VISIÓN ACADÉMICA
ISSN: 3103-1536

Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 445-460

DOI: https://doi.org/10.70577/0dpbb154

Análisis Comparativo de Estrategias de Pronóstico de Sequías Meteorológicas: De la Estadística Clásica a los Algoritmos de Inteligencia Artificial

Darío Xavier Zhiña [1]
Maestría en Estadística con mención en Ciencia de Datos e Inteligencia Artificial, Evaluación de Riesgos Ambientales en Sistemas de Producción y Servicios (RISKEN), Carrera de Ingeniería Ambiental, Facultad de Ciencias Químicas
https://orcid.org/0000-0001-9556-4025
dario.zhinia@espoch.edu.ec
dario.zhina0608@ucuenca.edu.ec
Escuela Superior Politécnica de Chimborazo (ESPOCH)- Universidad de Cuenca
Cuenca – Ecuador

Andrea Hernández Allauca [2]
Facultad de Recursos Naturales
https://orcid.org/0000-0001-6413-5607
andrea.hernandez@espoch.edu.ec
Escuela Superior Politécnica de Chimborazo (ESPOCH)
Riobamba – Ecuador

Cómo citar:

Análisis Comparativo de Estrategias de Pronóstico de Sequías Meteorológicas: De la Estadística Clásica a los Algoritmos de Inteligencia Artificial. (2026). *Visión Académica*, *4*(3), 445-460. https://doi.org/10.70577/0dpbb154

Fecha de recepción: 2026-07-06

Fecha de aceptación: 2026-07-20

Fecha de publicación: 2026-08-03

Resumen

El pronóstico de sequías meteorológicas en cuencas tropicales de montaña se ve limitado a la complejidad orográfica, variabilidad climática y a la escasez de datos. El presente estudio compara cinco modelos de pronóstico (ARIMA/SARIMA, ETS, Random Forest, XGBoost y SVM-RBF) aplicados al Índice Estandarizado de Precipitación (SPI) en escalas de 1, 3, 6 y 12 meses, utilizando registros mensuales de precipitación (1970–2023) de dos estaciones de la cuenca alta del río Paute al sur del Ecuador.

Los modelos fueron calibrados y validados mediante una división 80/20 de entrenamiento-prueba y evaluados con diferentes métricas de desempeño, así como con diferentes pruebas estadísticas. Los resultados indican que la escala temporal del índice SPI fue el principal factor que condiciona el desempeño predictivo en los modelos. Los modelos de aprendizaje automático superaron a los modelos estadísticos clásicos desde SPI-3 a SPI-12, sin embargo, en evaluaciones de pronóstico móvil en el SPI-12, los cinco modelos no presentaron diferencias significativas. La complejidad del modelo no garantizó mejoras en el pronóstico, por lo que se destaca la utilidad de modelos parsimoniosos para el monitoreo operativo de sequías en cuencas andinas las cuales enfrentan escasez de datos.

Palabras clave: Sequías meteorológicas, Índice Estandarizado de Precipitación (SPI), Pronóstico de sequías, Aprendizaje automático, Modelos de pronóstico, Cuenca del río Paute, Escasez de datos.

Comparative Analysis of Forecasting Strategies for Meteorological Droughts: From Classical Statistics to Artificial Intelligence Algorithms

Abstract

The forecasting of meteorological droughts in tropical mountain basins is limited by orographic complexity, climate variability, and data scarcity. This study compares five forecasting models (ARIMA/SARIMA, ETS, Random Forest, XGBoost, and SVM-RBF) applied to the Standardized Precipitation Index (SPI) at 1, 3, 6, and 12-month timescales, using monthly precipitation records (1970–2023) from two stations in the upper Paute River basin in southern Ecuador.
The models were calibrated and validated using an 80/20 training-test split and evaluated with different performance metrics and statistical tests. The results indicate that the timescale of the SPI was the main factor influencing the predictive performance of the models. Machine learning models outperformed classical statistical models from SPI-3 to SPI-12; however, in moving forecast assessments in SPI-12, the five models showed no significant differences. Model complexity did not guarantee improved forecasting, highlighting the usefulness of parsimonious models for operational drought monitoring in Andean basins, which face data scarcity.

Keywords: Meteorological droughts, Standardized Precipitation Index (SPI), Drought forecasting, Machine learning, Forecasting models, Paute River basin, Data scarcity.

Introducción

La variabilidad climática y la intensificación de los extremos hidrometeorológicos se encuentran entre los desafíos científicos y sociales más importantes del siglo XXI. Es así, que las sequías se han convertido en una de los fenómenos más extendidos y dañinos debido al cambio climático, con impactos en la disponibilidad de agua y la seguridad alimentaria (Orimoloye, 2022), el funcionamiento de los ecosistemas, la producción de energía (Kapica et al., 2024) y la economía. La literatura reciente muestra un incremento en la frecuencia, duración y la severidad de las sequías en las últimas décadas, y, además, muestra que las regiones montañosas y tropicales resultan especialmente vulnerables (Liu y Chen, 2021). Debido a esto, es necesario el desarrollo de herramientas de pronóstico eficientes, no solo en el ámbito académico sino también en la gestión operativa del riesgo climático y la gobernanza sostenible del agua, tomando una atención especial los territorios de elevada variabilidad hidroclimática como la región andina.

Entre estas regiones, los Andes tropicales constituyen un caso especial debido a los pronunciados gradientes altitudinales, los regímenes de precipitación convectiva y la heterogeneidad orográfica los cuales generan una variabilidad espacial y temporal (Espinoza et al., 2020). Estas características, junto con patrones estacionales no lineales, tendencias débiles de largo plazo y una elevada variabilidad estocástica, dificultan la modelación y la predicción de precipitación en cuencas andinas (Buytaert et al., 2006). Como ejemplo, se tiene la cuenca alta del río Paute, en el sur del Ecuador, la cual es de gran importancia estratégica por su contribución a la generación hidroeléctrica, la agricultura y el abastecimiento de agua a zonas urbanas. Su régimen de precipitación responde a las teleconexiones asociadas con El Niño-Oscilación del Sur (ENOS) y a otras oscilaciones climáticas de gran escala (Mora y Willems, 2012; Campozano et al., 2016).

Los índices de sequía proporcionan representaciones cuantitativas de la severidad, la magnitud y la duración del fenómeno a partir de variables climáticas, y permiten identificar anomalías respecto de las condiciones de referencia (Ndayiragije y Li, 2022). Entre estos índices, se tiene el Índice Estandarizado de Precipitación (SPI,

por sus siglas en inglés), propuesto por McKee et al. (1993), el cual se ha consolidado como el índice estándar para el monitoreo y el pronóstico de la sequía meteorológica, debido a su formulación probabilística, su estructura multiescalar y su dependencia exclusiva de datos de precipitación (World Meteorological Organization, 2012). Debido a su naturaleza multiescala, el SPI relaciona las agregaciones de corto plazo (1 y 3 meses) con las sequías meteorológicas y agrícolas, y las de largo plazo (6 y12 meses) con los procesos hidrológicos, lo que requiere una evaluación multiescala para obtener pronósticos más robustos.

El pronóstico de las series de SPI se lo ha realizado tradicionalmente con modelos clásicos, como ARIMA/SARIMA (Achite et al., 2022) y los modelos de suavizamiento exponencial (ETS). Estos métodos ofrecen un fundamento teórico transparente y formulaciones parsimoniosas, pero su capacidad para representar dependencias no lineales y cambios de régimen es limitada. En respuesta a ello, la literatura reciente ha recurrido cada vez más a algoritmos de aprendizaje automático (Random Forest, Extreme Gradient Boosting (XGBoost) y las máquinas de vectores de soporte), que son capaces de capturar relaciones no lineales complejas e interacciones de alto orden entre predictores rezagados (Hukkeri et al., 2023; Zeng et al., 2025). Diversas revisiones sistemáticas muestran un crecimiento del pronóstico de sequías basado en aprendizaje automático, al tiempo que advierten sobre la heterogeneidad metodológica y el uso limitado de la inferencia estadística formal para sustentar las decisiones de selección de modelos (Fung et al., 2020; Prodhan et al., 2022).

A pesar de la creciente evidencia en la literatura aun persisten tres brechas metodológicas en este tema. En primer lugar, la mayoría de los estudios comparativos se ha realizado en regiones con redes de monitoreo densas y registros relativamente homogéneos (Poudel et al., 2024), lo cual limita la transferencia de las conclusiones en áreas montañosas con datos escasos. En segundo lugar, las comparaciones de modelos suelen basarse únicamente en métricas descriptivas de error (por ejemplo, RMSE o MAE), sin evaluar si las diferencias observadas en el desempeño predictivo son estadísticamente significativas. En tercer lugar, la evaluación simultánea de modelos estadísticos clásicos y algoritmos de aprendizaje automático sobre múltiples escalas de agregación del SPI (1, 3, 6 y 12 meses), sigue siendo poco frecuente en la literatura andina.

Como respuesta a estas brechas de conocimiento, el presente estudio propone un marco reproducible y estadísticamente comparable para el pronóstico del índice SPI en la cuenca alta del río Paute. Los objetivos de esta investigación son: (i) calcular el SPI en las escalas de 1, 3, 6 y 12 meses en dos estaciones representativas; (ii) implementar, calibrar y validar los cinco modelos bajo una partición cronológica 80/20 y evaluarlos con seis métricas complementarias; (iii) determinar la significancia estadística de las diferencias mediante las pruebas de Diebold-Mariano y Harvey-Leybourne-Newbold con corrección para comparaciones múltiples; y (iv) contrastar la evaluación de origen fijo frente a la de origen móvil.

Materiales y métodos

Área de estudio y datos meteorológicos

El estudio se realizó en la cuenca alta del río Paute, en la región andina del sur del Ecuador, una de las cuencas hidrológicamente más estratégicas del país (Figura 1). La cuenca abastece al complejo hidroeléctrico Paute (que genera una fracción sustancial de la electricidad nacional) y a la ciudad de Cuenca, que depende de sus cuencas de cabecera para usos domésticos y agrícolas (Celleri et al., 2007). La climatología regional

se rige por un régimen bimodal de precipitación, con picos durante marzo-abril y octubre-noviembre y una estación relativamente más seca entre junio y agosto, modulada por la interacción entre el transporte de humedad del Pacífico, los flujos amazónicos del este y el forzamiento orográfico de la cordillera oriental. La topografía compleja, con elevaciones que varían aproximadamente entre 2.500 y más de 4.400 m s. n. m., produce fuertes gradientes altitudinales de temperatura y precipitación, lo que la hace sensible a la variabilidad interanual asociada con el ENOS y a los efectos de largo plazo del cambio climático (Morán-Tejeda et al., 2016).

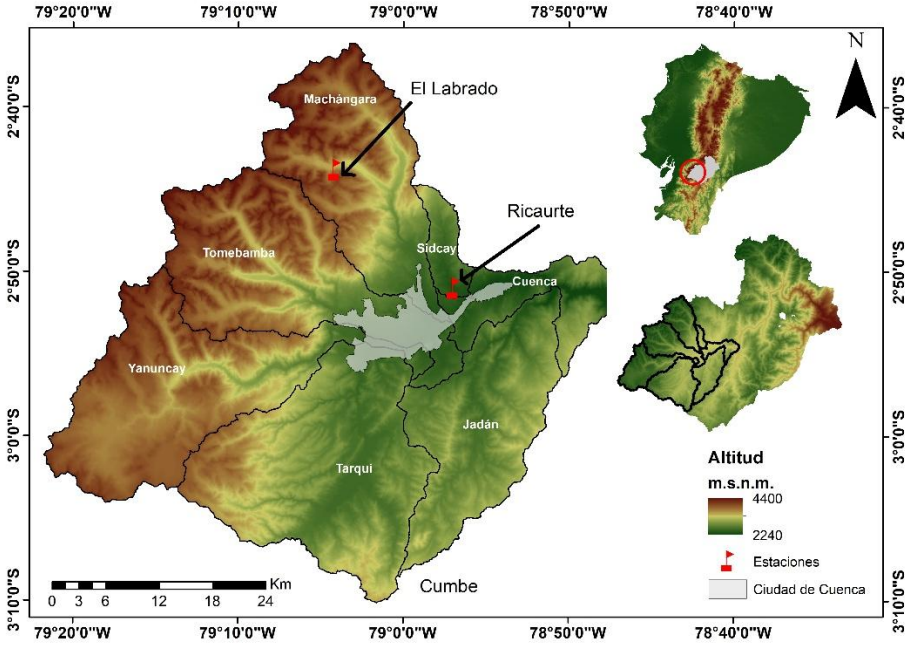

Figura 1. *Ubicación de la zona de estudio.*

El análisis se centró en dos estaciones meteorológicas (Tabla 1) operadas por el Instituto Nacional de Meteorología e Hidrología del Ecuador (INAMHI): El Labrado y Ricaurte. Ambas estaciones fueron seleccionadas por la calidad y continuidad de sus registros y por su contraste hidroclimático. El Labrado, situada en la parte alta de la cuenca y Ricaurte, ubicada a menor altitud y más próxima al área urbana de la ciudad de Cuenca. Las dos estaciones representan el gradiente altitudinal que gobierna la variabilidad de la precipitación en el sistema, y permite evaluar la robustez de los pronósticos en condiciones climáticas distintas. Los registros mensuales de precipitación (1970–2023) fueron procesados por el Grupo de Evaluación de Riesgos Ambientales en Sistemas de Producción y Servicios (RISKEN) de la Universidad de Cuenca.

Tabla 1. Estaciones meteorológicas utilizadas y sus coordenadas geográficas.

Estación	Código	Longitud (°)	Latitud (°)	Característica
El Labrado	M0141	-79,07	-2,73	Páramo de altura (cabecera)
Ricaurte	M0426	-78,95	-2,85	Valle interandino (baja)

Índice Estandarizado de Precipitación

El Índice Estandarizado de Precipitación (SPI), propuesto por McKee et al. 1993, se empleó como indicador de sequía debido a su simplicidad, naturaleza multiescala y amplio reconocimiento para el monitoreo de

sequías meteorológicas por parte de la literatura. El SPI mide la desviación estandarizada de la precipitación acumulada respecto a la climatología histórica para un período determinado. Se calcularon las escalas de 1, 3, 6 y 12 meses (SPI-1, SPI-3, SPI-6 y SPI-12), las cuales representan, anomalías meteorológicas de corto plazo, déficits de humedad del suelo a escala estacional y sequías hidrológicas de largo plazo (Vicente-Serrano et al., 2010). Los cálculos se realizaron en el entorno estadístico *R* mediante el paquete SPEI (Beguería et al., 2014).

Partición cronológica e ingeniería de variables

Para preservar la estructura temporal de las observaciones y evitar la fuga de información, se implementó una partición cronológica: el 80 % inicial de cada serie se destinó a la etapa de calibración y el 20% restante a la validación, lo que corresponde aproximadamente a 43 años de entrenamiento y 10 de prueba por estación. Para los modelos de aprendizaje automático, la serie univariada del SPI se reformuló como un problema de regresión supervisada. La matriz de predictores estuvo compuesta por doce rezagos autorregresivos (rezago-1 a rezago-12), que cubren un año hidrológico completo, y dos codificaciones cíclicas del mes calendario, sen(2πm/12) y cos(2πm/12), que preservan la naturaleza periódica del ciclo estacional sin introducir discontinuidades artificiales en los límites de año.

Modelos clásicos de series temporales

Los modelos ARIMA y su extensión estacional SARIMA constituyen un referente estadístico consolidado para el pronóstico de series de tiempo (Box et al., 2015). La identificación óptima del mejor modelo se lo realizó mediante una búsqueda no escalonada con selección basada en el criterio de información de Akaike corregido (AICc), y los residuos del modelo seleccionado se evaluaron con la prueba de Ljung-Box (Hyndman y Khandakar, 2008). Como alternativa, se emplearon modelos de suavizamiento exponencial de espacio de estados (ETS: error, tendencia, estacionalidad), cuyos componentes pueden ser aditivos, multiplicativos o nulos, con selección automática por AICc.

Modelos de aprendizaje automático

Se implementaron y compararon modelos de aprendizaje automático. Random Forest el cual combina las predicciones de numerosos árboles de regresión decorrelacionados, construidos a partir de muestras bootstrap, y reduce la varianza de predicción al promediar sus salidas (Breiman, 2001). XGBoost el cual es una implementación escalable y regularizada de árboles potenciados por gradiente que construye un ensamble aditivo de forma etapada, en el que cada árbol minimiza los errores residuales del ensamble vigente (Chen y Guestrin, 2016). La regresión de vectores de soporte con núcleo de base radial (SVM-RBF) estima relaciones no lineales mediante una función de regresión ε-insensible, y equilibra la planitud del modelo y la exactitud mediante un parámetro de regularización (Vapnik, 1995; Belayneh y Adamowski, 2012). Los hiperparámetros de los tres modelos se ajustaron mediante una búsqueda en malla con validación cruzada de origen móvil, lo que representa una adecuada estrategia para series temporales, debido a que evita el uso de información futura durante la calibración del modelo (Bergmeir y Benítez, 2012). Previamente, todos los predictores fueron centrados y escalados.

Métricas de evaluación del desempeño

El desempeño de los modelos (estadísticos y de machine learning) se evaluó sobre el conjunto de validación mediante el uso de métricas complementarias. La magnitud del error se cuantificó con la raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE). El coeficiente de determinación (R^2) y la eficiencia de Nash-Sutcliffe (NSE) miden el ajuste respecto de la media climatológica: NSE = 1 indica concordancia perfecta, NSE = 0 equivale a la media climatológica y NSE < 0 refleja un desempeño inferior a dicha referencia (Nash y Sutcliffe, 1970). La eficiencia de Kling-Gupta (KGE) descompone el desempeño en correlación, variabilidad y sesgo (Gupta et al., 2009). Por último, el estadístico U^2 de Theil compara el modelo con un pronóstico ingenuo de no cambio (persistencia): valores menores que 1 indican un desempeño superior al de referencia (Theil, 1966).

Comparación estadística de la exactitud de los pronósticos

Para determinar si las diferencias en el desempeño predictivo de los modelos son estadísticamente significativas, se realizaron comparaciones pareadas de los errores fuera de muestra con la prueba de Diebold-Mariano (DM) bajo pérdida cuadrática (Diebold y Mariano, 1995), con la corrección de muestra pequeña de Harvey-Leybourne-Newbold (HLN) (Harvey et al., 1997). Todas las comparaciones se sometieron a contraste bilateral al 5 %. Debido a que las comparaciones múltiples incrementan la probabilidad de falsos positivos, se aplicó la corrección de Holm-Bonferroni (Holm, 1979). El ordenamiento final de los modelos combinó las seis métricas de validación en un puntaje de rango promedio, para garantizar que el modelo seleccionado fuese robusto en múltiples dimensiones de desempeño. Todos los análisis se ejecutaron en R (versión 4.5.1) con los paquetes SPEI, forecast, tseries, caret, ranger, xgboost, kernlab y tidyverse.

Resultados

La escala de agregación como control dominante de la habilidad predictiva

En las dos estaciones de la cuenca alta del río Paute, la escala temporal de agregación del índice SPI controló la habilidad predictiva con mayor fuerza que la estación o el modelo. En el conjunto de validación, el desempeño aumentó desde SPI-1 hasta SPI-12 en ambas estaciones y en todas las familias de modelos, lo que indica que la escala del índice, y no el tipo de modelo, es el que controla los resultados de predicción.

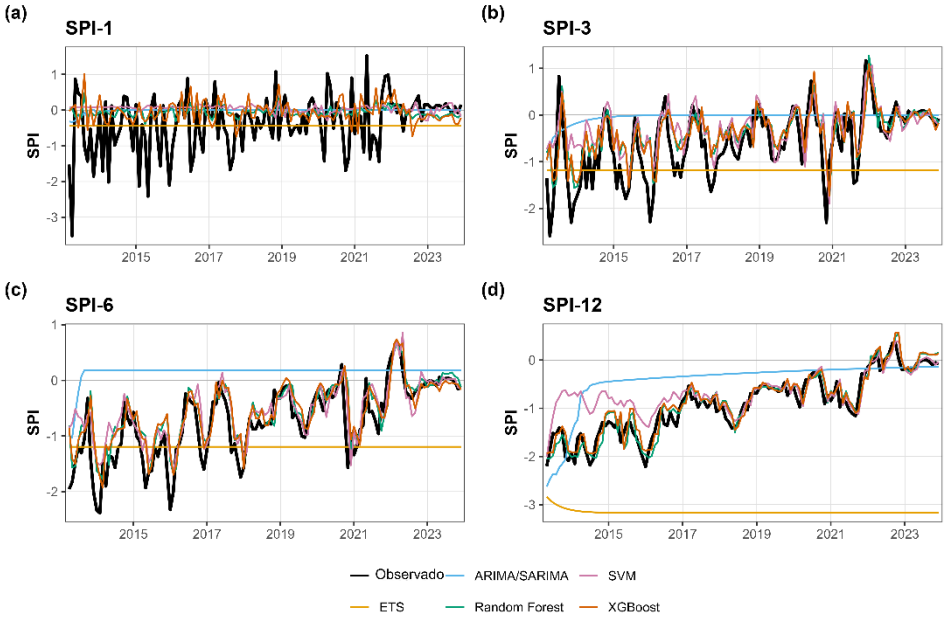

Figura 2. Series SPI observadas frente a las pronosticadas durante todo el período de validación en El Labrado para SPI-1, SPI-3, SPI-6 y SPI-12.

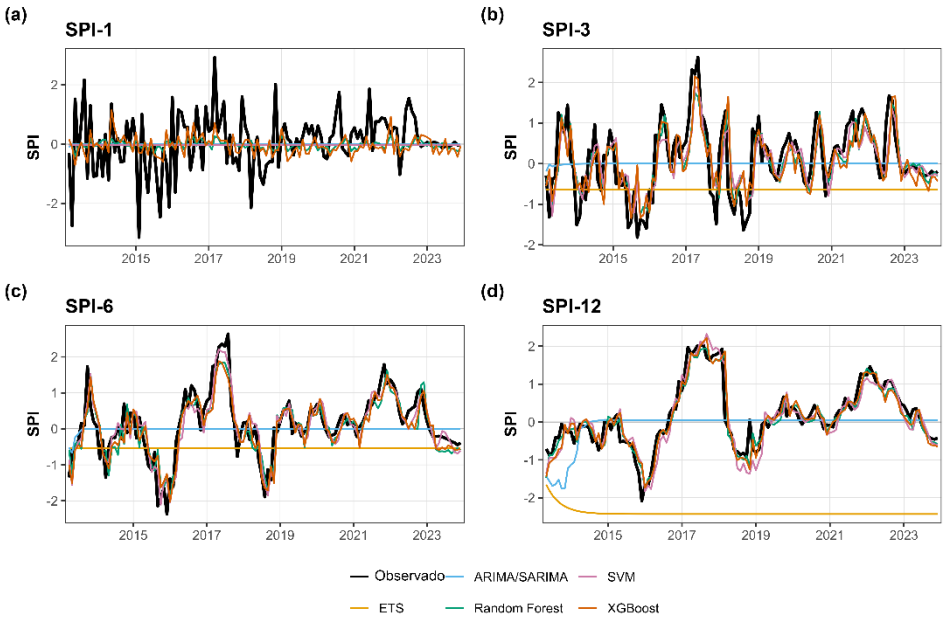

Figura 3. Series SPI observadas frente a las pronosticadas durante todo el período de validación en la estación Ricaurte para SPI-1, SPI-3, SPI-6 y SPI-12.

En la escala mensual (SPI-1), todos los modelos mostraron una exactitud débil: ninguna combinación de estación y modelo alcanzó un valor positivo de NSE. En El Labrado, ARIMA obtuvo el mejor resultado (NSE = -0,163), y en Ricaurte, ARIMA nuevamente lideró, aunque con habilidad prácticamente nula (NSE = -0,007). Dado que NSE = 0 equivale a la media climatológica, ningún modelo mostró una habilidad efectiva en la escala SPI-1. Este comportamiento responde principalmente a la naturaleza hidroclimática de la cuenca y no a las limitaciones de los modelos. A escala mensual, la precipitación está controlada por procesos convectivos, el efecto orográfico y la estacionalidad bimodal, como se mencionó previamente, lo que genera anomalías con baja persistencia temporal. Como resultado, los rezagos autorregresivos y la información

estacional aportan una capacidad predictiva limitada, por lo que los modelos tienden a converger hacia un pronóstico cercano a la media climatológica.

La predictibilidad de los modelos aumentó en las escalas estacionales y se fortaleció en la escala anual. En la escala SPI-3, los mejores modelos alcanzaron NSE de validación entre 0,35 y 0,53 (Random Forest en El Labrado, 0,351; SVM en Ricaurte, 0,527). En SPI-6, el desempeño mejoró hasta valores de 0,64 a 0,76, y en SPI-12 alcanzó su máximo, entre 0,86 y 0,87 (XGBoost en El Labrado, 0,872; Random Forest en Ricaurte, 0,856). Este comportamiento se explica por el filtrado del ruido de alta frecuencia mediante la ventana de acumulación, que resalta los componentes de baja frecuencia asociados a la memoria hidrológica de la cuenca, como la humedad del suelo, la regulación del páramo y la influencia del ENOS. Además, el solapamiento de las ventanas de acumulación incrementa la autocorrelación del índice, reforzada por la persistencia hidrológica, lo que genera series de SPI-6 y SPI-12 con mayor predictibilidad (Tabla 2).

El RMSE disminuyó de 0,86 y 0,96 en SPI-1 a 0,22 y 0,32 en SPI-12; sin embargo, esta mejora debe interpretarse con cautela, debido a que las escalas de agregación más largas suavizan la serie, reducen su varianza y conducen naturalmente a errores absolutos menores. El RMSE por lo tanto, no representa una base adecuada para la comparación entre escalas; la NSE normalizada por varianza describe el gradiente de escala de manera más fiable. No obstante, los valores de U^2 inferiores a 1 indican que, aun sin superar a la climatología según el NSE, los modelos presentan un mejor desempeño que un pronóstico basado únicamente en la persistencia

Tabla 2. Mejor modelo (por rango multimétrico consolidado) y habilidad de validación en cada estación y escala del SPI.

Estación	Escala	Mejor modelo	NSE	RMSE	KGE	U^2
El Labrado	SPI-1	ARIMA	-0,163	0,859	-0,558	0,786
El Labrado	SPI-3	Random Forest	0,351	0,581	0,400	1,061
El Labrado	SPI-6	Random Forest	0,637	0,396	0,725	1,102
El Labrado	SPI-12	XGBoost	0,872	0,224	0,898	1,105
Ricaurte	SPI-1	ARIMA	-0,007	0,959	-0,717	0,705
Ricaurte	SPI-3	SVM-RBF	0,527	0,605	0,377	0,897
Ricaurte	SPI-6	SVM-RBF	0,764	0,429	0,641	0,822
Ricaurte	SPI-12	Random Forest	0,856	0,319	0,691	1,006

Nota. NSE = eficiencia de Nash-Sutcliffe; RMSE = raíz del error cuadrático medio; KGE = eficiencia de Kling-Gupta; U^2 = estadístico de Theil relativo a un pronóstico de persistencia. La KGE debe interpretarse con cautela para índices de media cercana a cero (véase la Discusión).

Comportamiento por familia de modelos

La comparación entre la etapa de calibración y validación mostró diferencias en la capacidad de generalización de los modelos según la escala del SPI. En SPI-1, los modelos de aprendizaje automático alcanzaron altos valores de ajuste durante el entrenamiento, pero un bajo desempeño en la validación. Por ejemplo, en El Labrado, XGBoost obtuvo un R^2 de 0,99 en entrenamiento y un NSE negativo en validación, lo que indica un sobreajuste. En cambio, al aumentar la escala de agregación, la diferencia entre ambas etapas disminuyó. En SPI-12, los altos valores de R^2 (0,96-0,98) se acompañaron de NSE entre 0,86 y 0,87, reflejando una mayor capacidad de generalización.

El desempeño relativo de los modelos también varió con la escala temporal. En SPI-1, los modelos clásicos (ARIMA/SARIMA y ETS) presentaron los mejores resultados, aunque con una capacidad predictiva limitada. A partir de SPI-3, los modelos de aprendizaje automático mostraron un desempeño superior, siendo el modelo SVM-RBF el que obtuvo los mejores resultados en Ricaurte para SPI-3 y SPI-6, mientras que Random Forest destacó en El Labrado para las escalas intermedias. En SPI-12, los mejores resultados correspondieron a modelos basados en árboles, con XGBoost en El Labrado y Random Forest en Ricaurte. El modelo, ETS presentó un desempeño menos consistente en las escalas de mayor agregación, mientras que ARIMA/SARIMA mantuvo un comportamiento más estable, aunque con menor capacidad para representar las anomalías persistentes observadas en SPI-6 y SPI-12.

Significancia estadística de las diferencias de pronóstico

Las comparaciones pareadas mediante las pruebas DM y HLN muestran en la mayoría de los casos, los modelos de aprendizaje automático superaron a los clásicos en SPI-3, SPI-6 y SPI-12, con significancia frecuente en el nivel $p < 0,001$ y la mayor ventaja frente a ETS en las escalas largas (Tabla 3). Dada la magnitud de los conjuntos de validación, los estadísticos corregidos y sin corregir resultaron casi idénticos. Sin embargo, las diferencias dentro del grupo de modelos de aprendizaje automático fueron en general no significativas. Por ejemplo, Random Forest y XGBoost mostraron resultados comparables en casi todas las combinaciones, y el modelo de SVM-RBF fue superior en determinadas escalas intermedias sin sostener esa ventaja en SPI-12. Luego de aplicar la corrección de Holm-Bonferroni, las diferencias entre modelos de aprendizaje automático y clásicos desde SPI-3 hasta SPI-12 se mantuvieron significativas ($p < 0,001$), mientras que las diferencias internas del grupo de aprendizaje automático dejaron de serlo. En consecuencia, para las escalas de acumulación largas, la mejora principal proviene del uso de métodos no lineales frente a los lineales, y no de la elección de un algoritmo no lineal particular.

Tabla 3. Pruebas HLN que comparan el modelo mejor clasificado con cada modelo competidor, por estación y escala del SPI (período de validación).

Estación	Escala	Mejor	vs ARIMA	vs ETS	vs RF	vs XGB	vs SVM
El Labrado	SPI-1	ARIMA	-	-1,70	+0,76	+1,89	+3,31**
El Labrado	SPI-3	RF	+6,29***	+5,47***	-	+0,27	+0,97
El Labrado	SPI-6	RF	+9,08***	+6,91***	-	+1,72	+0,29
El Labrado	SPI-12	XGBoost	+8,79***	+21,18***	+0,37	-	+4,75***
Ricaurte	SPI-1	ARIMA	-	+1,20	+0,54	-0,13	+1,20
Ricaurte	SPI-3	SVM	+4,84***	+6,60***	+0,99	+1,34	-
Ricaurte	SPI-6	SVM	+5,14***	+6,64***	+3,38**	+2,52*	-
Ricaurte	SPI-12	RF	+6,75***	+16,76***	-	+1,56	+2,07*

Nota. Valores positivos indican que el modelo mejor clasificado presenta una pérdida de error cuadrático significativamente menor. Niveles de significancia: * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$. «—» indica la columna correspondiente al propio modelo de referencia.

Evaluación con origen móvil: habilidad de un paso y horizonte de predictibilidad

La evaluación de las Secciones 3.1 a 3.3 aplicó a los modelos clásicos un único origen fijo con pronósticos multipaso a lo largo de toda la ventana de prueba, mientras que los modelos de aprendizaje automático, al incorporar predictores rezagados, operaron efectivamente en un esquema de un paso adelante. Este desajuste afecta sobre todo a ARIMA/SARIMA y ETS, cuyos pronósticos multipaso revierten con el tiempo a la media. Para comparar la eficiencia de los pronósticos en igualdad de condiciones, se realizó un análisis adicional de avance progresivo (origen móvil) en la estación El Labrado, para SPI-1 y SPI-12, generando pronósticos de 1 a 12 meses de anticipación en cada origen y contrastándolos con las observaciones correspondientes.

En SPI-12 (Figura 4), los resultados difirieron notablemente de los del origen fijo. A diferencia de los pronósticos multipaso, que tendían a producir predicciones excesivamente suaves, los pronósticos de un paso de ARIMA/SARIMA y ETS mostraron una concordancia estrecha con las observaciones, con NSE de 0,891 y 0,887, R^2 próximos a 0,89 y RMSE cercanos a 0,21 (Tabla 4). Los ensambles de árboles los igualaron (Random Forest, NSE = 0,877; XGBoost, NSE = 0,883), y las pruebas HLN pareadas confirmaron que estos cuatro modelos son estadísticamente indistinguibles ($|HLN| \leq 0{,}91$; $p \geq 0{,}36$). La aparente inferioridad de los modelos clásicos en el análisis de origen fijo fue, por tanto, en gran medida un artefacto del protocolo multipaso: una vez que todos los modelos reciben el mismo conjunto de información, los modelos lineales y de suavizamiento exponencial parsimoniosos recuperan la fuerte autocorrelación a escala anual del SPI tan bien como los algoritmos más flexibles.

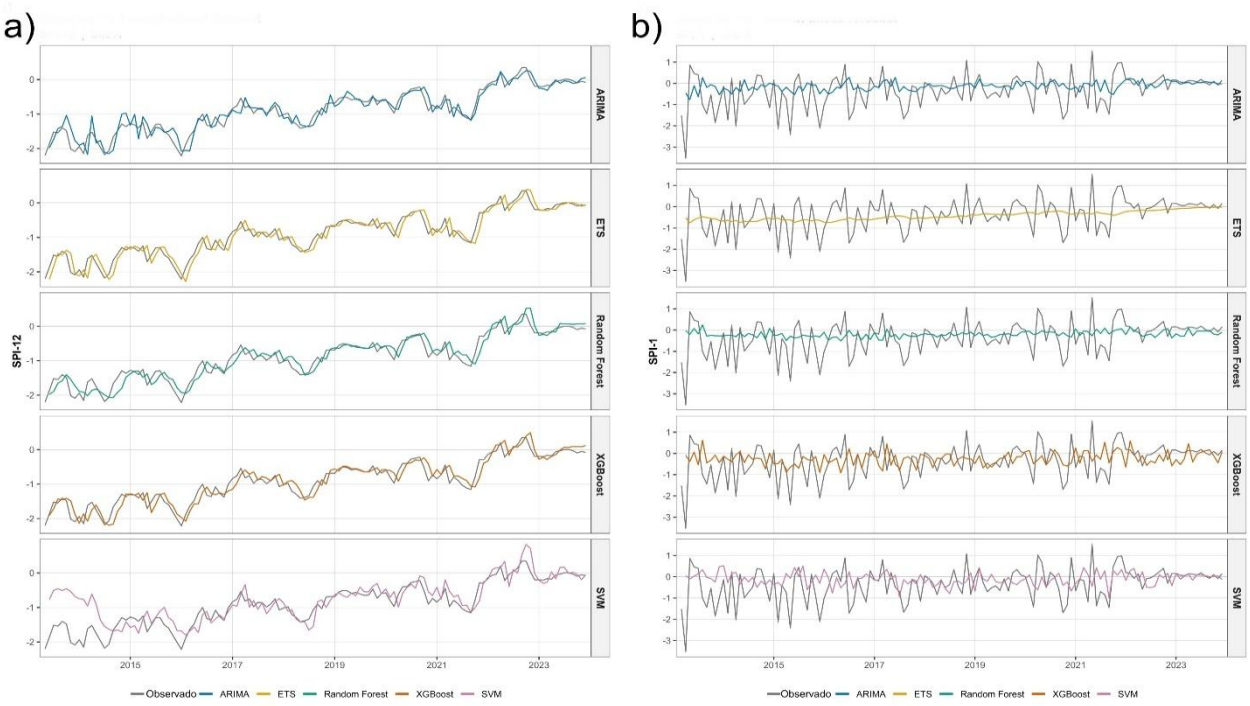

Figura 4. Comparación de datos observados y pronosticados con un avance de un paso (h = 1) en Labrado, por modelo: ARIMA, ETS, Random Forest y XGBoost.

En SPI-12, la única excepción fue el modelo SVM-RBF, que presentó el desempeño más bajo en todos los horizontes de pronóstico, incluso para un mes de anticipación (NSE = 0,538; R^2 = 0,613). Las comparaciones mediante la prueba HLN mostraron diferencias significativas frente a los demás modelos (estadísticos entre -3,9 y -4,1; $p < 0,001$). Este comportamiento sugiere que el núcleo de base radial y la función de pérdida ε-insensible tienden a suavizar en exceso las relaciones presentes en los predictores rezagados, reduciendo la capacidad del modelo para representar la variabilidad característica de SPI-12. En contraste, el análisis por horizonte evidenció dos comportamientos claramente diferenciados. Para SPI-1 (Figura 4), todos los modelos mostraron una capacidad predictiva muy limitada desde el primer horizonte, con valores de R^2 próximos a cero y NSE negativos, lo que indica que extender el horizonte de pronóstico no aporta información útil. En cambio, para SPI-12 la precisión disminuyó de manera progresiva a medida que aumentó el horizonte de predicción: el NSE descendió desde aproximadamente 0,89 para un paso hasta valores cercanos a cero en los horizontes más largos. Entre los modelos evaluados, Random Forest y XGBoost conservaron un NSE positivo hasta alrededor de 12 meses de anticipación, ETS hasta aproximadamente 11 meses, ARIMA/SARIMA hasta cerca de 8 meses, mientras que SVM-RBF únicamente mantuvo un desempeño aceptable en el primer horizonte.

Tabla 4. Habilidad de pronóstico de origen móvil de un paso (h=1) en El Labrado para las escalas más corta y más larga del SPI, con el horizonte de predictibilidad (máxima anticipación con NSE positivo).

Escala	Modelo	NSE (h=1)	RMSE (h=1)	R^2 (h=1)	U^2 (h=1)	NSE > 0 hasta (meses)
SPI-1	ARIMA	-0,094	0,829	0,009	0,756	—
SPI-1	ETS	0,031	0,780	0,042	0,712	≈7
SPI-1	Random Forest	-0,071	0,820	0,000	0,748	—
SPI-1	XGBoost	-0,162	0,854	0,000	0,779	—
SPI-1	SVM-RBF	-0,248	0,885	0,000	0,808	—
SPI-12	ARIMA	0,891	0,205	0,892	1,005	8
SPI-12	ETS	0,887	0,208	0,894	1,021	11
SPI-12	Random Forest	0,877	0,217	0,887	1,067	12
SPI-12	XGBoost	0,883	0,212	0,892	1,040	12
SPI-12	SVM-RBF	0,538	0,421	0,613	2,065	1

Nota. U^2 referenciado a un pronóstico de persistencia de h pasos; $U^2 < 1$ indica habilidad sobre la persistencia.

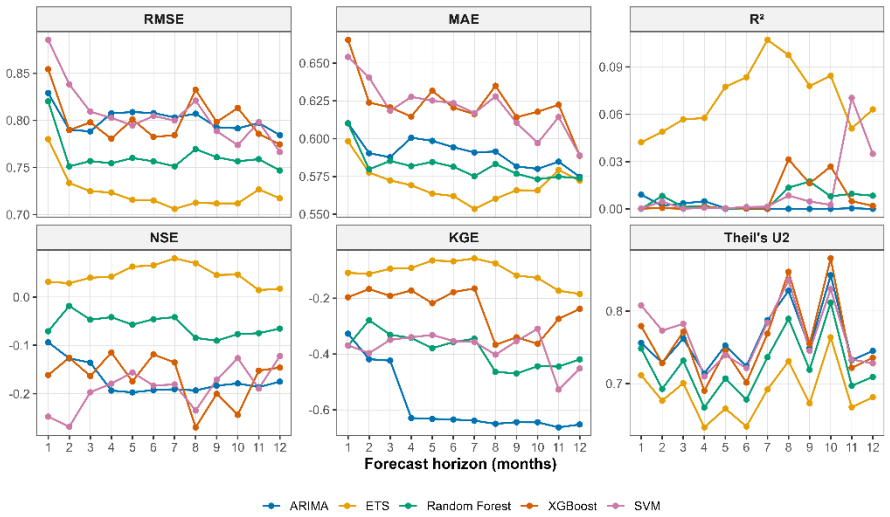

Figura 5. Desempeño del pronóstico en SPI-1 (origen móvil).

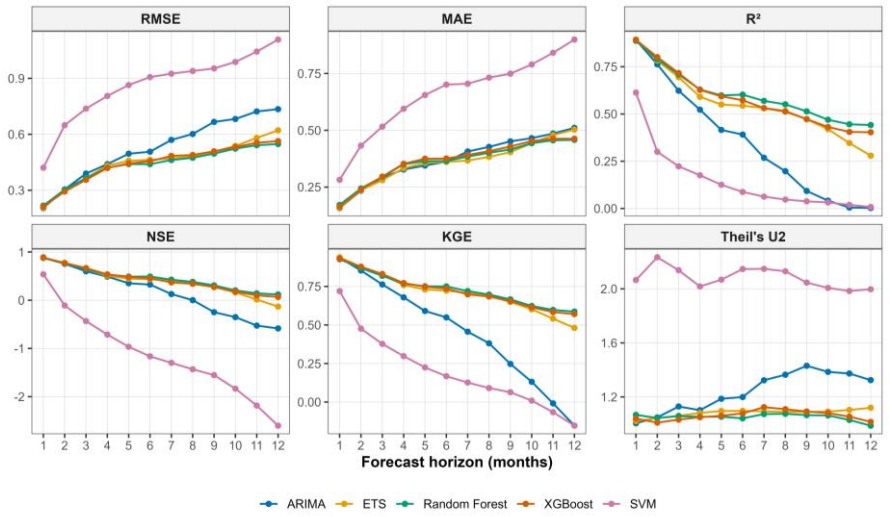

Figura 6. Desempeño del pronóstico en SPI-12 (origen móvil).

Discusión

Los resultados obtenidos en la investigación confirman que la escala temporal del SPI es el principal factor que determina la capacidad predictiva de los modelos. Tanto en El Labrado como en Ricaurte se observó un patrón consistente. La predicción fue limitada en SPI-1 y SPI-3, mientras que el desempeño mejoró de forma importante en SPI-6 y, especialmente, en SPI-12, o cual indica que la persistencia introducida por las escalas de acumulación más largas proporciona una señal más estable y predecible, favoreciendo tanto a los modelos estadísticos como a los de aprendizaje automático. Este comportamiento coincide con estudios previos que reportan un aumento de la habilidad predictiva conforme aumenta la escala del SPI (Belayneh y Adamowski, 2012; Poudel et al., 2024), así como con investigaciones recientes que muestran el buen desempeño del modelo XGBoost en índices de sequía de largo plazo (Zeng et al., 2025).

El experimento de validación mediante origen móvil aporta un resultado relevante a nuestro estudio. Aunque en la validación convencional Random Forest y XGBoost obtuvieron los mejores

indicadores para SPI-12, las diferencias desaparecieron cuando todos los modelos fueron comparados bajo un esquema de pronóstico de un paso. En estas condiciones, ARIMA/SARIMA, ETS, Random Forest y XGBoost presentaron desempeños estadísticamente equivalentes. Este resultado sugiere que, cuando la serie posee una elevada autocorrelación, el incremento en la complejidad del modelo no necesariamente se traduce en una mejora significativa de la capacidad predictiva. En consecuencia, parte de la superioridad atribuida a los algoritmos de aprendizaje automático puede depender del protocolo de validación empleado más que del algoritmo en sí, una situación señalada previamente en estudios metodológicos (Bergmeir y Benítez, 2012).

La comparación entre métricas también evidenció que la interpretación del desempeño depende del criterio de evaluación utilizado. En algunas combinaciones la eficiencia de Kling-Gupta (KGE) tomó valores muy negativos a pesar de que NSE, R^2 y RMSE indicaban un buen ajuste. Este comportamiento se explica debido a que el índice SPI está estandarizado alrededor de una media cercana a cero, condición que vuelve muy sensible el componente de sesgo de la métrica KGE. De manera similar, el coeficiente U^2 de Theil mostró una interpretación diferente de la NSE en SPI-12 debido a que utiliza como referencia un modelo de persistencia, el cual resulta particularmente competitivo cuando la autocorrelación aumenta con la escala temporal, o cual muestra que la evaluación del pronóstico de sequías debe basarse en un conjunto de métricas y no en un único indicador.

Desde un punto de vista práctico, los resultados muestran que los modelos evaluados ofrecen un desempeño suficiente para el pronóstico de sequías de mediano y largo plazo, especialmente en las escalas de SPI-6 y SPI-12, escalas que son de interés para la gestión de recursos hídricos y la planificación agrícola. Dado que los modelos clásicos alcanzaron un rendimiento comparable al de los algoritmos de aprendizaje automático en el experimento de origen móvil, ARIMA/SARIMA y ETS constituyen alternativas confiables cuando se requiere una implementación sencilla e interpretable en instituciones con recursos computacionales limitados.

Entre las principales limitaciones del estudio se encuentra el uso exclusivo de información univariada, sin incorporar índices climáticos de gran escala como los índices asociados al ENOS, que podrían mejorar la capacidad predictiva en las escalas de corto y mediano plazo. De igual forma, el análisis de origen móvil se realizó únicamente en una estación, por lo que su aplicación en otras estaciones permitiría evaluar la consistencia espacial de los resultados. Finalmente, ampliar el período de validación para incluir un mayor número de eventos climáticos extremos podría contribuir a fortalecer la evaluación de los modelos bajo diferentes condiciones hidroclimáticas.

Conclusiones

La comparación de cinco modelos de pronóstico para el Índice Estandarizado de Precipitación (SPI) en cuatro escalas de agregación (SPI-1, SPI-2, SPI-3 y SPI-12) y en dos estaciones representativas de la cuenca alta del río Paute, muestra que la escala de agregación fue el factor determinante de la exactitud predictiva, con mayor influencia que la ubicación de la estación o el

tipo de modelo; el pronóstico en SPI-1 con información puramente autorregresiva resultó inefectivo, mientras que las escalas de 6 y 12 meses ofrecieron una predictibilidad útil para la gestión operativa del agua. Bajo un protocolo de origen fijo, los modelos de aprendizaje automático superaron a los clásicos desde SPI-3 según las pruebas de Diebold-Mariano y Harvey-Leybourne-Newbold, sin embargo, bajo una evaluación de origen móvil de un paso, ARIMA/SARIMA, ETS, Random Forest y XGBoost no presentaron diferencias significativas en SPI-12, y SVM-RBF fue el modelo con el menor desempeño en ambos extremos de la escala. Se concluye que la ventaja del aprendizaje automático en escalas largas depende en parte del protocolo de evaluación y que la mayor complejidad no garantiza una mejora cuando la serie es suficientemente estable, por lo que los modelos parsimoniosos e interpretables constituyen una base adecuada para el monitoreo operativo de sequías y para futuros estudios en cuencas tropicales de altura con datos escasos. El trabajo futuro debería concentrarse en incorporar predictores climáticos como los indicadores del ENOS y en examinar el desempeño de los modelos bajo un clima cambiante.

Declaración de disponibilidad de datos: Los datos que respaldan los hallazgos de este estudio están disponibles a través del autor de correspondencia, previa solicitud.

Agradecimientos: Este manuscrito fue elaborado y presentado como parte del cumplimiento de los requisitos para la obtención del título de **Magíster en Estadística, Ciencia de Datos e Inteligencia Artificial** en la Escuela Superior Politécnica de Chimborazo (ESPOCH), donde la titulación se realiza mediante la publicación de un artículo científico. Darío Zhiña agradece el apoyo del proyecto de investigación ***"El Niño y las Sequías en el Ecuador: Un Análisis Futuro de sus Conexiones ante el Cambio Climático"***, financiado por el Vicerrectorado de Investigación e Innovación de la Universidad de Cuenca, Cuenca, Ecuador. El presente manuscrito se desarrolló en el marco de dicho proyecto de investigación.

Conflictos de interés: Los autores declaran no tener conflictos de interés.

Referencias Bibliográficas

Achite, M., Bazrafshan, O., Azhdari, Z., Wałęga, A., Krakauer, N., & Caloiero, T. (2022). Forecasting of SPI and SRI using multiplicative ARIMA under climate variability in a Mediterranean region: Wadi Ouahrane basin, Algeria. Climate, 10(3), 36. https://doi.org/10.3390/cli10030036

Belayneh, A., & Adamowski, J. (2012). Standard precipitation index drought forecasting using neural networks, wavelet neural networks, and support vector regression. Applied Computational Intelligence and Soft Computing, 2012, 1–13. https://doi.org/10.1155/2012/794061

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control. John Wiley & Sons.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Buytaert, W., Celleri, R., Willems, P., De Bièvre, B., & Wyseure, G. (2006). Spatial and temporal rainfall variability in mountainous areas: A case study from the south Ecuadorian Andes. Journal of Hydrology, 329(3–4), 413–421. https://doi.org/10.1016/j.jhydrol.2006.02.031

Campozano, L., Célleri, R., Trachte, K., Bendix, J., & Samaniego, E. (2016). Rainfall and cloud dynamics in the Andes: A southern Ecuador case study. Advances in Meteorology, 2016, 1–15. https://doi.org/10.1155/2016/3192765

Celleri, R., Willems, P., Buytaert, W., & Feyen, J. (2007). Space–time rainfall variability in the Paute basin, Ecuadorian Andes. Hydrological Processes, 21(24), 3316–3327. https://doi.org/10.1002/hyp.6575

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

Espinoza, J. C., Garreaud, R., Poveda, G., Arias, P. A., Molina-Carpio, J., & Masiokas, M. (2020). Hydroclimate of the Andes Part I: Main climatic features. Frontiers in Earth Science, 8, 64. https://doi.org/10.3389/feart.2020.00064

Fung, K. F., Huang, Y. F., Koo, C. H., & Soh, Y. W. (2020). Drought forecasting: A review of modelling approaches 2007–2017. Journal of Water and Climate Change, 11(3), 771–799. https://doi.org/10.2166/wcc.2019.236

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. International Journal of Forecasting, 13(2), 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.

Hukkeri, G. S., Naganna, S. R., Pruthviraja, D., Bhat, N., & Goudar, R. H. (2023). Drought forecasting: Application of ensemble and advanced machine learning approaches. IEEE Access, 11, 141375–141393. https://doi.org/10.1109/ACCESS.2023.3341587

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(3). https://doi.org/10.18637/jss.v027.i03

Kapica, J., Jurasz, J., Canales, F. A., Bloomfield, H., Guezgouz, M., & De Felice, M. (2024). The potential impact of climate change on European renewable energy droughts. Renewable and Sustainable Energy Reviews, 189, 114011. https://doi.org/10.1016/j.rser.2023.114011

Liu, Y., & Chen, J. (2021). Future global socioeconomic risk to droughts based on estimates of hazard, exposure, and vulnerability in a changing climate. Science of the Total Environment, 751, 142159. https://doi.org/10.1016/j.scitotenv.2020.142159

McKee, T. B., Doesken, N. J., & Kleist, J. (1993). The relationship of drought frequency and duration to time scales. In Proceedings of the 8th Conference on Applied Climatology (Vol. 17, pp. 179–184).

Mora, D. E., & Willems, P. (2012). Decadal oscillations in rainfall and air temperature in the Paute River basin - southern Andes of Ecuador. Theoretical and Applied Climatology, 108(1–2), 267–282. https://doi.org/10.1007/s00704-011-0527-4

Morán-Tejeda, E., Bazo, J., López-Moreno, J. I., Aguilar, E., Azorín-Molina, C., & Sanchez-Lorenzo, A. (2016). Climate trends and variability in Ecuador (1966–2011). International Journal of Climatology, 36(11), 3839–3855. https://doi.org/10.1002/joc.4597

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology, 10(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Ndayiragije, J. M., & Li, F. (2022). Effectiveness of drought indices in the assessment of different types of droughts, managing and mitigating their effects. Climate, 10(9), 125. https://doi.org/10.3390/cli10090125

Orimoloye, I. R. (2022). Agricultural drought and its potential impacts: Enabling decision-support for food security in vulnerable regions. Frontiers in Sustainable Food Systems, 6, 838824. https://doi.org/10.3389/fsufs.2022.838824

Poudel, B., Dahal, D., Banjara, M., & Kalra, A. (2024). Assessing meteorological drought patterns and forecasting accuracy with SPI and SPEI using machine learning models. Forecasting, 6(4), 1026–1044. https://doi.org/10.3390/forecast6040051

Prodhan, F. A., Zhang, J., Hasan, S. S., Pangali Sharma, T. P., & Mohana, H. P. (2022). A review of machine learning methods for drought hazard monitoring and forecasting: Current research trends, challenges, and future research directions. Environmental Modelling & Software, 149, 105327. https://doi.org/10.1016/j.envsoft.2022.105327

Theil, H. (1966). Applied economic forecasting. North-Holland Publishing Company.

Vapnik, V. N. (1995). The nature of statistical learning theory. Springer. https://doi.org/10.1007/978-1-4757-2440-0

Vicente-Serrano, S. M., Beguería, S., & López-Moreno, J. I. (2010). A multiscalar drought index sensitive to global warming: The Standardized Precipitation Evapotranspiration Index. Journal of Climate, 23(7), 1696–1718. https://doi.org/10.1175/2009JCLI2909.1

World Meteorological Organization. (2012). Standardized Precipitation Index user guide. WMO.

Zeng, F., Gao, Q., Wu, L., Rao, Z., Wang, Z., & Zhang, X. (2025). Modeling short-term drought for SPEI in mainland China using the XGBoost model. Atmosphere, 16(4), 419. https://doi.org/10.3390/atmos16040419