Desarrollo de una metodología para la predicción estacional de déficits y excesos hídricos en los departamentos de Quindío, Risaralda y Caldas, mediante técnicas de *machine learning*.

Diana Carolina Arias Sinisterra
Oscar Hernan Estrada Vargas

Universidad ICESI
Facultad de Ingeniería, Maestría en Ciencia de Datos
Santiago de Cali, Colombia
2024

Desarrollo de una metodología para la predicción estacional de déficits y excesos hídricos en los departamentos de Quindío, Risaralda y Caldas, mediante técnicas de *machine learning*.

Diana Carolina Arias Sinisterra
Oscar Hernan Estrada Vargas

Proyecto Aplicado I

Tutor:
Diego Fernando Agudelo, M.Sc.
Asesor:
Camilo Barrios Perez, Ph.D.

Universidad ICESI
Facultad de Ingeniería, Maestría en Ciencia de Datos
Santiago de Cali, Colombia
2024

Resumen

Este estudio presenta el desarrollo de una metodología para pronosticar condiciones de exceso o déficit hídrico en la región del eje cafetero de Colombia (departamentos de Quindío, Risaralda y Caldas), utilizando el Índice Estandarizado de Precipitación-Evapotranspiración (SPEI) como principal indicador. La primera fase de la investigación se centró en la consolidación y homogeneización de datos climáticos, la caracterización de las condiciones hídricas de la región y la construcción del SPEI-3, que estima el balance hídrico usando los datos de precipitación y evapotranspiración de los últimos 3 meses, y el SPEI-6, que lo hace con los datos de los últimos 6 meses, para entender las variaciones a corto y mediano plazo. En primer lugar, se realizó una recopilación y homogeneización de datos de diversas fuentes climáticas, ajustándolos a una resolución uniforme para su adecuado análisis. Posteriormente, se llevó a cabo la caracterización de la zona de estudio, identificando las particularidades climáticas. Además, se realizó una comparación de los índices SPEI con períodos históricos de los fenómenos de El Niño y La Niña para resaltar la capacidad del SPEI de reflejar la realidad climática de la zona de estudio. Se observó que los valores del SPEI coinciden con las temporadas en que estos fenómenos ocurrieron en Colombia, validando así su utilidad como indicador de sequías y excesos hídricos. Se utilizó el software CPT para generar la predicción del SPEI-3 y SPEI-6 para el mes de marzo de 2024. La segunda fase del proyecto consistió en probar otros predictores para realizar la predicción mediante CCA y mediante un modelo de machine learning, para realizar la comparación de los resultados obtenidos por ambos métodos. Finalmente se destaca la importancia de pronosticar el SPEI con mayor exactitud, ya que esto no solo reflejaría la realidad climática de manera más precisa, sino que también proporcionaría una herramienta valiosa para la planificación y toma de decisiones en sectores industriales y agrícolas.

Palabras clave: SPEI**,** Déficit hídrico, exceso hídrico, pronóstico climático

Contenido

Lista de figuras

Introducción

La predicción estacional de déficits y excesos hídricos es crucial para la planificación agrícola y la gestión de recursos en las áreas cultivables en Colombia, ya que la adecuada preparación de los productores para afrontar situaciones climáticas extremas es clave para lograr la mejor productividad de sus cultivos. A pesar de los esfuerzos del Instituto de Hidrología, Meteorología y Estudios Ambientales del país (IDEAM) para predecir estos escenarios, utilizando métodos estadísticos tradicionales como análisis de correlaciones canónicas y regresiones lineales múltiples, existe la necesidad de evaluar nuevas metodologías basadas en técnicas de *machine learning* o *deep learning* para mejorar la capacidad predictiva de fenómenos agroclimáticos.

El déficit y el exceso hídrico afectan significativamente la producción agrícola, y especialmente al cultivo de café, un sector clave para la economía colombiana que a nivel nacional produjo 10.6 millones de sacos de 60 kg durante el 2023 (FNC, 2023). La capacidad de predecir estas condiciones con mayor precisión puede ayudar a mitigar riesgos, optimizar el uso del agua y mejorar la resiliencia de las comunidades agrícolas. En este contexto, el Índice Estandarizado de Precipitación y Evapotranspiración (SPEI) se presenta como un indicador esencial para la identificación de regiones con condiciones de sequías y excesos hídricos.

A pesar de su relevancia, la predicción de déficits y excesos hídricos mediante métodos estadísticos tradicionales presenta limitaciones. Por ello, este estudio propone el uso de técnicas avanzadas de *machine learning* para mejorar la precisión de los pronósticos estacionales, utilizando el indicador SPEI. Estas técnicas permiten identificar patrones complejos en los datos climáticos, ofreciendo una mayor robustez frente a las variaciones climáticas.

Este trabajo se desarrolla en un contexto donde la falta de un sistema nacional de pronóstico estacional preciso de los indicadores de déficits o excesos hídricos limita la

capacidad de anticipación y respuesta a condiciones climáticas adversas. Se busca, por tanto, evaluar el desempeño de un modelo de *machine learning* en la predicción estacional del SPEI en las áreas productoras de café de los departamentos de Quindío, Risaralda y Caldas.

El alcance de este estudio incluye la revisión de la literatura, la obtención y estandarización de datos climáticos, la caracterización de las zonas de estudio, la implementación de una línea base con métodos estadísticos tradicionales y la generación de predicciones mediante un modelo de aprendizaje automático, evaluando el desempeño de este.

1. Planteamiento del problema

En este capítulo se presentan las razones que motivan la realización de este trabajo. De acuerdo con Esquivel et al. (2018), se evidencia que en el país no existe un servicio que suministre pronósticos estacionales de indicadores de sequía o exceso de humedad, y cómo esta limitación afecta la capacidad de identificar anticipadamente las regiones agrícolas expuestas a condiciones climáticas extremas. Además, se observa que actualmente no se utilizan metodologías de *machine learning* para la elaboración de pronósticos estacionales. Finalmente, se declaran los objetivos y el alcance del proyecto aplicado.

1.1. Definición del problema

Colombia es un país situado en América del Sur, conocido por su rica biodiversidad y variabilidad climática, con una superficie de 1.141.749 km² y una población aproximada de 48.258.494 habitantes, según el Departamento Administrativo Nacional de Estadística (DANE, 2018). Administrativamente, el país se divide en 32 departamentos descentralizados y un distrito capital (Bogotá D.C.), cada uno con una diversidad de productos y procesos económicos influenciados por factores específicos como la geografía, el clima, las políticas departamentales, y la calidad del sistema educativo. Estos factores generan variabilidad en los bienes y servicios ofrecidos por cada departamento.

La predicción climática es un desafío importante para Colombia, especialmente en las áreas productoras de café ubicadas en los departamentos de Quindío, Risaralda y Caldas, sin embargo, actualmente, la falta de un sistema nacional de pronóstico estacional preciso de indicadores relacionados con estas anomalías limita la capacidad del país para anticipar y gestionar adecuadamente las condiciones de sequías y excesos de humedad.

El IDEAM ha empleado métodos estadísticos tradicionales como análisis de correlaciones canónicas y regresión lineal para la predicción de indicadores climáticos, no obstante, estos enfoques presentan limitaciones significativas en términos de precisión y adaptabilidad a las complejas dinámicas climáticas de Colombia. Debido a esto, surge la necesidad de explorar nuevas metodologías basadas en técnicas avanzadas de *machine learning* o *deep learning* para mejorar las predicciones de los indicadores relacionados con el déficit o exceso hídrico.

El Índice Estandarizado de Precipitación y Evapotranspiración (SPEI) es un indicador clave para la caracterización de sequías y excesos hídricos. Este índice permite evaluar el balance hídrico en diferentes escalas temporales, proporcionando una medida integral del estrés hídrico en las regiones agrícolas. Sin embargo, la predicción precisa del SPEI requiere modelos más sofisticados que puedan capturar las complejas interacciones climáticas y adaptarse a las variabilidades regionales.

El objetivo global de este estudio es evaluar el desempeño de modelos de *machine learning* en la predicción estacional del SPEI en las áreas productoras de café de Quindío, Risaralda y Caldas. Para ello, esta primera etapa del proyecto se enfoca en el establecimiento de una línea base para las predicciones del indicador, por lo tanto, se deben consolidar y homogeneizar las fuentes de información disponibles, caracterizar detalladamente las diferentes zonas de interés, identificar los predictores más relevantes y efectivos para la modelación climática, y se deben generar pronósticos estacionales mediante la metodología convencional a través de la herramienta *Climate Predictability Tool* - CPT (IRI, 2008).

El estudio utiliza datos públicos disponibles en plataformas abiertas, incluyendo predictores como la temperatura superficial del mar, la precipitación y la evapotranspiración. Los datos se utilizan para calcular el SPEI en diferentes escalas temporales (SPEI-3 y SPEI-6) y generar pronósticos mediante CPT, que se pueden evaluar utilizando los datos calculados de SPEI.

Este enfoque metodológico permite identificar las relaciones entre los diferentes predictores y el SPEI, facilitando la interpretación de los datos y proporcionando una herramienta valiosa para la gestión climática en las regiones cafeteras de Colombia. En

última instancia, el estudio contribuirá a mejorar la resiliencia de las comunidades agrícolas frente a las variaciones climáticas y a optimizar el uso de los recursos hídricos en el país.

1.2. Justificación

En Colombia, la ausencia de un sistema nacional de pronóstico estacional de alta precisión para déficits y excesos hídricos representa una limitación significativa para la gestión anticipada de las condiciones climáticas extremas, como sequías e inundaciones (Esquivel et al, 2018). Esta carencia impide una respuesta oportuna y adecuada que podría mitigar los impactos negativos en las regiones agrícolas del país.

Actualmente el Centro Internacional de Agricultura Tropical -CIAT- está liderando el proyecto “Colombia Agroalimentaria Sostenible”, iniciativa proyectada a 5 años y patrocinada principalmente por el Fondo Verde del Clima, la cual involucra a 9 sistemas productivos agrícolas y 15 socios, entre gremios y otras instituciones de investigación agropecuaria, tales como Agrosavia, Cenicaña y Cenicafé. Dentro de las necesidades que se pretenden abordar en los próximos años para el sector cafetero, se encuentra la identificación de una metodología que permita predecir adecuadamente las temporadas de excesos o déficits hídricos, especialmente en áreas clave como las zonas productoras de café de Quindío, Risaralda y Caldas.

Actualmente, el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) realiza análisis de correlaciones canónicas, regresiones por componentes principales y regresiones lineales, apoyándose en aplicaciones como el *Climate Predictability Tool* (CPT) del Instituto Internacional de Investigación para el Clima y la Sociedad (IRI) de la Universidad de Columbia. Esta es una herramienta muy versátil y muy explicativa, que permite observar espacialmente diferentes patrones climáticos, característica muy valiosa para los meteorólogos y climatólogos, sin embargo, estas metodologías presentan limitaciones en su capacidad para capturar la complejidad y variabilidad inherentes al clima colombiano.

El avance de las tecnologías de la información y el desarrollo de nuevos enfoques metodológicos, particularmente en el campo del *machine learning* y el *deep learning*,

ofrecen una oportunidad para superar estas limitaciones. Los algoritmos de *machine learning* tienen la capacidad de procesar grandes volúmenes de datos y descubrir patrones complejos no lineales que los métodos estadísticos tradicionales pueden pasar por alto. De esta manera, pueden mejorar significativamente la precisión y la fiabilidad de los pronósticos climáticos estacionales.

Implementar y evaluar modelos de predicción de indicadores de déficits y excesos hídricos basados en aprendizaje automático puede transformar la capacidad predictiva del país, proporcionando herramientas más robustas y adaptativas para la gestión de riesgos climáticos. Estos modelos pueden integrarse con datos climáticos históricos y actuales, permitiendo predicciones más precisas del Índice Estandarizado de Precipitación y Evapotranspiración (SPEI). Esto no solo puede mejorar la planificación y toma de decisiones en la agricultura, sino que también puede fortalecer la resiliencia de las comunidades frente a eventos climáticos adversos, asegurar un futuro más sostenible y seguro para las regiones agrícolas del país, haciendo un uso adecuado de los recursos hídricos y mitigando los riesgos asociados a fenómenos climáticos extremos.

1.3. Objetivos

1.3.1. Objetivo Principal

Evaluar el desempeño de un modelo de machine learning en la predicción estacional del Índice Estandarizado de Precipitación y Evapotranspiración (SPEI) en los departamentos de Quindío, Risaralda y Caldas.

1.3.2. Objetivos específicos

•Consolidar y homogeneizar la información de clima disponible procedente de las diversas fuentes y realizar una caracterización climática detallada de la zona de interés.

•Construir el indicador SPEI para la zona de estudio y generar un pronóstico usando CCA mediante la herramienta CPT.

•Evaluar el desempeño del modelo de pronóstico estacional para el SPEI.

•Implementar y evaluar un modelo de machine learning para la predicción del índice SPEI.

•Realizar un análisis comparativo de la capacidad de predicción entre el método de CCA y el modelo de machine learning implementado.

2. Revisión de antecedentes

Se han desarrollado algunos estudios en torno a la predicción de sequías y su impacto en la agricultura y los ecosistemas, utilizando metodologías estadísticas tradicionales, como lo realizado en Colombia por Esquivel et al., 2018, así como la aplicación de modelos de inteligencia artificial y métodos empíricos en otras partes del mundo. La precisión en la predicción de sequías es crucial para mitigar sus efectos adversos en la economía, la agricultura y los sistemas de recursos hídricos. Según Dong et al. (2023), "el uso de modelos avanzados de aprendizaje profundo puede mejorar significativamente nuestra capacidad para prever las características de las sequías y así facilitar estrategias de mitigación más efectivas". Esta afirmación se ve respaldada por múltiples estudios que han explorado diferentes metodologías para mejorar la predicción del Índice Estandarizado de Precipitación-Evapotranspiración (SPEI).

En este capítulo se presentan investigaciones relacionadas con la predicción climática estacional mediante métodos estadísticos tradicionales y la predicción de sequías agrícolas utilizando índices climáticos y métodos de aprendizaje automático, que incluyen variantes avanzadas de redes neuronales para la estimación del SPEI en diferentes zonas climáticas. Los documentos se presentan en orden cronológico ascendente.

Esquivel et al. (2018) evaluaron la predictibilidad de la precipitación estacional y la habilidad de pronóstico en cinco departamentos clave de Colombia para la agricultura de arroz, maíz y frijol. Utilizaron el análisis de correlación canónica con datos observados y modelados de la temperatura de la superficie del mar (SST) y de la precipitación. Analizaron diferentes predictores y tiempos de adelanto, encontrando que una mayoría significativa de las situaciones analizadas mostró buenos índices de correlación.

Los resultados indicaron que la predictibilidad era limitada en el este de Colombia y durante los períodos húmedos en los valles interandinos. Los pronósticos impulsados

por dos diferentes conjuntos de datos (ERSST y CFSv2) fueron consistentes, sugiriendo que ambos son valiosos para los pronósticos climáticos en Colombia. El estudio representa un primer paso hacia el establecimiento de un servicio climático sostenible y exitoso para la agricultura en el país, pero se identificaron áreas para mejorar, como la habilidad del pronóstico y la vinculación con aplicaciones agrícolas.

Tian et al. (2018) realizan un análisis sobre la predicción de sequías agrícolas en la cuenca del río Xiangjiang utilizando un modelo de Regresión de Vectores de Soporte (SVR). El objetivo del estudio es analizar la relación entre la humedad del suelo y la sequía, así como predecir la sequía agrícola mediante el uso del Índice de Precipitación-Evapotranspiración (SPEI). Para ello, el autor incorpora índices climáticos, incluyendo El Niño Oscilación del Sur (ENOS) y la Alta Subtropical del Pacífico Occidental (WPSH). La metodología se basa en la creación de un modelo SVR que utiliza estos índices climáticos para mejorar la precisión de las predicciones de sequía.

Los resultados indican que el SPEI a seis meses (SPEI-6) es el más efectivo para representar la humedad del suelo comparado con SPEI-3 y SPEI-1. Además, se encontró que el Punto de Cresta de WPSH es un factor clave que influye en la sequía agrícola, controlando principalmente la temperatura regional. El modelo SVR mejora la precisión de la predicción en un 4.4% en entrenamiento y un 5.1% en prueba para un tiempo de anticipación de tres meses, medido por el coeficiente de eficiencia de Nash-Sutcliffe (NSE). La mejora es más significativa para predicciones con un tiempo de anticipación de un mes (15.8% en entrenamiento y 27.0% en prueba). No obstante, el autor destaca la importancia de una cuidadosa selección de los parámetros de entrada para evitar efectos negativos debido a la información redundante.

Soh et al. (2018) exploran el uso de modelos de inteligencia artificial para la predicción del Índice Estandarizado de Precipitación-Evapotranspiración (SPEI) en la cuenca del río Langat, Malasia. El objetivo del estudio es mitigar el impacto de las sequías en la economía, el turismo, la agricultura y los sistemas de recursos hídricos. Utiliza dos modelos avanzados: Wavelet-ARIMA-ANN (WAANN) y Wavelet-Adaptive Neuro-Fuzzy Inference System (WANFIS), para predecir el SPEI a diferentes escalas de tiempo (1

mes, 3 meses y 6 meses). Los datos históricos del SPEI, desde 1976 a 2007, se emplean para entrenar los modelos y luego predecir para el periodo de prueba de 2008 a 2015.

La evaluación de los modelos se realiza mediante el Coeficiente de Determinación Ajustado (R^2 adj), Error Cuadrático Medio (RMSE), Error Absoluto Medio (MAE), Índice de Acuerdo de Willmott (d) y el Coeficiente de Eficiencia de Nash-Sutcliffe (E). Los resultados muestran que la precisión de predicción mejora con la longitud de la escala de tiempo. El modelo WAANN es superior para la predicción de SPEI-3 y SPEI-6, mientras que el modelo WANFIS tiene un rendimiento satisfactorio en la predicción de sequías a mediano plazo. El estudio concluye que el modelo WAANN ofrece mejor precisión tanto para la predicción de sequías a corto como a mediano plazo.

Dikshit et al. (2021) investigan el uso de la red neuronal de memoria a corto y largo plazo (LSTM) para la predicción del Índice Estandarizado de Precipitación-Evapotranspiración (SPEI) en Nueva Gales del Sur (NSW), Australia. El objetivo es desarrollar un modelo de pronóstico de sequías preciso y entender mejor las características de las sequías. El estudio compara el modelo LSTM con otros métodos de aprendizaje automático como Bosques Aleatorios y Redes Neuronales Artificiales, utilizando variables hidrometeorológicas como predictores y datos del conjunto de datos de la Unidad de Investigación Climática (CRU) desde 1901 hasta 2018.

Los resultados muestran que el modelo LSTM logra un valor de R^2 superior a 0.99 para SPEI-1 y SPEI-3. Además, el análisis de la variación en los resultados pronosticados de la categoría de sequía utilizando Curvas de Operación del Receptor (ROC-AUC) revela valores de AUC de 0.83 y 0.82 para SPEI-1 y SPEI-3, respectivamente. La variación espacial entre los valores observados y pronosticados se analiza para los meses de verano de 2016 a 2018. El estudio concluye que el modelo LSTM mejora las predicciones en comparación con los modelos de aprendizaje automático para un período de anticipación de 1 mes y puede ser útil para la mitigación de sequías.

Dong et al. (2023) desarrollan un estudio sobre la estimación del Índice Estandarizado de Precipitación-Evapotranspiración (SPEI) utilizando variantes de la red neuronal de memoria a corto y largo plazo (LSTM) en cuatro zonas climáticas de China. El objetivo es evaluar el rendimiento de estas variantes con datos meteorológicos limitados a nivel

nacional. Se comparan métodos empíricos, SVM, RNN, LSTM, BiLSTM y CNN-LSTM para la estimación de SPEI en diferentes escalas de tiempo.

Los resultados indican que BiLSTM es el modelo más adecuado para la estimación de SPEI a 3 meses, con valores de R^2, NSE y RMSE en el rango de 0.916 a 0.997, 0.907 a 0.997 y 0.143 a 0.353, respectivamente. Para otras escalas de tiempo, CNN-LSTM muestra un mejor rendimiento. El estudio concluye que las variantes de LSTM presentan un excelente desempeño para la estimación de SPEI a escala múltiple, proporcionando predicciones precisas de sequías meteorológicas, agroecológicas e hidrológicas en toda China. Además, se encontró que la precisión de los modelos de aprendizaje automático disminuye con la reducción del número de variables independientes.

En la tabla 1 se presenta un resumen de los antecedentes revisados.

Tabla 1. Resumen de antecedentes

Autores	Título	Año de publicación	Ubicación del estudio	Metodología utilizada
Esquivel, A., Llanos, L., Agudelo, D., Prager, S. D., Fernandes, K., Rojas, A., ... & Ramirez-Villegas, J.	Predictability of seasonal precipitation across major crop growing areas in Colombia	2018	Colombia	Correlaciones canónicas
Tian, Y., Xu, Y. P., & Wang, G.	Agricultural drought prediction using climate indices based on Support Vector Regression in Xiangjiang River basin.	2018	China	SVM
Soh, Y. W., Koo, C. H., Huang, Y. F., & Fung, K. F.	Application of artificial intelligence models for the prediction of standardized precipitation evapotranspiration index (SPEI) at Langat River Basin, Malaysia.	2018	Malasia	WAANN y WANFIS

Dikshit, A., Pradhan, B., & Huete, A.	An improved SPEI drought forecasting approach using the long short-term memory neural network.	2021	Australia	NN LSTM
Dong, J., Xing, L., Cui, N., Zhao, L., Guo, L., & Gong, D.	Standardized precipitation evapotranspiration index (SPEI) estimated using variant long short-term memory network at four climatic zones of China.	2023	China	NN LSTM

De acuerdo con estos antecedentes, se observa que algunos métodos estadísticos tradicionales, como el análisis de correlación canónica (CCA) utilizado en Colombia, han demostrado un éxito moderado en la predicción de la precipitación, pero con un desempeño limitado en ciertas regiones, dada la variabilidad climática del país, esto genera un vacío metodológico dentro de entidades como el IDEAM, que actualmente no cuentan con herramientas de pronóstico de déficits y excesos de humedad precisas y confiables. Otros estudios realizados en diferentes regiones del mundo sobre la predicción del SPEI, que utilizan diversos modelos, incluidos métodos de machine learning, series temporales y deep learning, comparten el objetivo común de mejorar el pronóstico de sequías pero con diferentes enfoques: el estudio de la cuenca del río Xiangjiang en China se centra en índices climáticos como el WPSH para predecir la sequía, mientras que los estudios de Malasia y Australia aplican modelos híbridos avanzados (WAANN, LSTM) para mejorar el desempeño de series temporales. El estudio en China de 2023 destaca la solidez de los modelos basados en redes neuronales LSTM para las condiciones climáticas de ese país.

El enfoque de este proyecto es llenar el vacío metodológico existente en Colombia, abordando las limitaciones en el desempeño de las predicciones que actualmente son generadas con CCA y evaluar si nuevas metodologías, como los modelos de machine learning, pueden ser una herramienta más precisa y confiable para el pronóstico de condiciones de déficits y excesos hídricos en zonas de producción agrícola en el país, dadas las características variadas y complejas del clima en ciertas regiones, como es el caso del eje cafetero.

3. Marco teórico

En este capítulo se desarrolla un marco teórico conceptual, en el que se presentan los conceptos necesarios para comprender el desarrollo de este estudio. Se establecen las definiciones de evapotranspiración, déficits y excesos hídricos, el Índice Estandarizado de Precipitación y Evapotranspiración (SPEI), y una descripción del modelo estadístico tradicional aplicado a la predicción climática (CPT).

Evapotranspiración: se conoce como evapotranspiración (ET) a la combinación de dos procesos separados por los que el agua se pierde a través de la superficie del suelo: el primero es la evaporación, y, por otra parte, mediante la transpiración de las plantas (Allen et al., 2006). La evaporación corresponde al agua líquida que se convierte en vapor de agua y se retira de superficies como lagos, ríos, suelo, etc. La transpiración por su parte consiste en el agua que se pierde como vapor de agua de los tejidos de las plantas hacia la atmósfera, ocurre principalmente a través de las estomas de las hojas. Estos dos procesos dependen del aporte de energía (radiación), el gradiente de presión de vapor en el ambiente y la velocidad del viento.

Existen varios métodos para calcular la evapotranspiración, entre los que se destacan el método de Penman-Monteith, el método de Hargreaves y el método de Thornthwaite. Estos métodos utilizan datos meteorológicos como la temperatura, la humedad relativa, la radiación solar y la velocidad del viento para estimar la pérdida de agua por evapotranspiración. Gracias a su exactitud, el método de cálculo de Penman-Monteith es el estándar recomendado por FAO (Allen et al., 2006), sin embargo, la calidad de sus resultados se debe principalmente a que requiere información de mayor cantidad de variables climáticas, comparado con los otros métodos. La fórmula de este método para calcular la evapotranspiración es la siguiente:

$$ETo = \frac{0.408\,\Delta(Rn - G) + \gamma \frac{900}{T + 273} u2(es - ea)}{\Delta + \gamma(1 + 0.34\,u2)}$$

donde:

ETo: evapotranspiración de referencia (mm día^{-1})

Rn: radiación neta en la superficie del cultivo (MJ m^{-2} día^{-1})

G: flujo del calor de suelo (MJ m^{-2} día^{-1})

T: temperatura media del aire a 2 m de altura (°C)

u2: velocidad del viento a 2 m de altura (m s^{-1})

es: presión de vapor de saturación (kPa)

ea: presión real de vapor (kPa)

Δ: pendiente de la curva de presión de vapor (kPa °C^{-1})

γ: constante psicrométrica (kPa °C^{-1})

El flujo de calor de suelo “G” se ignora cuando se realiza el cálculo a escala diaria. Si la velocidad del viento con la que se cuenta no está medida a los 2 metros de altura, el dato puede ajustarse multiplicándolo por el siguiente factor de conversión:

$$\frac{487}{ln\,(67.8\,z - 5.42)}$$

donde z es la altura a la que se midió la velocidad del viento (m).

La presión de vapor de saturación “es” se deriva de la temperatura máxima y mínima mediante la ecuación:

$$es = \frac{\left(0.6108\; e^{\left(\frac{17.27\,Tmax}{Tmax+237.3}\right)}\right) + \left(0.6108\; e^{\left(\frac{17.27\,Tmin}{Tmin+237.3}\right)}\right)}{2}$$

La presión real de vapor “ea” se calcula a partir de la presión de vapor de saturación y la humedad relativa promedio:

$$ea = \frac{es\;HumRel}{100}$$

La pendiente de la curva de presión “Δ” se calcula con la siguiente ecuación:

$$\Delta = \frac{4098\,[0.6108\,e^{\frac{17.27\,T}{T+237.3}}]}{(T+237.3)^2}$$

La constante psicrométrica "γ" se calcula mediante la siguiente ecuación:

$$\gamma = \frac{C_P P}{\varepsilon \lambda}$$

donde:

C_P: 0.001013

Ɛ: 0.622

λ: 2.45 MJ kg^{-1}

P: presión atmosférica

La presión atmosférica "P" se puede hallar mediante la ecuación:

$$P = 101.3\left(\frac{293 - 0.0065\,z}{293}\right)^{5.26}$$

donde z es la altitud (msnm).

Déficit hídrico: Se refiere a la situación en la cual la disponibilidad de agua en una región es insuficiente para satisfacer las necesidades de la vegetación, los cultivos y la demanda humana. Este fenómeno puede ser causado por una disminución en las precipitaciones, un aumento en la evapotranspiración, o ambos. Los déficits hídricos pueden tener impactos negativos significativos en la agricultura, la economía y la ecología de una región (SGS, 2023).

Exceso hídrico: La presencia excesiva de agua en áreas de cultivo puede alterar significativamente la estructura del suelo. Este fenómeno puede llevar a la compactación del suelo, lo cual reduce la aireación necesaria para las raíces y dificulta la regulación de la temperatura del suelo. Como resultado, se ven afectados los procesos físicos,

químicos y microbiológicos del suelo, especialmente aquellos que dependen de condiciones aeróbicas. (SGS, 2023).

Índices de sequía multiescalares: los índices de sequía multiescalares reconocen que la sequía es un fenómeno complejo que puede variar en diferentes escalas temporales y afectar diversos recursos hídricos, como la humedad del suelo, el agua subterránea y los caudales de los ríos. La importancia de estas escalas temporales radica en la acumulación de déficits hídricos, que puede diferir significativamente según el recurso considerado. Por ello, estos índices deben adaptarse a escalas temporales específicas para ser efectivos en la gestión y monitoreo de la sequía en distintos contextos.

Índice de Precipitación Estandarizada (SPI): Es una herramienta utilizada en climatología para evaluar si en una región y período específicos existe un déficit o exceso de precipitación en comparación con las condiciones normales. Se basa en analizar únicamente la precipitación como variable principal. Para calcular el SPI, se utilizan registros de precipitación de largo plazo, preferiblemente de 50 a 60 años, aunque algunos investigadores sugieren un mínimo de 30 años.

El cálculo del SPI implica ajustar una función de distribución de probabilidad, típicamente la función de distribución Gamma, a los datos de precipitación mensual acumulada para cada escala de tiempo y cada ubicación de interés en el área en estudio. Los valores resultantes del SPI se obtienen al transformar los valores de la distribución Gamma en valores de la variable normal estándar. En resumen, el SPI permite cuantificar y comparar anomalías de precipitación en diferentes lugares y escalas temporales.

Un evento de sequía se caracteriza por la persistencia de valores negativos del Índice de Precipitación Estandarizada (SPI) a cualquier escala de tiempo, alcanzando un valor de -1.0 o menor de forma continua. Este estado de sequía finaliza cuando el SPI se vuelve positivo. Esta definición se respalda en trabajos como el de McKee et al. (1993) y Vrochidou y Tsanis (2012).

La utilidad del SPI radica en su capacidad para comparar valores de sequía de manera sencilla y simultánea en distintos lugares y períodos de tiempo, tal como señalan Lopez-Bustins et al. (2013). Además, el SPI permite clasificar los eventos de sequía según su

intensidad, en función del rango de valores que alcanzan. Los eventos de sequía se clasifican de ligeros a extremos, como se presenta en la Tabla 2.

Tabla 2. Clasificación del índice SPI

Valor SPI	Clasificación
≥ 2.00	Extremadamente húmedo
1.50 a 1.99	Severamente húmedo
1.00 a 1.49	Moderadamente húmedo
0 a 0.99	Ligeramente húmedo (cercano a lo normal)
0 a -0.99	Sequía ligera (cercano a lo normal)
-1.00 a -1.49	Sequía moderada
-1.50 a -1.99	Sequía severa
≤ -2.00	Sequía extrema

El SPI no puede identificar el papel del aumento de la temperatura en las condiciones de sequía futuras, e independientemente de los escenarios de calentamiento global no puede tener en cuenta la influencia de la variabilidad de la temperatura y el papel de las olas de calor.

El Índice de Precipitación Estándar (SPI) presenta una limitación significativa al basarse únicamente en la precipitación, omitiendo otras variables que influyen en la demanda de agua atmosférica, como la temperatura, la velocidad del viento, la radiación solar y el déficit de presión de vapor (McEvoy et al., 2012). Para superar esta deficiencia, Vicente-Serrano et al. (2010) desarrollaron el Índice de Precipitación Estandarizado y Evaporación (SPEI). Este índice integra la sensibilidad del Índice de Sequía de Palmer (PDSI) a las variaciones en la demanda de evaporación provocadas por fluctuaciones y tendencias en la temperatura, con la simplicidad de cálculo y la capacidad multiescalar del SPI (Vicente-Serrano et al., 2010; Banimahd & Khalili, 2013).

El SPEI se fundamenta en un balance hídrico climático mensual que se obtiene restando la evapotranspiración potencial (ETP) de la precipitación, formando así una serie denominada D. Este índice se calcula en diversas escalas temporales y es matemáticamente similar al SPI, pero incorpora la evapotranspiración potencial (López-Moreno et al., 2013). Para el cálculo del SPEI, se emplea una distribución de probabilidad log-logística de tres parámetros, en lugar de la distribución Gamma de dos parámetros

utilizada para el SPI. Además, la clasificación de sequía definida por el SPI puede aplicarse también al SPEI.

Para calcular el SPEI, se determina la diferencia mensual entre la precipitación y la ETP, creando la serie temporal D^{ij} en milímetros (mm) para cada mes ***j*** del año ***i***. Luego esta serie se agrega en la escala temporal K, generando la serie D_k^{ij}. Este enfoque permite evaluar la disponibilidad hídrica en diferentes periodos, ofreciendo una perspectiva más completa sobre las condiciones de sequía al incorporar tanto la precipitación como la demanda evaporativa.

Este índice, puede utilizarse para determinar el inicio, la duración y la magnitud de las condiciones de sequía con respecto a las condiciones normales en diversos sistemas naturales y gestionados, como cultivos, ecosistemas, ríos, recursos hídricos, etc. Puede calcularse a escala mensual o acumularse a más de un mes.

Ráster: En su forma más simple, un ráster consta de una matriz de celdas (o píxeles) organizadas en filas y columnas (o una cuadrícula) en la que cada celda contiene un valor que representa información, como por ejemplo la temperatura. Los rásters pueden ser fotografías aéreas digitales, imágenes de satélite, imágenes digitales o incluso mapas escaneados. Entre sus ventajas está la simplicidad de su estructura: una matriz de celdas con valores que representan una coordenada, su formato potente para análisis espacial y estadístico avanzado y su capacidad de representar superficies continuas para llevar a cabo análisis de superficie (ESRI, 2024).

CPT: *Climate Predictability Tool* es un programa informático que permite construir un modelo de previsión climática estacional, validar el modelo y elaborar previsiones con datos actualizados. Su diseño se ha adaptado para producir previsiones climáticas estacionales utilizando correcciones de estadísticas de salida del modelo (MOS) a las predicciones climáticas del modelo de circulación general (GCM), o para producir previsiones utilizando campos de temperaturas de la superficie del mar o predictores similares. Aunque el software está específicamente diseñado para estas aplicaciones, puede utilizarse en entornos más generales para realizar análisis de correlación

canónica, regresión de componentes principales (PCR) o regresión lineal múltiple (MLR) sobre cualquier dato y para cualquier aplicación (Mason et al., 2024).

Correlación canónica (CCA): es un tipo de análisis estadístico lineal de múltiples variables, actualmente se usa en química, biología, meteorología, demografía, inteligencia artificial y ciencias de administración, entre otros campos, para analizar relaciones multidimensionales entre múltiples variables independientes y múltiples variables dependientes. El análisis de correlación canónica es el método más generalizado de la familia de las técnicas estadísticas multivariante. Se relaciona directamente con varios métodos de dependencia. Al igual que en la regresión, el objetivo de la correlación canónica es cuantificar la validez de la relación, en este caso entre los dos conjuntos de variables (dependiente e independiente).

El análisis de correlación canónica (CCA) busca identificar las combinaciones lineales de cada conjunto de variables que maximicen la correlación entre ellas, con el fin de evaluar la existencia de alguna asociación entre ambos conjuntos de variables. En situaciones donde uno de los conjuntos se define como variables predictoras o independientes y el otro como variables dependientes o de respuesta, ya sea por fundamentos teóricos o por el objetivo del estudio, el propósito del CCA es establecer si las variables predictoras influyen o explican las variables de respuesta. (Badii y Castillo, 2007).

Partiendo de n observaciones en dos conjuntos de variables $x_1, x_2, x_3 \dots x_p$ y $y_1, y_2, y_3 \dots y_q$, el primero representa a las variables explicativas(independientes) y el segundo a las variables de respuesta (dependientes). Por conveniencia $p \geq q$. Sea Z la matriz de datos se puede representar como:

$$Z = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix}$$

A partir de esta estructura la matriz de covarianza se puede estructurar de la siguiente forma:

$$S = \begin{pmatrix} S_{xx} & \vdots & S_{xy} \\ \dots & & \dots \\ S_{yx} & \vdots & S_{yy} \end{pmatrix}$$

El propósito del análisis de correlación canónica es identificar una combinación lineal de las variables predictoras X (independientes) que maximice la correlación con una

combinación lineal de las variables respuesta Y (dependientes). Específicamente, el objetivo es determinar la combinación lineal de las p variables independientes y las q variables dependientes que presenten la mayor correlación entre sí. Consiste entonces en encontrar las siguientes relaciones lineales

$$U = a_1x_1 + a_2x_2 + \cdots + a_px_p = xa$$

que tenga la correlación más alta con la siguiente combinación lineal

$$V = b_1y_1 + b_2y_2 + \cdots + b_qy_q = yb$$

Suponiendo que $U_1 y\, V_1$ son las variables con la máxima correlación $r_1 = cor(U_1, V_1)$. A continuación, se busca $U_2 y\, V_2$ con la máxima correlación $r_2 = cor(U_2, V_2)$, sin correlación con $U_1 y\, V_1$. Este proceso se repite para encontrar $U_3 y\, V_3$, y así sucesivamente, hasta obtener $m = min\,(p, q)$ parejas de variables canónicas. Cada par sucesivo tiene una correlación decreciente $r_1 > r_2 > r_3 \ldots > r_m$.

El procedimiento para maximizar la correlación es un problema de optimización, el cual al resolverse se consigue que:

$$\text{A} \left\{ S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx}a = \lambda a \right. \qquad \text{B} \left\{ S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}b = \lambda b \right.$$

Donde se puede probar que las matrices A y B tienen los mismos valores propios. La raíz cuadrada de los valores propios se denominan **correlaciones canónicas**.

Línea base: En el contexto del proyecto, la línea base es un punto de referencia utilizado para comparar los resultados obtenidos mediante distintas metodologías de predicción. En este estudio, se establece a partir de las predicciones generadas utilizando el método tradicional de correlaciones canónicas, las cuales se comparan con los resultados obtenidos mediante el modelo de Machine Learning.

El objetivo de definir una línea base es evaluar si el modelo de Machine Learning ofrece una mejora significativa respecto al enfoque convencional, en términos de precisión, eficiencia y capacidad predictiva. Así, la línea base permite realizar una comparación entre el enfoque tradicional y el modelo de machine learning propuesto en este estudio.

Métodos ensambladores: En el campo del aprendizaje automático, los métodos ensambladores (o *ensemble methods*) se han consolidado como una estrategia poderosa para mejorar la precisión y la robustez de los modelos predictivos. Estos métodos combinan múltiples modelos base, también conocidos como aprendices débiles, para generar un modelo final más sólido y preciso (López, 2018).

Existen dos enfoques principales en los métodos ensambladores:

- Métodos secuenciales: Construyen los modelos base de manera progresiva, como es el caso de algoritmos como *Boosting*. En este enfoque, cada modelo intenta corregir los errores del modelo anterior.
- Métodos paralelos: Generan varios modelos de manera independiente y combinan sus resultados. Este enfoque, utilizado en algoritmos como *Bagging* y **Random Forest**, busca reducir la varianza al promediar o votar entre múltiples predicciones.

Entre las ventajas de los métodos ensambladores se encuentran su capacidad para reducir el sobreajuste, mejorar la precisión de los modelos y manejar datos complejos con ruido o desequilibrios.

Random Forest: Dentro de los métodos ensambladores, Random Forest sobresale por su capacidad para combinar múltiples árboles de decisión que se entrenan de manera independiente utilizando el método de Bagging (agregación por muestreo). Cada árbol del modelo se construye a partir de una muestra aleatoria con reemplazo, conocida como muestra de arranque, tomada del conjunto de entrenamiento (Figura 1).

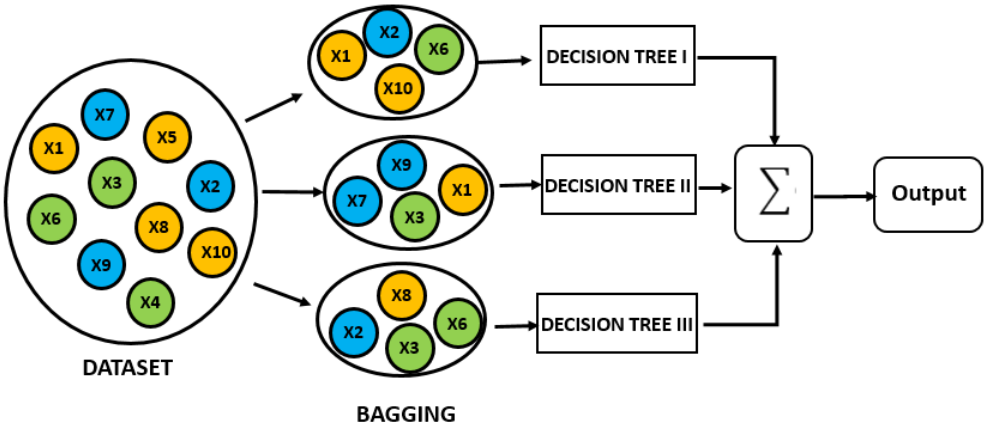


Figura 1. Ejemplo de Bagging. Tomada de Ensambladores: Random Forest - Parte I por M. López, 2018, Bookdown.

El bagging es una estrategia utilizada para disminuir la varianza en las predicciones al combinar los resultados de múltiples clasificadores, cada uno entrenado con distintos subconjuntos extraídos de la misma población de datos.

Para construir un modelo de random forest, se debe tener un conjunto de entrenamiento con N instancias, se selecciona aleatoriamente una muestra de estas instancias con reemplazo, lo que significa que una misma instancia puede ser elegida varias veces. Esta muestra constituye el conjunto de datos utilizado para entrenar cada árbol i.

Si el modelo cuenta con M variables de entrada, se define un número m<M que indica cuántas variables se seleccionan al azar en cada nodo del árbol. De estas m variables seleccionadas, se utiliza la mejor división posible para generar las ramas del árbol. Es importante destacar que m se mantiene constante durante la construcción de todo el bosque.

Cada árbol crece sin límites hasta alcanzar su máxima profundidad, ya que no se aplica un proceso de poda. Las predicciones para nuevas instancias se obtienen al combinar las salidas de los árboles: mediante el voto mayoritario en problemas de clasificación o el promedio en problemas de regresión.

El hiperparámetro más relevante para ajustar en el modelo Random Forest es el número de variables candidatas consideradas para dividir cada nodo (mtry). Sin embargo, también existen otros parámetros importantes que deben tenerse en cuenta:

- **Número de árboles:** Define la cantidad de árboles en el bosque. Este parámetro se ajusta para estabilizar el error del modelo. Sin embargo, aumentar excesivamente el número de árboles puede ser ineficiente en términos de tiempo y recursos computacionales.
- **Número de variables candidatas:** Especifica cuántas variables aleatorias se consideran en cada nodo para evaluar las posibles divisiones.
- **Tamaño de muestra:** Indica la proporción de datos utilizada para entrenar cada árbol.
- **Tamaño mínimo del nodo:** Determina el número mínimo de muestras requeridas en un nodo terminal. Este parámetro es clave para encontrar un equilibrio entre sesgo y varianza en el modelo.
- **Número máximo de nodos:** Establece el límite en la cantidad de nodos terminales que puede tener un árbol.

Estos hiperparámetros permiten ajustar el modelo para mejorar su precisión, estabilidad y eficiencia, adaptándolo a las características específicas del conjunto de datos (IBM, s.f.).

Modelo XGBoost: Es un algoritmo de aprendizaje automático supervisado que se utiliza tanto para tareas de clasificación como de regresión. Su nombre proviene de las siglas en inglés "Extreme Gradient Boosting", lo que refleja su enfoque en el proceso de *boosting* (refuerzo) mediante gradientes.

Este modelo está basado en árboles de decisión y se considera una mejora respecto a otros enfoques, como el de los bosques aleatorios y el refuerzo de gradientes tradicional. Una de sus principales ventajas es su capacidad para manejar grandes volúmenes de datos complejos, ya que incorpora varias técnicas de optimización, como la regularización y la reducción del sobreajuste, lo que le permite alcanzar mejor rendimiento en comparación con otros modelos de aprendizaje automático.

El proceso comienza ajustando un modelo inicial, como un árbol de regresión o clasificación, a los datos. Luego, se construye un segundo modelo que se enfoca en predecir con precisión las observaciones que el primer modelo predijo incorrectamente. Se espera que la combinación de estos modelos sea más precisa que cada uno individualmente. Este proceso de refuerzo se repite varias veces, con cada nuevo modelo intentando corregir los errores de los modelos anteriores (Figura *2*).

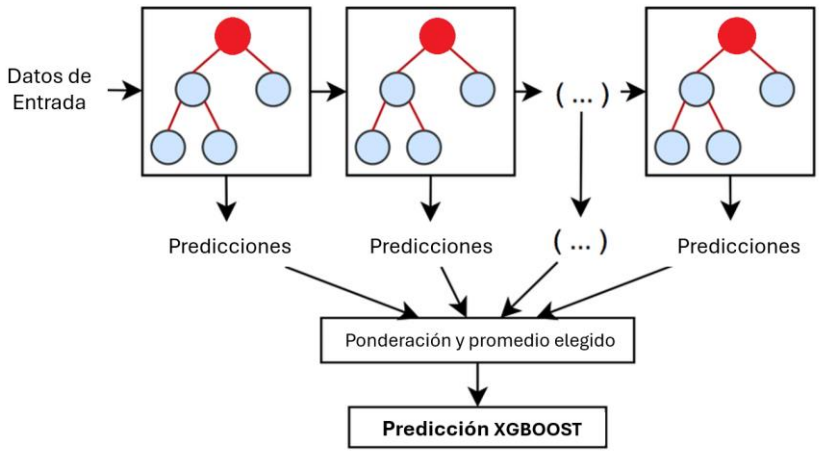


Figura 2 . Ejemplo ilustrativo de XGBoost. Tomado de Demajo, Lara. (2020). Explainable AI for Interpretable Credit Scoring.

El refuerzo de gradiente es una variante del boosting que se basa en la idea de minimizar el error de predicción global. En este enfoque, en cada paso, se calcula un peso para cada observación, que se utiliza en el siguiente modelo para reducir el error. Este peso se determina en función del gradiente del error con respecto a las predicciones del modelo. Es decir, cada nuevo modelo se ajusta para mejorar la predicción al "dar un paso" en la dirección que minimiza el error, lo que refuerza el aprendizaje de los modelos anteriores.

A diferencia del método de Random Forest, en el cual cada árbol es independiente, en XGBoost, cada árbol sucesivo aprende de los anteriores, y los árboles no tienen el mismo peso. Para la predicción final, la salida de cada árbol se multiplica por una tasa de aprendizaje y se suma a la predicción inicial, generando así un valor final o una clasificación.

4. Metodología

En este capítulo se presenta la metodología propuesta para cumplir con los objetivos planteados en este trabajo, específicamente para el cálculo del Índice Estandarizado de Precipitación y Evapotranspiración (SPEI), el desarrollo de la línea base utilizando el *Climate Predictability Tool* (CPT) y la construcción de un modelo de machine learning para realizar predicciones del índice. Primero, se describen en detalle los datos utilizados, definiendo los periodos, las variables y las condiciones del estudio, así como las fuentes de los datos. Posteriormente, se realiza una caracterización de la zona de interés y se realizan los cálculos correspondientes para construir la línea base del estudio. Finalmente se describe la construcción del modelo de machine learning y la evaluación de las métricas de desempeño. A continuación, se presenta en la Figura 3 y Figura 4 las fases de desarrollo del trabajo.

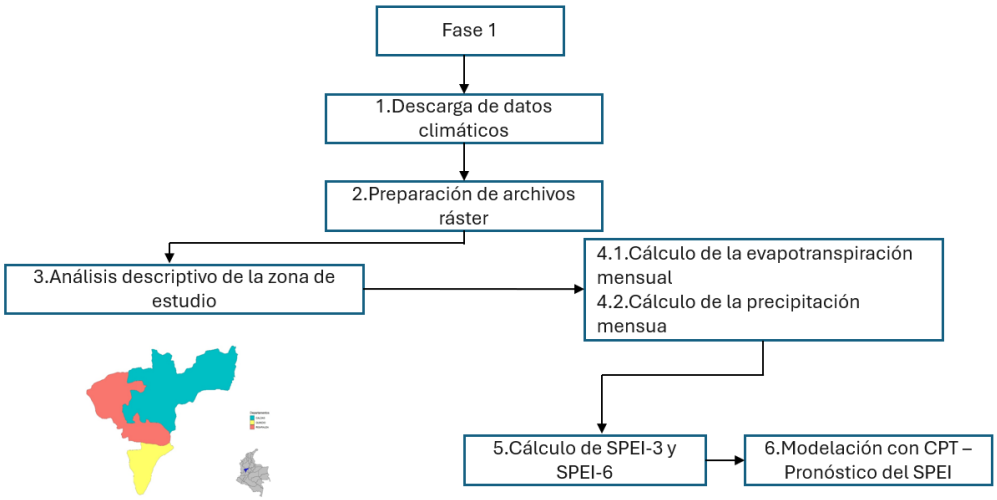


Figura 3. Diseño metodológico propuesto para la Fase 1.

La metodología propuesta en este trabajo consta de 2 fases. La primera fase consiste en la descarga de datos climáticos, recopilando la información necesaria para los análisis

posteriores. Una vez obtenidos los datos, el siguiente paso, implica la preparación de los archivos en formato ráster, que es el formato adecuado para los análisis geoespaciales.

Posteriormente, se realiza un análisis descriptivo de la zona de estudio, permitiendo caracterizar las particularidades de la región. En el cuarto paso se realizan dos cálculos clave para realizar la siguiente fase: la evapotranspiración mensual (4.1) y la precipitación mensual (4.2).

Luego, se procede a calcular los índices SPEI-3 y SPEI-6, que representan el Índice Estandarizado de Precipitación y Evapotranspiración a escalas de 3 y 6 meses, respectivamente.

Finalmente, en el último paso de esta fase, se utiliza el Climate Predictability Tool (CPT) para modelar y generar pronósticos del SPEI mediante CCA, lo que permite obtener una línea base de referencia para futuros estudios.

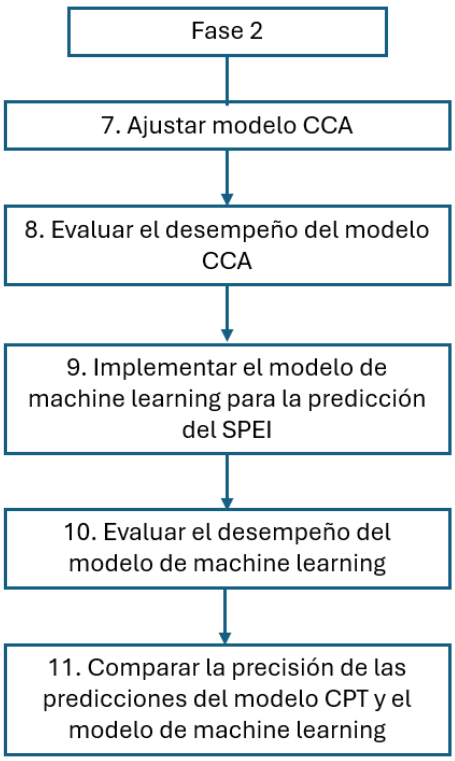


Figura 4. Diseño metodológico propuesto para la Fase 2.

Como se presenta en la Figura 4, la Fase 2 del proyecto se enfoca en el desarrollo y comparación de modelos predictivos.

La primera parte se centra en el refinamiento del modelo de CCA, ajustando los predictores y predictandos para obtener una predicción de los índices SPEI3 y SPEI6 para marzo de 2024, mes históricamente lluvioso en la zona de estudio pero que tuvo condiciones climáticas anómalas en el 2024:

- Se inicia con el ajuste del modelo CCA, utilizando como predictores los datos de temperatura superficial del mar (SST) correspondientes al mes pronosticado para todos los años disponibles (1982-2024).
- Luego, se evalúa el desempeño del modelo CCA ajustado, analizando su precisión y capacidad predictiva.

A continuación, se aborda la etapa relacionada con el modelo de machine learning:

- Se implementan modelos de machine learning para la predicción del índice SPEI, en este caso se utilizaron los algoritmos Random Forest y XGBoost incluidos en la librería CAST (Caret Applications for Spatio-Temporal models) de R, la cual permite el uso de estos algoritmos con datos espacio-temporales mediante técnicas de vectorización, evitando posibles autocorrelaciones en los datos espaciales. Se evalúa el desempeño de los modelos basándose en las métricas predefinidas.

Finalmente, se compara la precisión de las predicciones del modelo CCA ajustado con las de los modelos de machine learning utilizando un "*Goodness Index*" que en este caso corresponde a la correlación de Kendall (Wilks, 2006). Esta comparación permite evaluar las fortalezas y debilidades de cada enfoque, proporcionando puntos clave sobre qué metodología podría ser más adecuada para la predicción del SPEI en la zona de estudio.

Esta fase busca no solo mejorar la precisión de las predicciones del SPEI, sino también contrastar métodos tradicionales (CCA) con técnicas más modernas de machine learning, ofreciendo una perspectiva completa sobre las opciones de modelado disponibles para este tipo de predicciones climáticas.

4.1. Zona de estudio

La zona de estudio comprende los departamentos de Quindío, Risaralda y Caldas, todos tres pertenecientes a la ecorregión del Eje cafetero colombiano (Figura 5) y de interés para el proyecto “Colombia Agroalimentaria Sostenible – CAS-”.

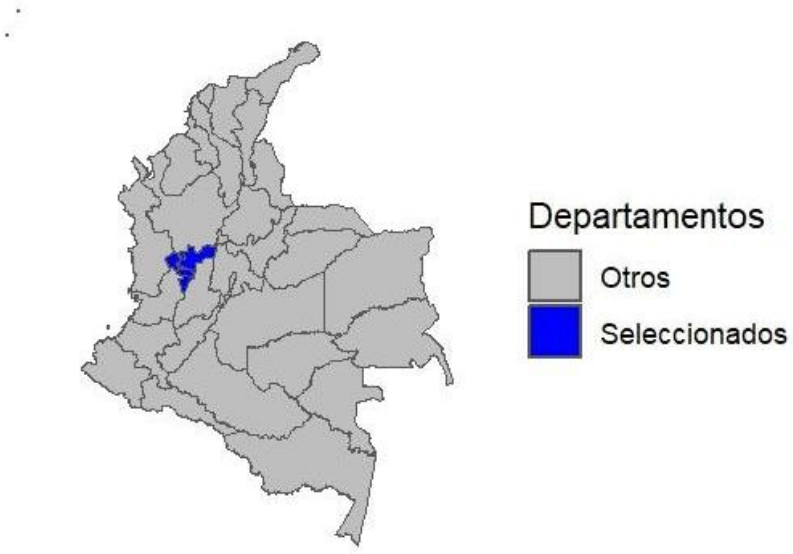


Figura 5 Mapa de Colombia – Departamentos seleccionados.

4.2. Software y hardware utilizado

Para la obtención y procesamiento de los datos necesarios en este estudio, se empleó el lenguaje de programación Python ver. 3.11.6 (VS Code) con las librerías “cdsapi”, “pandas”, “numpy”, “matplotlib” y “seaborn”; y el lenguaje R ver. 4.4.0 (RStudio), utilizando las librerías “CAST”, “raster”, “terra”, “sp”, “dplyr”, ”NetCDF”, entre otras, para el manejo de archivos tipo ráster. Para la generación de predicciones del indicador climático en la primera fase del proyecto se utilizó el software CPT (*Climate Predictability Tool*) ver. 18.3.1.

El hardware utilizado para las tareas de descarga y procesamiento de los datos estuvo compuesto por un servidor dedicado con 2 procesadores Intel Xeon Gold (2.7 GHz) con

256 GB de RAM y 2 computadores personales con procesadores Intel I7 de 12ª generación (2.1 GHz) con 16 GB de RAM.

4.3. Información climática

4.3.1. Fuentes de los datos: la información utilizada es proveniente de plataformas satelitales abiertas reconocidas, de amplio uso en investigación agroclimática, incluyendo CHIRPS para la variable de precipitación y Copernicus para las variables de radiación solar, humedad relativa, temperatura mínima y máxima, así como velocidad del viento. La temperatura superficial del mar SST se obtuvo a través de CPT directamente. El uso de datos satelitales garantiza una cobertura espacial y temporal adecuada, permitiendo un análisis integral y detallado de las condiciones ambientales en la zona de interés.

CHIRPS pertenece al Climate Hazards Center de la Universidad de California y cuenta con el soporte de USAID, NASA y NOAA, y la colaboración del Centro de Observación y Ciencia de los Recursos Terrestres (EROS) del Servicio Geológico de Estados Unidos (USGS). CHIRPS proporciona datos de precipitación a escala mundial con una resolución de 0.05º, que proviene de la combinación de series climáticas calibradas a nivel local con información satelital (Climate Hazards Center, 2024) (ver Tabla 3).

Tabla 3. Variables climáticas descargadas de CHIRPS.

VARIABLE	UNIDAD	ESCALA	RANGO
Precipitación (ver. 2)	mm	Diaria	1981-01-01 a 2024-03-31

Copernicus es un programa de la Unión Europea para la observación de la tierra, el cual cuenta con una red propia de satélites (Sentinel), y del que hacen parte la Agencia Espacial Europea (ESA), la Organización Europea para la Explotación de Satélites Meteorológicos (Eumetsat) y el Centro Europeo de Previsiones Meteorológicas a Plazo Medio (ECMWF), entre otras agencias e instituciones de la Unión Europea. Copernicus proporciona datos globales basados en observaciones procedentes de los satélites y de observaciones in situ terrestres, aéreas y marítimas, con una resolución de 0.1º, los cuales pasan por correcciones topográficas y métodos de interpolación (Copernicus,

2024). Se obtuvieron datos de quinta generación de reanálisis de datos climáticos globales (AgERA5, revisión 1.1) para las variables presentadas en la Tabla 4.

Tabla 4. Variables climáticas descargadas de Copernicus.

VARIABLE	UNIDAD	ESCALA	RANGO
Temperatura máxima	ºK a 2 m	Diaria	1981-01-01 a 2024-03-31
Temperatura mínima	ºK a 2 m	Diaria	1981-01-01 a 2024-03-31
Humedad relativa media	% a 2 m	Diaria	1981-01-01 a 2024-03-31
Radiación solar	$J.m^{-2}.d^{-1}$	Diaria	1981-01-01 a 2024-03-31
Velocidad del viento	$m.s^{-1}$ a 10 m	Diaria	1981-01-01 a 2024-03-31

4.3.2 Descarga de datos de clima: los datos se descargaron en formato ráster con frecuencia diaria, directamente desde los repositorios de cada plataforma mediante scripts de Python, utilizando la API “cdsapi” requerida por Copernicus, y a través de servidores FTP en el caso de CHIRPS. Toda la información se almacenó en uno de los servidores de CIAT (ALLIANCEDFS) desde donde pueden ser accedidos por los miembros de diferentes equipos de investigación del centro. Parte del objetivo de la descarga de la información fue reemplazar la información previamente guardada en el servidor con las nuevas versiones corregidas disponibles en las plataformas, así como completar los datos hasta el mes de marzo de 2024.

4.3.3 Preparación de archivos ráster: los archivos ráster descargados contienen la información diaria de las variables climáticas para el mundo entero, con una resolución espacial de aproximadamente 5 Km en el caso de precipitación (CHIRPS) y de aproximadamente 10 Km para las demás variables (AgERA5), por lo tanto, se hizo necesario hacer un remuestreo de estas últimas variables para estandarizar los datos con una resolución de 5 Km, y además recortar todos los archivos mediante un archivo *shapefile* (polígonos de los departamentos de Colombia) para solo conservar la información del área de estudio. Este procedimiento significó un reto técnico ya que, si se realiza primero el corte del ráster y se hace el remuestreo posteriormente, los bordes de los mapas quedan con poca definición debido al tamaño de los pixeles (10x10 Km) y no coinciden espacialmente con los rásters de precipitación; por otro lado, si se hace el remuestreo primero antes del corte, los bordes quedaran mejor definidos (pixeles 5x5

Km) pero el proceso requiere demasiado tiempo y recursos de máquina porque se estarían procesando rásters con los datos del mundo completo.

La solución adoptada fue aprovechar el archivo *shapefile* seleccionando departamentos adicionales y realizar un corte preliminar de los rásters con un área un poco más grande que el área de estudio, manteniendo los departamentos de Quindío, Risaralda, Caldas, Chocó, Antioquia, Tolima, Valle del Cauca, Cundinamarca y Boyacá. Este procedimiento se realizó en el software R con las librerías “raster”, “terra” y “sp”.

Posteriormente se realizó el remuestreo de todos los ráster recortados utilizando la función “resample” de la librería “ráster” y teniendo como referencia un ráster de precipitación de CHIRPS (5 Km de resolución).

Finalmente, se hizo un segundo recorte a los rásters, incluyendo en esta etapa los archivos de precipitación y el modelo digital de elevación, para conservar solo los departamentos de Quindío, Risaralda y Caldas, asegurando así un análisis detallado y la compatibilidad de todos los archivos.

4.4 Cálculo de evapotranspiración mensual

En este estudio se decidió utilizar el método de Penman-Monteith dado que es el método más exacto y consistente tanto para climas áridos como para los húmedos (Allen et al., 2006), además se cuenta con todos los datos necesarios para su cálculo.

Mediante el software R, con ayuda de las librerías “raster”, “terra” y “sp” se calculó el valor de la evapotranspiración diaria para cada píxel del ráster del área de estudio, aplicando la ecuación de Penman-Monteith a partir de los rásters recortados del modelo digital de elevación y de las variables climáticas provenientes de AgERA5 (a escala diaria), con datos comprendidos entre enero 01 de 1981 hasta marzo 31 de 2024.

Posteriormente, mediante otro script en R, se calculó la evapotranspiración mensual desde 1981 hasta el 2024, agregando la información diaria calculada previamente, y se guardó la información en nuevos rásters.

4.5 Cálculo de la precipitación mensual

Para el cálculo de la precipitación mensual acumulada se tomó la información contenida en los rásters recortados de precipitación diaria, proveniente de CHIRPS, y mediante el software R se realizó la agregación a nivel mensual desde el año 1981 hasta el 2024 para cada píxel del área de estudio, guardando los datos en nuevos rásters.

4.6 Cálculo de SPEI3 y SPEI6

Como paso previo al cálculo del índice SPEI se debe hallar el balance hídrico para cada píxel del área de estudio. Para esto, nuevamente con la ayuda del software R, se cargan los rásters con información a escala mensual de evapotranspiración y precipitación, obtenidos en las etapas anteriores, y se realiza una resta píxel a píxel para obtener un nuevo archivo ráster temporal. Los datos de este nuevo ráster que contiene el balance hídrico mensual para cada píxel se convierten en una tabla que será el insumo principal para el cálculo del SPEI.

El indicador SPEI se calcula mediante la librería “SPEI” de R, enviándole los datos en formato tabla de cada ráster de balance hídrico en escala mensual y especificando el número de meses que debe tener en cuenta para el cálculo del indicador. En este caso se realizan los cálculos teniendo en cuenta los 3 y 6 meses anteriores para la construcción del índice. La librería “SPEI” retorna los índices en escala mensual, calculados para cada píxel en formato tabla, estas tablas se transforman a formato ráster nuevamente y se guardan para uso posterior.

4.7 Modelación con CPT y obtención de pronósticos del SPEI

Uno de los puntos fuertes de CPT es que permite descargar información climática directamente desde el menú de la aplicación, guardando los datos con el formato que requiere la herramienta. De acuerdo con Córdoba-Machado et al. (2015), la temperatura superficial del mar (SST) en la zona tropical del pacifico tiene gran efecto sobre las precipitaciones en Colombia, por lo tanto, y de acuerdo con las recomendaciones de Esquivel et al. (2018), se eligió la variable SST como predictor del modelo CCA en CPT, tomando como punto de partida todo el trópico a nivel global, limitado entre las latitudes 30° y -30°, y con datos generados por el modelo CFSv2 con resolución de 1° (ver Figura

6), para los meses de marzo desde el año 1982 hasta el año 2024, predichos por el modelo a partir de la información disponible en febrero, esto permite hacer las predicciones del SPEI a futuro.

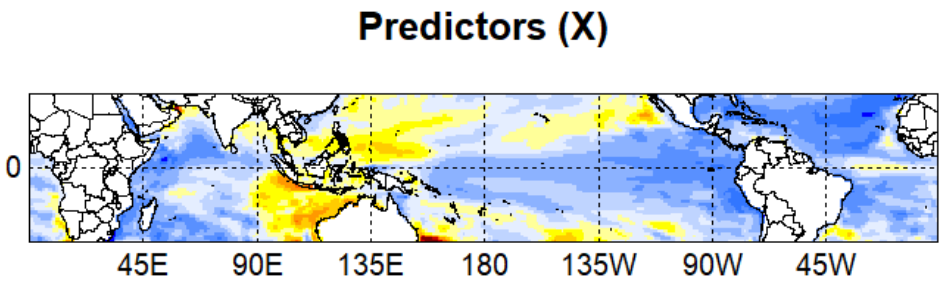


Figura 6. Predictores para el CCA, temperatura superficial del mar SST

Los predictandos para el modelo CCA son los valores de SPEI-3 y SPEI-6 calculados para la zona de estudio a partir de la evapotranspiración y precipitación. Con la ayuda del software R y las librerías “raster”, “lubridate”, “terra”, “sp” y “zoo” se lee cada uno de los rásters con el SPEI-3 y SPEI-6 mensual, se transforman en tablas y se genera un archivo tipo .txt con el formato requerido por CPT para su lectura. En la Figura 7 se presentan ejemplos de los archivos ráster con el SPEI-3 y SPEI-6 para el mes de marzo de 2024.

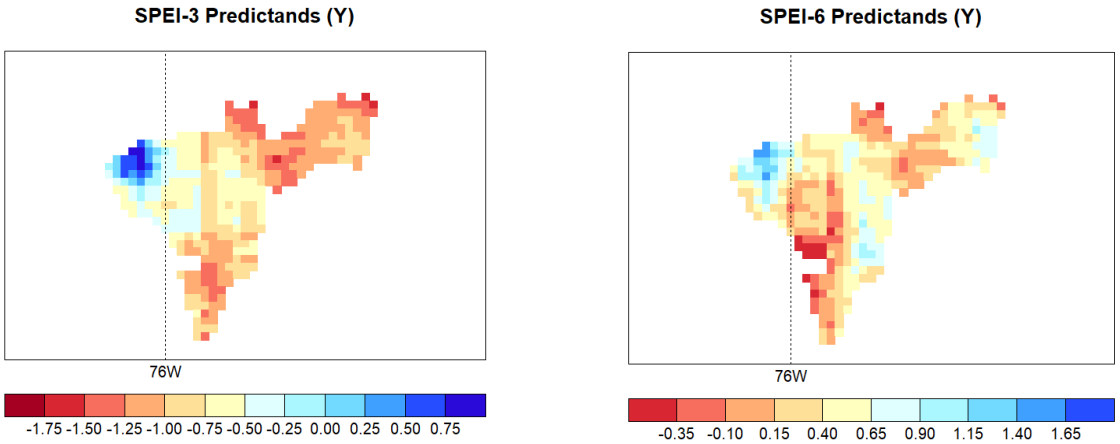


Figura 7 (a) Índice SPEI-3 del mes de marzo de 2024. (b) Índice SPEI-6 del mes de marzo de 2024.

Se implementa el Análisis de Correlación Canónica (CCA) para identificar los modos de variabilidad que maximizan la correlación entre las funciones empíricas ortogonales del predictor y del predictando. De acuerdo con las recomendaciones de Esquivel et al. (2018) para la configuración del análisis CCA, se tomaron como punto de partida entre 1 y 10 modos (componentes) para los predictores y entre 1 y 3 modos para los predictandos, evaluando las posibles combinaciones mediante el protocolo de validación cruzada integrada en la herramienta.

4.8 Evaluación de desempeño de los modelos construidos en CPT

Para cada modelo (SPEI-3 y SPEI-6), se calculó el promedio espacial de la correlación de Kendall entre el SPEI observado y el pronosticado. Este promedio, conocido como Goodness Index (Wilks, 2006), se empleó para evaluar la precisión de los modelos.

La correlación de Kendall, también denominada coeficiente de concordancia de Kendall o tau de Kendall, es una medida no paramétrica que evalúa la fuerza y la dirección de la relación entre dos variables. En este estudio, se utilizó esta medida para analizar la asociación entre los datos observados y los pronosticados y para validar el desempeño de los modelos, se aplicó una validación cruzada.

La fórmula para calcular la correlación de Kendall (τ) entre dos variables X e Y con n pares de datos es la siguiente:

$$\tau = \frac{Número\ de\ pares\ concordantes - número\ de\ pares\ discordantes}{\frac{n * (n-1)}{2}}$$

Valores resultantes:

$$\tau = 1\text{: } Correlación\ perfecta\ positiva$$

$$\tau = -1\text{: } Correlación\ perfecta\ negativa$$

$$\tau = 1 : \ Ausente\ de\ correlación$$

4.9 Implementación de los modelos de machine learning

Para la implementación del modelo de machine learning se construyó un modelo Random Forest y un modelo XGBoost, gestionados por la librería CAST, que como se mencionó anteriormente, permite el uso de este tipo de algoritmos con datos espacio-temporales con múltiples variables de entrada y salida a través de técnicas de vectorización, evitando las autocorrelaciones típicas de los datos espaciales. Se eligieron estos dos modelos debido a su capacidad para procesar grandes cantidades de datos y por su disponibilidad en la librería CAST, como se mencionó anteriormente.

En este caso los predictores utilizados fueron las variables climáticas para la zona de estudio: precipitación, temperaturas máximas y mínimas, humedad relativa, radiación solar y velocidad del viento, pero con un rezago de un mes (LT-1) y dos meses (LT-2) respecto al mes pronosticado. Este cambio de predictores obedece a que los modelos implementados a través de la librería CAST requieren que tanto predictores como predictandos correspondan a la misma área geográfica y los archivos raster contengan el mismo número de píxeles.

Con el objetivo de identificar las variables que aportan más información a los modelos, se realizó un análisis preliminar utilizando todas las variables climáticas disponibles para la construcción del modelo. Posteriormente, a través de la evaluación del atributo de importancia de las variables en los modelos, se seleccionaron únicamente los predictores más relevantes y se llevó a cabo un nuevo entrenamiento del modelo, utilizando únicamente estas variables seleccionadas.

Finalmente, se llevó a cabo una optimización por búsqueda en cuadrícula (grid search) de los hiperparámetros de los modelos con el objetivo de identificar la combinación que ofreciera los mejores resultados. En el modelo Random Forest, el hiperparámetro optimizado fue mtry (número de variables consideradas en cada división de nodo). Para el modelo XGBoost, los hiperparámetros optimizados incluyeron: n_estimators (número de árboles), max_depth (profundidad máxima), learning_rate (tasa de aprendizaje), gamma (reducción mínima de pérdida), colsample_bytree (fracción de columnas

utilizadas por árbol), min_child_weight (peso mínimo de las hojas), y subsample (fracción de muestras utilizadas por iteración).

4.10 Evaluación de desempeño del modelo de machine learning

Al igual que en el caso del modelo CCA, la métrica de evaluación de desempeño de los modelos Random Forest y XGBoost entrenados con datos espacio-temporales es el promedio espacial del coeficiente de correlación de Kendall de todos los píxeles. Teniendo en cuenta el desempeño de esta métrica para los modelos evaluados se puede llegar a una conclusión sobre el método recomendado para la predicción del índice en la zona de estudio.

5. Resultados

En este capítulo se presenta en dos secciones el análisis de los datos climáticos y la generación de pronóstico, alineado con los objetivos planteados. La primera fase aborda la consolidación y homogeneización de las fuentes de información climática disponibles y la caracterización detallada de las zonas de interés, así como la construcción del Índice Estandarizado de Precipitación y Evapotranspiración (SPEI) para la zona de estudio, seguido de los resultados de los pronósticos del SPEI generados mediante la herramienta CPT (*Climate Predictability Tool*) para marzo de 2024. La segunda fase se enfoca en la generación de la predicción del SPEI mediante CCA y un modelo de machine learning utilizando otros predictores climáticos.

5.1 Homogeneización de datos

A continuación, en la Figura 8, se presenta un ejemplo de la visualización de un ráster sin recortar para la zona de interés. En este caso se puede observar la variable temperatura máxima para todo el mundo.

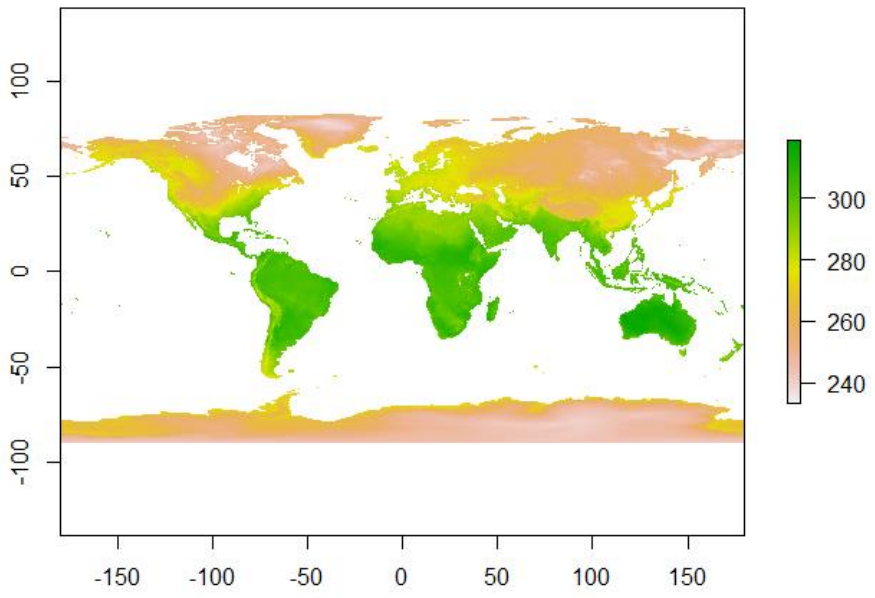
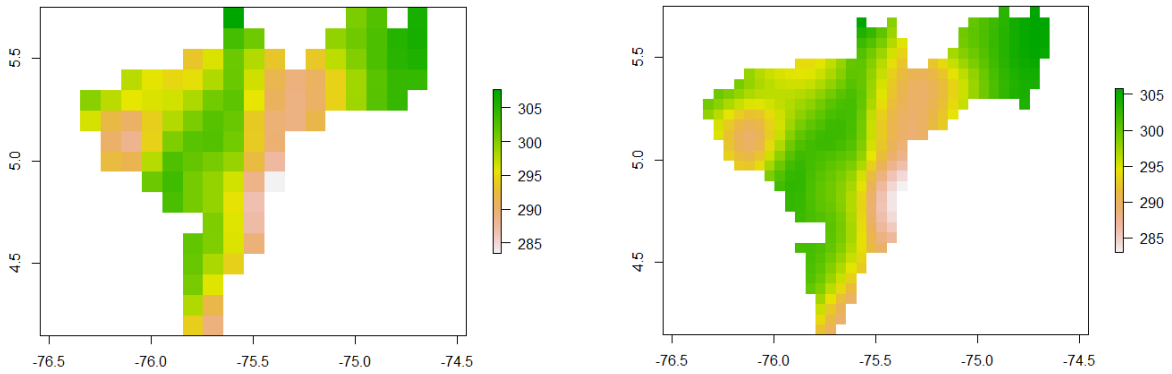


Figura 8. Temperatura máxima global.



(a) (b)

Figura 9 (a) Ráster de temperatura máxima con resolución de 10 Km. (b) Ráster de temperatura máxima con resolución de 5 Km.

Como se puede observar en la Figura 9, el gráfico (a) muestra un ráster correspondiente a los datos de la fuente AgERA5. En este gráfico, los píxeles son relativamente grandes. Después de realizar el proceso de *downscaling* a 5 kilómetros, los píxeles se reducen y se vuelven más finos, como se muestra en el gráfico (b). Este procedimiento se realizó para cada una de las variables presentes en este estudio.

Este proceso de *downscaling* o escalado, es esencial para homogeneizar los datos provenientes de diferentes fuentes y resoluciones. En este caso, ajustar los datos de AgERA5 a la resolución de CHIRPS permite una comparabilidad y análisis más precisos, lo cual es crucial para la evaluación y predicción climática en las zonas de estudio.

5.2 Caracterización de la zona de estudio

El eje cafetero es una región geográfica, cultural, económica y ecológica de Colombia, comprendida por los departamentos de Quindío, Risaralda y Caldas (ver Figura 10).

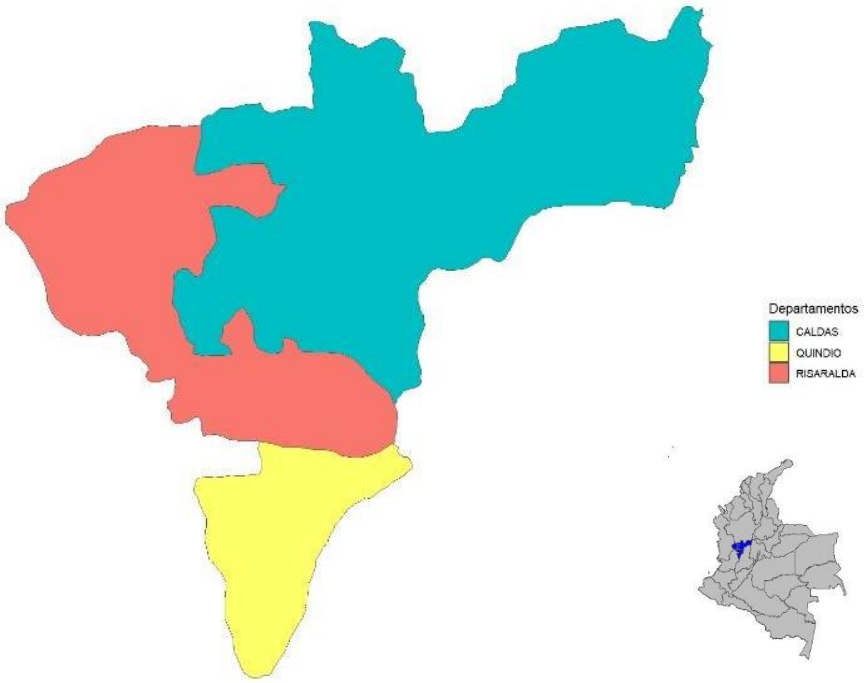


Figura 10. Mapa de los departamentos de Quindío, Risaralda y Caldas.

Caldas

El Departamento de Caldas cuenta con una superficie de 7.888 km^2 y está ubicado en el centro occidente de Colombia, en la región Andina. Hace parte de la Ecorregión del Eje Cafetero junto con Risaralda y Quindío, y es el más extenso de los departamentos que la conforman. Tiene un área de 63.051 ha en producción de café distribuidas en 25 municipios, donde es el sustento principal de alrededor de 32.000 familias (FNC, 2023).

Risaralda

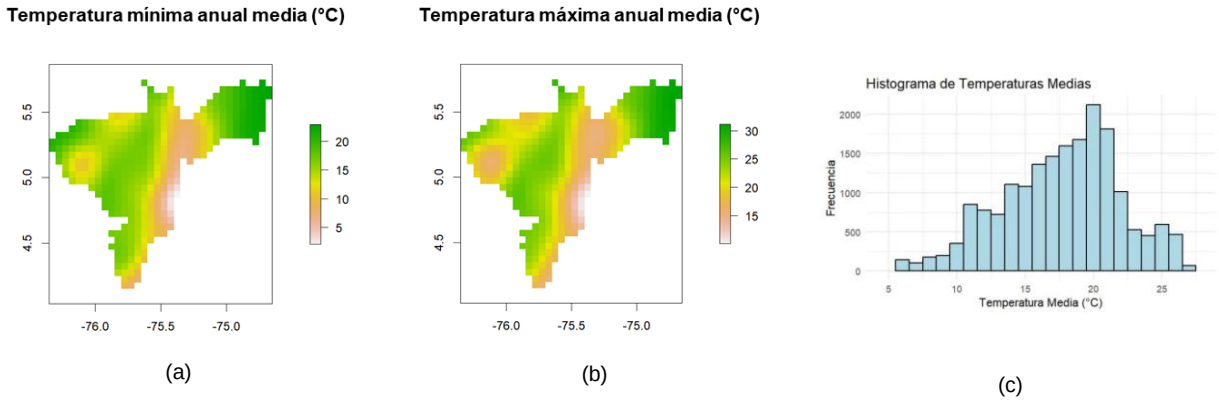
Risaralda tiene una superficie de 4.140 km^2 y es el cuarto departamento más densamente poblado del país. Se tiene registro de 45.127 ha dedicadas a la producción de café distribuidas en 14 municipios, de las que dependen económicamente alrededor de 19.000 familias (FNC, 2023).

Quindío

El departamento del Quindío es el segundo menos extenso del país, con una extensión de 1.845 km^2. Sus principales productos agrícolas son el café, el plátano y el banano. El

cultivo de café ocupa 18.559 ha distribuidas en sus 12 municipios siendo la actividad económica principal de más de 5.000 familias (FNC, 2023).

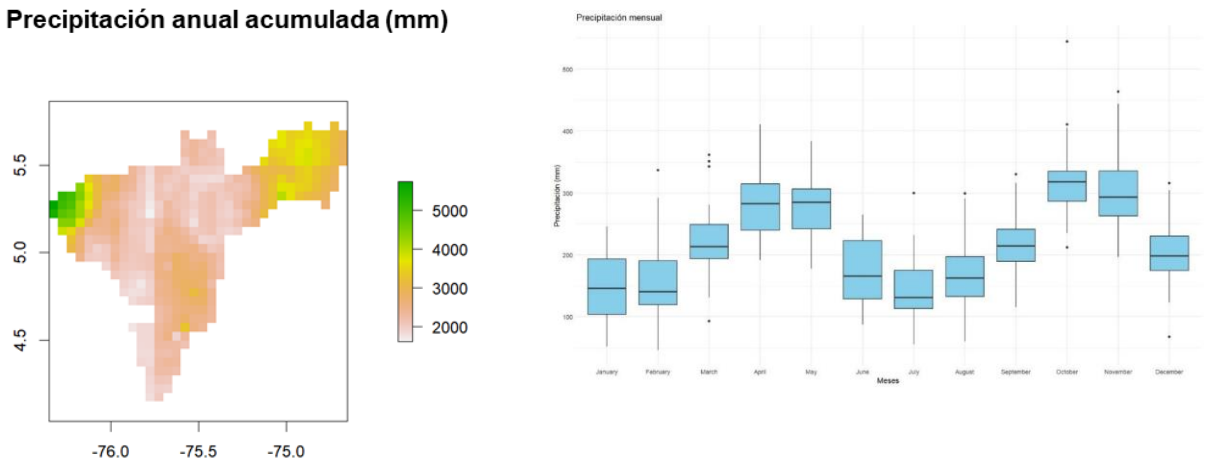
Análisis descriptivo de la Zona de interés



(a) (b) (c)

Figura 11. (a) Temperatura mínima anual media de 2023 de la zona de estudio. (b) Temperatura máxima anual media de 2023 de la zona de estudio. (c) Histograma de Temperatura media de la zona de estudio (1981-2024).

La Figura 11a presenta la temperatura mínima anual media para el año 2023 en la zona de estudio, se observa que la mayoría de los pixeles se encuentran en el rango de 5 a 20 -°C, en contraste con la Figura 11b que presenta la temperatura máxima anual media con valores que se mueven alrededor de 15 y 30 °C. El histograma de la Figura 11c muestra que la mayoría del área del Eje Cafetero tiene una temperatura media anual que se encuentra entre 18 y 22 °C. Sin embargo, también hay algunos puntos con valores más bajos o altos. Por ejemplo, hay zonas con una temperatura media anual de menos de 10 °C, y otros puntos de 5 donde la temperatura media anual es de más de 25 °C.



(a) (b)

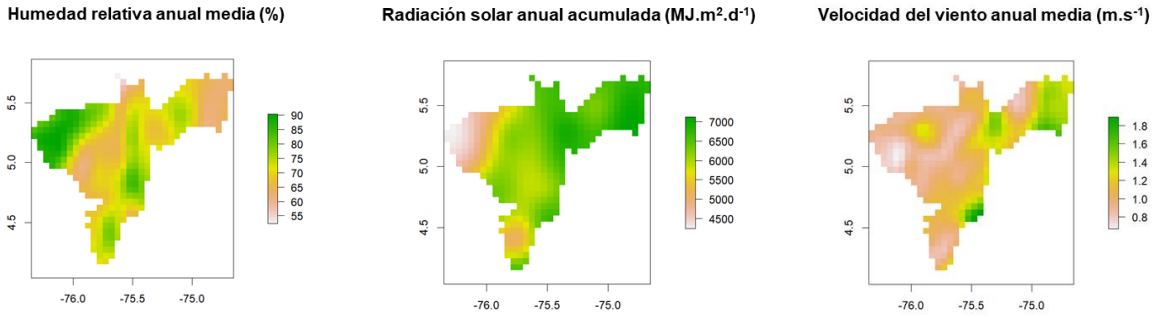
Figura 12 (a) Precipitación anual acumulada para 2023 en la zona de estudio. (b) Distribución de las lluvias a nivel mensual en la zona de estudio (promedio de 44 años).

En la Figura *12*a se presenta la precipitación acumulada para la zona de estudio correspondiente al año 2023, donde se observan las lluvias más abundantes donde el departamento de Risaralda limita con el Chocó. En la Figura 12b se presenta un gráfico de cajas que muestra la precipitación mensual promedio en el eje cafetero colombiano a lo largo de 44 años, revelando un régimen bimodal característico de la región andina, con dos períodos principales de lluvias. El primer pico de lluvias ocurre en abril y mayo, mientras que el segundo y más intenso pico se presenta en octubre y noviembre.

Las temporadas más secas se observan en enero y febrero, con una precipitación mediana ligeramente superior a los 150 mm, y en julio y agosto, con una mediana alrededor de los 150 mm. Octubre y noviembre son los meses con mayor precipitación, presentando medianas alrededor de los 300 mm, seguidos por abril y mayo con medianas cercanas a los 250-300 mm.

La mayoría de los meses muestran una considerable variabilidad en la precipitación, evidenciada por el tamaño de las cajas y la extensión de los bigotes en el gráfico, siendo octubre y noviembre no solo los más lluviosos, sino también los que presentan mayor variabilidad. Además, se observan varios valores atípicos, especialmente en los meses más lluviosos, indicando eventos de precipitación excepcionalmente alta.

Esta distribución de lluvias es crucial para el ciclo del café y las prácticas agrícolas en la región del Eje Cafetero. Influye en las épocas de floración, cosecha y beneficio del café, siendo determinante para la planificación y el éxito de las actividades agrícolas en esta zona (FNC, 2023).



(a) (b) (c)

Figura 13. (a) Humedad relativa anual media para 2023 en la zona de estudio. (b) Radiación solar anual acumulada para 2023 en la zona de estudio. (c) Velocidad del viento anual media para 2023 en la zona de estudio.

En la Figura 13a se presenta la humedad relativa anual media, Figura *13*b radiación solar anual acumulada y Figura *13*c velocidad del viento anual media para el año 2023 en la zona de estudio. Se observan especialmente patrones en la humedad relativa y la radiación acumulada relacionados con las zonas atravesadas por las cordilleras central y occidental, así como la cercanía con el departamento del Chocó.

Estas características sugieren un clima típico del eje cafetero colombiano: temperaturas moderadas y niveles de humedad relativamente altos, condiciones favorables para el cultivo del café.

5.3 Indicador SPEI

Mediante la librería “SPEI” del software R se construyeron los rásters con los índices SPEI para la zona del eje cafetero a nivel mensual, iniciando desde marzo de 1981 para el caso del SPEI-3, y a partir de junio para el SPEI-6. La pérdida de los 2 y 5 primeros meses de la serie, respectivamente, se debe a la naturaleza del indicador que requiere acumular estos meses para la construcción del primer valor de la serie. Con el objetivo de comprobar la consistencia de los indicadores se verificaron los resultados obtenidos

para meses en los que se presentó Fenómeno de la Niña y Fenómeno del Niño en Colombia.

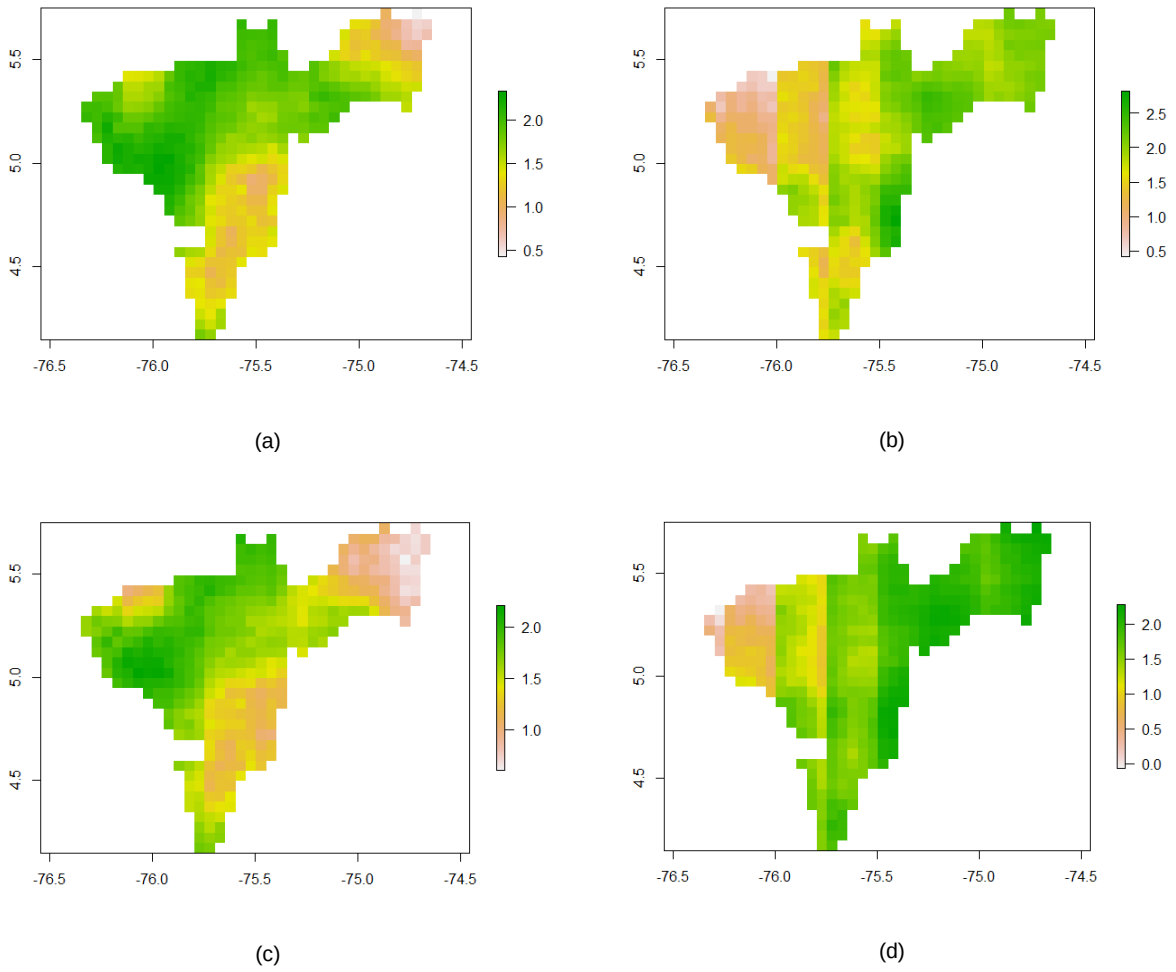


Figura 14 Índices SPEI para la zona de estudio durante temporada de ocurrencia del Fenómeno de La Niña: (a) SPEI-3 en enero de 2011; (b) SPEI-3 en noviembre de 2022; (c) SPEI-6 en enero de 2011 y (d) SPEI-6 en noviembre de 2022.

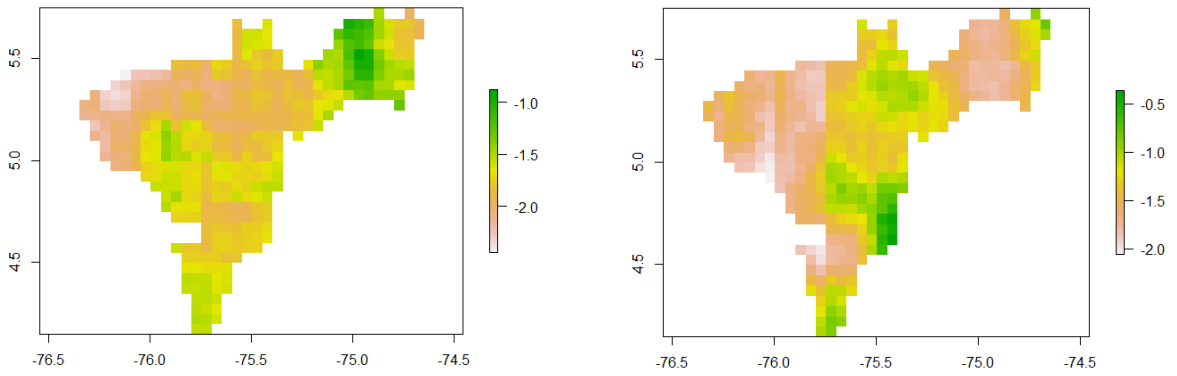
En la Figura 14 se presentan mapas que muestran los índices SPEI para la zona del eje cafetero colombiano durante períodos de ocurrencia del fenómeno de La Niña. El SPEI es un indicador de sequía que considera tanto la precipitación como la evapotranspiración. Valores positivos (verdes) indican condiciones más húmedas, mientras que valores negativos (amarillos a rojos) indican condiciones más secas.

En enero de 2011, el mapa del SPEI-3 (Figura 14a) muestra condiciones predominantemente húmedas (verdes) en la mayor parte de la región, con algunas áreas

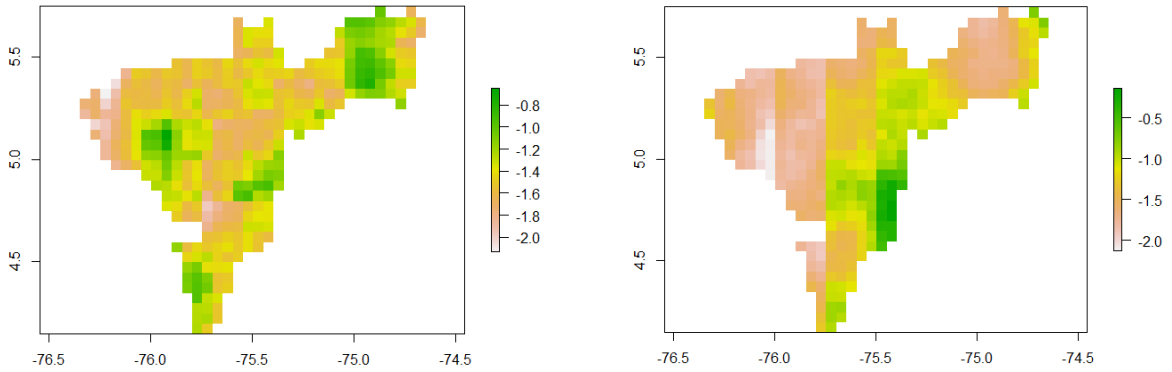
en el sur y este presentando condiciones normales a ligeramente secas (amarillo). Por su parte, el SPEI-6 (Figura 14b) de ese mismo período presenta un patrón similar, pero con condiciones húmedas más extendidas y una menor área con condiciones normales o ligeramente secas.

Para noviembre de 2022, el mapa del SPEI-3 (Figura 14b) muestra condiciones muy húmedas (verde oscuro) en el centro y norte de la región, mientras que el sur muestra condiciones normales a ligeramente húmedas (verde claro a amarillo). El SPEI-6 (Figura 14d) de noviembre de 2022 revela condiciones extremadamente húmedas (verde oscuro) en gran parte de la región, con algunas áreas en el oeste mostrando condiciones normales a ligeramente húmedas.

En una comparación general, los mapas de 2022 (14b y 14d) muestran condiciones más húmedas que los de 2011 (14a y 14c). Los índices SPEI-6 (14c y 14d) presentan condiciones más húmedas y homogéneas que los SPEI-3 (14a y 14b) para los mismos períodos. En todos los casos, se observa el efecto de La Niña con condiciones en general húmedas, especialmente en 2022. Hay variabilidad espacial en la intensidad de las condiciones húmedas, probablemente debido a la topografía y otros factores locales. Estos patrones son consistentes con el fenómeno de La Niña, que típicamente trae condiciones más húmedas a esta región de Colombia.



(a) (b)



(c) (d)

Figura 15. Índices SPEI para la zona de estudio durante temporada de ocurrencia del Fenómeno de El Niño: (a) SPEI-3 en febrero de 1998; (b) SPEI-3 en enero de 2024; (c) SPEI-6 en febrero de 1998 y (d) SPEI-6 en enero de 2024.

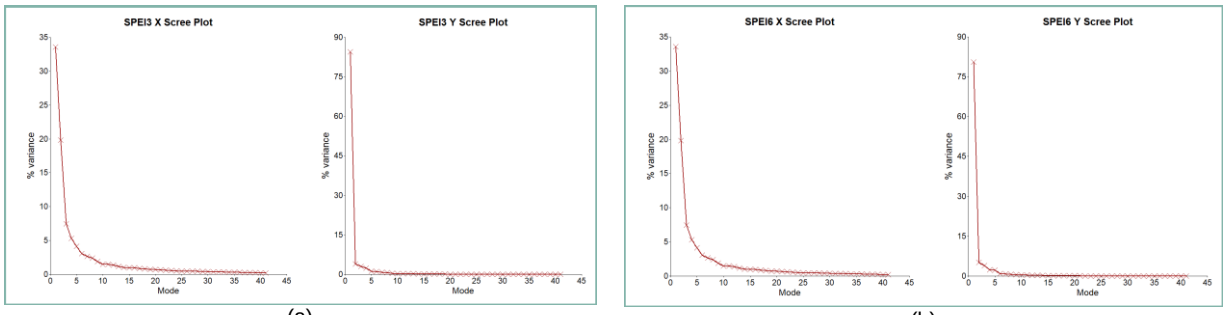
La Figura 15, muestra los índices SPEI para la zona de estudio, durante períodos de ocurrencia del fenómeno de El Niño. En la figura (15a) se presenta el mes de febrero de 1998, el SPEI-3 muestra condiciones predominantemente secas (amarillo a naranja) en la mayor parte de la región. Sin embargo, algunas áreas pequeñas en el noreste muestran condiciones normales a ligeramente húmedas (verde claro). En la Figura (15c), se tiene a febrero del mismo periodo, pero para el SPEI-6, este revela condiciones secas más intensas y extendidas que en el SPEI-3. La mayoría de la región presenta tonos naranjas, indicando sequía moderada a severa, aunque pequeñas áreas en el noreste mantienen condiciones normales.

Por otra parte, en la figura (15b) se presenta el mes de enero de 2024, el SPEI-3 presenta un patrón variado con condiciones secas (naranja) en el oeste y noroeste, mientras que el centro-este de la región muestra condiciones húmedas (verde). El sur tiene condiciones normales a ligeramente secas (amarillo). En la Figura (15d) se presenta enero del mismo periodo per para el SPEI-6, esta muestra condiciones secas (naranja) más pronunciadas en el oeste y noroeste, mientras que el centro-esta muestra condiciones normales a ligeramente húmedas (amarillo a verde claro). En general, presenta un patrón similar al SPEI-3, pero con una sequía más intensa en el oeste.

En una comparación general, el evento de El Niño en 1998 (15a y 15c) parece haber sido más intenso y generalizado en términos de sequía que el de 2024 (15b y 15d). Los mapas de 2024 muestran una mayor variabilidad espacial, con algunas áreas experimentando condiciones húmedas a pesar del fenómeno de El Niño. Los índices SPEI-6 (15c y 15d) generalmente muestran condiciones más extremas que los SPEI-3 (15a y 15b) para los mismos períodos. En ambos eventos, se observa el efecto de El Niño con condiciones en general más secas, especialmente en 1998. Estas diferencias podrían indicar variaciones en la intensidad de El Niño entre los dos períodos, o cambios en los patrones climáticos regionales a lo largo del tiempo.

5.4 Modelo CCA mediante CPT

La métrica utilizada para evaluar el desempeño de los modelos mediante la validación cruzada realizada por CPT es el promedio espacial de la correlación Tau de Kendall. La herramienta presenta los resultados a través de diferentes gráficas y tablas para su interpretación.



(a) (b)

Figura 16. Porcentaje de varianza explicada por el número de modos del CCA para (a) SPEI-3 y (b) SPEI-6.

La Figura 16 muestra las gráficas del porcentaje de varianza explicada por el número de modos del Análisis de Correlación Canónica (CCA) para los índices SPEI-3 (16a) y SPEI-6 (16b) en la zona de estudio. Para el SPEI-3 X Scree plot, muestra una caída pronunciada inicial, con el primer modo explicando alrededor del 35% de la varianza. La curva decrece rápidamente hasta aproximadamente el modo 5, después de lo cual se estabiliza. Para el SPEI-3 Y Scree plot, se puede observar una caída aún más

pronunciada, con el primer modo explicando cerca del 90% de la varianza. Después del primer modo, la curva cae drásticamente y se estabiliza rápidamente.

Por otra parte, en la Figura 16b se presenta el SPEI-6 por lo que para el X Scree Plot: se puede observar una similitud con respecto al SPEI-3 X, pero con el primer modo explicando un porcentaje ligeramente menor, alrededor del 30% de la varianza. La curva también decrece rápidamente hasta aproximadamente el modo 5. En cuanto al SPEI-6 Y Scree Plot se puede ver la caída más pronunciada de todas, con el primer modo explicando casi el 80% de la varianza. Después del primer modo, la curva cae abruptamente y se estabiliza.

Para ambos índices, después de los primeros 5-10 modos, la contribución adicional a la varianza explicada es mínima, lo que sugiere que un número relativamente pequeño de modos en el CCA puede capturar eficientemente la mayor parte de la variabilidad en los datos SPEI para la región estudiada.

El índice de bondad de ajuste en CPT utiliza la correlación tau de Kendall como métrica predeterminada. Esta elección se debe a que la tau de Kendall es adecuada para maximizar el poder discriminatorio de las previsiones y es insensible a la distribución de los datos. Además, Kendall's tau-c, una versión corregida de la tau de Kendall, es utilizada cuando existen empates en los datos, lo que puede ocurrir, por ejemplo, si los datos son recuentos o contienen múltiples ceros. Esta métrica es especialmente útil en situaciones donde los datos no siguen una distribución normal.

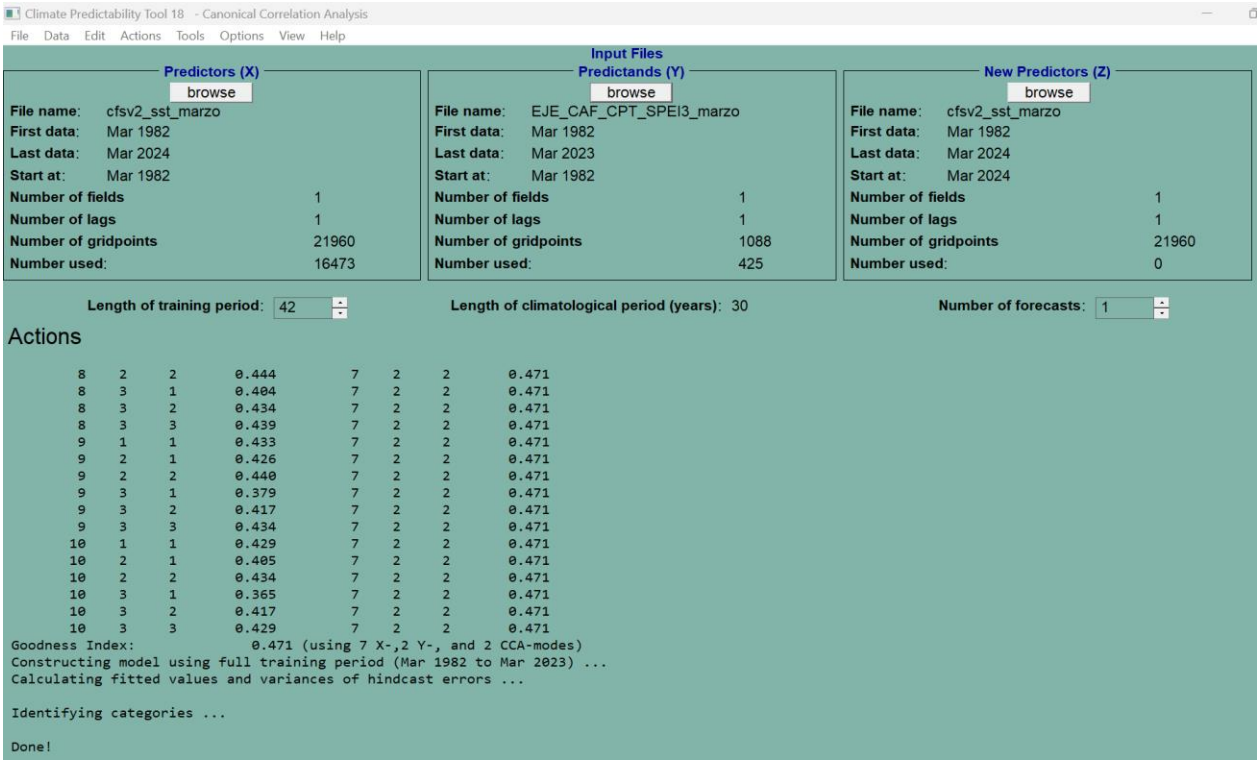


Figura 17. Resultados de la construcción de los modelos de predicción en CPT para SPEI-3

En la Figura 17 se presentan los resultados de los modelos SPEI-3 y SPEI-6. El índice de Kendall sugiere que el modelo para SPEI-3 tiene una relación moderadamente positiva con los datos observados, con un valor de 0.47, lo que indica una asociación relativamente fuerte y positiva entre las predicciones y los valores reales.

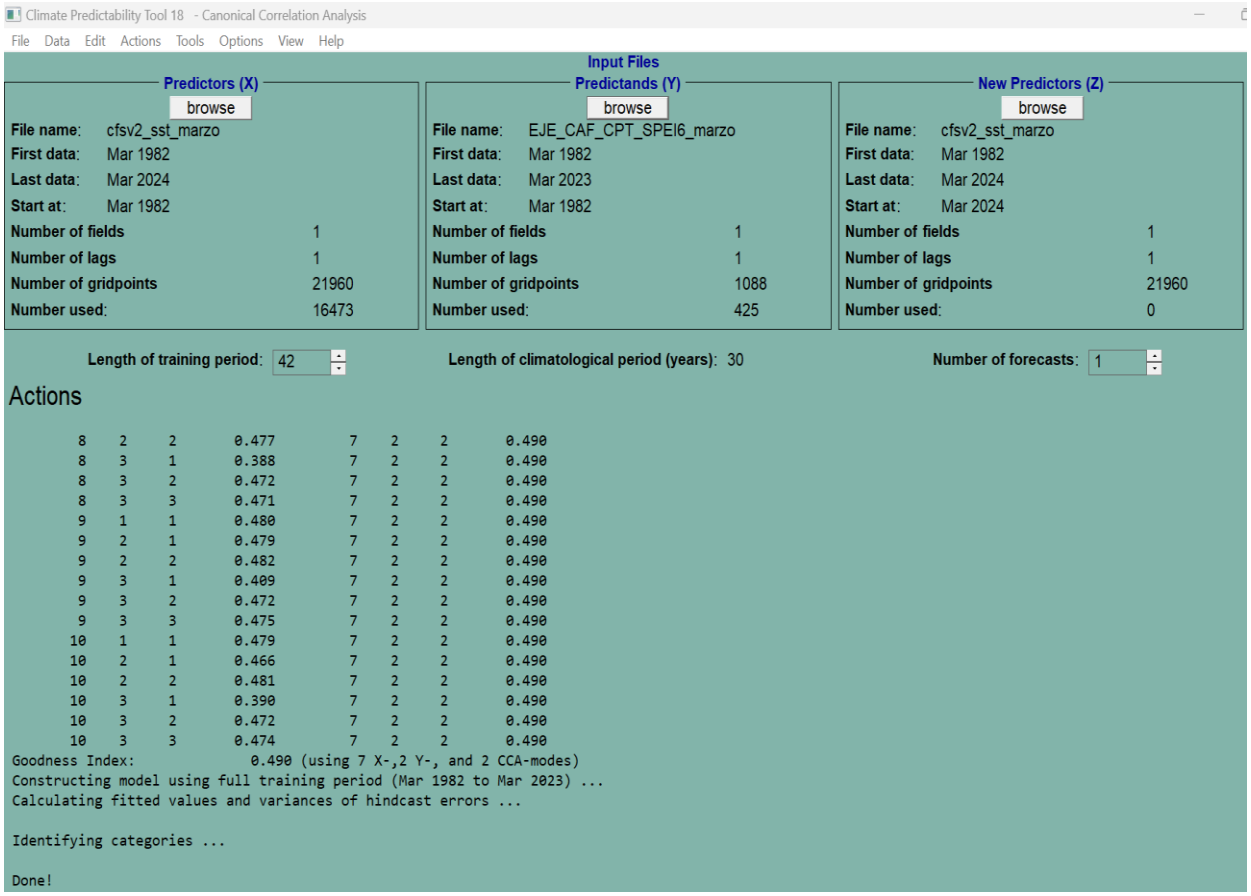


Figura 18. Resultados de la construcción de los modelos de predicción en CPT para SPEI-6

Por otra parte, en la Figura 18 se presenta el modelo SPEI-6 , el cual tiene un índice de 0.49, lo que refleja una relación ligeramente más fuerte, sugiriendo un ajuste superior con respecto al SPEI-3. Esto muestra que el modelo SPEI-6 tiene un mejor rendimiento predictivo.

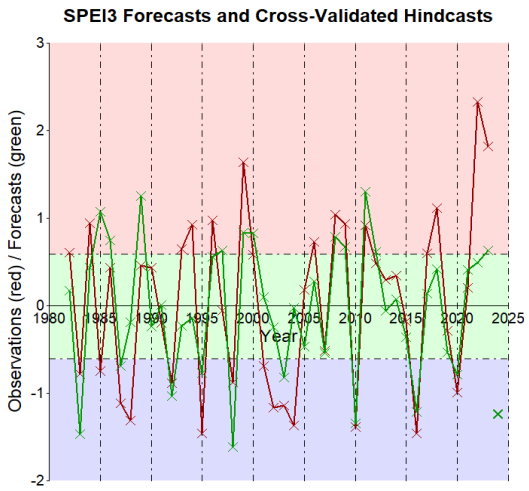


Figura 19. Resultados del proceso de validación cruzada para la tarea de regresión del modelo de predicción de CPT para SPEI-3. El análisis se refiere a un punto de cuadrícula específico en latitud 5.1°N y longitud 75.7°W

En la Figura 19 se presentan los resultados gráficos del proceso de validación cruzada para un píxel en el modelo de predicción CPT del índice SPEI-3. Se observa una similitud en la tendencia de los valores observados y los calculados para todos los años. Esto indica una buena correlación y habilidad predictiva del modelo, sugiriendo que puede predecir eficazmente las variaciones en los datos observados.

En general los resultados sugieren que el modelo CPT para SPEI-3 tiene un buen rendimiento en tareas de regresión (predicción de valores continuos). El modelo demuestra una sólida habilidad predictiva y una buena capacidad para discriminar entre diferentes categorías de eventos climáticos.

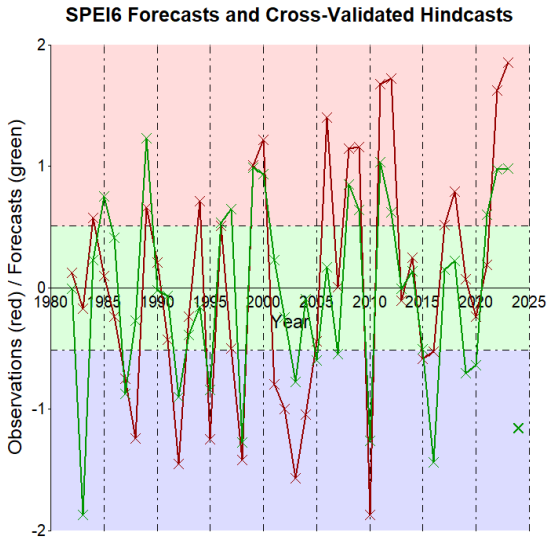
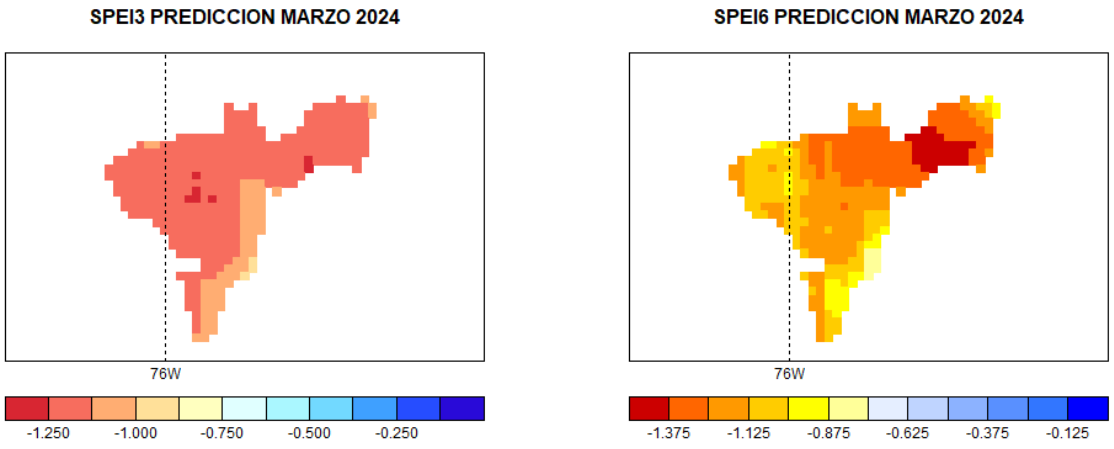


Figura 20. Resultados del proceso de validación cruzada para las tareas de regresión del modelo de predicción de CPT para SPEI-6. El análisis se refiere a un punto de cuadrícula específico en latitud 5.1°N y longitud 75.7°W.

Por otra parte, en la Figura 20 se presentan los resultados gráficos del proceso de validación cruzada para un píxel en el modelo de predicción CPT del índice SPEI-6. En este caso se observa, al igual que para el SPEI-3, una semejanza en las líneas de datos observados y predichos, lo que indica un buen desempeño en la tarea de predicción.



(a)

Figura 21. Pronóstico para marzo de 2024 para la zona de estudio del índice (a) SPEI-3 y (b) SPEI-6.

En la Figura 21 se presentan los pronósticos del SPEI a partir de 3 y 6 meses para marzo 2024 en la zona de estudio. En el mapa 21a que presenta el pronóstico del SPEI-3, se puede observar que la mayor parte del área de estudio está representada en tonos naranjas, lo que indica condiciones secas con valores negativos intermedios. Sin embargo, hay una pequeña región en la zona sureste que presenta tonos más cercanos al amarillo, sugiriendo la presencia de condiciones un poco más neutrales en esa zona específica. Esta distribución sugiere que, para ese mes, la mayor parte de la región pudo experimentar un déficit de humedad, con algunas áreas localizadas enfrentando sequía.

Por otra parte, el mapa 21b, presenta el SPEI a 6 meses, en este se puede destacar que hay mayor variabilidad comparado con el SPEI-3. En este mapa, el centro y el noreste de la región están dominados por tonos naranjas y rojos, lo que indica condiciones secas. En contraste, el oeste y sur de la región muestra tonos más amarillos, sugiriendo la presencia de una sequía ligera, como ocurrió en la realidad.

En las gráficas se puede notar que el SPEI-3 sugiere una distribución muy homogénea con una tendencia hacia condiciones de sequía en la mayor parte del área de estudio, mientras que el SPEI-6 presenta contrastes más marcados entre zonas, sin embargo, se observa que las predicciones son muy generalizadas para la región.

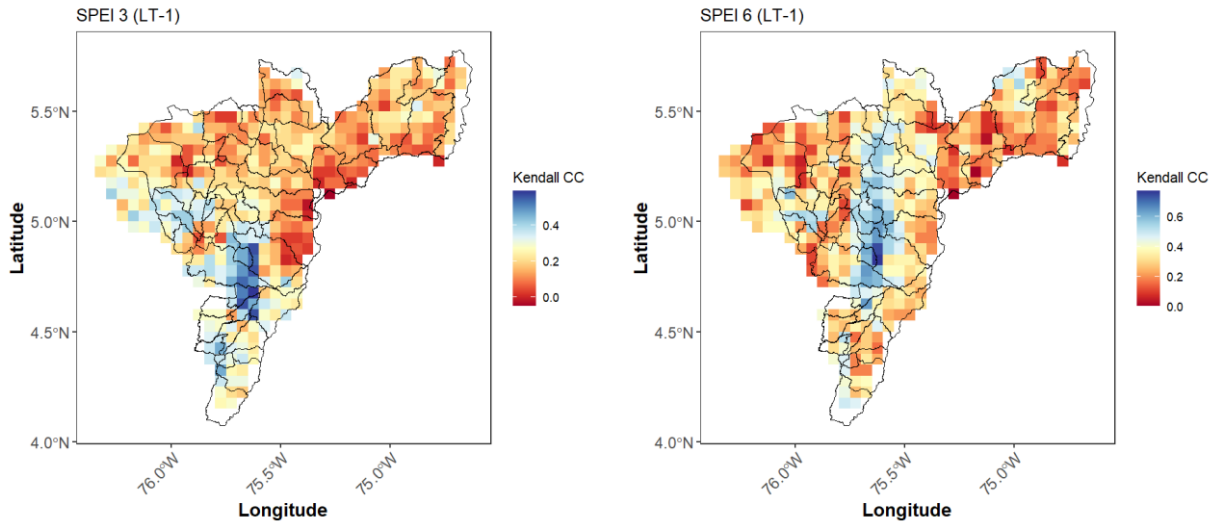
5.5 Modelo Random Forest (CAST)

Con el objetivo de seleccionar las variables con mayor aporte a los modelos de machine learning se construyó un primer modelo tipo Random Forest con todas las variables climáticas y mediante el atributo Importancia de variables del modelo se seleccionaron Precipitación, Radiación solar y Temperatura mínima para continuar la modelación, ya que según los resultados las demás variables no realizan ningún aporte al modelo. Con las tres variables seleccionadas se construyeron nuevamente los modelos a evaluar.

Los mapas de la Figura 22 muestran la correlación espacial obtenida mediante el coeficiente de Kendall (τ\tauτ) para evaluar el desempeño del modelo Random Forest en la predicción del Índice Estandarizado de Precipitación-Evapotranspiración (SPEI) en la región de interés. Este análisis permite identificar las áreas donde las predicciones del

modelo tienen mayor concordancia con los valores observados y cómo esta relación varía dependiendo del rezago temporal y la escala del índice.

Se analizaron las dos escalas temporales el **SPEI-3**, y **SPEI-6**. En ambos casos, se utilizaron como variables predictoras los valores rezagados de precipitación, radiación solar y temperatura mínima, considerando rezagos de uno (LT-1) y dos meses (LT-2). Los mapas están codificados mediante una escala de color, donde tonos cálidos (naranja y rojo) indican correlaciones más bajas o negativas y tonos fríos (azul) representan correlaciones positivas más fuertes.



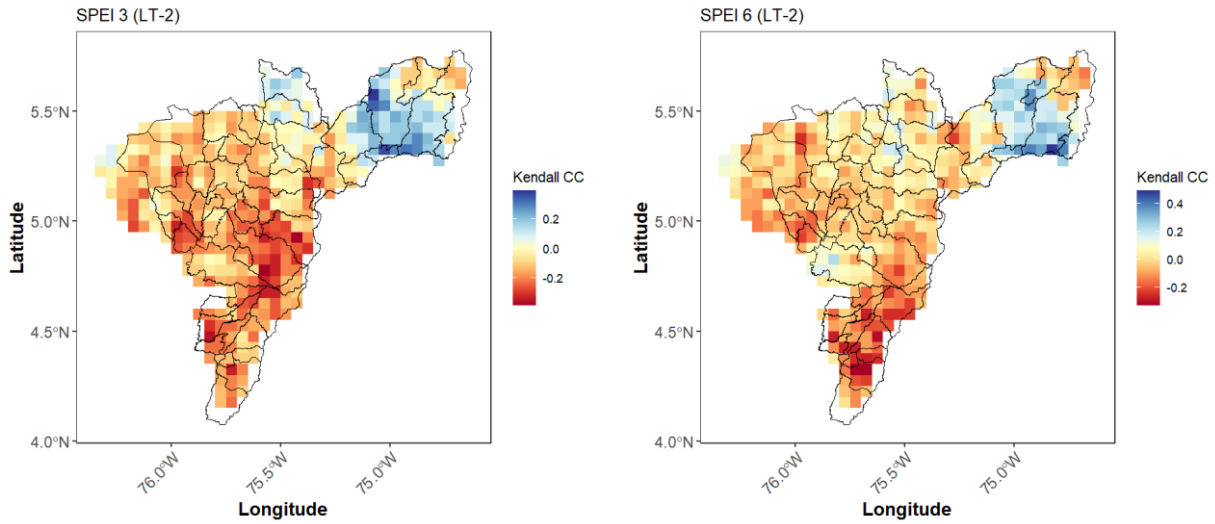
(a) (b)

Figura 22. (a) Mapa de correlaciones de Kendall para SPEI-3 pronosticado con un rezago de 1 mes. (b) Mapa de correlaciones de Kendall para SPEI-6 pronosticado con un rezago de 1 mes.

La Figura 22a presenta el SPEI-3 con un mes de rezago (valores de febrero) con la siguiente optimización de hiperparámetros: mtry=2. El coeficiente de Kendall fue de 0.22, que indica una correlación positiva moderada entre los valores observados y predichos. En el mapa, las zonas con menor correlación se encuentran representadas en tonos cálidos (naranja a rojo), mientras que las áreas con mayor correlación aparecen en tonos fríos (azul).

Por otra parte, en la Figura 22b se presenta el SPEI-6 con un mes de rezago (valores de febrero) con la siguiente optimización de hiperparámetros: mtry=2**.** Se obtuvo un coeficiente de Kendall de 0.31, indicando una correlación positiva más fuerte en

comparación con el SPEI-3. Este resultado destaca que el modelo logra capturar mejor el balance hídrico a mediano plazo utilizando un rezago de un mes.



(a) (b)

Figura 23. (a) Mapa de correlaciones de Kendall para SPEI-3 pronosticado con un rezago de 2 meses. (b) Mapa de correlaciones de Kendall para SPEI-6 pronosticado con un rezago de 2 meses

En la Figura 23a se presenta el SPEI3 con dos meses de rezago (valores de enero) con la siguiente optimización de hiperparámetros: mtry=2**.** El coeficiente de Kendall disminuyó a -0.09, lo que implica una correlación débil y negativa. Esto sugiere que, al aumentar el rezago temporal, la capacidad predictiva del modelo disminuye para esta escala.

En el caso del SPEI-6 presentado en la Figura 23b con dos meses de rezago (valores de enero) con la siguiente optimización de hiperparámetros: mtry=2, el coeficiente de Kendall fue de -0.01, lo que evidencia prácticamente nula correlación entre los valores observados y predichos. Esto resalta la pérdida de información predictiva al trabajar con un rezago mayor.

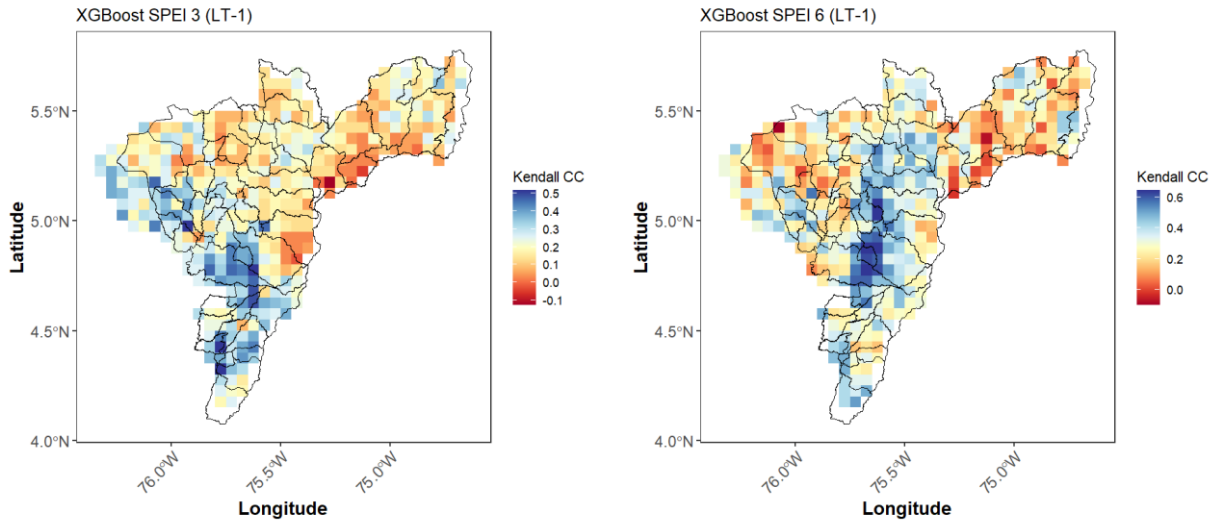
En general en ambas escalas temporales tanto en el SPEI-3 y el SPEI-6 , se observa que las áreas con mejor desempeño predictivo tienden a concentrarse en ciertas regiones. Esto puede estar relacionado con características climáticas locales y los resultados muestran que los rezagos más cortos (1 mes) proporcionan mejores

correlaciones, mientras que los rezagos mayores (2 meses) resultan en una pérdida significativa de la capacidad predictiva, especialmente para el SPEI-3.

5.6 Modelo XGBoost

Los mapas presentados en la Figura 24 muestran la correlación espacial obtenida mediante el coeficiente de Kendall (τ) para evaluar el desempeño del modelo XGBoost en la predicción del SPEI en la región de interés. Este modelo se corrió bajo las mismas condiciones que el Random Forest, utilizando como variables predictoras los valores rezagados de precipitación, radiación solar y temperatura mínima, y considerando rezagos de uno y dos meses.

Al igual que en el Random Forest, los resultados del modelo XGBoost se presentan para las escalas temporales SPEI-3 y SPEI-6. Este análisis permite identificar las áreas donde el modelo logra mayor concordancia con los valores observados, destacando patrones de desempeño que podrían relacionarse con características climáticas locales y variaciones temporales del índice.



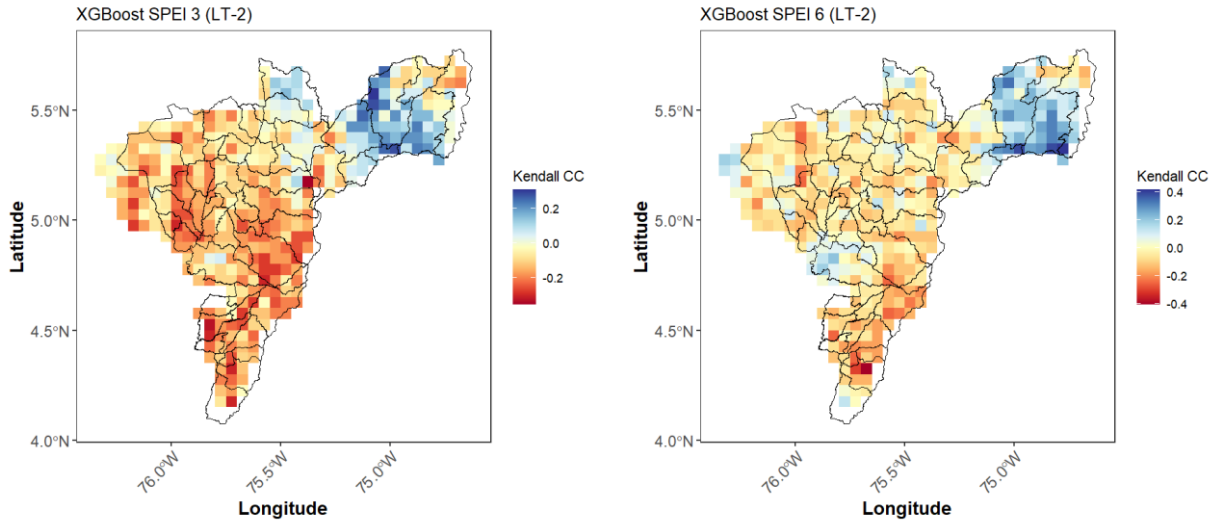
(a) (b)

Figura 24. (a) Mapa de correlaciones de Kendall para SPEI-3 pronosticado con un rezago de 1 mes. (b) Mapa de correlaciones de Kendall para SPEI-6 pronosticado con un rezago de 1 mes.

La Figura 24a muestra el mapa de correlaciones de Kendall obtenidas para el SPEI-3 con un mes de rezago (valores de febrero) utilizando el modelo XGBoost con la siguiente optimización de hiperparámetros: n_estimators=100, max_depth=9, learning_rate=0.1, gamma=0, colsample_bytree=0.8, min_child_weight=1 y subsample=0.8. El coeficiente de Kendall obtenido fue de 0.21, indicando una correlación positiva moderada entre los valores observados y predichos. En el mapa predomina una correlación positiva de moderada a alta (representada en tonos azules claros y oscuros), especialmente en la región central y suroeste del área analizada. Sin embargo, también se observan áreas con correlaciones más bajas (amarillo) e incluso negativas (rojo), principalmente en la región norte y noroeste, lo que sugiere un desempeño variable del modelo dependiendo de la ubicación.

Por otra parte, la Figura 24b presenta el mapa de correlaciones de Kendall para el SPEI-6 con un mes de rezago (valores de febrero), utilizando el modelo XGBoost con la siguiente optimización de hiperparámetros: n_estimators=100, max_depth=9, learning_rate=0.1, gamma=0, colsample_bytree=0.8, min_child_weight=5 y subsample=0.8. El coeficiente de Kendall obtenido fue de 0.29, lo que indica una correlación positiva más alta en comparación con el SPEI-3, destacando un mejor ajuste general del modelo para esta escala temporal.

En el mapa predomina una correlación positiva alta y más extendida (mayor presencia de tonos azules oscuros), lo que evidencia un desempeño predictivo superior para el SPEI-6. Las áreas con correlaciones negativas o bajas (tonos amarillos y rojos) son significativamente menos frecuentes, lo que sugiere que el modelo logra capturar de manera más precisa las tendencias climáticas observadas a mediano plazo. Este resultado resalta la capacidad del modelo para representar el balance hídrico con mayor precisión en esta escala temporal.



(a) (b)

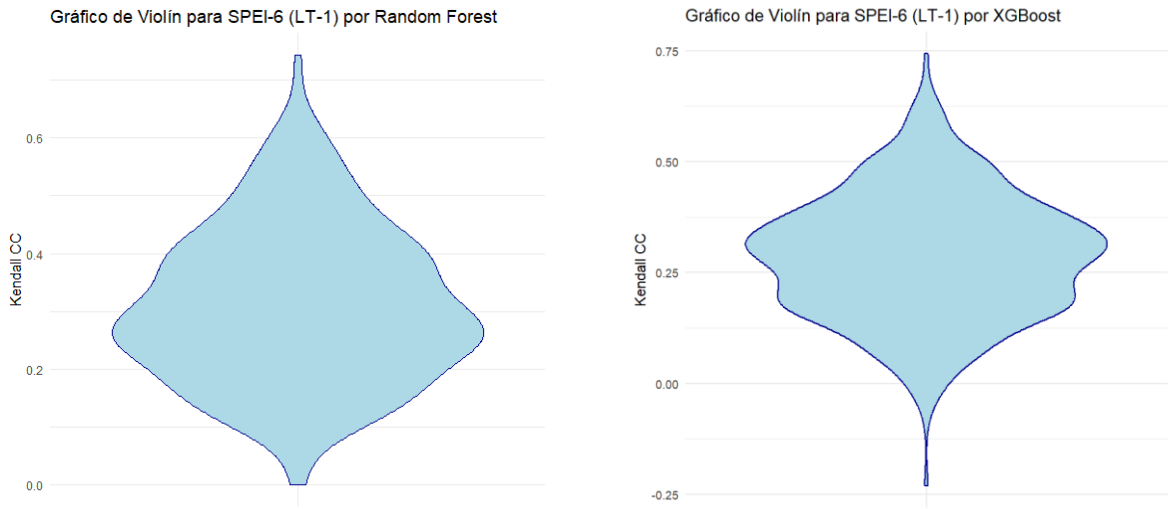
Figura 25. (a) Mapa de correlaciones de Kendall para SPEI-3 pronosticado con un rezago de 2 meses. (b) Mapa de correlaciones de Kendall para SPEI-6 pronosticado con un rezago de 2 meses.

La Figura 25 muestra las correlaciones espaciales obtenidas mediante el coeficiente de Kendall para las predicciones del modelo XGBoost con un rezago de dos meses (valores de enero). En el caso del SPEI-3 (Figura 25a), con hiperparámetros optimizados (n_estimators=100, max_depth=9, learning_rate=0.1, gamma=0, colsample_bytree=0.8, min_child_weight=1, subsample=0.8), se obtuvo un coeficiente de Kendall de -0.02, indicando una correlación muy baja entre los valores predichos y observados. Predominan los tonos cálidos (amarillos y rojos) en el centro y sur de la región, señalando correlaciones bajas, mientras que las áreas con correlaciones positivas (tonos azules) son escasas y dispersas en la zona norte correspondiente a la cordillera central.

Por su parte, en el SPEI-6 (Figura 25b), con hiperparámetros similares, pero ajustando min_child_weight a 5, el coeficiente de Kendall fue de -0.08, lo que evidencia una correlación baja y negativa más marcada. En este caso, los tonos cálidos son aún más prevalentes, indicando un desempeño inferior en comparación con el SPEI-3. Estos resultados reflejan un bajo desempeño del modelo XGBoost en la predicción de ambas escalas temporales con dos meses de rezago.

En este análisis, el valor global de la métrica corresponde al promedio de las correlaciones de Kendall calculadas para todos los píxeles del mapa. Sin embargo, este promedio puede verse afectado por zonas con correlaciones extremadamente bajas o altas, lo que podría limitar la capacidad para reflejar el desempeño general del modelo. Por ejemplo, al analizar los mapas de correlación, se observa que el modelo XGBoost presenta correlaciones altas en las zonas centrales, menos montañosas, a pesar de que el promedio global es penalizado por áreas con correlaciones bajas de las zonas correspondientes a las zonas de cordillera.

Para entender mejor el comportamiento de las correlaciones espaciales, se generaron gráficos de violín para el SPEI-6 con un rezago de un mes (LT-1) utilizando los modelos Random Forest (Figura 26a) y XGBoost (Figura 26b). Se escogió el SPEI-6 para estos dos modelos porque la correlación global de Kendall fue alta para el SPEI-6 LT-1, lo que lo hace representativo para analizar la distribución y dispersión de las correlaciones espaciales en cada caso.



(a) (b)

Figura 26. Distribución de las predicciones de los modelos de machine learning mediante gráficos tipo violín. (a) Modelo Random Forest. (b) Modelo XGBoost.

Los gráficos de violín muestran la distribución y dispersión de los coeficientes de correlación de Kendall en las diferentes áreas del mapa. En el caso de Random Forest,

se observa una mayor concentración de valores entre 0.2 y 0.4, lo que indica un desempeño moderado en la mayoría de las regiones, aunque con menor dispersión hacia valores extremos. Por otro lado, el gráfico de violín para XGBoost presenta una distribución más amplia, con valores que se extienden hasta 0.75 en las zonas de mejor desempeño, pero también alcanza valores negativos en las áreas de menor ajuste. Esta mayor dispersión sugiere que el modelo XGBoost captura mejor las correlaciones en las regiones centrales, a pesar de tener un desempeño más variable en otras áreas.

Por lo tanto, aunque la métrica global de Kendall resume el desempeño promedio del modelo, los gráficos de violín y los mapas de correlación permiten identificar las fortalezas y debilidades espaciales, proporcionando una visión más detallada del comportamiento de los modelos evaluados.

En comparación con el modelo Random Forest, el modelo XGBoost ofrece resultados que destacan por su capacidad de optimización y ajuste a los datos. Esto permite una mejor identificación de las áreas con mayor concordancia entre los valores observados y predichos, especialmente en escalas temporales más amplias como el SPEI-6 y con rezagos más cortos (1 mes). Sin embargo, al igual que en el caso del Random Forest, se observa una disminución en el desempeño predictivo al trabajar con rezagos mayores (2 meses).

Un hallazgo adicional es que, aunque ambos modelos muestran un desempeño general más bajo en términos de correlaciones espaciales para el rezago de 2 meses, en la zona norte se identifica una mejora notable en las correlaciones en comparación con el rezago de 1 mes, donde estas son consistentemente bajas. Esto sugiere que, para esta región, la información rezagada a dos meses podría contribuir positivamente a mejorar las correlaciones, posiblemente debido a características geográficas específicas de montaña en esta zona que influyen en el comportamiento climático.

Tabla 5. Resumen de desempeño de los modelos evaluados.

Modelo	Predictores	Goodness Index (Kendall tau) SPEI3		Goodness Index (Kendall tau) SPEI6	
CCA (CPT)	SST	0.47		0.49	
RF (CAST)	Precipitación, Radiación solar y Temperatura mínima	**LT-1** 0.22	**LT-2** -0.09	**LT-1** 0.31	**LT-2** -0.01
XGBoost (CAST)	Precipitación, Radiación solar y Temperatura mínima	**LT-1** 0.21	**LT-2** -0.02	**LT-1** 0.29	**LT-2** -0.08

Como se observa en la Tabla 5, los resultados de los modelos evaluados para predecir el Índice Estandarizado de Precipitación-Evapotranspiración (SPEI) muestran variaciones en el desempeño dependiendo de la escala temporal (SPEI-3 o SPEI-6) y el rezago (LT-1 y LT-2). El modelo CCA (CPT), basado en anomalías de la temperatura superficial del mar (SST), logra las correlaciones promedio más altas con valores de Kendall tau de 0.47 y 0.49 para SPEI-3 y SPEI-6, respectivamente. En comparación, los modelos RF (Random Forest) y XGBoost, que utilizan variables climáticas locales (precipitación, radiación solar y temperatura mínima), muestran un desempeño inferior. Para RF, el coeficiente Kendall tau alcanza un máximo de 0.31 (SPEI-6, LT-1), mientras que XGBoost obtiene un valor de 0.29 en la misma configuración. Sin embargo, ambos modelos presentan una disminución notable en el desempeño con un rezago de dos meses (LT-2), especialmente para SPEI-3, donde las correlaciones son cercanas a cero o negativas, sugiriendo que la capacidad predictiva se debilita con mayor rezago.

A pesar de que el modelo CCA con CPT presenta un Kendall tau promedio más alto, es importante destacar que utiliza una covariable global, la temperatura superficial del mar, lo que puede conducir a una generalización de las correlaciones en la zona de interés. Por el contrario, los modelos de machine learning muestran correlaciones más altas en áreas específicas, indicando que logran capturar patrones locales de manera más precisa. Esto confirma que, a nivel geográfico, las regiones de interés son diversas, con características que los modelos basados en variables locales pueden interpretar mejor,

mientras que el modelo CPT no logra alcanzar este nivel de detalle y no puede capturar los patrones específicos de déficit o exceso de humedad para la zona. Estos hallazgos resaltan no solo la importancia de seleccionar rezagos y escalas temporales adecuadas para maximizar la capacidad predictiva del modelo, sino también el potencial de los modelos de machine learning para abordar la heterogeneidad espacial de las regiones analizadas. Además, destacan la utilidad del **SPEI-6** para capturar variaciones a mediano plazo, siendo más robusto frente a rezagos en comparación con el **SPEI-3**.

6. Conclusiones

Este estudio permitió describir de manera detallada el comportamiento climático de la zona de estudio, proporcionando una base sólida para entender las condiciones hídricas de la región y su impacto en actividades agrícolas y en la gestión de recursos hídricos. La caracterización climática realizada permitió identificar particularidades de las diferentes zonas, destacando su relación con la variabilidad climática y fenómenos como El Niño y La Niña. A través del uso de los índices SPEI-3 y SPEI-6, construidos con datos provenientes de plataformas abiertas, se logró una buena correspondencia con las temporadas históricas de estos fenómenos en Colombia. Esto valida el SPEI como un indicador confiable para evaluar sequías y excesos hídricos, mostrando su utilidad en la caracterización de la variabilidad climática.

En cuanto a la evaluación de los modelos, aunque el modelo CCA (CPT) basado en la temperatura superficial del mar logró las correlaciones globales más altas (Kendall tau de 0.47 para SPEI-3 y 0.49 para SPEI-6), se debe resaltar que su uso generaliza las predicciones al no permitir capturar las variaciones espaciales en la región. Esto posiblemente se deba a que el modelo CPT se basa en una covariable global, lo que limita su capacidad para representar patrones locales en zonas específicas.

Por el contrario, los modelos de machine learning (Random Forest y XGBoost), aunque con un desempeño global menor (máximo Kendall tau de 0.31 para SPEI-6, LT-1 con RF), lograron identificar áreas específicas con correlaciones más altas, lo que evidencia su capacidad para capturar patrones locales y reflejar la diversidad geográfica de la región. En particular, el modelo XGBoost destaca porque, aunque su desempeño global es ligeramente inferior al de Random Forest debido a una mayor dispersión en los datos (con valores extremos más altos o más bajos), logra correlaciones espaciales altas en un mayor número de regiones. Esto sugiere que XGBoost tiene un mejor entendimiento de patrones específicos en ciertas zonas, lo que lo hace especialmente útil para análisis espaciales detallados.

Además, se encontró que el rezago de dos meses (LT-2) mejoró las correlaciones en la zona norte, particularmente en áreas montañosas, un hallazgo que destaca la importancia de ajustar los rezagos según la ubicación espacial para maximizar el desempeño predictivo.

Con los resultados logrados se pudo identificar que la mayor limitante del desempeño de los modelos de machine learning es la restricción que plantean librerías como CAST en R o XCast en Python, donde los predictores y los predictandos deben pertenecer a la misma área geográfica, en este aspecto el modelo de correlaciones canónicas tiene ventaja al utilizar predictores procedentes de áreas más grandes que la zona de estudio y que tienen efecto directo sobre la climatología de esa región.

Finalmente, dado que Random Forest mostró un desempeño global promedio ligeramente superior, pero XGBoost logró mejores correlaciones en regiones específicas, sería interesante explorar combinaciones o modelos híbridos que integren las fortalezas de ambos enfoques. También sería valioso evaluar modelos que combinen predictores globales y locales, lo que podría maximizar la capacidad predictiva y ofrecer un nivel de detalle más adaptado a las necesidades regionales.

La capacidad de pronosticar el SPEI con mayor precisión no solo refleja de manera más fiel la realidad climática, sino que también proporcionará una herramienta invaluable para la planificación y toma de decisiones en sectores industriales y agrícolas. Estos avances no solo fortalecerán la capacidad de predicción, sino que ampliarán las aplicaciones de los modelos en la planificación y gestión de recursos hídricos. De este modo, se permitirá una adaptación más eficiente a las variaciones climáticas, mejorando la resiliencia de las actividades económicas frente a eventos extremos y contribuyendo de manera significativa a la sostenibilidad y seguridad de la región.

6.1. Recomendaciones

Para futuros estudios se recomienda ampliar el conjunto de predictores utilizados, incorporando tanto variables climáticas locales como globales que permitan capturar de manera más precisa la variabilidad de la zona de estudio. Explorar predictores adicionales podría reflejar fenómenos climáticos locales y regionales con mayor detalle, fortaleciendo la capacidad predictiva de los modelos implementados.

Asimismo, ampliar la zona de análisis sería un paso clave para capturar una mayor diversidad en los patrones climáticos y así enriquecer la calidad de los predictores. Este enfoque permitiría mejorar el desempeño de los modelos de machine learning al considerar variaciones espaciales más amplias y complejas.

La optimización de rezagos y escalas temporales se presenta como otra área crucial de mejora. Continuar evaluando diferentes configuraciones de estos parámetros puede ayudar a identificar las combinaciones más adecuadas para cada región y fenómeno climático, potenciando la precisión de las predicciones, especialmente en áreas de alta variabilidad geográfica como las montañosas.

Para facilitar el trabajo con información espacio-temporal, se sugiere utilizar herramientas como R, que ofrece librerías específicas para manejar datos climáticos de múltiples áreas geográficas. Estas herramientas pueden ser fundamentales para integrar de manera eficiente predictores provenientes de diferentes fuentes y áreas de estudio.

Referencias

Agudelo, D., Mendez, A., Llanos, L., Barrios Pérez, C., Montes, C., Ayes, I., & Ramírez Villegas, J. A. (2023). Análisis retrospectivo del pronóstico estacional en Honduras. https://cgspace.cgiar.org/items/b977a12b-032e-4700-b8a9-8e1f8d5f5bad

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (2006). Evapotranspiración del cultivo: guías para la determinación de los requerimientos de agua de los cultivos. Roma: FAO, 298(0). https://openknowledge.fao.org/server/api/core/bitstreams/8802ddc9-86b6-4f13-96b7-4871dd3aee65/content

Angelidis, P., Maris, F. P., Kotsovinos, N., & Hrissanthou, V. (2009). Computation of Drought Index SPI with Alternative Distribution Functions. *Water Resources Management, 26*(23). https://doi.org/10.1007/s11269-012-0026-0

Badii, M. H., & Castillo, J. (2007). Análisis de correlación canónica (ACC) e investigación científica. Innovaciones de negocios, 4(8), 405-422.

Banimahd, S. A., Khalili, D. (2013). Factors influencing Markov chains predictability characteristics, utilizing SPI, RDI, EDI and SPEI drought indices in different climatic zones. *Water Resources Management, 27*(11), 3525-3538. https://doi.org/10.1007/s11269-013-0387-z

Climate Hazards Center. (2024). UC. Santa Barbara. https://www.chc.ucsb.edu/data/chirps

Copernicus. (2024). https://www.copernicus.eu/

Córdoba-Machado, S., Palomino-Lemus, R., Gámiz-Fortis, S.R., Castro-Díez, Y., Esteban-Parra, M.J. (2015). Influence of tropical Pacific SST on seasonal precipitation in Colombia: prediction using El Niño and El Niño Modoki. Clim. Dyn. 44, 1293–1310. https://doi.org/10.1007/s00382-014-2232-3

DANE. (2018). Censo Nacional de Población y Vivienda 2018. https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018/cuantos-somos

Demajo, Lara. (2020). Explainable AI for Interpretable Credit Scoring.

Dikshit, A., Pradhan, B., & Huete, A. (2021). An improved SPEI drought forecasting approach using the long short-term memory neural network. Journal of environmental management, 283, 111979.

Dong, J., Xing, L., Cui, N., Zhao, L., Guo, L., & Gong, D. (2023). Standardized precipitation evapotranspiration index (SPEI) estimated using variant long short-term

memory network at four climatic zones of China. Computers and Electronics in Agriculture, 213, 108253.

Environmental Systems Research Institute. (n.d.). *How XGBoost works*. Recuperado de https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm

Esquivel, A., Llanos-Herrera, L., Agudelo, D., Prager, S. D., Fernandes, K., Rojas, A., ... & Ramirez-Villegas, J. (2018). Predictability of seasonal precipitation across major crop growing areas in Colombia. Climate Services, 12, 36-47.

ESRI. (2024). Qué son los datos ráster?. https://desktop.arcgis.com/es/arcmap/latest/manage-data/raster-and-images/what-is-raster-data.htm

FNC. (2023). Informe del gerente 2023. 92 Congreso Nacional de Cafeteros. https://federaciondecafeteros.org/app/uploads/2023/11/IG-92-CNC-DIGITAL.pdf

IBM. (s.f.). *Random Forest*. Recuperado de https://www.ibm.com/mx-es/topics/random-forest

IRI. (2008). Climate Predictability Tool (CPT) User Guide. https://iri.columbia.edu/~awr/wiki/Downscaling/HydroOutlooks/Documents/cpttutorial_june08.pdf

Lopez-Bustins, J. A., Pascual, D., Pla, E., & Retana, J. (2013). Future variability of droughts in three Mediterranean catchments. *Climatic Change*, 119(2), 455-469. https://doi.org/10.1007/s10584-013-0771-6

López-Moreno, J. I., Lorenzo-Lacruz, J., Vicente-Serrano, S. M., González-Hidalgo, J. C.,& Cortesi, N. (2013). Hydrological drought response to meteorological drought in the Iberian Peninsula. Climate Research, 58(2), 117-131. https://doi.org/10.3354/cr01177

López, M. (2018). *Ensambladores: Random Forest - Parte I*. Bookdown. Recuperado de https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html

McEvoy, P. M., & Mahoney, A. E. J. (2012). To Be Sure, To Be Sure: Intolerance of Uncertainty Mediates Symptoms of Various Anxiety Disorders and Depression. Behavior Therapy, 43(3), 533–545. doi:10.1016/j.beth.2011.02.007

McKee, T. B., Doesken, N. J., & Kleist, J. (1993, January 17-22). The relationship of drought frequency and duration to time scales. In *Eighth Conference on Applied Climatology*. Anaheim, California: American Meteorological Society. Retrieved from https://climate.colostate.edu/pdfs/relationshipofdroughtfrequency.pdf

Mason, S. J., Tippett, M. K., Song, L., Muñoz, Á. G. (2023). Climate Predictability Tool version 18.3.1. Columbia University Academic Commons. https://doi.org/10.7916/x4yc-vc61.

Ortiz-Gómez, R., Cardona-Díaz, J. C., Ortiz-Robles, F. A., & Alvarado-Medellin, P. (2018). Caracterización de las sequías mediante la comparación de tres índices multiescalares en Zacatecas, México [Characterization of droughts by comparing three multiscale indices in Zacatecas, Mexico]. *Tecnología y Ciencias del Agua, 9*(3). https://doi.org/10.24850/j-tyca-2018-03-03

Santa Cruz-Suárez, A., Nápoles-García, M. C., & Morales-Guevara, D. (2022). El déficit hídrico en los cultivos y la acción de los microorganismos. *Cultivos Tropicales*, 43(3), 1-8. Ediciones INCA. https://www.redalyc.org/journal/1932/193275342013/html/#B6

SGS. (2023, diciembre). Sostenibilidad: Uso del agua en la agricultura. https://www.sgs.com/es-pe/noticias/2023/12/sostenibilidad-uso-agua-agricultura.

Soh, Y. W., Koo, C. H., Huang, Y. F., & Fung, K. F. (2018). Application of artificial intelligence models for the prediction of standardized precipitation evapotranspiration index (SPEI) at Langat River Basin, Malaysia. Computers and electronics in agriculture, 144, 164-173.

Tian, Y., Xu, Y. P., & Wang, G. (2018). Agricultural drought prediction using climate indices based on Support Vector Regression in Xiangjiang River basin. Science of the Total Environment, 622, 710-720.

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

Vrochidou, A.-E. K., & Tsanis, I. K. (2012). Assessing precipitation distribution impacts on droughts on the island of Crete. *Hydrology and Earth System Sciences*, 16, 1441-1455. https://doi.org/10.5194/hess-16-1441-2012