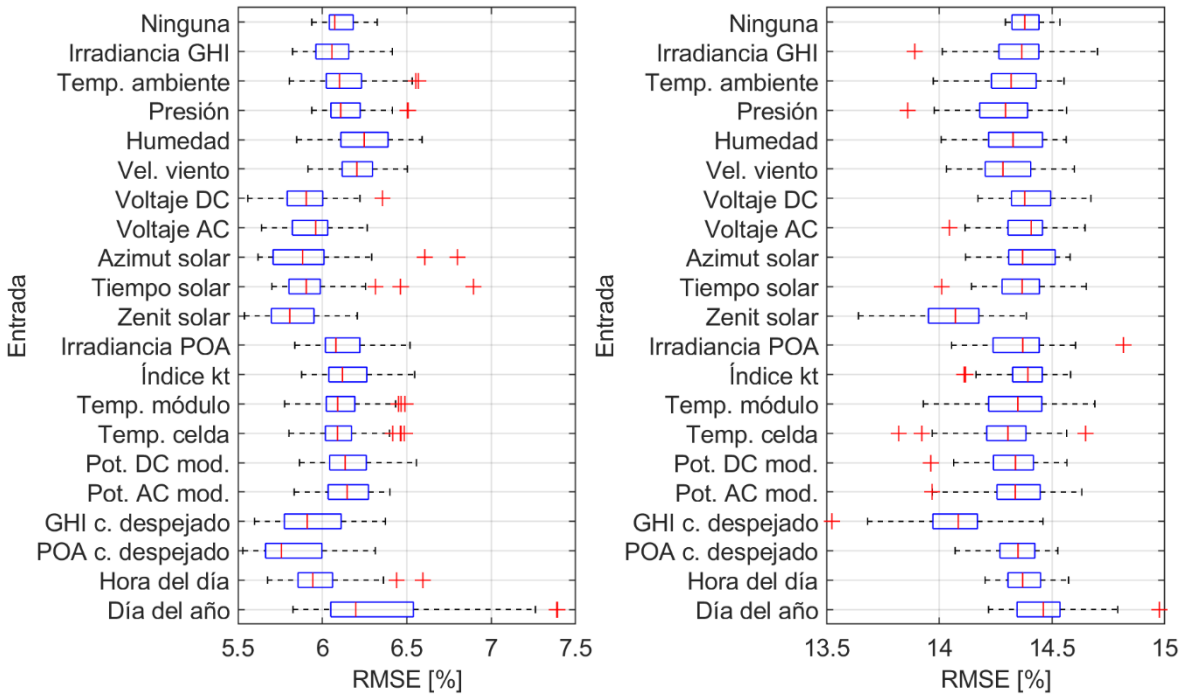


Fig. 5.7 Efecto de agregar una única variable exógena a los modelos NAR.
(a) horizonte de 1-hora; **(b)** horizonte de 24-horas; se eliminaron los peores valores atípicos para mejorar la legibilidad.

6 Conclusiones

6.1 Sumario

El objetivo principal de este estudio fue la utilización de modelos basados en redes neuronales artificiales para la creación de pronósticos de generación de energía fotovoltaica, su comparación y mejora mediante incorporación de información física, utilizando datos locales de la planta experimental ubicada en la Universidad de Concepción. Para esto se realizó una expansión de la base de datos incorporando variables obtenidas con modelos de posición solar, de cielo despejado, y de planta PV. La base de datos fue dividida en bloques para la evaluación del error de generalización. Luego se entrenaron y evaluaron los modelos utilizando procesos de selección de hiperparámetros, selección de características y validación cruzada, utilizando distintos conjuntos de variables. Además, se tomó como referencia algunos modelos de la literatura para utilizar como referencia.

Para asegurar la validez de las comparaciones se entrenaron múltiples redes neuronales para cada configuración y se evaluó la mediana de sus indicadores de error, mitigando el efecto que tiene la inicialización de parámetros en el desempeño. Los modelos más utilizados fueron modelos NARX, los cuales fueron sometidos a los mismos procesos de selección de hiperparámetros y de características, permitiendo atribuir las diferencias en el desempeño a la expansión de la base de datos.

El estudio se limitó a evaluar predicciones con resolución horaria y en horizontes de 1-6 horas, y 24-horas o día-adelante. Se optó por utilizar redes poco profundas y elegir modelos priorizando su simplicidad y capacidad de generalización. Para esto se hizo uso del Criterio de Información de Akaike durante la validación cruzada.

6.2 Conclusiones

La inclusión de variables de modelos físicos puede aumentar la precisión de los pronósticos de manera leve pero consistente. En términos de RMSE, el error disminuyó entre un 0.2% y un 0.4% de la potencia nominal de la planta en los distintos horizontes de tiempo.

Algunas de las variables más comúnmente seleccionadas fueron el ángulo de zénit (o elevación) y las temperaturas de módulo o celda, lo que sugiere que la inclusión de modelos de posición solar y del estado de la planta son importantes para la mejora.

A pesar de su alta correlación con la potencia, las variables de irradiancia, tanto medida como calculada, rara vez causaron una mejora del desempeño. Una posible explicación es que la

información que contienen es redundante con la serie de tiempo de la potencia. Por otro lado, las variables de irradiancia de cielo despejado pueden ayudar a mejorar las predicciones.

La comparación final de los modelos en el horizonte de 24 horas arrojó resultados similares a los que se obtienen replicando modelos de la literatura. Esto indica que no es tan simple mejorar los resultados con la metodología propuesta. Cabe señalar que en este horizonte los pronósticos suelen beneficiarse del uso de modelos de predicción meteorológica numérica.

6.3 Trabajo futuro

La investigación realizada deja varias áreas abiertas para futuras investigaciones. Un enfoque común es investigar el uso de modelos con mayor complejidad para minimizar el error, algunos ejemplos pueden ser el desarrollo de modelos específicos por estación del año o el uso de redes neuronales profundas. Sin embargo, existe una alta necesidad de profundizar en la implementación práctica de los pronósticos, así como en la explicabilidad de los métodos para mejorar la confiabilidad de las predicciones. Esto es de especial importancia porque en la práctica es necesario justificar la toma de decisiones [8]. Por último, también existe la posibilidad de variar los horizontes de tiempo y tiempo de muestreo, a fin de abarcar predicciones de muy corto plazo.

7 Bibliografía

[1] "What is Climate Change?" (https://www.un.org/en/climatechange/what-is-climate-change), United Nations. (accessed Mar/18, 2023).

[2] "Climate Change." (https://www.un.org/en/global-issues/climate-change), United Nations. (accessed Mar/18, 2023).

[3] R. Ahmed, V. Sreeram, Y. Mishra, and M. D. Arif, "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization," *Renewable and Sustainable Energy Reviews,* vol. 124, p. 109792, 2020/05/01/ 2020, doi: https://doi.org/10.1016/j.rser.2020.109792.

[4] "Causes and Effects of Climate Change." (https://www.un.org/en/climatechange/science/causes-effects-climate-change), United Nations. (accessed Mar/18, 2023).

[5] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Conversion and Management,* vol. 198, p. 111799, 2019/10/15/ 2019, doi: https://doi.org/10.1016/j.enconman.2019.111799.

[6] "Energy." (https://www.un.org/sustainabledevelopment/es/energy/), United Nations. (accessed Dec/26, 2023).

[7] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Conversion and Management,* vol. 156, pp. 459-497, 2018/01/15/ 2018, doi: https://doi.org/10.1016/j.enconman.2017.11.019.

[8] A. Alcañiz, D. Grzebyk, H. Ziar, and O. Isabella, "Trends and gaps in photovoltaic power forecasting with machine learning," *Energy Reports,* vol. 9, pp. 447-471, 2023/12/01/ 2023, doi: https://doi.org/10.1016/j.egyr.2022.11.208.

[9] G. Notton *et al.*, "Intermittent and stochastic character of renewable energy sources: Consequences, cost of intermittence and benefit of forecasting," *Renewable and Sustainable Energy Reviews,* vol. 87, pp. 96-105, 2018/05/01/ 2018, doi: https://doi.org/10.1016/j.rser.2018.02.007.

[10] D. V. Pombo, M. J. Rincón, P. Bacher, H. W. Bindner, S. V. Spataru, and P. E. Sørensen, "Assessing stacked physics-informed machine learning models for co-located wind–solar power forecasting," *Sustainable Energy, Grids and Networks,* vol. 32, p. 100943, 2022/12/01/ 2022, doi: https://doi.org/10.1016/j.segan.2022.100943.

[11] C. Brancucci Martinez-Anido *et al.*, "The value of day-ahead solar power forecasting improvement," *Solar Energy,* vol. 129, pp. 192-203, 2016/05/01/ 2016, doi: https://doi.org/10.1016/j.solener.2016.01.049.

[12] L. Gigoni *et al.*, "Day-Ahead Hourly Forecasting of Power Generation From Photovoltaic Plants," *IEEE Transactions on Sustainable Energy,* vol. 9, no. 2, pp. 831-842, 2018, doi: 10.1109/TSTE.2017.2762435.

[13] A. Parasyris, G. Alexandrakis, G. V. Kozyrakis, K. Spanoudaki, and N. A. Kampanis, "Predicting Meteorological Variables on Local Level with SARIMA, LSTM and Hybrid Techniques," *Atmosphere,* vol. 13, no. 6, p. 878, 2022. [Online]. Available: https://www.mdpi.com/2073-4433/13/6/878.

[14] R. A. Rajagukguk, R. Kamil, and H.-J. Lee, "A Deep Learning Model to Forecast Solar Irradiance Using a Sky Camera," *Applied Sciences,* vol. 11, no. 11, p. 5049, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/11/5049.

[15] S. Mishra and P. Palanisamy, "Multi-time-horizon Solar Forecasting Using Recurrent Neural Network," in *2018 IEEE Energy Conversion Congress and Exposition (ECCE)*, 23-27 Sept. 2018 2018, pp. 18-24, doi: 10.1109/ECCE.2018.8558187.

[16] M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz, "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," *Renewable and Sustainable Energy Reviews,* vol. 27, pp. 65-76, 2013/11/01/ 2013, doi: https://doi.org/10.1016/j.rser.2013.06.042.

[17] A. Nespoli *et al.*, "Day-Ahead Photovoltaic Forecasting: A Comparison of the Most Effective Techniques," *Energies,* vol. 12, no. 9, p. 1621, 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/9/1621.

[18] M. H. Alomari, O. Younis, and S. Hayajneh, "A predictive model for solar photovoltaic power using the Levenberg-Marquardt and Bayesian regularization algorithms and real-time weather data," *International Journal of Advanced Computer Science and Applications,* vol. 9, no. 1, pp. 347-353, 2018.

[19] D. V. Pombo, P. Bacher, C. Ziras, H. W. Bindner, S. V. Spataru, and P. E. Sørensen, "Benchmarking physics-informed machine learning-based short term PV-power forecasting tools," *Energy Reports,* vol. 8, pp. 6512-6520, 2022/11/01/ 2022, doi: https://doi.org/10.1016/j.egyr.2022.05.006.

[20] A. Sabadus *et al.*, "A cross-sectional survey of deterministic PV power forecasting: Progress and limitations in current approaches," *Renewable Energy,* vol. 226, p. 120385, 2024/05/01/ 2024, doi: https://doi.org/10.1016/j.renene.2024.120385.

[21] F. Wang *et al.*, "Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting," *Energy Conversion and Management,* vol. 181, pp. 443-462, 2019/02/01/ 2019, doi: https://doi.org/10.1016/j.enconman.2018.11.074.

[22] M. S. Hossain and H. Mahmood, "Short-Term Photovoltaic Power Forecasting Using an LSTM Neural Network and Synthetic Weather Forecast," *IEEE Access,* vol. 8, pp. 172524-172533, 2020, doi: 10.1109/ACCESS.2020.3024901.

[23] M. Husein and I.-Y. Chung, "Day-Ahead Solar Irradiance Forecasting for Microgrids Using a Long Short-Term Memory Recurrent Neural Network: A Deep Learning Approach," *Energies,* vol. 12, no. 10, p. 1856, 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/10/1856.

[24] M. T. Hagan, H. B. Demuth, and M. Beale, *Neural network design*. PWS Publishing Co., 1997.

[25] "PV Performance Modeling Collaborative (PVPMC) Modeling Guide." (https://pvpmc.sandia.gov/modeling-guide/), Sandia National Laboratories. (accessed 09/Abril, 2024).

[26] J. S. Stein, W. F. Holmgren, J. Forbess, and C. W. Hansen, "PVLIB: Open source photovoltaic performance modeling functions for Matlab and Python," in *2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC)*, 5-10 June 2016 2016, pp. 3425-3430, doi: 10.1109/PVSC.2016.7750303.

[27] D. L. King, J. A. Kratochvil, and W. E. Boyson, *Photovoltaic array performance model*. Citeseer, 2004.

[28] H. D. Tran, N. Muttil, and B. J. C. Perera, "Selection of significant input variables for time series forecasting," *Environmental Modelling & Software,* vol. 64, pp. 156-163, 2015/02/01/ 2015, doi: https://doi.org/10.1016/j.envsoft.2014.11.018.

[29] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics,* vol. 35, no. 6, pp. 2769-2794, 26, 2007. [Online]. Available: https://doi.org/10.1214/009053607000000505.

[30] S. Liu. "Distance correlation." (https://www.mathworks.com/matlabcentral/fileexchange/39905-distance-correlation), MATLAB Central File Exchange. (accessed 08/May, 2024).

[31] M. R. Arahal, A. Cepeda, and E. F. Camacho, "INPUT VARIABLE SELECTION FOR FORECASTING MODELS," *IFAC Proceedings Volumes,* vol. 35, no. 1, pp. 463-468, 2002/01/01/ 2002, doi: https://doi.org/10.3182/20020721-6-ES-1901.00730.

[32] D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," in *1990 IJCNN International Joint Conference on Neural Networks*, 17-21 June 1990 1990, pp. 21-26 vol.3, doi: 10.1109/IJCNN.1990.137819.

[33] "Deep Learning Toolbox Documentation." (https://www.mathworks.com/help/deeplearning/index.html), The Mathworks, Inc. (accessed 7/Dec, 2021).

[34] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Transactions on Neural Networks,* vol. 5, no. 6, pp. 989-993, 1994, doi: 10.1109/72.329697.