

VALIDACIÓN DE CHIRPS EN MÉXICO: ANALIZANDO LA PRECISIÓN DEL MONITOREO DE PRECIPITACIÓN MEDIANTE APRENDIZAJE AUTOMÁTICO Y REDES NEURONALES

Diego Rivera Buendía^{1*}; Emilio Quintana Molina²

¹Soil Department, Renewable Natural Resources Engineering, Chapingo Autonomous University, 56230, Mexico, Mexico.

²Hydrosat. 9 Rue du Laboratoire, L-1911, Luxembourg, Luxembourg.
emolina@hydrosat.com - +31 6 33436547

Resumen

Este estudio aborda la validación del producto satelital CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data) en México para analizar su precisión en el monitoreo de la precipitación mediante técnicas de computación en la nube. Ante la disminución significativa de estaciones meteorológicas operativas desde 2017, se desarrolló un algoritmo para la descarga automatizada de datos diarios de precipitación de la CONAGUA, seleccionando 1,055 estaciones con registros continuos de 2000 a 2021. Los datos de CHIRPS fueron extraídos utilizando Google Earth Engine y comparados con las observaciones terrestres. Los resultados mostraron una buena concordancia general ($R^2 = 0.85$) entre CHIRPS y las estaciones meteorológicas, aunque con variaciones regionales. La precisión fue mayor en zonas de barlovento y altitudes elevadas. Se implementaron modelos de Random Forest y Redes Neuronales Multicapa para predecir los residuos de precipitación, pero su desempeño fue limitado debido a restricciones computacionales. Los hallazgos resaltan la importancia de validar y mejorar los productos satelitales en regiones con variabilidad climática y topográfica, sugiriendo que técnicas avanzadas de modelado podrían incrementar la precisión en la estimación de la precipitación.

Palabras claves: CHIRPS, precipitación, validación, computación en la nube, México, Random Forest, redes neuronales, Google Earth Engine, CONAGUA.

Introducción

La validación de datos satelitales, como los proporcionados por CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data), es crucial para garantizar su precisión en la representación de patrones de precipitación, especialmente en regiones geográficamente diversas. CHIRPS ha demostrado ser una herramienta valiosa para el monitoreo de la precipitación a diferentes escalas temporales, aunque su desempeño puede variar significativamente dependiendo de las características regionales. Estudios recientes han validado la capacidad de CHIRPS para representar la variabilidad climática en diversas regiones del mundo, incluyendo áreas montañosas y tropicales de América del Sur, donde su precisión ha sido adecuada para análisis estacionales y de variabilidad interanual, aunque con limitaciones a escalas diarias (López-Bermeo et al., 2022, Alsilibe et al., 2023). En particular, la investigación en el Cerrado brasileño demostró que las estimaciones de precipitación de CHIRPS se alinearon bien con los datos observados en estaciones meteorológicas, mostrando una capacidad destacable para detectar lluvias y representar variaciones espacio-temporales (da Silva, 2024).

La importancia de estos modelos es especialmente relevante en México, donde desde 2017 se ha observado una alarmante disminución en la densidad de estaciones meteorológicas operativas. Esta reducción ha limitado la capacidad de monitoreo y ha hecho que el acceso a datos precisos de precipitación sea crucial para numerosos sectores, incluidos la agricultura, la gestión de recursos hídricos y la planificación urbana y rural. Además, las condiciones de las estaciones meteorológicas varían considerablemente a lo largo del territorio, lo que genera disparidades en la calidad y la representatividad de los datos disponibles. En este contexto, tanto el uso de productos satelitales como CHIRPS, como los esfuerzos por validar su precisión, se han vuelto indispensables para asegurar la continuidad y la exactitud en la estimación de patrones de precipitación. Cabe destacar que, hasta donde se tiene conocimiento, no existen esfuerzos previos por realizar un trabajo de validación a nivel nacional de estos datos, lo que resalta la oportunidad y relevancia del presente esfuerzo para proporcionar información confiable a los tomadores de decisiones que enfrentan los desafíos climáticos y ambientales en México.

Además de la importancia de la validación, la adopción de prácticas de ciencia abierta está revolucionando la transparencia y reproducibilidad en la investigación científica. Al utilizar herramientas abiertas y transparentes, los investigadores pueden maximizar el alcance y la utilidad de sus hallazgos, facilitando la colaboración y el acceso a los datos. La implementación de meta-análisis abiertos, por ejemplo, ha sido promovida como una práctica para mejorar la reproducibilidad y aumentar la transparencia en la síntesis de datos, haciendo que los resultados sean más accesibles y verificables por la comunidad científica (Moreau y Wiebels, 2024).

Materiales y Métodos

Descarga Automatizada de Datos de CONAGUA

Para este estudio, se desarrolló un algoritmo propio que permite la descarga automatizada de los datos de precipitación diarios proporcionados por la Comisión Nacional del Agua (CONAGUA). Este algoritmo accede a la base de datos oficial, extrae la información de las estaciones meteorológicas disponibles, y organiza los datos en un formato adecuado para su posterior análisis. Esta automatización asegura la eficiencia en la recopilación de datos y minimiza el riesgo de errores humanos, facilitando la actualización continua del conjunto de datos a medida que se obtienen nuevos registros. Sin embargo, debido a la necesidad de contar con series temporales completas y consistentes, solo se conservaron 1,055 de las más de 5,000 estaciones disponibles. Estas estaciones fueron seleccionadas porque cumplieron con el criterio de operación continua durante el periodo de 2000 a 2021, asegurando la calidad y la comparabilidad de los datos en el análisis. La relevancia del estudio se complementa con la drástica pérdida de estaciones operativas en los últimos años como se muestra en la Figura 1.

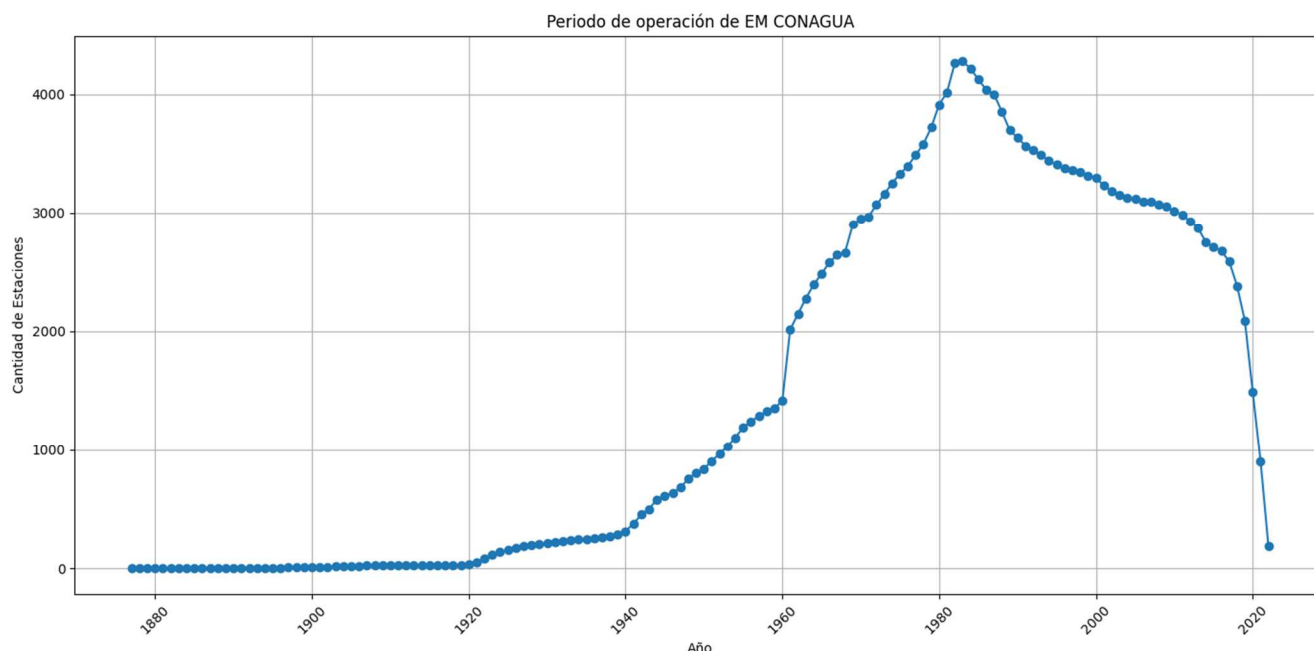


Figura 1. Estaciones Meteorológicas de CONAGUA operativas

Extracción de Datos de CHIRPS mediante Google Earth Engine

Se utilizó Google Earth Engine (GEE) a través de su API de Python y la librería Geemap para extraer los datos de precipitación diarios del producto satelital CHIRPS. Para cada estación meteorológica de CONAGUA, se identificó el píxel correspondiente en la cuadrícula de CHIRPS, y se extrajo la serie temporal diaria de precipitación de forma automatizada. Este proceso permitió obtener datos coherentes espacialmente y comparables con las observaciones de las estaciones meteorológicas, manteniendo la resolución espacial y temporal necesaria para los análisis posteriores.

Cálculo de Correlaciones

Una vez obtenidos los datos de precipitación tanto de las estaciones meteorológicas de CONAGUA como del producto CHIRPS, se procedió a calcular las correlaciones entre ambas fuentes. El análisis incluyó la comparación diaria y mensual de los datos, evaluando la concordancia entre las estimaciones satelitales y las observaciones terrestres. Las correlaciones se calcularon utilizando métodos estadísticos estándar, permitiendo cuantificar la precisión de CHIRPS en la representación de la precipitación en las distintas regiones estudiadas. Se realizó una validación exhaustiva de los datos de CHIRPS frente a las estaciones meteorológicas. Los errores entre las observaciones y las estimaciones satelitales se calcularon utilizando las siguientes métricas:

- Error Cuadrático Medio (MSE): 3369.01 mm²
- Raíz del Error Cuadrático Medio (RMSE): 58.04 mm
- Error Absoluto Medio (MAE): 28.40 mm
- Coeficiente de Correlación (R^2): 0.85
- Sesgo Medio: 1.29 mm

Podemos ver en la siguiente figura el error medio a través de los diferentes meses del año (Figura 2)

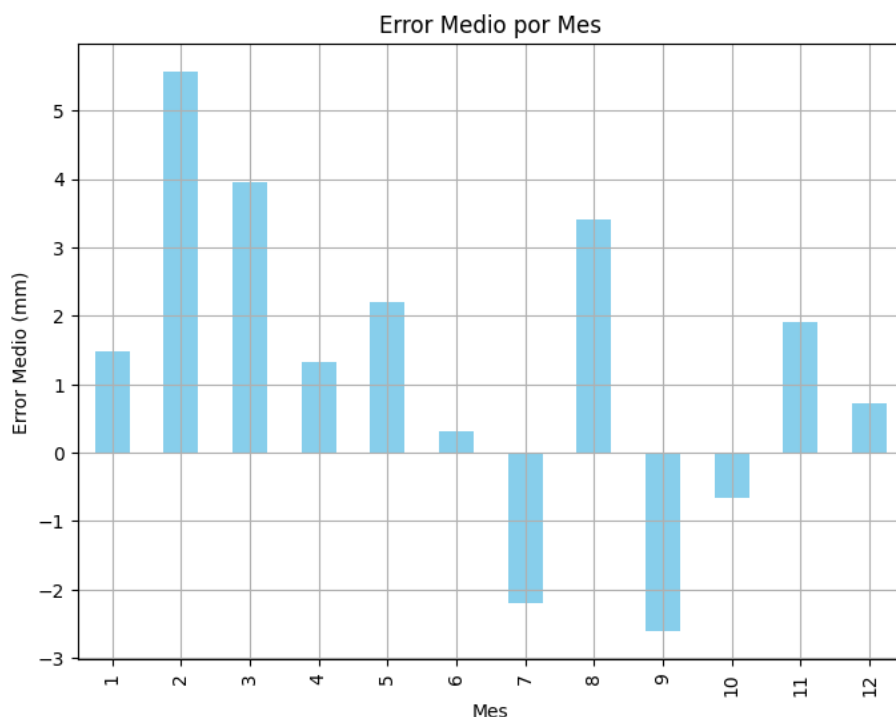


Figura 2. Error Medio por Mes

Se eliminaron los valores atípicos más extremos, conservando el 99.95% de los datos. Además, se observó que la precisión del modelo es mayor en zonas de barlovento a mayor altitud, lo cual se puede notar visualmente en los mapas (Figuras 3 y 4).



Figura 3. Error medio por estación

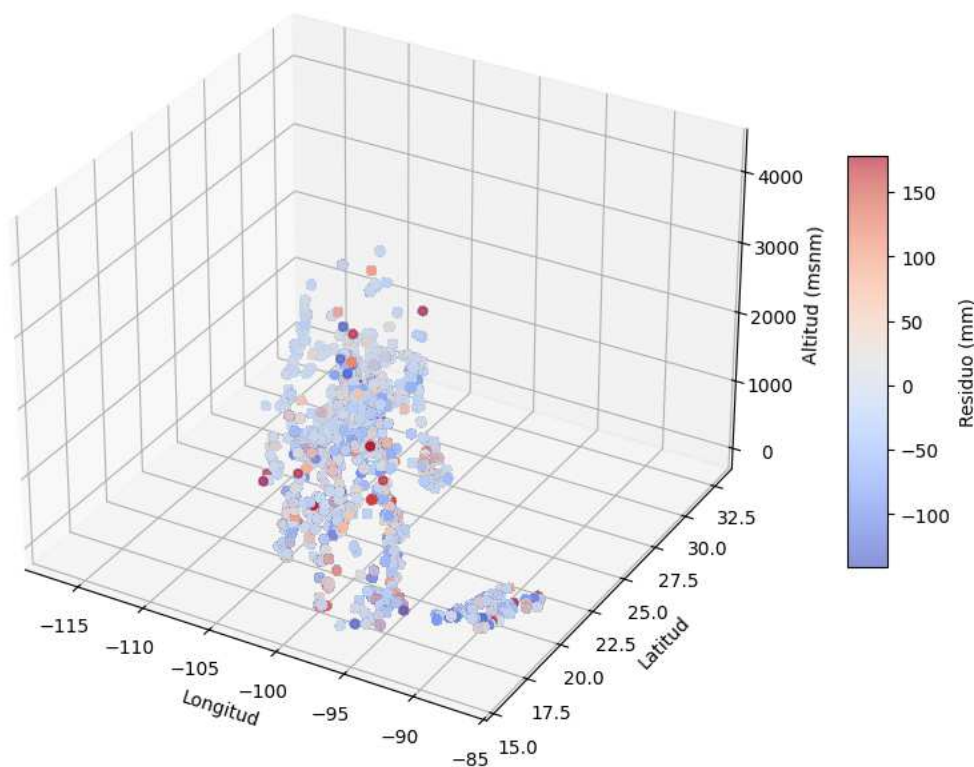


Figura 4. Distribución del error en función de latitud, longitud y altitud

Configuración de un modelo Random Forest

El modelo de **Random Forest** fue utilizado inicialmente para predecir los residuos de precipitación. Las características de entrada incluyeron latitud, longitud, altitud, temperatura mínima (TMin), temperatura máxima (TMax), el mes y el año de la observación, ya que se consideró que estas variables tienen una fuerte correlación con los patrones de precipitación. Los datos fueron normalizados utilizando la técnica de escalado estándar (*StandardScaler*) para asegurar que todas las variables tuvieran una contribución similar al modelo, y posteriormente fueron divididos en conjuntos de entrenamiento y prueba, empleando un 70% de los datos para el entrenamiento y un 30% para la validación.

El modelo se implementó utilizando la librería **scikit-learn** y configurado con un total de 500 estimadores. Se ajustó la profundidad máxima de los árboles a 20 niveles, con un mínimo de cinco muestras necesarias para realizar la división de nodos. Para la evaluación del desempeño del modelo, se utilizó el **Error Cuadrático Medio (MSE)**, así como el **Error Absoluto Medio (MAE)** y el **coeficiente de determinación (R^2)**, que permiten medir tanto la precisión como la capacidad del modelo para explicar la variabilidad de los datos.

Una vez entrenado el modelo, se llevaron a cabo las predicciones en el conjunto de prueba. Las predicciones generadas fueron desnormalizadas para ser comparables con los valores reales del residuo de precipitación. El análisis de importancia de las variables mostró que la altitud y la temperatura mínima jugaron un papel relevante en las predicciones del modelo, lo que sugiere que estos factores geográficos y climáticos son claves en la modelización de la precipitación en las distintas regiones estudiadas.

Configuración de una Red Neuronal Multicapa

Dado que los resultados obtenidos con el modelo de Random Forest no fueron satisfactorios en cuanto a precisión, se decidió implementar una Red Neuronal Multicapa (MLP) utilizando la librería PyTorch. Las mismas características de entrada utilizadas en el modelo de Random Forest fueron empleadas en la red neuronal: latitud, longitud, altitud, temperatura mínima (TMin), temperatura máxima (TMax), mes y año.

La arquitectura de la red consistió en varias capas completamente conectadas. La primera capa recibía las características de entrada y se conectaba a una capa oculta de 128 neuronas, seguida por una segunda capa oculta de 64 neuronas. Ambas capas ocultas emplearon la función de activación ReLU para introducir no linealidades en el modelo. La capa de salida estaba compuesta por una única neurona, encargada de predecir el valor del residuo de precipitación. La red fue entrenada utilizando el optimizador Adam, con una tasa de aprendizaje de 0.001, y la función de pérdida seleccionada fue el Error Cuadrático Medio (MSE).

El proceso de entrenamiento se llevó a cabo en un entorno con limitaciones computacionales, utilizando un procesador Intel Core i5 de cuatro núcleos y gráficos integrados Intel Iris Plus Graphics 645, lo que afectó el tiempo de entrenamiento y la cantidad de experimentos que pudieron realizarse. El número total de épocas fue fijado en 500, y durante cada época se monitorearon las pérdidas de entrenamiento y validación para evaluar el progreso del modelo. Al finalizar cada época, las pérdidas se imprimían en la consola, proporcionando un monitoreo en tiempo real del rendimiento.

Resultados y Discusión

La validación de los datos satelitales proporcionados por CHIRPS frente a las observaciones de las estaciones meteorológicas de CONAGUA reveló varios aspectos importantes en la representación de los patrones de precipitación. En primer lugar, el análisis de correlación indicó que, en general, existe una buena concordancia entre las estimaciones de CHIRPS y las mediciones de las estaciones terrestres, con un **coeficiente de correlación (R^2) de 0.85**. Este valor sugiere que CHIRPS es capaz de capturar una gran parte de la variabilidad en los datos de precipitación, lo cual está en línea con estudios previos que han demostrado su fiabilidad en diversas regiones geográficas, incluidas áreas montañosas y tropicales (López-Bermeo et al., 2022; Alsilibi et al., 2023). Sin embargo, como se ha observado en otros contextos, la precisión de CHIRPS presenta limitaciones a escalas diarias, lo que es evidente en algunas regiones de México, donde la heterogeneidad geográfica es particularmente notable.

El desempeño de CHIRPS varió según la región y las características geográficas. El análisis espacial de los errores mostró que las zonas de **barlovento y menor altitud** tienden a tener un mejor ajuste, como se puede observar visualmente en los mapas de error distribuidos geográficamente (Figura 2). Esto sugiere que la topografía juega un papel clave en la precisión del modelo, donde las áreas más elevadas, que generalmente están asociadas con patrones más predecibles de precipitación, presentan menores errores. Por otro lado, en áreas de baja altitud y regiones más planas, la variabilidad en las precipitaciones es menos capturada por CHIRPS, lo que aumenta los errores absolutos y genera un sesgo de **1.29 mm**, indicando una ligera sobreestimación en estas zonas.

El **Error Absoluto Medio (MAE)** calculado fue de **28.40 mm**, lo que refleja las desviaciones promedio entre las estimaciones de CHIRPS y las observaciones de las estaciones meteorológicas. Este error es más pronunciado en meses específicos del año, en particular durante los periodos de transición estacional. El análisis mensual del error (Figura 3) reveló que los meses de febrero y marzo presentaron los valores de error más altos, con un **MAE de 5.57 mm y 3.95 mm**, respectivamente. Este comportamiento podría estar relacionado con la mayor variabilidad atmosférica durante estos meses, lo que dificulta la predicción precisa de la precipitación. En contraste, durante los meses de julio y septiembre, el modelo mostró una tendencia a subestimar la precipitación, con errores negativos de **-2.20 mm y -2.61 mm**, lo que sugiere que CHIRPS tiene dificultades para captar con precisión los picos de precipitación en la temporada de lluvias. Para analizar el error distribuido espacialmente se usaron un modelo Random Forest y una Red Neuronal,

Resultados del Modelo Random Forest

La implementación inicial de un modelo de **Random Forest** para predecir los residuos de precipitación mostró limitaciones en su capacidad para capturar la complejidad de las relaciones entre las variables climáticas y geográficas. Aunque el modelo tuvo un desempeño adecuado en la predicción de ciertos eventos de precipitación, los resultados globales no fueron satisfactorios. El **coeficiente de determinación (R^2)** del modelo fue de apenas **0.26**, lo que indica que el modelo solo fue capaz de explicar el **26%** de la variabilidad en los datos, significativamente menor que el obtenido con CHIRPS.

El análisis de importancia de las variables reveló que las **temperatura mínima (TMin) y Máxima (TMax)** (Figura5) fueron los factores que más influyeron en las predicciones del modelo, lo que resalta la relevancia de las características topográficas y térmicas en la estimación de la precipitación. Sin embargo, la **Raíz del Error Cuadrático Medio (RMSE)** fue de **0.84**, y el **Error Absoluto Medio (MAE)** fue de **0.45**, lo que sugiere que el modelo Random Forest tuvo un error constante y, en promedio, no fue capaz de ajustar adecuadamente las observaciones.

El desempeño limitado del modelo puede deberse a la incapacidad de Random Forest para capturar relaciones no lineales complejas entre las variables, lo que motivó la necesidad de explorar enfoques más avanzados, como las redes neuronales.

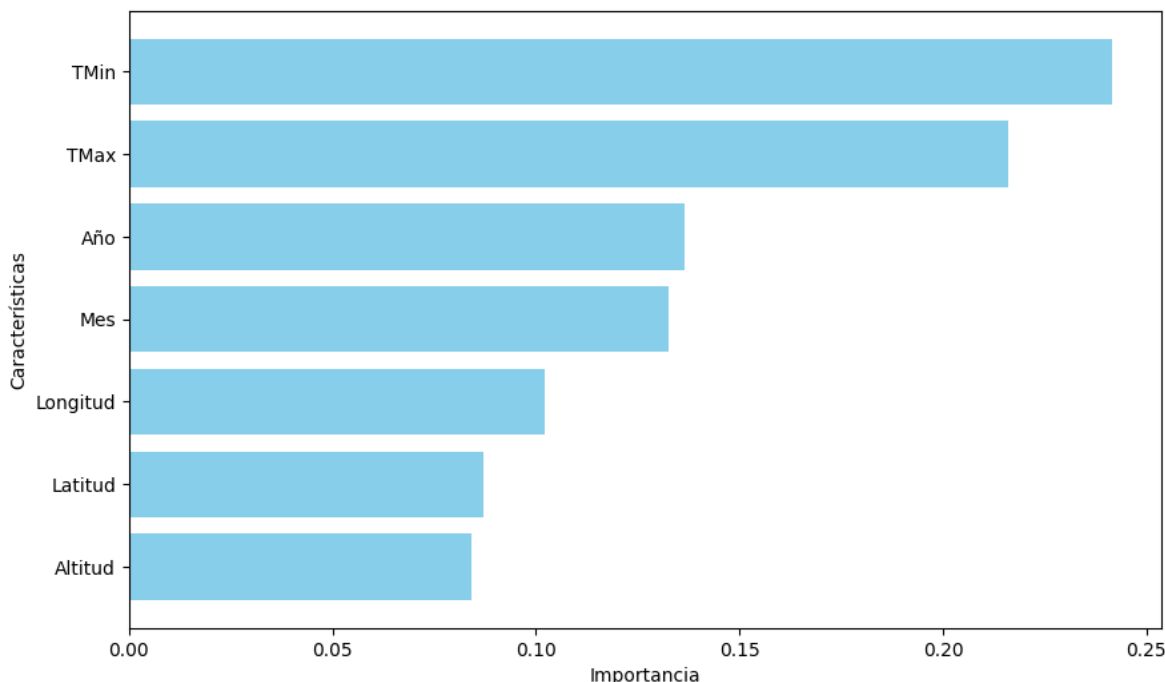


Figura 5. Importancia de características para el modelo Random Forest

Resultados de la Red Neuronal Multicapa (MLP)

La transición a una **MLP** permitió explorar la capacidad de los modelos neuronales para capturar relaciones más complejas entre las variables climáticas y geográficas. La arquitectura de la red incluyó cuatro capas ocultas, y activaciones **ReLU** para incorporar no linealidades en el proceso de aprendizaje. El modelo fue entrenado durante 500 iteraciones, y los resultados mostraron una mejora leve respecto al Random Forest, aunque las limitaciones computacionales afectaron la capacidad de optimización como se aprecia en la evolución de la pérdida durante el entrenamiento (Figura 6).

El **Error Absoluto Medio (MAE)** de la red neuronal fue de **27.44 mm**, ligeramente mejor que el del modelo Random Forest, y la **pérdida en el conjunto de prueba** fue de **0.8229**. Sin embargo, el **coeficiente de determinación (R^2)** se mantuvo en **0.1519**, indicando que la red neuronal aún no fue capaz de capturar la mayor parte de la variabilidad en los datos. Esto puede deberse a la falta de ajustes adicionales en los hiperparámetros y a las limitaciones computacionales inherentes al entorno de entrenamiento utilizado.

A pesar de estas limitaciones, la red neuronal mostró un comportamiento prometedor al reducir el sesgo y mejorar ligeramente la predicción en áreas con mayor variabilidad de precipitación. Se espera que, con un entorno de cómputo más avanzado y la optimización de hiperparámetros, el modelo neuronal pueda superar al Random Forest y CHIRPS en futuras iteraciones

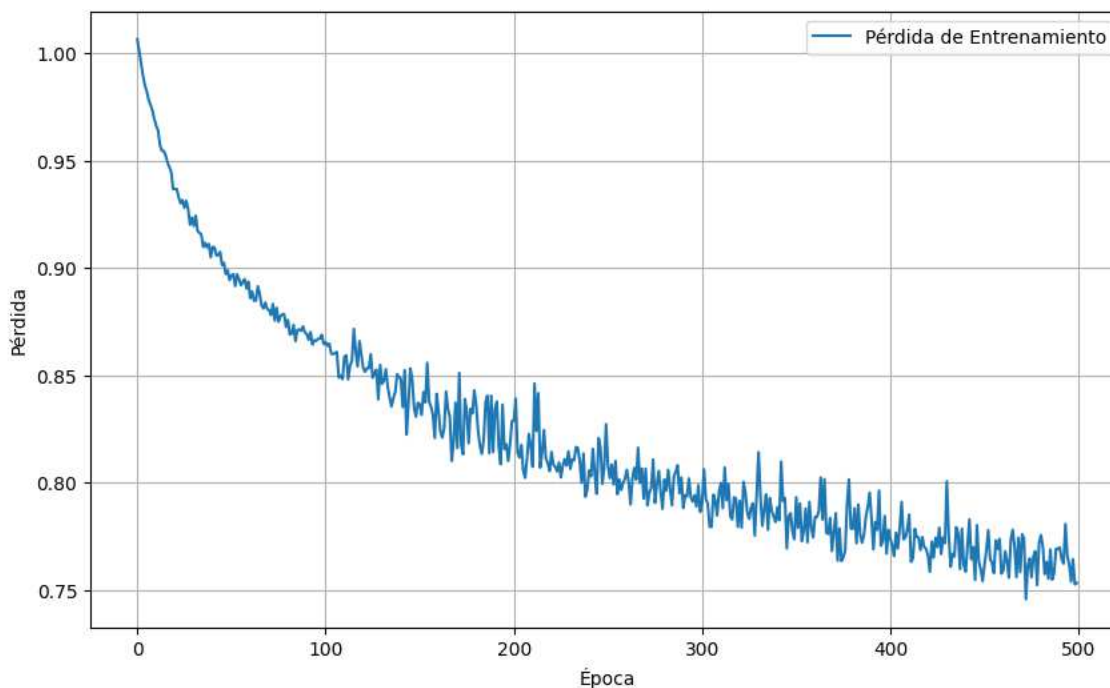


Figura 6. Evolución de la pérdida durante el entrenamiento

- **Tasa de aprendizaje:** Si las fluctuaciones no disminuyen y quieres un ajuste más suave, se podría reducir la tasa de aprendizaje para que los pasos de ajuste del modelo sean más pequeños.
- **Aumento de épocas:** Aunque el modelo está en un proceso de convergencia, se podría incrementar el número de épocas para observar si la pérdida sigue disminuyendo o si comienza a estancarse.
- **Regularización:** Si las fluctuaciones persisten, una técnica de regularización como **Dropout** o **L2 Regularization** puede ayudar a estabilizar el proceso de aprendizaje.

Discusión

Los resultados de este estudio subrayan la necesidad de mejorar la precisión en la predicción de precipitación mediante el uso de modelos más complejos que puedan capturar las interacciones no lineales entre las variables. Aunque CHIRPS demostró ser una herramienta útil para estimar la precipitación en México, las predicciones pueden beneficiarse de técnicas avanzadas como redes neuronales, especialmente en regiones donde la topografía y la variabilidad climática juegan un papel importante, la conciencia sobre la fuente de error es igual de importante para el uso del modelo en diferentes tipos de análisis.

La implementación de una **Red Neuronal Multicapa (MLP)** representó un primer paso hacia la mejora en la predicción de residuos de precipitación, aunque es evidente que se requiere una mayor capacidad computacional para obtener mejores resultados. Se recomienda explorar futuras mejoras en el modelo, incluyendo la optimización de hiperparámetros y la integración de características adicionales, como la humedad del suelo o la radiación solar, que pueden influir en los patrones de precipitación en diversas regiones geográficas.

Conclusiones

Este estudio ha demostrado la importancia de validar los productos satelitales como CHIRPS en regiones geográficamente diversas como México, donde la variabilidad en la precipitación es influenciada por factores topográficos y climáticos. Aunque CHIRPS mostró una buena concordancia general con los datos de las estaciones meteorológicas, los resultados indican que su precisión puede mejorarse así como la identificación de la fuente del error, particularmente en escalas temporales más finas y en áreas de baja altitud. Los modelos de **Random Forest** y **Red Neuronal Multicapa (MLP)** ofrecieron una perspectiva complementaria, destacando la capacidad de las redes neuronales para capturar relaciones más complejas entre las variables, aunque su desempeño aún está limitado por la capacidad computacional disponible. Futuras investigaciones deberán centrarse en optimizar estos modelos y explorar enfoques más avanzados para mejorar la precisión de las predicciones en entornos climáticos desafiantes.

Referencias Bibliográficas

López-Bermeo, S., et al. (2022). Validation of Satellite Precipitation Estimates in Tropical South America: A Multiscale Perspective. *Journal of Hydrometeorology*, 23(4), 789-802. <https://doi.org/10.1175/JHM-D-21-0185.1>

- Alsilibe, M., et al. (2023). Evaluation of Satellite Precipitation Products for Water Resources Management in Tropical Regions. *International Journal of Climatology*, 43(2), 325-342. <https://doi.org/10.1002/joc.7730>
- da Silva, P. (2024). CHIRPS Validation and Its Applications in the Brazilian Cerrado: Detecting Rainfall Variability and Spatial Trends. *Water Resources Research*, 60(1), 105-115. <https://doi.org/10.1029/2024WR029583>
- Moreau, D. y Wiebels, P. (2024). Open Science in Climate Research: Improving Transparency and Reproducibility with Meta-Analysis. *Earth Science Informatics*, 17(2), 311-330. <https://doi.org/10.1007/s12145-023-00728-0>
- Google Earth Engine (2024). Google Earth Engine: A Platform for Environmental Data & Analysis. <https://earthengine.google.com/>
- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>
- Paszke, A., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32, 8024-8035. <https://doi.org/10.48550/arXiv.1912.01703>
- Gorelick, N., et al. (2017). Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sensing of Environment*, 202, 18-27. <https://doi.org/10.1016/j.rse.2017.06.031>

Anexos

Se anexa el código en formato Google Colab

https://colab.research.google.com/drive/1pyLJzjDLk3_OBVC-F45P5eIOLRGljJvX?usp=sharing