Ingenio Magno
ISSN 2145-9282 - ISSN 2422-2399 (En línea)
UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA
T U N J A
REVISTA INGENIO MAGNO ISSN digital 2422-2399, Volumen 16 No 1, año 2025

RED NEURONAL ARTIFICIAL PARA LA ESTIMACIÓN DE DATOS FALTANTES DE PRECIPITACIÓN EN LA ESTACIÓN EL ENCANTO DEL MUNICIPIO DE TUNJA, BOYACÁ

Artificial Neural Network for the Estimation of Missing Precipitation Data at the EL ENCANTO Station in the Municipality of Tunja, Boyacá

David Felipe Bermúdez Duarte[a], Laura Natalia Garavito Rincón[b], Sergio David Torres Piraquive[c]

[a,b] Facultad Ingeniería Civil, Especialización en Ingeniería Hidroambiental Universidad Santo Tomás – Seccional Tunja; david.bermudez@usantoto.edu.co, laura.garavito@usantoto.edu.co
[c]Instituto de Hidrología, Meteorología y Estudios Ambientales; sdtorres@ideam.gov.co

Recibido: 02 de marzo 2026. Aceptado: 23 de abril 2026 Publicado: 05 de agosto 2026.

***Resumen*—** La ausencia y discontinuidad de registros de precipitación limitan la evaluación hidrológica, la planificación del recurso hídrico y la gestión del riesgo climático en zonas de montaña como Tunja, Boyacá. Esta investigación desarrolló y evaluó un modelo de red neuronal artificial para estimar datos faltantes de precipitación mensual en la estación El Encanto, usando como insumos registros de estaciones vecinas con comportamiento pluviométrico correlacionado, variables geográficas y el Índice Oceánico Niño (ONI). La metodología incluyó la depuración de registros del DHIME, la clasificación de los meses según patrones de completitud, la normalización de variables, el entrenamiento con TensorFlow/Keras y la validación mediante MAE, RMSE, R^2, eficiencia de Nash, PBIAS y pruebas estadísticas de comparación. Los resultados muestran un desempeño aceptable para reconstruir la tendencia y estacionalidad mensual (MAE = 19.3 mm; RMSE = 31.6 mm; R^2 = 0.44; PBIAS = 13.6 %), aunque con limitaciones para reproducir eventos extremos y la variabilidad de los meses lluviosos. El enfoque es útil como apoyo técnico para completar series, siempre que sus estimaciones se interpreten con control de calidad hidrológico.

***Palabras clave*—** Redes neuronales artificiales, precipitación mensual, datos faltantes.

***Abstract*—** Missing and discontinuous precipitation records limit hydrological assessment, water-resource planning, and climaterisk management in mountainous areas such as Tunja, Boyacá. This study developed and evaluated an artificial neural network model to estimate missing monthly precipitation data at the El Encanto station, using neighboring rainfall stations with correlated behavior, geographic variables, and the Oceanic Niño Index (ONI) as predictors. The methodology included DHIME data cleaning, classification of monthly completeness patterns, variable normalization, TensorFlow/Keras training, and validation through MAE, RMSE, R^2, Nash efficiency, PBIAS, and statistical comparison tests. The results show acceptable performance in reconstructing monthly trend and seasonality (MAE = 19.3 mm; RMSE = 31.6 mm; R^2 = 0.44; PBIAS = 13.6 %), although the model is limited in reproducing extreme events and high-rainfall variability. The approach is useful as a technical support tool for completing series, provided that its estimates are interpreted with hydrological quality control.

***Keywords*—** Artificial neural networks, monthly precipitation, missing data.

I. INTRODUCCIÓN

La disponibilidad de datos climáticos es un insumo propio de diferentes áreas de ingeniería como la ingeniería hidráulica, civil, ambiental y agrícola, siendo la información climatológica esencial para el análisis y el diseño de obras hidráulicas, la

planificación de sistemas de drenaje, la evaluación de riesgos hidrológicos y la gestión sostenible del recurso hídrico en el caso de fenómenos climáticos extremos (Kisi, 2004). Sin embargo, la falta de registros o la discontinuidad de las estaciones climatológicas son condiciones que aumentan la incertidumbre de los análisis técnicos, limitando la fiabilidad de los estudios hidrológicos y la toma de decisiones en los mismos.

En este contexto, el avance del aprendizaje automático ha permitido proponer estrategias para estimar información hidrometeorológica en escenarios con series incompletas o fenómenos extremos de baja frecuencia. Los modelos de perceptrón multicapa (MLP) han demostrado ser capaces de capturar relaciones no lineales entre estaciones y variables meteorológicas, y enfoques recientes de aprendizaje automático en hidrología destacan su utilidad, pero también advierten sobre la necesidad de evaluar la robustez, interpretabilidad y tamaño muestral (Xu y Liang, 2021; Nearing et al., 2021). Papailiou et al. (2022) contrastaron ensambles de redes neuronales con regresión lineal múltiple en el problema particular de completar lluvia faltante, mientras que Wangwongchai et al. (2023) compararon técnicas estadísticas con técnicas de inteligencia artificial para lluvia diaria; ambos trabajos demuestran que la elección del método debe tener en cuenta un equilibrio entre precisión, transparencia, disponibilidad de estaciones vecinas y costo computacional.

Las redes neuronales recurrentes, como las redes Elman, también pueden modelar dependencias temporales en series hidrológicas (Elman, 1990; Coulibaly et al., 2005). También se han estudiado las redes de función de base radial como alternativas para la estimación de precipitación (Nourani y Hosseini, 2011), ya que utilizan funciones de base gaussiana en sus unidades ocultas y han demostrado capacidad de aproximación en problemas hidrológicos. No obstante, la literatura reciente sobre imputación mensual de lluvia también muestra que los métodos de aprendizaje automático deben compararse con alternativas de interpolación espacial, razón normal, promedios o regresión, porque los métodos tradicionales pueden ser competitivos cuando existe alta correlación entre estaciones vecinas (Pinthong et al., 2024; Wangwongchai et al., 2023).

El monitoreo de la precipitación es un componente clave para la gestión de los recursos hídricos, la planificación del sector agrícola y la evaluación del riesgo climático (Congreso de la República de Colombia, 2012). No obstante, la presencia de datos faltantes en estaciones meteorológicas, asociada a fallas técnicas o limitaciones operativas, compromete la calidad de los análisis hidroclimáticos (Gupta et al., 1999). Esta problemática adquiere especial relevancia en regiones como Boyacá, Colombia, donde la compleja topografía y la alta variabilidad climática exigen series temporales completas y confiables.

Esta problemática adquiere especial relevancia en regiones como Boyacá, Colombia, donde la compleja topografía y la variabilidad climática demandan series de datos completas y de calidad. De acuerdo con el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM, 2020), más del 20 % de las estaciones pluviométricas en zonas de montaña del centro del país presentan discontinuidades significativas en sus registros históricos, con periodos de datos faltantes que superan los cinco años en algunas estaciones. Las redes neuronales artificiales (RNA) se han presentado como un instrumento para estimar datos meteorológicos faltantes gracias a su capacidad para modelar relaciones lineales, no lineales y patrones espaciotemporales.

Por lo tanto, esta investigación tiene su origen en la necesidad de subsanar los vacíos de información en los registros pluviométricos de Boyacá, los cuales limitan la toma de decisiones ambientales acertadas. En este contexto, el IDEAM (2019) reconoce que la discontinuidad de registros en estaciones pluviométricas limita la caracterización de amenazas hidrológicas y la formulación de estrategias de gestión del riesgo. Estudios previos han demostrado la eficacia de las RNA en modelación hidrológica (Smith et al., 2018; Coulibaly et al., 2005), aunque siguen siendo necesarios análisis locales que contrasten el potencial de estos modelos con sus restricciones ante datos escasos, variabilidad espacial y eventos extremos. La presente investigación se basa en marcos teóricos de aprendizaje automático e hidrología, empleando perceptrones multicapa, variables climáticas y validación estadística para mejorar la estimación de precipitación mensual ausente.

La razón de ser de este estudio se basa en tres puntos fundamentales: (1) optimizar el uso de bases de datos incompletas con fines científicos y operativos, (2) proporcionar una metodología que pueda replicarse en regiones que presentan problemas similares y (3) contribuir a la gestión sostenible del agua en situaciones de variabilidad climática (Ordoñez et al., 2024). El objetivo principal es crear un modelo que utilice RNA para reconstruir series mensuales de lluvia en cuatro estaciones representativas (Tunja) con datos de estaciones aledañas como insumo. Aunque el modelo está limitado a agregaciones mensuales (por disponibilidad de datos), sus resultados apoyarán aplicaciones en agricultura, planificación urbana y reducción del riesgo de desastres.

Metodológicamente, la investigación se desarrolla en un enfoque de tres fases: la recopilación y el preprocesamiento de la información climatológica; en la segunda fase se aborda el desarrollo del modelo y, finalmente, en la fase de validación, se evalúa el desempeño del modelo con indicadores como el

coeficiente de determinación (R^2) y el error cuadrático medio (MSE) y validación cruzada frente a datos observados. La investigación presenta avances significativos al adaptar técnicas que emergieron de las redes neuronales artificiales a climas tropicales de alta montaña con características de la dinámica de las precipitaciones marcadamente distintas de las de otras regiones templadas; también plantea un marco metodológico transferible a regiones donde la información hidrometeorológica es escasa en este tipo de países en vías de desarrollo.

II. MATERIALES Y MÉTODOS

La metodología del presente trabajo se desarrolló en cuatro etapas articuladas (Figura 1). Primero, se recopilaron registros históricos de precipitación mensual del DHIME para la estación El Encanto, tomada como referencia de Tunja, y para estaciones vecinas con similitud altitudinal y correlación pluviométrica (Cómbita, Panelas y UPTC). También se incorporó el Índice Oceánico Niño (ONI) trimestral para explorar la relación entre la variabilidad ENSO y la precipitación local. Segundo, se realizó control de calidad para identificar registros ausentes, posibles valores atípicos e inconsistencias de formato. Tercero, cada mes se clasificó según la disponibilidad de información en la estación objetivo y en las estaciones auxiliares. Cuarto, se entrenó y validó el modelo de red neuronal o, cuando no existían insumos suficientes, se aplicó un procedimiento estadístico de imputación, según los cuatro casos siguientes:

Caso 1: Todas las estaciones contaron con datos disponibles para el mes (i), por lo que no fue necesaria ninguna imputación, pero se usaron como datos de entrenamiento y testeo para los algoritmos.

Caso 2: Se encontraron datos disponibles en todas las estaciones excepto en la estación de referencia para el mes (i). En este caso, los datos se imputaron mediante el algoritmo que mejores resultados presentó de acuerdo con la evaluación realizada tanto en el entrenamiento como en el testeo.

Caso 3: La estación de referencia presentó datos disponibles, pero algunas estaciones cercanas registraron valores faltantes. En este caso, no es necesaria ninguna imputación de datos para la estación de referencia.

Caso 4: La estación de referencia tuvo datos faltantes para el mes (i) y una o varias estaciones auxiliares tampoco contaron con información suficiente. En este caso no se forzó la red neuronal; se aplicó una imputación estadística mensual basada en la distribución histórica del mismo mes, calculando media y desviación estándar, generando valores aleatorios no negativos dentro del rango plausible y verificando que el valor imputado conservara la estacionalidad de la serie.

Para desarrollar el modelo, los registros de precipitación del DHIME y el ONI se organizaron en una matriz mensual por fecha y estación. En los casos 1 y 2, donde existía información suficiente para aprender relaciones entre la estación objetivo y las estaciones vecinas, se añadieron variables derivadas como mes, distancia entre estaciones, diferencias altitudinales y valores simultáneos de precipitación en estaciones cercanas. Las variables fueron normalizadas mediante MinMaxScaler antes del entrenamiento, con el fin de evitar que magnitudes distintas dominaran el aprendizaje de la red. La red se implementó en Keras/TensorFlow para un problema de regresión, utilizando Adam como optimizador por su estabilidad en gradientes ruidosos y funciones de activación que restringen la salida a valores físicamente posibles de precipitación.

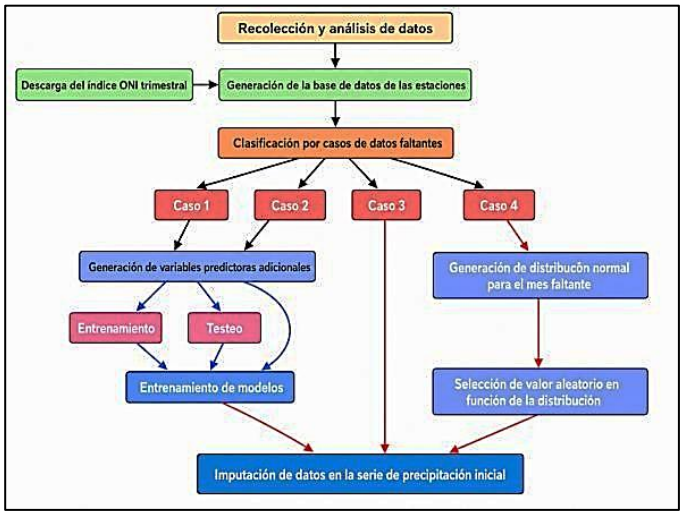

Fig. 1. Ruta metodológica desarrollada para la imputación de datos mensuales de precipitación. Fuente: Autor.

Los datos disponibles se dividieron en conjuntos de entrenamiento y prueba (80 % y 20 %), preservando la estructura temporal para evaluar la capacidad de generalización antes de imputar los valores faltantes reales. En el caso 3, cuando la estación de referencia sí contaba con dato observado pero alguna estación auxiliar no lo tenía, no se imputó el valor de la estación de referencia; el registro se conservó para caracterización y contraste. En el caso 4, al no existir una combinación robusta de entradas para entrenar el modelo, se empleó el procedimiento estadístico descrito anteriormente. Este criterio evita extrapolar la red neuronal fuera de su dominio de entrenamiento y reduce el riesgo de generar valores artificiales sin soporte hidrológico.

A. Redes Perceptrón Multicapa (MLP)

Una MLP estándar está compuesta por tres elementos: la capa de entrada, la oculta y la de salida. Una capa es un conjunto de neuronas (o unidades de procesamiento) que poseen

idénticas conexiones de entrada y salida. En la red MLP convencional, la información (o señal de entrada) solo se propaga hacia adelante. Por lo tanto, la red MLP es también conocida como red de alimentación hacia adelante con múltiples capas. Se calcula la salida de la red MLP, teniendo en cuenta una única capa oculta con h nodos ocultos sigmoideos, una neurona de salida lineal j y una variable de entrada xtxt, de esta forma:

$$Y_k = F\left(\sum_{j=1}^{h} w_j G(S_i) + b_k\right) \quad (1)$$

Donde bk es el sesgo (o umbral) de la neurona de salida k y (F) es su función de activación lineal; (G) es la función sigmoide tangente hiperbólica que se emplea como función de activación para los nodos; (wj) son los pesos de conexión entre las unidades ocultas y las unidades de salida y puede expresarse de la siguiente manera:

$$G(S_i) = \frac{e^{s_i} - e^{-s_i}}{e^{s_i} + e^{-s_i}} \quad (2)$$

Donde (si) es la suma ponderada de toda la información entrante y también se le conoce como la señal de entrada.

$$S_i = \sum_{i=1}^{n} W_i X_i \quad (3)$$

Donde xi son las entradas a la red y wi son los pesos de conexión entre los nodos de las capas de entrada y oculta. Una ventaja importante del MLP es que presenta menor complejidad que algunas arquitecturas temporales y conserva capacidad de mapeo entrada-salida, lo cual resulta adecuado cuando el objetivo principal es estimar una variable mensual a partir de predictores simultáneos y metadatos de estaciones cercanas. Además, el MLP puede entrenarse con retropropagación (Rumelhart et al., 1986) y ha sido aplicado previamente en modelación de lluvia (French et al., 1992). En este estudio, su selección se justifica por la disponibilidad limitada de datos mensuales, la necesidad de evitar modelos excesivamente complejos y la evidencia de correlación espacial entre estaciones vecinas.

B. Red neuronal feedforward con retardos temporales (TLFN)

Para explotar la estructura temporal en los datos, la red neuronal debe tener acceso a esta dimensión temporal. Si bien las MLP son populares en muchas áreas de aplicación, no son adecuadas para procesar secuencias temporales debido a la falta de conexiones de retardo temporal y/o retroalimentaciones necesarias para proporcionar un modelo dinámico. Una forma de introducir "memoria" en la MLP es reemplazando las neuronas en la capa de entrada con una estructura de memoria llamada "línea de retardo por muestreo". Este tipo de MLP se llama red neuronal feedforward con retardos temporales (TLFN) y también se conoce como red neuronal pseudo-dinámica.

La topología TLFN ha sido utilizada con éxito en la predicción de series temporales hidrológicas (por ejemplo, Coulibaly et al., 2001a, b), reconocimiento de patrones temporales (por ejemplo, Principe et al., 2000) y modelado híbrido paralelo (Anctil et al., 2003). Recientemente, se ha demostrado que la TLFN supera a las redes impulsadas dinámicamente en la desagregación de series diarias de precipitación y temperatura (Coulibaly et al., 2005).

C. Red neuronal recurrente (RNN)

La RNN utilizada en este estudio es el tipo básico de RNN de Elman (1990), también conocida como RNN conectada globalmente. La red consta de cuatro capas: la capa de entrada, la oculta, la de contexto y la de salida. Cada unidad de entrada está conectada a cada unidad oculta, al igual que cada unidad de contexto. A su vez, existen conexiones descendentes uno a uno entre los nodos ocultos y las unidades de contexto, lo que conduce a un número igual de unidades ocultas y de contexto. Por lo tanto, las conexiones recurrentes permiten que las unidades ocultas reciclen la información a lo largo de múltiples pasos de tiempo y así descubran la información temporal contenida en la entrada secuencial y relevante para la función objetivo. En la práctica, este método se considera como la mejor técnica en línea para problemas prácticos (Pearlmutter, 1995) y ha sido ampliamente utilizado para el entrenamiento de RNN en diversas aplicaciones (Williams y Zipser, 1995; Feldkamp y Puskorius, 1998; Coulibaly et al., 2001).

D. Red neuronal de función de base radial (RBF)

La red RBF considerada en este estudio también se conoce como la red de función de base radial generalizada, que representa la forma general y típica de la red RBF. La red RBF asume funciones de base gaussiana (o radial) para sus unidades ocultas y combina la ventaja de la generalidad y la reducida complejidad computacional (Specht, 1991), pero no es buena para ignorar entradas irrelevantes o ruidosas. Además, se ha demostrado que las redes RBF con una sola capa oculta de unidades gaussianas son aproximadores universales (Park y Sandberg, 1991).

Básicamente, una red RBF consta de tres capas similares a la red MLP estándar. La capa de entrada pasa las variables a una capa oculta que transforma no linealmente el espacio de entrada; luego, una capa de salida lineal proporciona la respuesta de la red. Aunque la red RBF típica se parece a una red feedforward convencional de tres capas, su funcionamiento

es diferente (Hsu et al., 1995): el aprendizaje de una red de retropropagación puede entenderse como una aproximación estocástica, mientras que el aprendizaje de una red RBF busca una superficie óptima en un espacio multidimensional.

III. DESARROLLO DEL TRABAJO

A. Adquisición de datos

Los datos de precipitación empleados en esta investigación fueron obtenidos a partir del DHIME (Descarga de Datos Hidrometeorológicos), una de las bases de datos oficiales del IDEAM (Instituto de Hidrología, Meteorología y Estudios

Ambientales de Colombia), disponible en https://www.dhime.ideam.gov.co/webgis/home/. Esta plataforma proporciona acceso a series temporales de variables hidroclimáticas, incluyendo precipitación diaria y mensual, registradas por estaciones meteorológicas distribuidas a lo largo del territorio nacional. La selección de las estaciones se realizó considerando dos criterios principales: similitud altitudinal con el área de estudio y disponibilidad de series históricas extensas y continuas, con el fin de garantizar la representatividad climática y la robustez estadística del análisis. El intervalo de estudio fue definido en el rango temporal compartido con mayor cobertura de datos entre las estaciones elegidas, con el objetivo de maximizar la longitud de las series y minimizar la incertidumbre asociada a los registros incompletos, y se extendió desde 1993 hasta 2021. Los datos se descargaron en formato CSV y posteriormente se depuraron, validaron y sometieron a control de calidad para su posterior análisis y modelización.

B. Código preprocesamiento

El código desarrolla un flujo sistemático de procesamiento de datos de precipitación para el municipio de Tunja, comenzando por la definición explícita de los parámetros de entrada, como el período de análisis (1993-2021), la estación de referencia y el conjunto de estaciones pluviométricas seleccionadas. Después se leen los datos crudos de precipitación de archivos en formato CSV y los metadatos asociados a cada estación (coordenadas geográficas y altitud). Los registros se filtran por el intervalo temporal definido y se someten a un preprocesado que incluye la detección de los valores ausentes (NaN), la combinación de los datos a escala mensual y el cálculo de estadísticos básicos por estación.

Finalmente, con esta información se monta una matriz de precipitación mensual, estructurada por fechas y estaciones, la cual se completa con variables geográficas derivadas (distancia entre estaciones, mediante la fórmula de Haversine y la altitud); y, según la configuración de los valores ausentes, se clasifican los datos en cuatro tipos de cobertura y se permite clasificar entre series observadas parcialmente, así como tramos de ausencia sin información. Al final, se generan estadísticas descriptivas (media, percentiles) y productos gráficos como series de tiempo y BOXPLOTS, y se exportan los resultados en un archivo de Excel estructurado y en un zip, listos para su posterior utilización en el entrenamiento de la red neuronal artificial. El diseño modular del código permite su aplicación a otros municipios mediante la modificación de los parámetros iniciales, sin alterar la estructura general del flujo de procesamiento.

C. Desarrollo del modelo de red neuronal artificial

El código establece un entorno de trabajo importando todas las bibliotecas necesarias para el procesamiento de datos (NumPy, Pandas), visualización (Matplotlib, Seaborn) y modelado con redes neuronales (TensorFlow/Keras). Se configura un estilo visual uniforme para los gráficos y se fija una semilla aleatoria (42) para garantizar reproducibilidad en todos los modelos. La conexión con Google Drive permite acceder a los datasets almacenados, mientras que la creación automática de directorios organiza los resultados por municipio, seleccionando la estación de referencia correspondiente (EL ENCANTO para Tunja).

El proceso de adquisición de datos integra múltiples fuentes:

- Registros mensuales de precipitación
- Metadatos de las estaciones meteorológicas
- Datos climáticos del Índice ONI descargados directamente desde la NOAA mediante web scraping.

Esta etapa incluye la limpieza inicial de datos, la fusión de tablas relacionadas y el ajuste de formatos temporales. Los metadatos se procesan para eliminar columnas innecesarias y estandarizar nombres, creando una base consistente para el análisis.

El EDA se centra en caracterizar los patrones de datos faltantes mediante una clasificación en cuatro categorías definidas por la disponibilidad de registros en las estaciones. Un gráfico de barras detallado muestra el porcentaje de casos en cada categoría. Este análisis revela que la mayoría de los registros completos provienen de situaciones donde todas las estaciones reportan datos (Caso 1), mientras que los casos con estaciones faltantes (Casos 2-4) representan desafíos específicos para la imputación.

Caso 1: Se utilizará como insumo para el entrenamiento de la RED.

Caso 2: Será el principal insumo para generar el llenado de datos faltantes en la RED.

Caso 3: Es el caso en el cual se presenta un estado idóneo donde no se requiere el llenado de datos ya que se cuentan con todos los registros.

Caso 4: Por medio de modelos estadísticos se hace el completado de datos siguiendo una distribución normal donde se destacan pasos como; calcular la desviación estándar y la media mensualmente, crear la distribución normal y finalizando se eligen números de esa distribución normal para el completado de datos.

D. Reprocesamiento y transformación.

La preparación de los datos para el modelado incluyó la normalización mediante MinMaxScaler, la codificación de la variable mes y la incorporación de variables geográficas para representar proximidad y diferencia altitudinal. La división 80 % entrenamiento y 20 % prueba se aplicó con control temporal, de modo que la evaluación no dependiera únicamente de una partición aleatoria. El algoritmo quedó configurado para manejar automáticamente los patrones de datos faltantes detectados en el análisis exploratorio, manteniendo la estructura mensual de la serie antes de alimentar la red neuronal artificial.

E. Diseño de la arquitectura neuronal

La arquitectura seleccionada para Tunja fue definitivamente fijada a partir de pruebas controladas de desempeño y de criterios de parsimonia. Se eligió una red lo suficientemente flexible como para captar relaciones no lineales entre las estaciones, pero a la vez regularizada, mediante dropout, para reducir el riesgo de sobreajuste. La combinación de capas de 132, 64 y 32 neuronas, se justificó a partir del número de variables de entrada disponibles y de la necesidad de ir reduciendo progresivamente la dimensionalidad hasta conseguir una salida única - la precipitación estimada en El Encanto. Esta elección no se entiende como si fuera una arquitectura general, sino como una arquitectura ajustada al tamaño muestral, la correlación espacial y la variabilidad mensual que presenta Tunja.

La Figura 2 de barras muestra la distribución porcentual de los cuatro casos: CASO1, CASO2, CASO3 y CASO4. Cada barra representa el porcentaje de ocurrencias de cada caso con respecto al total e incluye el valor porcentual exacto sobre ella. Esta gráfica proporciona el resumen de porcentajes de casos y la cantidad de datos que la red neuronal deberá estimar en el caso correspondiente.

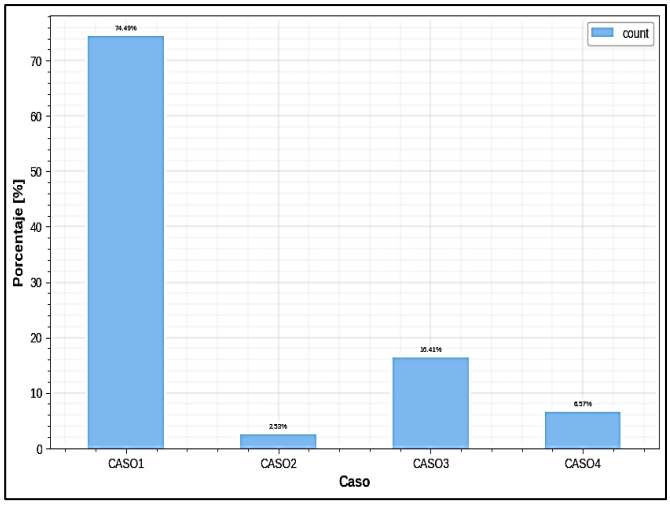

Fig. 2. Porcentaje de casos. Fuente: Autor.

La Figura 3 representa las series temporales mensuales multianuales de la precipitación registradas entre 1993 y 2021 en las seis estaciones meteorológicas elegidas para Tunja. Cada subgráfico corresponde a una estación diferente y se puede observar la evolución de los valores mensuales a lo largo del tiempo.

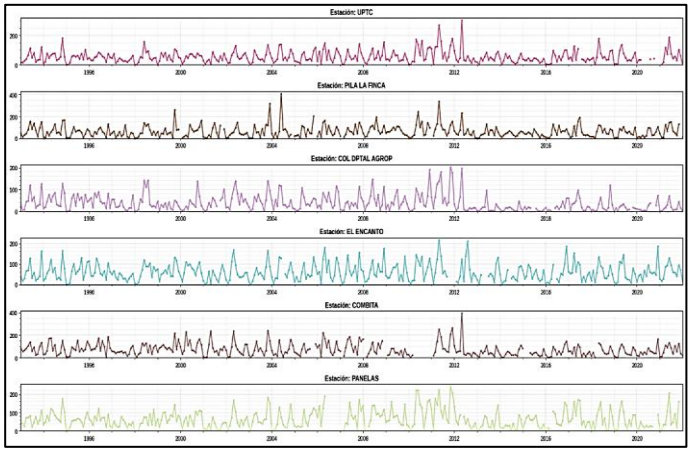

Fig. 3. Grafica de las series temporales mensuales. Fuente: Autor.

El diagrama de caja (boxplots), Fig. **4**, de la precipitación mensual registrada a lo largo del año permite observar una clara estacionalidad bimodal con dos periodos de mayor precipitación: uno entre mayo y abril, y otro entre noviembre y octubre, que son propios del régimen climático de la región. Los meses de enero, febrero, julio y agosto tienen niveles de precipitación más bajos.

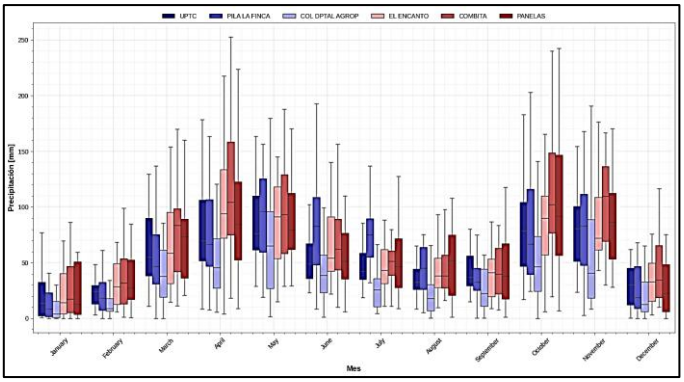

Fig. 4. Diagramas de cajas de las estaciones. Fuente: Autor.

La Fig. 5 presenta la relación entre el Índice Oceánico Niño (ONI) y la precipitación mensual registrada en la estación El Encanto tomada como referencia para la ciudad de Tunja.

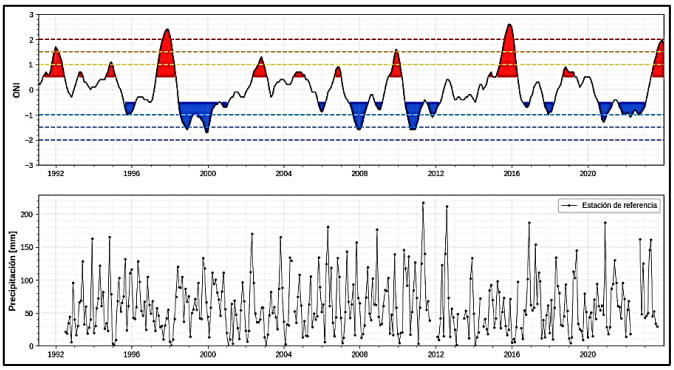

Fig. 5. Graficas del ONI y la estación de referencia. Fuente: Autor.

El código define y entrena una red neuronal en Keras para un problema de regresión. La función create_model construye un modelo secuencial con tres capas ocultas que usan activación softplus y regularización mediante Dropout para reducir el sobreajuste. La capa de salida usa activación ReLU, lo que restringe las predicciones a valores no negativos, condición coherente con la naturaleza física de la precipitación. El modelo se entrena con el optimizador Adam, tasa de aprendizaje inicial de 0.001, función de pérdida Huber —robusta frente a valores atípicos— y validación interna del 20 % de los datos de entrenamiento. Además, se emplea ReduceLROnPlateau para disminuir automáticamente la tasa de aprendizaje si la pérdida no mejora durante cinco épocas.

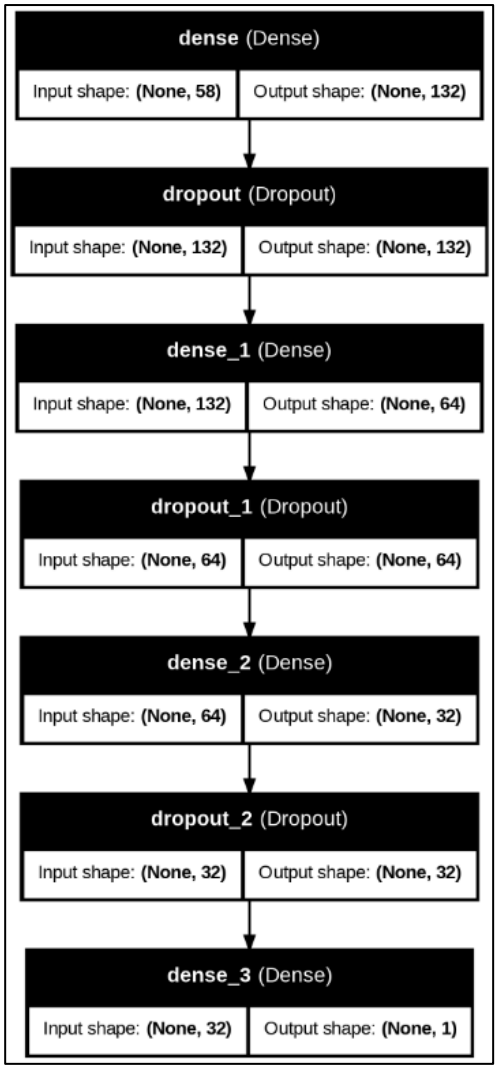

Fig. 6. Arquitectura empleada. Fuente: Autor.

La arquitectura de la red neuronal artificial utilizada está compuesta por una capa de entrada con 58 variables, capas ocultas de 132, 64 y 32 neuronas, capas intercaladas de dropout y una neurona de salida correspondiente a la precipitación estimada. Las capas de dropout eliminan aleatoriamente conexiones durante el entrenamiento para reducir la dependencia de patrones específicos de la muestra. La selección de esta arquitectura se sustentó en el desempeño de validación, la estabilidad de la curva de pérdida y el equilibrio entre capacidad predictiva e interpretabilidad. En función de esto se obtuvieron los siguientes resultados:

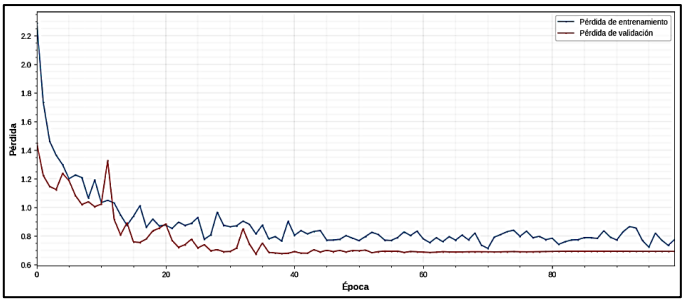

Fig. 7. Curva de perdidas. Fuente: Autor.

El descenso gradual de la pérdida, tanto en validación como en entrenamiento, durante los periodos de tiempo, indica que el modelo está aprendiendo correctamente y no está sobreajustando. La pérdida de validación se mantiene estable y cercana a la de entrenamiento sin presentar sobreajuste, lo que sugiere un buen desempeño general y una adecuada capacidad de generalización del modelo.

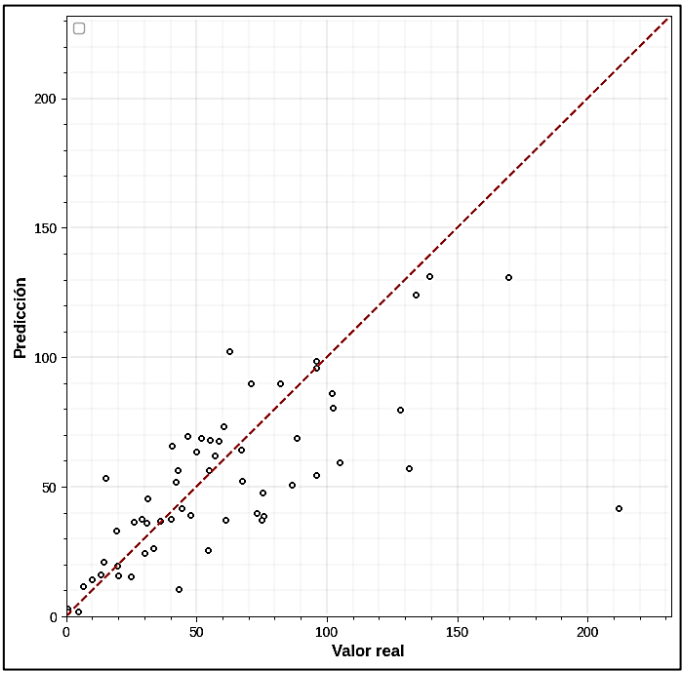

Fig. 8. Model testing. Fuente: Autor

La dispersión muestra la relación entre los valores reales y las predicciones del modelo.

Tabla 1.
Métricas

Métrica	Valor
MAE	19.315
MSE	998.292
RMSE	31.596
R2	0.442
NASH	0.442
PBIAS	13.641

Fuente: Autor.

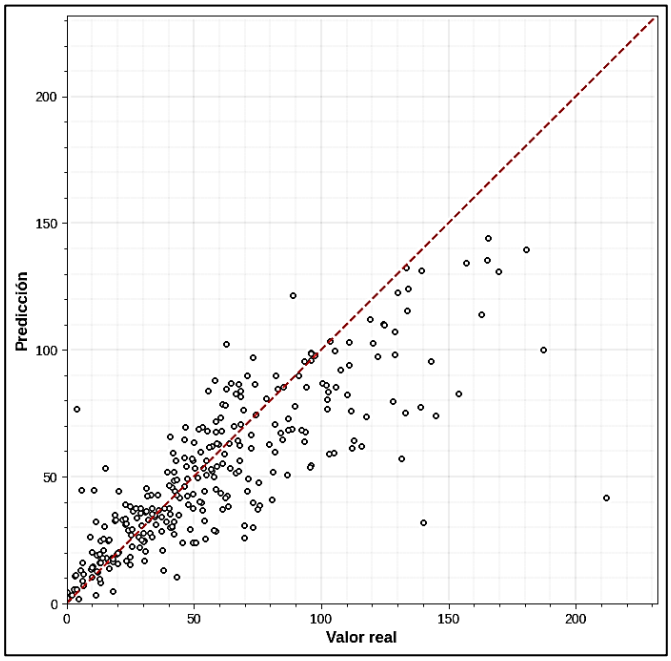

Fig. 9. Scatter plot simulando los datos de la estación de referencia. Fuente: Autor.

Se simuló la red neuronal planteando la hipótesis de que la estación no existe utilizando como insumo netamente la información de referencia, comparando a continuación los datos reales con los simulados por la red evidenciando su eficiencia de predicción.

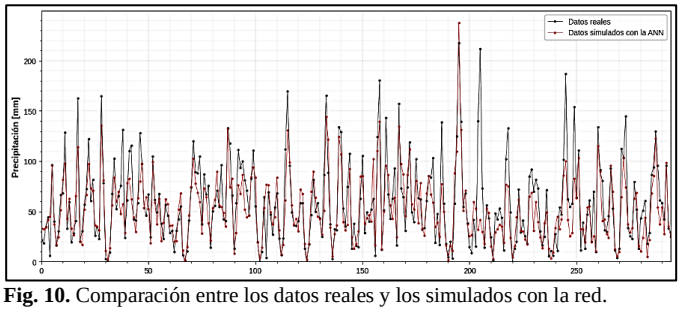

Fig. 10. Comparación entre los datos reales y los simulados con la red. Fuente: Autor.

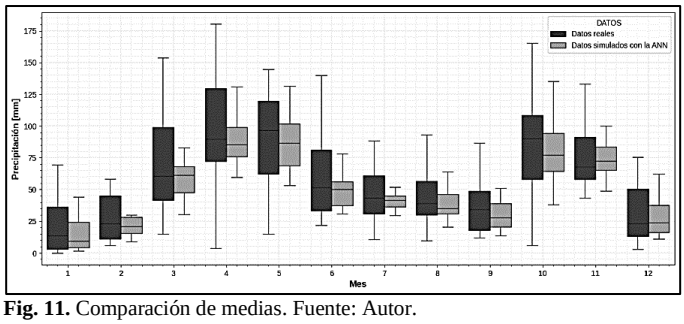

Fig. 11. Comparación de medias. Fuente: Autor.

El modelo logra reproducir el patrón estacional de la precipitación, especialmente en los meses lluviosos como abril, mayo y octubre. Sin embargo, las cajas simuladas tienden a ser más estrechas y con menores valores extremos que las observadas, lo que indica subestimación de la variabilidad y dificultad para capturar eventos de precipitación intensa. Dicha conducta es acorde con estudios recientes de imputación de lluvia que indican que la estimación mejora usando métodos basados en el aprendizaje automático en comparación con enfoques simples, pero requieren de series representativas, ajuste de hiperparámetros y se comprueban contra métodos estadísticos de referencia (Papailiou et al., 2022; Pinthong et al., 2024; Wangwongchai et al., 2023). La **tabla 2** desglosa las pruebas realizadas, usando t-test para meses con distribución aproximadamente normal y Mann-Whitney U para los meses sin normalidad, con significancia del 5 %.

Tabla 2.
Comparación de medias.

Mes	Estadístico	p-valor	Prueba	Diferencia significativa
1	346	0.89	Mann-Whitney U	FALSO
2	378	0.82	Mann-Whitney U	FALSO
3	275	0.826	Mann-Whitney U	FALSO
4	306	0.718	Mann-Whitney U	FALSO
5	0.55	0.585	T-test	FALSO
6	324	0.46	Mann-Whitney U	FALSO
7	443	0.732	Mann-Whitney U	FALSO
8	423	0.316	Mann-Whitney U	FALSO
9	308	0.687	Mann-Whitney U	FALSO
10	0.661	0.512043	T-test	FALSO
11	222	0.647156	Mann-Whitney U	FALSO
12	281	0.725192	Mann-Whitney U	FALSO

Fuente: Autor.

En el entrenamiento se han aplicado algunas técnicas de optimización: EarlyStopping (patience=10) que supervisa la validación de la pérdida para parar el entrenamiento en el caso que no mejore; los callbacks guardan los pesos finales del modelo automáticamente; el learning rate se optimiza en el entrenamiento; la validación cruzada temporal se realiza para probar que el modelo aprende bien a periodos no visto; se graba todo el proceso, incluyendo las métricas por época y las visualizaciones de convergencia.

Se emplearon también varias métricas estadísticas: coeficiente de determinación (R^2), error absoluto medio (MAE: mean absolute error), error cuadrático medio (MSE: mean squared error), raíz del error cuadrático medio (RMSE: root mean squared error), índice de Nash-Sutcliffe y sesgo porcentual (PBIAS: percentage bias). Estas métricas fueron cuantificadas a nivel global y complementadas con contrastes mensuales observados/imputados. Agregando a lo anterior, se incluyó la comparación gráfica de observados y estimados, el análisis de residuos en escenarios de datos faltantes y las pruebas t-test y de Mann-Whitney para comprobar si las imputaciones diferían significativamente de los registros observados. Este esquema de evaluación permite valorar no sólo el ajuste medio, sino la estabilidad estacional, el sesgo y la habilidad para representar eventos de mayor envergadura.

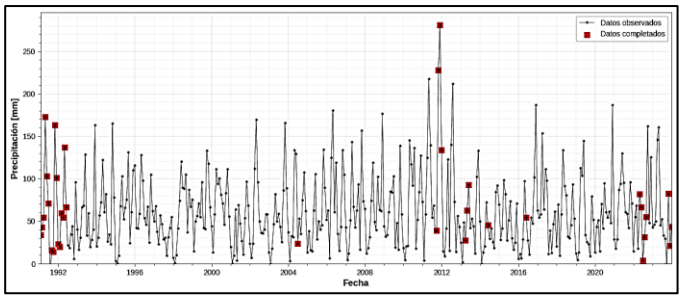

Fig. 12. Serie temporal con los datos completados para Tunja. Fuente: Autor.

Se identificaron y rellenaron ciertos vacíos en la serie mediante estimaciones, estos valores completados se distribuyen principalmente en los extremos de la serie donde las ausencias eran más frecuentes. Se puede ver como la RNA en 2011-2012 completó datos faltantes en un contexto de precipitación extrema y el valor estimado se alinea con relación a la magnitud del evento.

IV. DISCUSIÓN

La presencia de datos faltantes en varias estaciones confirma una cobertura incompleta o irregular que debe considerarse en el análisis hidrológico. La estación El Encanto se tomó como referencia para Tunja, y las estaciones Cómbita, Panelas y UPTC aportaron información auxiliar por su coherencia espacial y altitudinal. El comportamiento bimodal de la precipitación y la escasa frecuencia de los eventos extremos hacen que el modelo sea más capaz de aprender los patrones medios y estacionales que de hacerlo con los valores extremos. Esta situación coincide con algunos estudios recientes en el sentido de que los modelos de aprendizaje automático pueden aprender relaciones no lineales que resultan ser útiles, pero su desempeño se debilitó cuando los episodios extremos en el conjunto de entrenamiento eran escasos o cuando las estaciones vecinas eran incapaces de capturar de forma oportuna la variabilidad local (Xu y Liang, 2021; Nearing et al., 2021).

La matriz de correlación muestra una fuerte relación positiva entre las estaciones de precipitación entre UPTC y El Encanto (r = 0.80). También para El Encanto y Cómbita (r = 0.73) y entre El Encanto y Panelas (r = 0.73), lo que hace suponer que se comportan de forma parecida en la variabilidad de las alturas de lluvia en el área de estudio. Las estaciones en general presentan correlaciones moderadas a altas, lo que hace pensar en la existencia de coherencia espacial entre las mediciones de precipitación en el área de estudio, uno de los criterios fundamentales de la elección de las estaciones.

A pesar de que hay una tendencia general positiva acorde a la línea ideal de predicción, la dispersión en los valores más

altos hace que el modelo aún no sea preciso para precipitaciones altas. El MAE de 19.3 mm y el RMSE de 31.6 mm concuerdan con errores no de una magnitud extrema para una serie mensual. Asimismo, el R^2 y el índice de Nash ofrecido de 0.44 indican que el modelo sólo reproduce una fracción importante, pero no mayoritaria, de la variabilidad observada. En cuanto al PBIAS de 13.6 % vamos a suponer una tendencia ligera a la sobreestimación. Todo esto ha de ser interpretado como un desempeño aceptable a efectos de completar series y apoyar análisis exploratorios y no como un sustituto pleno de la medición instrumental. Frente a métodos tradicionales como promedios, razón normal, interpolación o regresión, la red neuronal ofrece mayor flexibilidad para relaciones no lineales; sin embargo, los métodos lineales o espaciales siguen siendo comparadores necesarios por su transparencia y menor costo computacional (Papailiou et al., 2022; Wangwongchai et al., 2023).

La comparación entre las precipitaciones reales y las simuladas pone de manifiesto que la red obedece la tendencia y la estacionalidad de la serie, pero, a cambio, presenta las principales discrepancias en la estimación de los eventos intensos. Esta limitación se podría explicar, en parte, por la baja frecuencia de precipitaciones extremas en el conjunto de entrenamiento y la dificultad para llevar las relaciones aprendidas a eventos poco específicos o representados. Por ello, las estimaciones generadas por la RNA han de ser utilizadas junto con controles de plausibilidad hidrológica, revisiones de consistencia mensual e, idealmente, con la comprobación frente a métodos de imputación tradicionales. En aplicaciones operativas sí es cierto que el modelo es más fiable en términos de reconstruir continuidad y patrones medios que en la estimación de máximos extremos relacionados con el diseño hidráulico o con alertas de emergencia.

V. LIMITACIONES DEL ESTUDIO

La principal limitación de este estudio se basa en el tamaño y en la completitud de las series temporales disponibles. Dado que se ha trabajado con datos dispuestos de forma mensual y estaciones con vacíos de información, la red neuronal tiene menos ejemplos con los que aprender eventos infrecuentes, hecho que afecta más concretamente la estimación de precipitaciones extremas; estas son muy importantes a la hora de realizar análisis de amenaza, drenaje urbano y del riesgo.

Otra limitación es que el modelo ha sido ajustado para la estación El Encanto y su entorno más inmediato y, por tanto, no debe extrapolarse a otro municipio sin ajustar de nuevo las variables, las estaciones auxiliares, la arquitectura y los hiperparámetros. A fin de cuentas, también hemos de tener en cuenta que la comparación cuantitativa de la RNA con respecto a métodos tradicionales se ha tratado como un referente metodológico y de discusión, pero que en un futuro se debería seguir una línea base local mediante regresión múltiple, razón normal, IDW o kriging, para así estimar el verdadero gain real de la RNA versus las alternativas más simples.

Por último, la introducción del ONI permite incluir una nueva fuente de variabilidad climática regional, aunque no agota otras fuentes de información que pueden ser consideradas para explicar la precipitación local, como en el caso de la propia circulación atmosférica, la temperatura, la humedad, la cobertura del suelo, etc., o variables orográficas de mayor resolución. La incorporación de otras fuentes podría mejorar la captura de la variabilidad mensual prolongando la media de la incertidumbre de las imputaciones.

VI. CONCLUSIONES

Esta investigación demuestra que las redes neuronales artificiales pueden ser una útil alternativa para conseguir completar los registros mensuales de precipitación desde estaciones que presentan una inexistencia de información, siempre que sean utilizadas bajo validación estadística y criterios hidrológicos. En el caso de Tunja, el modelo reproducía bien la estacionalidad y tendencia de las precipitaciones, aunque con errores moderados y con un poder explicativo parcial ($R^2 = 0.44$). Por lo que hemos de considerar el resultado como una herramienta para la reconstrucción de series y análisis técnicos, no como un sustituto de la observación meteorológica directa ni como una solución para poder evaluar los extremos.

El ONI, también analizado en este contexto, mostró que existe una importante relación climática: eventos de El Niño (ONI > +0.5) como los de las campañas 1997-1998, 2015-2016 y 2023 presentan como comportamiento la disminución de la precipitación, en tanto que los eventos de La Niña (ONI < -0.5) tales como los de las campañas 1999-2000, 2010-2011 parecen mostrar incrementos en los niveles de lluvia. La señal del ONI puede estar en consonancia con la interpretación de las imputaciones, pero no puede sustituir la caracterización de cada estación y la revisión de la consistencia de cada uno de los meses estimados.

En la práctica, una serie mensual de registros puede contribuir a la toma de decisiones agrícolas como la planificación de calendarios de siembra, la programación de riego, la estimación de la disponibilidad hídrica en condiciones específicas, como clima seco o clima húmedo. En el contexto de la planificación urbana y de acuerdo con aplicaciones recientes de machine learning en sistemas de drenaje urbano (Kwon y Kim, 2021), puede servir como un insumo para un diagnóstico de drenaje de forma temporal, priorizar el mantenimiento de sumideros, actualizar un análisis de amenaza de inundaciones y la formulación de medidas de adaptación climática. Trabajos futuros deberían propiciar mejorar el

número de estaciones, explorar variables climáticas adicionales, comparar directamente con métodos de imputación convencionales y evaluar modelos para eventos extremos.

1. Referencias

[1] Congreso de la República de Colombia. (2012). Ley 1523 de 2012. Por la cual se adopta la Política Nacional de Gestión del Riesgo de Desastres y se establece el Sistema Nacional de Gestión del Riesgo de Desastres y se dictan otras disposiciones. Diario Oficial No.48.411 del 24 de abril de 2012. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=47141

[2] Coulibaly, P., Anctil, F., y Bobée, B. (2000). Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. Journal of Hydrology, 301(1-4), 244-257.

[3] Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179-211.

[4] Feldkamp, L. A., & Puskorius, G. V. (1998). A signal processing framework based on dynamic neural networks with application to problems in adaptation, filtering, and classification. Proceedings of the IEEE, 86(11), 2259–2277. https://doi.org/10.1109/5.726790

[5] French, M. N., Krajewski, W. F., & Cuykendall, R. R. (1992). Rainfall forecasting in space and time using a neural network. Journal of Hydrology, 137(1–4), 1–31. https://doi.org/10.1016/0022-1694(92)90046-X

[6] Gupta, H. V., Sorooshian, S., y Yapo, P. O. (1999). Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. Journal of Hydrologic Engineering, 4(2), 135-143.

[7] Hsu, K.-L., Gupta, H. V., y Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. Water Resources Research, 31(10), 2517-2530.

[8] IDEAM (2019). Evaluación de la calidad y continuidad de la red hidrometeorológica nacional. Instituto de Hidrología, Meteorología y Estudios Ambientales, Colombia.

[9] Kisi, O. (2004). Streamflow forecasting using different artificial neural network algorithms. Journal of Hydrologic Engineering, 9(6), 491-496.

[10] Kwon, S. H., & Kim, J. H. (2021). Machine learning and urban drainage systems: State-of-the-art review. Water, 13(24), 3545. https://doi.org/10.3390/w13243545

[11] Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning? Water Resources

[12] Nourani, V., y Hosseini B. (2011). Monthly rainfall-runoff modeling using a wavelet-based neural network conjunction model. Journal of Hydrology, 401(3-4), 190-199.

[13] Ordoñez, L., Muñoz, S., Tineo, P., & Mejía, I. (2024). Aplicación de redes neuronales artificiales a la modelación de lluvia-escorrentía en la cuenca del río Chancay Lambayeque. Tecnología Y Ciencias Del Agua , 15 (6), 95-141. https://doi.org/10.24850/j-tyca-2024-06-03

[14] Papailiou, I., Spyropoulos, F., Trichakis, I., & Karatzas, G. P. (2022). Artificial neural networks and multiple linear regression for filling in missing daily rainfall data. Water, 14(18), 2892. https://doi.org/10.3390/w14182892

[15] Pasquier Jaramillo, C. A. (2025) Modelado bayesiano de extremos espacio-temporales de precipitación: aplicaciones en estimación de modelos con covariables.

[16] Pearlmutter, B. (1995). Gradient calculations for dynamic recurrent neural networks: A survey. IEEE Transactions on Neural Networks, 6(5), 1212-1228.

[17] Pinthong, S., Ditthakit, P., Salaeh, N., Hasan, M. A., Son, C. T., Linh, N. T. T., Islam, S., & Yadav, K. K. (2024). Imputation of missing monthly rainfall data using machine learning and spatial interpolation approaches in Thale Sap Songkhla River Basin, Thailand. Environmental Science and Pollution Research, 31, 54044-54060. https://doi.org/10.1007/s11356-022-23022-8

[18] Principe, J. C., Euliano, N. R., & Lefebvre, W. C. (2000). Sistemas neuronales y adaptativos: Fundamentos a través de simulaciones. John Wiley.

[19] Retropropagación: Teoría, arquitecturas y aplicaciones (pp. 433-486). Lawrence Erlbaum Associates.

[20] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536. https://doi.org/10.1038/323533a0

[21] Smith, A. B., Jones, C. D., y Johnson, E. F. (2018). Application of artificial neural networks for precipitation estimation in mountainous regions. Journal of Hydrology, 543, 765-776.

[22] Specht, D. F. (1991). A general regression neural network. IEEE Transactions on Neural Networks, 2(6), 568–576. https://doi.org/10.1109/72.97934

[23] Wang, L., y Robertson, A. (2020). Improving precipitation estimation using neural networks in complex terrain. Water Resources Research, 56(2), e2019WR026101.

[24] Wangwongchai, A., Waqas, M., Dechpichai, P., Hlaing, P. T., Ahmad, S., & Humphries, U. W. (2023). Imputation of missing daily rainfall data; A comparison between artificial intelligence and statistical techniques.

[25] Williams, R. J., & Zipser, D. (1995). Algoritmos de aprendizaje basados en gradiente para redes recurrentes y su complejidad computacional. En Y. Chauvin & D. E. Rumelhart (Eds.), Retropropagación: Teoría, arquitecturas y aplicaciones (pp. 433–486). Lawrence Erlbaum Associates, Inc.

[26] Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. Wiley

[27] Yescas, I. A. O. (2016). Elaboración de un protocolo para construir bibliotecas de RNAs pequeños y su implementación para el estudio de la respuesta a sequía crónica en Brachypodium distachyon L (Doctoral dissertation, Universidad del Papaloapan).