Universidad del Valle

PRONÓSTICO DEL ÍNDICE ESTANDARIZADO DE PRECIPITACIÓN (SPI) EN LA REGIÓN DE NARIÑO UTILIZANDO TÉCNICAS DE INTELIGENCIA ARTIFICIAL

Cristhian Eduardo Fernández Álvarez

Documento presentado al Programa de Posgrado en Ingeniería Eléctrica y Electrónica como requisito para optar por el título de Magíster en Ingeniería Énfasis en Automática

Directores

Wilfredo Alfonso Morales, Eng.D.

Teresita del Rocio Canchala Nastar, Ph.D.

UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA
Diciembre 2022

Resumen

En este trabajo de investigación se presenta un modelo computacional para el pronóstico del índice estandarizado de precipitación (SPI) en el departamento de Nariño. El modelo está conformado por las siguientes etapas: fuente de datos, metodología, modelo y finalmente el pronóstico del SPI. Para la etapa fuente de datos, se utilizó la base de datos satelital CHIRPS *(Climate Hazards Group InfraRed Precipitation with Station data)* y variables macroclimáticas, de CHIRPS se extrajeron datos de precipitación mensual. Posteriormente en la etapa de metodología, se calculó el SPI para diferentes escalas temporales (1, 3, 6 y 12 meses); luego, se realizo una reducción de dimensionalidad empleando el método de análisis de componentes principales no lineales *(ACPNL)* en conjunto con un análisis de cluster, con la finalidad de encontrar regiones espaciales donde los datos SPI presentaran un comportamiento similar; determinando las regiones SPI similares, se realizó un análisis de correlación entre las regiones y los índices macroclimáticos. Los índices que reportaron correlaciones altas fueron seleccionadas como variables predictoras del modelo de pronóstico, el modelo general, es la unión de cuatro modelos NNARMAX creados a partir de las regiones identificadas y la arquitectura ACPNL que dan como resultado el SPI pronosticado.

El modelo general de pronostico resultó en un modelo robusto con una ventana temporal de pronóstico de 6 meses, errores relativamente bajos de 0.13 (0.37) en el ECM, 0.85 (0.79) en el CPP y de 0.85 (0.78) en BICOR para la región Este (Oeste). El pronostico del SPI permite evaluar la sequía meteorológica, como punto de partida en la implementación de estrategias de adaptación y mitigación a la variabilidad climática. La reducción de la vulnerabilidad a la sequía mediante la generación de nuevo conocimiento y construcción de herramientas interdisciplinarias se constituye como el principal aporte de esta investigación a la región.

Palabras claves - *Sequía, SPI, Pronóstico, ACPNL, RNA, NNARMAX*

Abstract

This research work presents a computational model for the forecast of the standardized precipitation index (SPI) in the department of Nariño. The model is composed of the following stages: data source, methodology, model and finally the SPI forecast. For the data source stage, the CHIRPS satellite database (Climate Hazards Group InfraRed Precipitation with Station data) and macroclimatic variables were used; monthly precipitation data were extracted from CHIRPS. Subsequently, in the methodology stage, the SPI was calculated for different time scales (1, 3, 6 and 12 months); then, a dimensionality reduction was performed using the nonlinear principal component analysis method (ACPNL) together with a cluster analysis, in order to find spatial regions where the SPI data presented similar behavior; determining the similar SPI regions, a correlation analysis was performed between the regions and the macroclimatic indices. The indices that reported high correlations were selected as predictor variables of the forecast model. The general model is the union of four NNARMAX models created from the identified regions and the ACPNL architecture that result in the forecast SPI.

The overall forecast model resulted in a robust model with a 6-month forecast time window, relatively low errors of 0.13 (0.37) in the ECM, 0.85 (0.79) in the CPP and 0.85 (0.78) in BICOR for the East (West) region. The SPI forecast allows the assessment of meteorological drought, as a starting point in the implementation of adaptation and mitigation strategies to climate variability. The reduction of vulnerability to drought through the generation of new knowledge and the construction of interdisciplinary tools is the main contribution of this research to the region.

Keywords *- Drought, SPI, Forecast, NLPCA, ANN, NNARMAX*

Agradecimientos

Agradezco, inicialmente a mis directores: Wilfredo Alfonso Morales y Teresita Canchala Nastar por la confianza, dedicación y acompañamiento que me brindaron durante este proceso, también a los profesores Yesid Carvajal Escobar, Wilmar Loaiza Cerón y Wilmar Torres López, por su tiempo, aportes, observaciones y conocimiento compartido.

Gracias al grupo de investigación Percepción y Sistemas Inteligentes (PSI) y al grupo de investigación de Ingeniería de Recursos Hídricos y Suelos (IREHISA) por permitirme aprender en conjunto.

Agradezco a las entidades financiadoras del proyecto de investigación: Ministerio de Ciencias de Colombia (Minciencias) y Gobernación de Nariño mediante la convocatoria 818 en alianza con la Universidad de Nariño (Udenar). Proyecto titulado: "Análisis de eventos extremos de precipitación asociados a variabilidad y cambio climático para la implementación de estrategias de adaptación en sistemas productivos agrícolas de Nariño".

Finalmente, agradezco a la Universidad del Valle por formarme en este campo de estudio y a todos los docentes que hicieron parte de mi proceso de formación.

Índice

Lista de Tablas

Lista de Figuras

Listado de abreviaturas

A

ACP: Análisis de Componentes Principales.
ACPNL: Análisis de Componentes Principales No Lineales.
AMM: Modo Meridional del Atlántico.
AMO: Oscilación del Atlántico Norte.
ARMA: Modelo Autorregresivo de Media Móvil.
AR: Procesos autoregresivo.

B

BEST: Serie de tiempo bivariada del ENSO.
BMDI: Índice de sequía de Bhalme y Mooly.
BICOR: La correlación media biponderada.

C

CC: Cambio climático.
CMI: Índice de humedad de los cultivos.
CP: Componente principal.
CHIRPS: Grupo de Riesgos Climáticos de Precipitación Infrarrojos con datos de Estaciones.

E

EN: El Niño.
ENOS: El Niño Oscilación del Sur.
ENSO: Índice de Precipitación del ENOS.

F

FAO: Organización de las Naciones Unidas para la Alimentación y la Agricultura.
FUZZY: Agrupamiento difuso.

H

HMM: Modelo de Cadenas ocultas de Markov.
HC_AGLO: Cluster jerarquico aglomerativo.
HC_DIV: Cluster jerarquico divisivo.

I

IDEAM: Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia.

K

KMEANS: Agrupamiento por la media.

L

LN: La Niña.

M

MAE: Error Absoluto Medio.
MAPE: Error Porcentual Absoluto Medio.
MLP: Perceptrón Multicapa.
MA: Proceso de media móvil.
MEI: Índice Multivariado del ENSO.
MLR: Regresión Linear Multiple.
MC: Modelo completo.
MSP1: Modelo simplificado-1.
MSP2: Modelo simplificado-2.
MSP3: Modelo simplificado-3.

N

NAO: Oscilación del Atlántico Norte.
NRI: Índice de precipitación nacional.
NAO: Oscilación del Atlántico Norte.
NNARMAX: Modelo Autorregresivo de Media Móvil usando variables eXógenas con enfoque RNA.
NTA: Índice de la Temperatura Superficial del Mar del Atlántico Norte Tropical.

O

ONI: Índice Oceánico del Niño.
OMM: Organización Mundial de Meteorología.
OCB: Oscilación Cuasi Bienal.

P

PDO: Oscilación Decadal del Pacífico.
PDSI: Índice de severidad de la sequía de Palmer.
PDO: Oscilación Decadal del Pacífico.
PERSIANN: Estimación de la Precipitación de Información de Sensores Remotos.

Q

QBO: Oscilación Quiasibienal.

R

RNA: Red Neuronal Artificial.
RMSE: Raíz del Error Cuadrático Medio.
CCP: Coeficiente de determinación de Pearson.
RAI: Índice de anomalías pluviale.
R.E: Región este.
R.OE: Región oeste.

S

SOI: Índice de Oscilación Sur.
SPI: Indice Estandarizado de Precipitación.
SWSI: Índice de suministro de agua superficial.
SOM: Mapas Autorganizados de Kohonen.

T

TSM: Temperatura Superficial del Mar.
TNA: Índice del Atlántico Norte Tropical.
TNI: Índice del Niño Trans.
TSA: Índice del Atlántico Sur Tropical.
TSM: Temperatura Superficial del Mar.

V

VC: Variabilidad climática.
VM: Variables Macroclimáticas.
VIF: Factor de inflación de varianza.

WP

WP: Índice del Pacífico Occidental.

Z

ZCIT: Zona de Convergencia Intertropical.

Capítulo 1

Introducción

La variabilidad climática (VC) y el cambio climático (CC) son el principal problema ambiental del siglo XXI que están afectando negativamente a la sociedad y los ecosistemas [5], ya que impactan ámbitos sociales, económicos, ambientales, y ecológicos. El CC y la VC es un problema global que afecta de manera diferenciada a distintas zonas del planeta, como la región tropical. Esta región ubicada entre los trópicos de Cáncer y de Capricornio o entre los $27.5°S$ y $27.5°N$ [5], tiene la particularidad de disponer de la mayor cantidad de energía solar que recibe el planeta. Se caracteriza por presentar altas temperaturas donde predomina la circulación de los vientos alisios (vientos que soplan del Este, del Noreste o del Sureste) los cuales interactúan en la Zona de Convergencia Intertropical (ZCIT) [5]. La ZCIT, se caracteriza por ser una franja o cinturón de baja presión constituido por corrientes de aire ascendente, donde convergen grandes masas de aire cálido y húmedo provenientes de los hemisferios norte y sur; lo cual inducen en el comportamiento hidroclimatológico de la región.

En Colombia, el ciclo hidroclimatológico anual está controlado por la oscilación de la ZCIT, superpuesta a patrones regionales causados por la influencia orográfica de los Andes, la evapotranspiración en la cuenca amazónica, las interacciones continente-atmósfera y la dinámica de las corrientes de viento del oeste de Colombia (corriente de chorro del Oeste Colombiano y corriente de chorro del Chocó) [6, 5, 7]. En escalas de tiempo más largas, se experimentan importantes anomalías hidrológicas debidas a los ciclos del fenómeno El Niño Oscilación del Sur (ENOS) y sus dos fases extremas, la fase fría denominada "La Niña" (LN) y la fase cálida denominada "El Niño" (EN) [5]. Varios estudios afirman que los episodios de EN (LN) tienen relación con la disminución (aumento) de las precipitaciones, caudales y otras variables hidroclimáticas, predominantemente en el centro, norte y occidente del país [8, 9, 10, 11, 12]. Además de ENOS, otros fenómenos macroclimáticos como la Oscilación del Atlántico Norte (NAO) y la Oscilación Decadal del Pacífico (PDO) [6], son importantes en el comportamiento hidroclimatológico de Colombia.

La ocurrencia de ENOS es cíclica, no periódica en escalas de tiempo que van desde meses hasta varios años. Este fenómeno es determinante en la influencia sobre el recurso hídrico en Colombia. En el Océano Pacífico, existe una conexión entre la Temperatura Superficial del Mar (TSM), la presión atmosférica, la dirección de los vientos y corrientes oceánicas, variables que se retroalimentan y determinan la intensidad del fenómeno ENOS.

Estas variables macroclimáticas (VM) pueden ser usadas para la construcción de índices que permite evaluar la severidad y duración de un evento ENOS.

Existe variedad de índices que son usados para evaluar la fuerza de un evento ENOS particular, el Índice Oceánico del Niño (ONI) es uno de los más usados, se calcula como la media de 3 meses consecutivos de las anomalías de la TSM en la región Niño 3.4 del océano pacifico ($50°N$-$5°S$, $120°W$-$170°W$). Además del índice ONI, también es usado el Índice de Oscilación Sur (SOI), este índice es calculado como la diferencia normalizada de la presión atmosférica en superficie entre Darwin, Australia (Pacífico Occidental) y Tahití, Polinesia Francesa (Pacífico Oriental). En la actualidad, hay diferentes índices desarrollados para medir la variabilidad de los parámetros oceánicos y atmosféricos de gran escala y son usados para estudiar la relación o incidencia en el comportamiento espacio-temporal de la precipitación, de la nubosidad y de otras variables climatológicas locales [5].

Los impactos del ENOS en las precipitaciones colombianas, han dado lugar a consecuencias ambientales socioeconómicas y medioambientales [7, 13]. Durante el evento EN de 1997-1998, Colombia registró una disminución del 10 % en la producción de café y se registraron cerca de 12.000 incendios forestales [14]. Durante el evento EN de 2015-2016, uno de los más fuertes de los que se tiene registro [15], el déficit de precipitaciones fue del 40 % en la región andina y del 37 % en el Caribe, afectando a más de un millón de hectáreas de cultivos, principalmente en los departamentos de Atlántico, Córdoba y Nariño [14]. Además, en las zonas endémicas de Colombia, EN se ha asociado también a un aumento de los casos de malaria [16], debido al aumento de la temperatura y la disminución de las precipitaciones, que son condiciones favorables para la formación de criaderos del mosquito transmisor.

La precipitación es una de las principales variables que definen el comportamiento climático en la mayoría de regiones del mundo, su variabilidad temporal y espacial tiene un efecto directo en la disponibilidad de los recursos hídricos. Por lo tanto, el análisis de la información histórica de la precipitación y su relación con VM permiten identificar variaciones temporales y espaciales para mitigar los riesgos relacionados con el agua, como la sequía.

El agua es esencial para aprovechar el potencial de la tierra y para permitir que las variedades mejoradas tanto de plantas como de animales utilicen plenamente factores de producción que elevan los rendimientos. Al incrementar la productividad, la gestión sostenible del agua (especialmente si va unida a una gestión adecuada del suelo) contribuye a asegurar una mejor producción tanto para el consumo directo como para el comercio. Desde los años sesenta, la producción mundial de alimentos ha seguido la tendencia del crecimiento demográfico mundial, suministrando más alimentos per cápita a precios cada vez más bajos en general, pero a costa de los recursos hídricos. Al final del siglo XX, la agricultura empleaba por término medio el 70 % de toda el agua utilizada en el mundo, y la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) estima que el agua destinada al riego aumentará un 14 por ciento para 2030 [17]. Aunque este aumento es muy inferior al registrado en los años noventa, según las proyecciones, la escasez de agua será cada vez mayor en algunas regiones, lo que limitará la producción local de alimentos [17].

La sequía, es provocada por la fase cálida de ENOS (EN) y afecta el margen occidental

de América Central, México, la cuenca del Amazonas y el norte de América del Sur (es decir, Colombia, Ecuador, el norte de Perú y el noreste de Brasil) [6]. La sequía es definida como los sucesos prolongados y regionalmente extensos de disponibilidad de agua natural por debajo del promedio [2], que afecta los recursos hídricos superficiales y subterráneos conllevando a dificultades en actividades económicas y sociales, tales como: reducción del suministro de agua, deterioro de la calidad del agua, pérdida de cosechas, reducción de la productividad en sectores agrícolas y disminución de la generación de energía [7].

En las ultimas décadas [18, 19, 4], las sequías han tenido graves repercusiones en el sector rural de los países en desarrollo, debido a que el 75 % de los países pobres del mundo habitan en zonas rurales, dependiendo de la agricultura de secano[1] para su subsistencia [20], tal como sucede en Colombia, donde una de las fuentes de su economía es la agricultura [7].

El departamento de Nariño localizado en el suroeste de Colombia tiene una economía basada en actividades agropecuarias, debido a las condiciones topográficas e hidrográficas, seguida por la industria, el comercio y el transporte [21]. Un departamento esencialmente rural con predominio de producción minifundista[2] [7]. La agricultura es un factor fundamental para el desarrollo y la seguridad alimentaria [21], el uso del agua y su gestión han sido un factor esencial para elevar la productividad y asegurar una producción previsible.

El pronóstico de eventos futuros de sequía en el departamento de Nariño es relevante para reducir los impactos especialmente en la gestión y mitigación del riesgo ante eventos de sequía [22]. Los eventos de sequía se explican usando indicadores que determinan características como la magnitud, frecuencia, duración y extensión espacial; el índice de sequía meteorológica recomendado por la Organización Mundial de Meteorología (OMM), es el Índice de Precipitación Estandarizado (SPI) [3], el cual es utilizado para abordar problemas relacionados con la sequía en todo el mundo [23].

Dada la necesidad del departamento de Nariño en predecir los futuros eventos de sequía, en esta investigación se propone la siguiente pregunta: *¿Cómo desarrollar un modelo computacional para pronosticar eventos de sequía a partir del índice SPI en el departamento de Nariño, usando técnicas de inteligencia artificial?.* Para responder a la pregunta de investigación se propusieron los siguientes objetivos.

1.1. Objetivos

1.1.1. Objetivo general

- Desarrollar un modelo computacional para el pronóstico del índice estandarizado de precipitación (SPI) en el departamento de Nariño, utilizando técnicas de inteligencia artificial.

[1] Actividad agrícola en la que el ser humano no contribuye a la irrigación del suelo, sino que utiliza únicamente la que proviene de las lluvias.

[2] El agricultor no obtiene una producción suficiente para comercializar, obligando al auto abastecimiento y la agricultura de subsistencia.

[3] En 2009, la OMM recomendó a los países que utilizaran el SPI, para vigilar y seguir las condiciones de sequía.

1.1.2. Objetivos específicos

- Identificar las técnicas de pronóstico del SPI usando índices macroclimáticos y métricas de desempeño preponderantes en la estimación de los pronósticos.
- Proponer un modelo computacional para el pronóstico del SPI usando técnicas de inteligencia artificial con un sistema de correlación robusta.
- Implementar el modelo computacional de pronóstico del SPI usando técnicas de inteligencia artificial con un sistema de correlación robusta.
- Validar el modelo computacional de pronóstico a través de pruebas funcionales y de integración, usando métricas de desempeño.

1.2. Descripción general de la solución propuesta

En esta investigación se diseñó e implementó un modelo computacional de pronóstico del índice SPI, utilizando series de tiempo de precipitación tomadas a partir de un conjunto de datos satelitales en el departamento de Nariño. Para el modelo de pronóstico se utilizaron técnicas de inteligencia artificial usando como referencia la metodología propuesta en [24].

La descripción general de la solución propuesta, se muestra en el diagrama de bloques de la Fig. 1.1, donde se inicia con la fuente de datos, seguido del modelo computacional de pronóstico, que está dividido en los sub-bloques de metodología y modelo; finalmente, como resultado se tiene el SPI pronosticado.

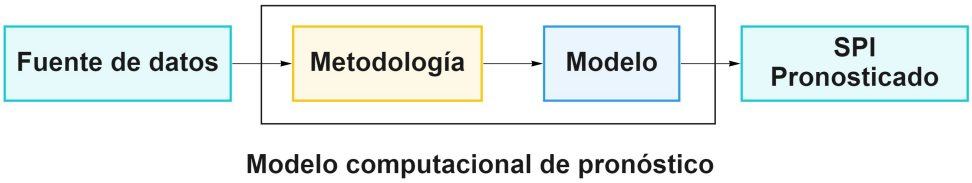


Figura 1.1: Diagrama general de la solución propuesta.

El bloque de **fuente de datos** contiene las series de tiempo de precipitación de estaciones meteorológicas *in situ* y series de tiempo que fueron extraídas de datos satelitales. Por su parte, las series de tiempo correspondientes a las variables macroclimáticas (VM) fueron tomadas de la Oficina Nacional de Administración Oceánica y Atmosférica (NOAA)[4]. Las series temporales descritas anteriormente son las variables de entrada al bloque de **metodología** (Ver Fig. 1.1), en donde inicialmente se realiza la estimación del SPI siguiendo la metodología propuesta por McKee, Doesken y Kleist [25], con el propósito de evaluar la sequía para diferentes escalas temporales (1, 3, 6 y 12 meses). Con el SPI calculado, se realiza una reducción de dimensionalidad temporal, para luego emplear métodos de *clustering* y determinar regiones espaciales con datos de similar comportamiento. Una vez las regiones de SPI han sido identificadas, se realiza una reducción espacial de los datos SPI calculados inicialmente y separados por regiones, con el propósito de analizar las

[4] Datos de la NOAA: `https://www.aoml.noaa.gov/es/data-products/`

teleconexiones existentes con las VM. Finalizado el análisis de teleconexiones se seleccionan las mejores variables predictoras para usarlas en el modelo de pronóstico.

Teniendo los mejores predictores, se ingresa en el bloque de **Modelo**, donde se realiza el entrenamiento al modelo de pronóstico, el proceso de entrenamiento tiene métricas de desempeño y una metodología definida, como resultado se tiene el modelo final de pronóstico, que permite determinar el **SPI pronosticado** como salida del bloque de **Modelo**.

1.3. Estructura del documento

En el documento, se presentan los capítulos que abordan los objetivos específicos propuestos, en donde se expone el desarrollo de estos siguiendo las etapas presentadas en la solución propuesta (Ver Fig. 1.1), cada capítulo tiene el siguiente contenido.

- **Capítulo 1:** Se presenta la introducción, descripción del problema abordado, los objetivos planteados y la solución propuesta, la cual se describe con mayor detalle en el capítulo 3 y 4.
- **Capítulo 2:** Se presenta la revisión del estado del arte respecto al pronóstico de índices meteorológicos de sequía usando modelos de inteligencia artificial y se definen conceptos que permiten asociar ideas que rodean el desarrollo del modelo computacional de pronóstico del índice SPI.
- **Capítulo 3:** Se define la zona de estudio y se explica el bloque de fuente de datos, posteriormente se aborda la metodología empleada para determinar los mejores predictores del modelo computacional del pronóstico.
- **Capítulo 4:** Se detalla el bloque de Modelo, en donde se registra el proceso de entrenamiento y mejoras del modelo de pronóstico . Además, se presentan las pruebas y resultados del modelo de final de pronóstico.

Capítulo 2

Antecedentes y marco teórico

Este capítulo, tiene como objetivo identificar las técnicas de pronóstico de índices de sequía usando información exógena y endógena. Además, identificar las métricas de desempeño mas usadas para la evaluación de modelos de pronóstico y definir conceptos que permiten comprender el desarrollo del modelo computacional de pronóstico del SPI.

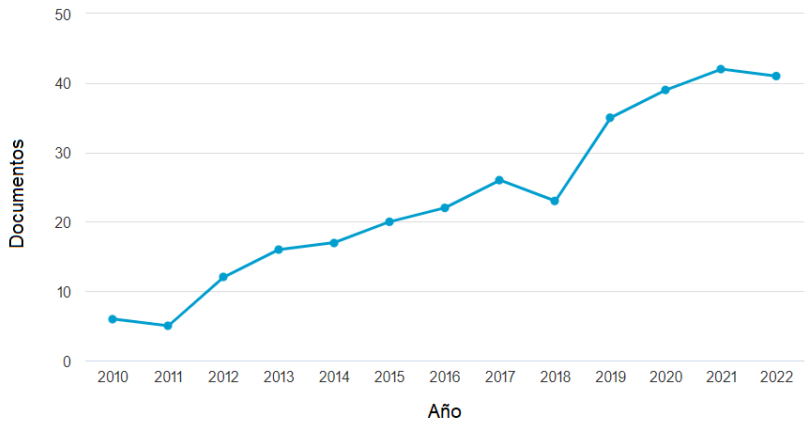
2.1. Antecedentes

A continuación, se presentan las investigaciones sobre el pronóstico de índices meteorológicos de sequía en los últimos 10 años. El proceso de búsqueda que se llevo a cabo utilizó dos ecuaciones de búsqueda las cuales se describen a continuación:

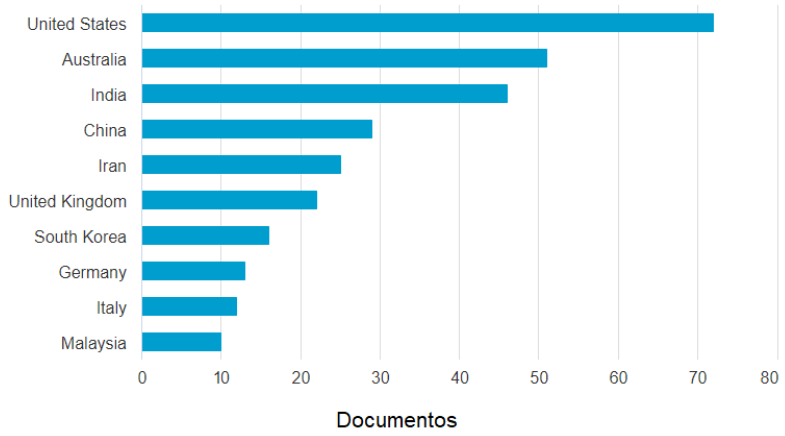
- TITLE-ABS-KEY (forecasting AND drought AND index AND rainfall) AND PUBYEAR 2009 AND PUBYEAR 2023.
- TITLE-ABS-KEY (forecasting AND drought AND index AND rainfall AND colombia) AND PUBYEAR 2009 AND PUBYEAR 2023.

Como se observa en la Fig. 2.1(a), las investigaciones han incrementado desde el año 2011 hasta la actualidad (Año 2022). El aumento en número de investigaciones está asociado al déficit de lluvias como consecuencia del calentamiento global y el cambio climático [8, 26, 7], cambios que actualmente se están experimentando y en un futuro sus efectos serán más evidentes [5].

Estados Unidos, Australia e India son los países que lideran las producciones científicas relacionadas con pronóstico de índices meteorológicos de sequía. En la Fig. 2.1(b), se puede evidenciar las producciones científicas discriminadas por país, y se observa que Latinoamérica tiene la menor producción científica en esta área de investigación. En este sentido, se evidencia la necesidad de investigar y generar aportes en esta temática.



(a) Tendencia de las investigaciones sobre el pronóstico de índices meteorológicos de sequía.



(b) Países líderes en producción de artículos sobre el pronóstico de índices meteorológicos de sequía.

Figura 2.1: Estado del arte, investigaciones y países lideres en productos académicos.

A continuación, se presenta una tabla comparativa de las características más relevantes encontradas en la revisión de antecedentes. Las características que se registran en la Tabla 2.1 son:

- **Título del documento/artículo:** título de la investigación.
- **Ref:** referencia del documento/artículo.
- **País:** país donde se desarrolló la investigación.
- **Área de estudio:** lugar en donde se realizó la investigación.
- **Año:** año de publicación de la investigación.
- **Modelo:** técnicas o métodos usados para el modelo de pronóstico.
- **Métricas:** índices que permiten evaluar el desempeño del modelo usado para el pronóstico.
- **Datos usados:** descripción de los datos usados para el pronóstico.
- **Datos pronosticados:** descripción de los datos pronosticados.
- **Datos exógenos:** descripción de las variables exógenas usadas en el modelo de pronóstico.

Tabla 2.1: Características más relevantes de investigaciones sobre el pronóstico de índices meteorológicos de sequía

Título	Ref	País	Área de estudio	Año	Modelo	Métricas	Datos usados	Datos pronosticados	Datos exógenos
Drought Forecasting Using a Hybrid Stochastic and Neural Network Model.	[27]	India	Cuenca del rio Kansabati	2010	Modelo híbrido: modelo lineal estocástico – ANN	RMSE, MAE, R^2	Precipitación mensual (1965-2001)	SPI 1, 3, 6 y 9 meses	-
Bayesian Neural Network Ensemble Model based on Partial Least Squares Regression and Its Application in Rainfall Forecasting.	[28]	China	Región de Guangxi	2010	Técnicas bayesianas y ANN usando diferentes algoritmos de entrenamiento	RMSE, MAPE, MAE, R^2	Precipitación mensual (1954-2008)	Precipitación mensual	-
North Tropical Atlantic influence on western Amazon fire season variability.	[29]	Perú	Región cercana al Amazonas	2011	Modelos de regresión lineal	RMSE, PRESS	Precipitación mensual (1970-1990)	SPI 1 y 6 meses	-
MLP-based drought forecasting in different climatic regions.	[30]	Irán	Región de Iran	2012	ANN	RMSE, R^2	Precipitación mensual (1972-2012)	SPI 3, 6, 9, 12 y 24 meses	-
Application of Artificial Neural Networks to Rainfall Forecasting in Queensland, Australia.	[31]	Australia	Estado de Queensland	2012	ANN	RMSE, R^2	Precipitación mensual (1900-1993)	Precipitación mensual	Temperatura atmosférica, SOI, PDO, Niño 3.4
Study of Droughts and Floods Predicting SystemBased on Spatial-Temporal Data Mining	[32]	China	Toda la región de China	2013	PCA y ANN	RMSE, MAE, R^2	Precipitación mensual (1951-2010)	SPI 1, 3 y 6 meses	-
Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models	[33]	Etiopía	Cuenca del río Awash	2014	ARIMA, ANN, SVR además se usó transformada wavelet para el pre-procesamiento de los datos	RMSE, MAE, R^2	Precipitación mensual (1970-2005)	SPI 12 y 24 meses	-
Forecasting of meteorological drought using Hidden Markov Model (case study: The upper Blue Nile river basin, Ethiopia).	[34]	Etiopía	Cuenca superior del río Nilo Azul	2015	Modelos ocultos de Márkov	RMSE, MAE, R^2	Precipitación mensual (1960-2007)	SPI 3, 6, 9 y 12 meses	-
Drought Forecasting Using MLP Neural Networks	[35]	Malasia	Cuenca del río Selangor	2015	MLP	RMSE, MAE, R^2	Precipitación mensual (1948-1993)	SPI 3, 6 y 9 meses	-

Continua en la siguiente página

Tabla 2.1 – Continua de la página anterior

Título	Ref	País	Área de estudio	Año	Modelo	Métricas	Datos usados	Datos pronosticados	Datos exógenos
A hybrid evolutionary probabilistic forecasting model applied for rainfall and wind power forecast	[36]	Brasil	Ciudad de Victoria	2016	Modelos difusos sintonizados con técnicas meta-heurísticas bio-inspiradas	RMSE, R^2	Precipitación mensual (2013-2014)	Precipitación mensual	-
Drought forecasting using novel heuristic methods in a semi-arid environment	[37]	Irán	Provincia de Semnan	2019	Métodos neuro-difusos, con algoritmos de partículas (ANFIS-PSO), con algoritmos genéticos (ANFIS-PSO)	RMSE, MAE, IA, R^2	Precipitación mensual (1985-2015)	SPI 3, 6, y 12 meses	-
Drought prediction based on SPI and SPEI with varying timescales using LSTM recurrent neural network	[38]	India	Región de Hyderabad	2019	Modelo estadístico ARIMA, redes neuronales recurrentes LSTM	RMSE, R^2	Precipitación mensual (1958-2013)	SPI 1, 2, 3, 4 y 6 meses	Humedad y temperatura atmosférica
Forecasting standardized precipitation index using data intelligence models: regional investigation of Bangladesh	[39]	Bangladesh	Todo el territorio	2020	Regresión de máquina de probabilidad mínima (MPMR), Árbol M5 (M5tree), Máquina de aprendizaje extremo (ELM) y ELM-secuencial (OSELM)	MSE, MAE, RMSE, R^2	Precipitación mensual (1949-2013)	SPI 1, 2, 3, 4 y 6 meses	-
Rainfall and Meteorological Drought Forecasting in Albay, Philippines Using Artificial Neural Network	[40]	Filipinas	Provincia de Albay	2021	Regresor usando redes neuronales MLP	RMSE, R^2	Precipitación mensual (1970-2000)	SPI mensual	precipitación del mes actual, humedad relativa, temperatura media
Assessment of Drought Forecasting Model for Northeastern, Thailand	[41]	Tailandia	Región noreste	2022	Regresor usando redes neuronales MLP, regresión lineal, regresión aditiva, árboles de decisión	MAE, RMSE, R^2	Precipitación mensual (2018-2021)	Humedad del suelo	Índice de Vegetación de Diferencia Normalizada (NDVI)

Los datos usados en la mayoría de las investigaciones son series de precipitación utilizadas para calcular los índices meteorológicos de sequía como lo hizo Rezaeianzadeh et al.[30], Mishra et al.[27], Khadr [34] y otros autores [31, 32, 37, 38, 29]. El índice de sequía más utilizado para determinar y evaluar un evento de sequía, es el índice estandarizado de precipitación (SPI). Se utiliza frecuentemente el SPI debido a que en el año 2009 la Organización Mundial de Meteorología (OMM) recomendó a los países el uso de este índice, para vigilar y monitorear las condiciones de sequía en todo el mundo [42]. Esto evidencia que los datos pronosticados en la mayoría de investigaciones corresponden al SPI usando diferentes escalas temporales; 3, 6, 9 y 12 meses en la mayoría de casos. Las métricas de desempeño más usadas para evaluar los modelos de pronóstico son: raíz del error cuadrático medio (RMSE), error absoluto medio (MAE), error porcentual absoluto medio (MAPE) y el coeficiente de determinación (R^2).

En algunos artículos se tuvieron en cuenta datos exógenos (variables adicionales a la variable respuesta usada en los modelos de predicción). Abbot et al.[31] y Poornima et al.[38] usaron variables exógenas: temperatura atmosférica, presión atmosférica y temperatura del océano. Con dichas variables exógenas y las series de tiempo del SPI (índice calculado a partir de los datos de precipitación a diferente escala temporal) fueron entrenados los respectivos modelos de pronóstico. No se logró encontrar un punto de comparación entre un modelo que usa variables exógenas en su entrenamiento y un modelo que no; pero se puede intuir que el modelo que tiene mayor información en su entrenamiento puede tener un mejor desempeño, dado que puede capturar relaciones entre las variables exógenas y las endógenas del sistema estudiado, logrando mejorar el pronóstico.

Los métodos usados para el regresor, se dividieron en tres grupos según Priyamvada et al.[43], modelos estadísticos (ARMA y ARIMA), modelos de inteligencia artificial (ANN, SVM y ANFIS) que utilizan diferentes algoritmos para el entrenamiento y modelos híbridos, que combinan los modelos estadísticos y de inteligencia artificial. El modelo híbrido fue el más usado en las investigaciones consultadas. El análisis anterior se resume en la Fig. 2.2, que incluye:

- Las principales técnicas de pronóstico de las series temporales.
- Las métricas de desempeño usadas para evaluar el modelo de pronóstico.
- Las técnicas de correlación empleadas para determinar el nivel de influencia entre las variables exógenas y las endógenas del modelo de pronóstico.
- Los índices meteorológicos más usados. Los índices se dividieron en dos grupos dependiendo del grado de complejidad en el cálculo del índice y en la interpretación según la OMM [42].
- La información de los índices meteorológicos de sequía se complementó consultando el manual de índices meteorológicos publicado por la OMM.

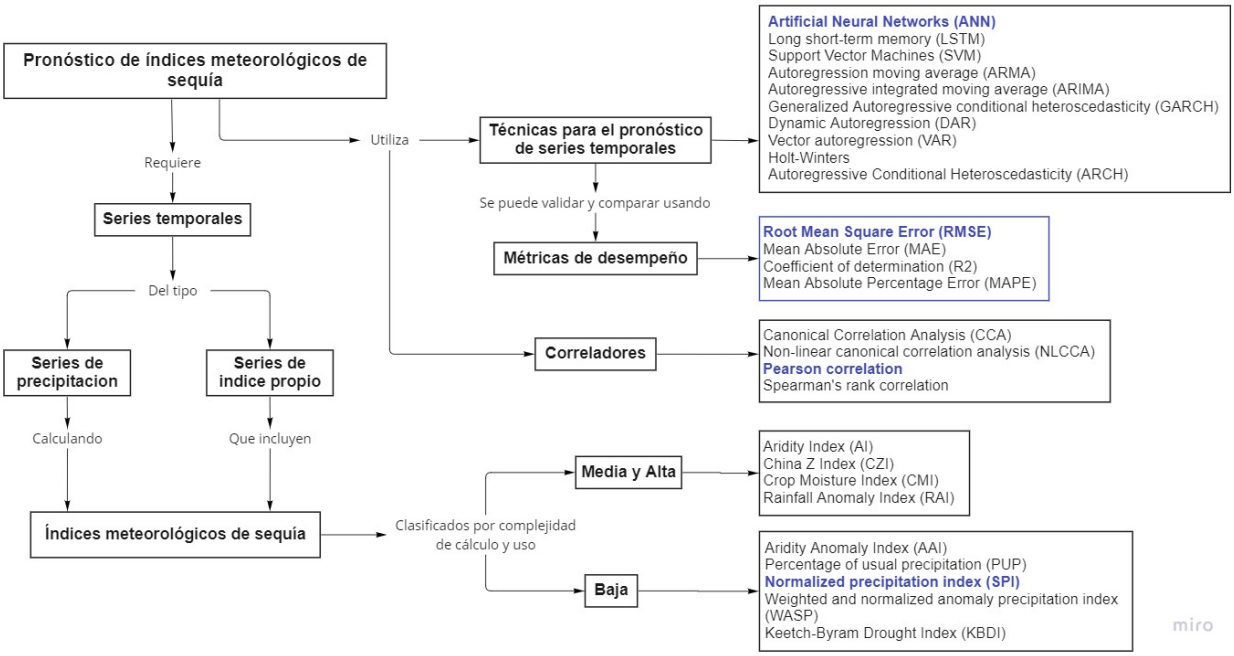


Figura 2.2: Diagrama de flujo para el pronóstico de índices meteorológicos de sequía, los métodos o índices en color azul son los más usados según la consulta de investigaciones previas.

2.2. Marco teórico

2.2.1. Clima

El clima es un conjunto de condiciones atmosféricas predominantes durante un período de tiempo determinado sobre un lugar o una región (días, semanas, meses, años, siglos) [5]. Las condiciones predominantes generalmente se cuantifican con el promedio de variables como la temperatura, acumulado de precipitación y número de fenómenos extremos durante el periodo definido [5].

Las estaciones del año son la expresión más común del concepto clima. En algunas regiones del mundo las estaciones se manifiestan en la variación de la temperatura promedio durante el año: verano, otoño, invierno, primavera. En regiones tropicales las estaciones del año están definidas por la precipitación, considerando periodos lluviosos y periodos secos [5].

Colombia se encuentra ubicada muy cerca de la línea ecuatorial, su clima es predominantemente tropical y las estaciones están marcadas por la precipitación: periodo lluvioso y periodo seco. Las temperaturas promedio son estables durante todo el año por lo que el distintivo entre las dos estaciones lo marca la precipitación promedio [5].

El clima regula la distribución de las condiciones meteorológicas y los fenómenos atmosféricos extremos. Debido a condiciones climáticas, determinado tipo de fenómeno extremo se registra o es más frecuente en una región, dado que es controlado por la estacionalidad [5] y esto conlleva a las distinciones estacionales de una región determinada.

2.2.2. Patrón climatológico

En el estudio del clima se utiliza el concepto patrón climatológico, el cual representa las condiciones climáticas que predominan durante un período largo de tiempo, generalmente 30 años, con el que se caracterizan el clima de una región [5, 26]. Este se cuantifica mediante el cálculo de promedios de las observaciones o mediciones realizadas a las variables climatológicas (temperatura del aire, presión atmosférica, humedad relativa, precipitación, etc.) y de la frecuencia de los fenómenos meteorológicos extremos registrados.

Los valores que representan lo predominante en un período de 30 años (por ejemplo, en 1971-2000 o 1989-2019), se denomina norma (normal) climatológica. Con las normales climatológicas se establece el patrón de la región, es decir la distribución espacial y la estacionalidad de las variables climatológicas. La situación que se diferencia del patrón establecido en un lugar o región se denomina anomalía climática [26].

2.2.3. Cambio climático y variabilidad climática

El CC se define como la modificación a largo plazo de las condiciones meteorológicas mundiales predominantes (patrón climático) debido a la variabilidad natural o como resultado de actividades humanas, estas condiciones pueden darse en periodos de tiempo largos (años) y espacialmente (hemisférico o global), los cuales pueden representar una amenaza natural, produciendo o intensificando eventos meteorológicos extremos como sequías, inundaciones, olas de frío o de calor, tormentas, entre otros [9].

Por su parte, la VC se refiere a variaciones en las condiciones medias del estado del tiempo (días, meses, años) y otras estadísticas (como las desviaciones típicas y los fenómenos extremos, entre otros) en todas las escalas temporales y espaciales que se extienden más allá de la escala de un fenómeno meteorológico en particular. La variabilidad puede deberse a procesos naturales internos que ocurren dentro del sistema climático [5].

2.2.4. Fenómenos de macroescala en Colombia

La VC se refiere a las fluctuaciones del clima durante períodos de tiempo cortos (años, meses), que genera un comportamiento anormal y puede ocasionar excesos o déficit en variables meteorológicas. Los fenómenos de macroescala que rigen la hidroclimatología en Colombia son [11]:

- Eventos convectivos de mesoescala conectados a la zona de convergencia intertropical.
- La Oscilación Decadal del Pacífico (PDO).
- El fenómeno El Niño Oscilación del Sur (ENOS).
- Oscilación Atlántico Norte (NAO).
- Oscilación Cuasi Bienal (OCB).

El fenómeno ENOS es el de mayor influencia sobre la variabilidad interanual del clima y del recurso hídrico en Colombia [11]. La magnitud de los cambios en los balances de agua y energía globales que ocurren durante las dos fases de ENOS, ocasionan fuertes perturbaciones

hidroclimáticas, particularmente en los cinturones tropicales y subtropicales de la tierra, con amplias repercusiones sociales, ambientales, ecológicas y económicas [11]. Estos fenómenos se presentan con mayor detalle en Capítulo 3.

2.2.5. Fenómeno El Niño Oscilación del Sur – ENOS

ENOS es el conjunto de variaciones océano-atmosféricas más influyentes a escala planetaria [5, 11], que altera condiciones normales intertropicales con impactos asociados que afentan el clima planetario, siendo una de las principales causas de VC interanual a nivel mundial [44] que resulta de las interacciones océano-atmósfera a gran escala sobre el Pacífico ecuatorial en escalas de tiempo de meses a varios años [45]. Este fenómeno presenta dos componentes esenciales para su desarrollo e intensidad: la componente oceánica y la componente atmosférica. El Niño generalmente es asociado con la aparición de una corriente oceánica cálida a lo largo del Océano Pacífico central y oriental, mientras que La Niña se refiere a una anomalía de la TSM inusualmente fría en el Pacífico tropical. Durante el fenómeno El Niño, la presión del nivel del mar tiende a ser más baja en el Pacífico oriental y más alta en el Pacífico occidental, mientras que durante La Niña ocurre lo contrario. La oscilación en la presión atmosférica entre el Pacífico tropical oriental y occidental se llama Oscilación del Sur, lo que podría influir en el transporte de humedad desde los océanos a los continentes. Por lo tanto, ENOS se genera principalmente a través del movimiento de la ZCIT, causando anomalías climáticas en todo el mundo [46, 7].

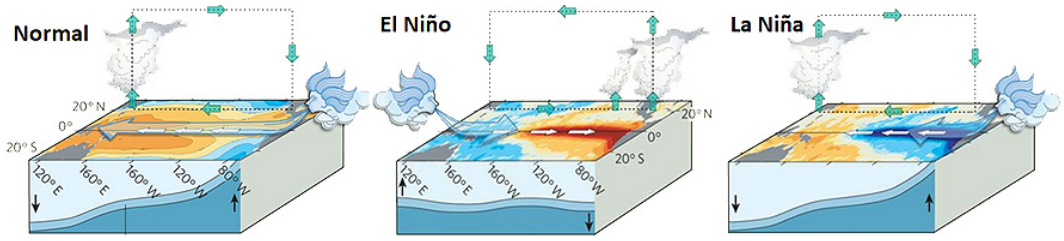


Figura 2.3: Condiciones ENOS. El Niño y La Niña. Adaptado de [2].

2.2.5.1. Fase cálida – El Niño

El fenómeno de El Niño (EN) se refiere a temperaturas oceánicas por encima del promedio a lo largo de las costas de Ecuador, Perú y el norte de Chile, así como a lo largo de la zona ecuatorial del Pacífico oriental, que persiste en promedio 5 meses, con una recurrencia de 2 a 7 años [14]. En condiciones normales, la superficie del océano en la zona oeste del Pacífico Tropical usualmente es caliente, la presión atmosférica es baja y las lluvias son frecuentes e intensas. Mientras que en el extremo opuesto del Pacífico (cerca de Suramérica), el agua es relativamente fría, la presión atmosférica es alta y hay poca lluvia. Cuando se presenta EN (Fig. 2.3), las aguas cálidas de Asia se dispersan hacia el Este (Suramérica). Las regiones de baja presión y las lluvias torrenciales también migran hacia el Este. En consecuencia, la zona central y oriental del Pacífico se calienta y se torna más lluviosa, mientras que en el extremo occidental del Pacífico las condiciones son más secas y frescas [46].

2.2.5.2. Fase fría – La Niña

El termino La Niña (LN) se usa para designar el evento extremo opuesto a El Niño. Es decir, el enfriamiento más allá de lo normal, de la temperatura del océano tropical (Fig. 2.3). Las características de La Niña son más frías que las condiciones normales del Pacífico Oriental y más húmedas y cálidas que las condiciones normales en el Pacífico de la zona oeste. Durante La Niña, las lluvias y tormentas disminuyen en el Pacífico ecuatorial central, y se concentran en Indonesia y el Pacífico occidental, produciéndose inviernos menos fríos en Canadá y desastres por inundaciones en Suramérica y Asia. En la Fig. 2.3, los colores más rojos denotan aguas cálidas, mientras que los azules oscuros muestran aguas más frías. En un evento El Niño (EN), temperaturas más cálidas en la superficie del mar en el este del océano que alteran la termoclina (línea imaginaria que separa 2 masas de agua de temperatura diferente), aumentan la humedad en la circulación atmosférica y el tiempo seco en el occidente [46].

2.2.6. Efectos socioeconómicos de los eventos ENOS cálidos en Colombia

No hay duda de que los fenómenos EN y LN, a través de su efecto climático, impactan en diverso grado el sistema socioeconómico en diferentes regiones del territorio colombiano. Los eventos de sequía causados por EN han generado efectos perjudiciales a lo largo de los años en muchos sectores económicos. Estos efectos se resumen de la siguiente manera [7]:

- Los eventos extremos de EN desencadenan en la región problemas en la disponibilidad de agua para diferentes propósitos (abastecimiento a la población, agricultura), lo cual afecta a la población por limitaciones en el suministro de agua y de productos agrícolas. Esto último ha llevado al incremento temporal de precios al consumidor.
- Bajo el efecto climático del fenómeno EN, se incrementa la frecuencia de los incendios de la cobertura vegetal, lo que afecta de manera especial los ecosistemas de la región.
- Los eventos extremos de EN puede elevar la cantidad de productos importados y disminuyen los exportados (petróleo, café, carbón), causando depreciación del PIB, asociado a mayores tasas de desempleo urbano en las principales áreas metropolitanas.
- Indirectamente se inducen cambios en las condiciones de las aguas costeras del Pacífico, que tienen efecto apreciable en el ambiente marino, repercutiendo en la distribución de especies planctónicas y bentónicas como el camarón, los peces y aquellas de rutas migratorias largas como las tortugas marinas y las ballenas.
- Los corales sufren blanqueamiento debido a las altas temperaturas. La sucesión de especies también es observada en períodos largos de sequías, donde migran poblaciones de aves, peces y artrópodos, algunas de las cuales no soportan los choques térmicos, reduciéndose drásticamente sus poblaciones afectando actividades económicas asociadas.
- El crecimiento económico y el ingreso generado durante eventos ENOS son menores que en épocas previas a estos, afectando la producción petrolera, la construcción, la producción agrícola y pesquera.

A pesar de conocerse los efectos socioeconómicos que los eventos ENOS pueden generar, existen vacíos de información económica sectorial, de monitoreo y seguimiento de las condiciones climáticas en varias regiones del país, haciendo compleja la cuantificación de los costos de prevención y/o recuperación de daños debido a que implica la trazabilidad y medición de la productividad de cada sector y de los procesos administrativos locales y regionales que gestionan la disponibilidad y finalidad de los recursos económicos.

2.2.7. Sequía

Entre los impactos de la VC y CC que más afectan al sector agrícola del país, esta la sequía [42]. Este concepto se relaciona con la reducción de la precipitación por debajo de lo normal o previsto, es decir, respecto a una media estadística o un promedio a largo plazo con las subsiguientes dificultades, asociado principalmente a la reducción de la oferta hídrica. Además, puede tener un impacto significativo en la generación de energía eléctrica, la sociedad, y los ecosistemas, entre otros [47].

La sequía se diferencia de otros desastres naturales principalmente por su extensión y sus efectos se hacen notar de manera lenta y sigilosa, de manera que el fenómeno se presenta durante semanas o meses de manera latente. Esta evolución de la sequía diferencia los tipos: (i) sequía meteorológica que es una desviación estadística de la precipitación normal; (ii) sequía agrícola que se asocia a los bajos niveles de humedad en el suelo, sustento para el crecimiento de las plantas; posteriormente aparece (iii) la sequía hidrológica, asociada a la disminución del caudal de los ríos, los niveles de los lagos y abatimiento del nivel freático; finalmente se manifiesta (iv) la sequía socioeconómica, donde afecta directamente a las sociedades humanas, aparece la hambruna y la falta de agua para potabilizar o distribuir a las personas [47].

2.2.7.1. Tipos de sequía

Sequía meteorológica. Escasez de precipitación. Frecuentemente agravada por altas temperaturas que inducen altas tasas de evapotranspiración. Este tipo de sequía está basada en datos climáticos y se expresa mediante la desviación de la precipitación respecto a la media durante un periodo de tiempo determinado [47].

Sequía hidrológica. Hace referencia a una deficiencia en el caudal o volumen de aguas superficiales o subterráneas (ríos, embalses, lagos, etc.). La frecuencia y severidad de una sequía hidrológica es usualmente definida sobre la base de su influencia en cuencas hidrográficas [47].

Sequía agrícola. Se produce cuando no hay suficiente humedad en el suelo para permitir el desarrollo de un determinado cultivo en cualquiera de sus fases de crecimiento. La demanda hídrica de una planta depende de condiciones meteorológicas, características biológicas específicas de la planta, su etapa de crecimiento, y de propiedades físicas y biológicas del suelo [47].

Sequía socioeconómica. Ocurre cuando la disponibilidad de agua disminuye hasta el punto de producir daños (económicos o sociales) a la población de la zona afectada por la

escasez de lluvias o cuando la demanda de agua excede la disponibilidad afectando de esta forma a una comunidad. La sequía socioeconómica asocia el suministro y demanda de algún bien económico o servicio con la sequía meteorológica, hidrológica y agrícola [47].

2.2.7.2. Índices de sequía

Se han obtenido varios índices de sequía en las últimas décadas. Por lo general, un índice de sequía es una variable principal para evaluar el efecto de una sequía y definir diferentes parámetros de sequía, que incluyen intensidad, duración, gravedad y extensión espacial. Una variable de sequía debería poder cuantificar la sequía para diferentes escalas de tiempo. La escala de tiempo comúnmente utilizada para el análisis de la sequía es un año, seguido de un mes. Aunque la escala de tiempo anual es larga, también se puede utilizar para abstraer información sobre el comportamiento regional de las sequías. La escala de tiempo mensual parece ser más apropiada para monitorear los efectos de una sequía en situaciones relacionadas con la agricultura, el suministro de agua y las extracciones de agua subterránea [4]. Una serie temporal de índices de sequía proporciona un marco para evaluar los parámetros de sequía de interés. Se han desarrollado varios índices diferentes para cuantificar una sequía, cada uno con sus propias fortalezas y debilidades, como:

- Índice de anomalías pluviales (RAI), Van Rooy año 1965 [48].
- Índice de severidad de la sequía de Palmer (PDSI), Palmer año 1965 [49].
- Los deciles, Gibbs y Maher año 1967 [50].
- Índice de humedad de los cultivos (CMI), Palmer año 1968 [51].
- Índice de sequía de Bhalme y Mooly (BMDI), Bhalme y Mooley año 1980 [52].
- Índice de suministro de agua superficial (SWSI), Shafer y Dezman año 1982 [53].
- Índice de precipitación nacional (NRI), Gommes y Petrassi año 1994 [48].
- Índice de precipitación estandarizado (SPI), McKee, Doesken y Kleist año 1993 [25].

Sobre la base de los estudios para los índices de sequía, prácticamente todos los índices de sequía utilizan la precipitación individualmente o en combinación con otros elementos meteorológicos, como insumo principal para el calculo del índice. El índice SPI es el más usado a nivel mundial [4, 42].

2.2.8. Índice Estandarizado de Precipitación (SPI)

El índice estandarizado de precipitación (SPI), es un índice ideado para cuantificar el déficit de precipitación para varias escalas temporales, las cuales reflejan el impacto de la sequía en la disponibilidad de los diferentes recursos hídricos. El SPI es calculado a partir del registro de datos de precipitación (idealmente se debe disponer de período continuo entre 20 y 30 años de valores mensuales) [42]. Dicho registro puede ser agrupado para distintas escalas temporales, usualmente en 3, 6, 12, 24 y 48 meses; las cuales permiten analizar los distintos tipos de sequía, pues, las condiciones meteorológicas y de humedad del suelo responden a anomalías de precipitación en escalas temporales relativamente cortas (entre 1 y 6 meses), mientras que los caudales fluviales, el almacenamiento en reservorios y las aguas subterráneas

responden a anomalías de precipitación a más largo plazo (6 meses y hasta 24 meses o más) [42].

El SPI fue desarrollado por McKee, Doesken y Kleist en el año de 1993 [25]. "Un índice potente y flexible, sencillo de calcular, el único parámetro necesario para su cálculo es la precipitación" [25] este índice es usado para determinar si en una región y en un periodo determinado hay déficit o exceso de precipitación respecto a las condiciones normales. El cálculo del SPI asume que los datos de precipitación se ajustan a una distribución gamma y fue diseñado para ser un indicador de la disponibilidad y uso del agua en el tiempo por medio de un proceso simple de medias móviles [25].

La distribución gamma tiene algunas dificultades en las zonas de muy poca precipitación, debido a que no se encuentra definida para valores de la variable iguales a 0, una estrategia para superar esta limitante es asignar lluvias con magnitud de 1mm, para que no sea indeterminada matemáticamente la función y continúe representando la sequía. La función de densidad es luego transformada a una distribución normal estandarizada (con media igual a 0 y varianza igual a 1), siendo el SPI el valor resultante de esta transformación [54].

La siguiente tabla presenta la clasificación del SPI propuesta por McKee [25]:

Tabla 2.2: Clasificación del índice SPI. Adaptado de [1].

Valor SPI	**Categoría**
Mayor a 2.0	Humedad extrema
1.5 a 1.99	Humedad severa
1.0 a 1.49	Humedad moderada
0.99 a -0.99	Normal
-1.0 a -1.49	Sequía moderada
-1.5 a -1.99	Sequía severa
Menor a -2.0	Sequía extrema

2.2.9. Pronóstico de series temporales

A diferencia de los problemas más sencillos de clasificación y regresión, los problemas de series temporales añaden la complejidad del orden o la dependencia temporal entre las observaciones. Es de mayor dificultad, porque se requiere un manejo especializado de los datos cuando se necesitan ajustar y evaluar los modelos. Esta estructura temporal también puede ayudar en la modelización, ya que proporciona una estructura adicional, como tendencias y estacionalidad, que puede aprovecharse para mejorar la habilidad del modelo. Tradicionalmente, la previsión de las series temporales ha estado dominada por métodos lineales como el ARIMA, porque son bien conocidos y eficaces en muchos problemas [43]. Pero estos métodos clásicos tienen algunas limitaciones [43], como:

- Se centran en datos completos: los datos que faltan o están corruptos no suelen ser admitidos.

- Se centran en relaciones lineales: asumir una relación lineal excluye distribuciones conjuntas más complejas.
- Se enfocan en la dependencia temporal fija: se debe diagnosticar y especificar la relación entre las observaciones en diferentes momentos y el número de observaciones retardadas que se proporcionan como entrada.
- Se enfocan en datos univariantes: muchos problemas del mundo real tienen múltiples variables de entrada.
- Se centran en previsiones de un solo paso: muchos problemas del mundo real requieren previsiones con un horizonte temporal largo.

Los métodos de aprendizaje automático pueden ser eficaces en problemas de previsión de series temporales más complejos con múltiples variables de entrada, relaciones no lineales complejas y datos ausentes. Para el desarrollo e implementación de estos métodos se requiere características diseñadas manualmente y preparadas por expertos en la materia o por profesionales con experiencia en el procesamiento de señales o datos [43].

2.2.10. Análisis de Componentes Principales (ACP)

El Análisis de Componentes Principales (ACP), es una técnica multivariada que permite describir un conjunto de datos en términos de nuevas variables (componentes). Técnicamente, el ACP busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. Esta técnica convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas componentes principales (CP). Los componentes se ordenan por la cantidad de varianza original que describen, esta técnica es muy usada para reducir la dimensionalidad de un conjunto de datos [10].

2.2.11. Análisis de Componentes Principales No Lineal (ACPNL)

El Análisis de Componentes Principales No Lineal (ACPNL), se basa en una red neuronal *feedforward* con una topología autoasociativa, también conocida como *deep-autoencoder*. ACPNL es una generalización no lineal del método clásico de PCA que busca superar las limitaciones de PCA cuando se involucran procesos no lineales [55]. Esta topología realiza un mapeo de identidad, donde el proceso de entrenamiento busca que las entradas (X) y salidas ($\hat{X}$) de la red sean las mismas. La Fig. 2.4 muestra la arquitectura de ACPNL, donde se observan las etapas de codificación y decodificación, la capa latente (z_1) indica el número de componentes principales no lineales (CPNLs) de este tipo de arquitectura.

Este método reduce con éxito la dimensionalidad y crea un mapa espacial de características similar a la distribución real de los parámetros del sistema subyacente [24]. En otras palabras, el ACPNL ayuda a establecer los modos dominantes de variabilidad en los datos [24]. Los detalles sobre el modelo están disponibles en [3] y [56].

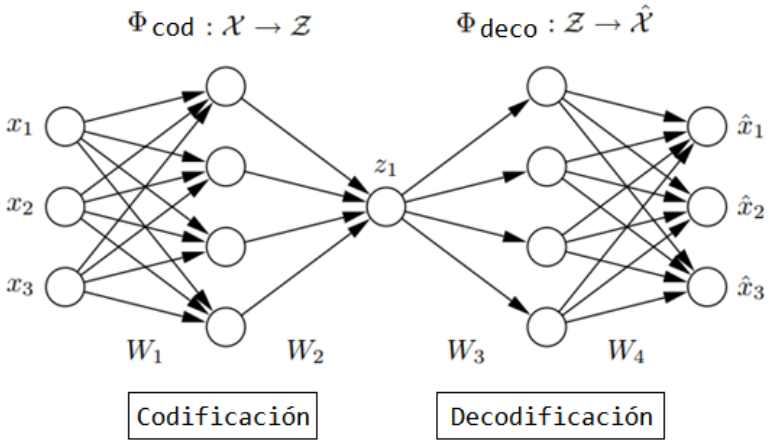


Figura 2.4: Topología de red neuronal de ACPNL. Adaptado de [3].

2.2.12. ACPNL inverso

ACPNL inverso es el segundo paso del enfoque completo del ACPNL; se utiliza la etapa de decodificación $\Phi_{deco} : Z \to \hat{X}$ (ver Fig. 2.4). Esta etapa se construye usando una red *feedforward*. La ecuación 2.1 muestra que la salida $\hat{X}$ depende de la entrada Z y de los pesos sinápticos $W \in W_3, W_4$.

$$\hat{X} = \Phi_{deco}(W, Z) = W_4 g(W_3 Z) \tag{2.1}$$

Donde, Φ_{deco} reconstruye el conjunto de datos $\hat{X}$, que debe ser cercano al conjunto de datos objetivo X minimizando el error cuadrado $\|x - \hat{x}\|^2$. Esta etapa de decodificación se detalla en [3] y [56].

2.2.13. Construcción de un modelo usando redes neuronales artificiales

La redes neuronales artificiales (RNA) es un modelo matemático basado en datos que emula una red neuronal del cerebro humano, que se ha utilizado para resolver problemas como la previsión y la clasificación [57]. Existen diferentes arquitecturas de RNA; sin embargo, el modelo más común es una red neuronal Perceptrón Multicapa (MLP), que tiene una estructura con una capa de entrada, capas ocultas y una capa de salida. El MLP se ha utilizado ampliamente para predecir varios fenómenos en meteorología e hidroclimatología [33, 41, 24]. La expresión matemática típica para una RNA es:

$$y_k = f_o \left[\sum_{j=1}^{n} W_{kj} f_h \left(\sum_{j=1}^{m} W_{ji} x_i + W_{jb} \right) + W_{kb} \right] \tag{2.2}$$

Donde, y_k es la salida, para el tiempo t; x_i es la i-ésima entrada; w_{ji} son los pesos sinápticos que conectan la capa de entrada y la capa oculta; w_{kj} son los pesos sinápticos que conectan las neuronas de las capas oculta y de salida; f_0 y f_h son las funciones de activación en las capas de salida y oculta, respectivamente; n y m son el número de capas de salida y capas

ocultas, respectivamente. w_{jb} y w_{kb} son el *bias* (sesgo) para las capas oculta y de salida [57].

Para el entrenamiento de las redes, se utiliza un algoritmo de aprendizaje que itera para encontrar la mejor combinación de pesos sinápticos con el menor error posible [57]. La actualización de los pesos sinápticos se realiza mediante el algoritmo de retropropagación (*Backpropagation*), que propaga el error entre la salida real y la calculada a través de la red. La expresión matemática se describe en la ecuación 2.3:

$$E = \sum_{p=1}^{P} Ep = \sum_{p=1}^{P} \sum_{k=1}^{n} (y_{pk} - \hat{y}_{pk})^2 \tag{2.3}$$

Donde, E es el error total, P es el número de datos de entrada, y E_p es el error de la diferencia al cuadrado entre las salidas reales y_{pk} y las salidas previstas $\widehat{y_{pk}}$ para el patrón p [58].

2.2.14. Modelo neuronal NNARMAX

Los modelos clásicos de series temporales, bajo la bandera de los modelos ARIMA (Box-Jenkins) [59], basan su predicción en la hipótesis de existencia de una relación de tipo lineal entre los regresores o entradas del modelo (valores pasados de la salida del modelo, distintos retardos de variables exógenas) y la variable tomada como salida. Esta hipótesis de linealidad limita en muchos casos la precisión de la estimación, debido a la existencia de relaciones no lineales en el proceso subyacente que se quiere modelar.

2.2.14.1. Procesos autoregresivos de media móvil (ARMA)

Es frecuente en la predicción de series temporales con modelos lineales que el análisis de correlaciones residuales muestre un alto contenido de información en la serie del error de predicción. Esta información, enmascarada bajo forma de ruido, puede ser realimentada al estimador de tal forma que su utilización mejore considerablemente la calidad de la estimación.

Esta presencia de información en los residuos puede ser debida a una elección indebida de la estructura del modelo o a una estimación incorrecta de los parámetros del mismo. La estructura del modelo puede ser a su vez inadecuada en dos sentidos:

- Cuando habiendo escogido un conjunto correcto de entradas, el modelo no tiene la capacidad de aproximación suficiente para realizar la transformación de entrada/salida requerida.
- Cuando el conjunto de entradas del modelo es incompleto. Si una variable que tiene una influencia relevante en la salida no es suministrada como entrada al modelo, la calidad de la estimación se verá gravemente afectada.

En muchos procesos del mundo real, la unión de diversos fenómenos puede tener una influencia aleatoria impredecible en el valor instantáneo de la salida del proceso, pero predecible en valores futuros. Se puede formular matemáticamente esta afirmación:

$$y(k) = f\left(x(k), ..., x(k-r), \varepsilon(k-l), ..., \varepsilon(k-s) + \varepsilon(k)\right) \tag{2.4}$$

Donde, $y(k) \in \mathbb{R}^m$ es la salida del proceso en el instante k, $x(k) \in \mathbb{R}^n$ es la entrada externa del proceso en el instante k. Si el proceso tiene propiedades autorregresivas, algunas de las componentes de x podrían corresponder a valores retrasados de las salidas, por ejemplo, $x1(k) = y1(k-1)$. El proceso de ruido blanco $\varepsilon(k) \in \mathbb{R}^m$ (la contribución de $\varepsilon(k)$ a $y(k)$ es impredecible).

El modelado de este tipo de procesos requiere la realimentación de las señales de error para poder estimar valores pasados del ruido. En el caso lineal se utilizan modelos autorregresivos de media móvil (ARMA) para tratar series temporales estacionarias y modelos autorregresivos integrados de media móvil (ARIMA) para el caso no estacionario [59]. La forma general de un proceso ARMA monovariable está definida por:

$$y(k) = \sum_{i=1}^{r} \alpha_i y(k-i) + \sum_{j=1}^{r} \beta_j \varepsilon(k-j) + \varepsilon(k) \tag{2.5}$$

Donde, $\varepsilon(k)$ es un proceso de ruido blanco estacionario. Los procesos autoregresivos (AR) son un caso particular del proceso ARMA con $\beta_1 = \beta_2 = ... = \beta_s = 0$. Los procesos de media móvil (MA) son un caso particular del proceso ARMA con $\alpha_1 = \alpha_2 = ... = \alpha_r = 0$.

2.2.14.2. Estructura del modelo NNARMAX

La extensión de los modelos lineales enunciados anteriormente al campo no lineal puede ser llevada a cabo mediante la aproximación de la función f (ver Ecu. 2.4) por un aproximador funcional no lineal fW, de parámetros W (teniendo en cuenta la propiedad de las RNA como aproximador universal de funciones [57]). La estructura resultante, se denomina NNARMAX (ver Fig. 2.5). El modelo NNARMAX produce la estimación vectorial de la salida $\hat{y}(k) \in \mathbb{R}^m$ en el instante k, tomando como entradas la serie temporal multivariable $u(k) \in \mathbb{R}^n$ y valores pasados de la salida deseada $y(k) \in \mathbb{R}^n$.

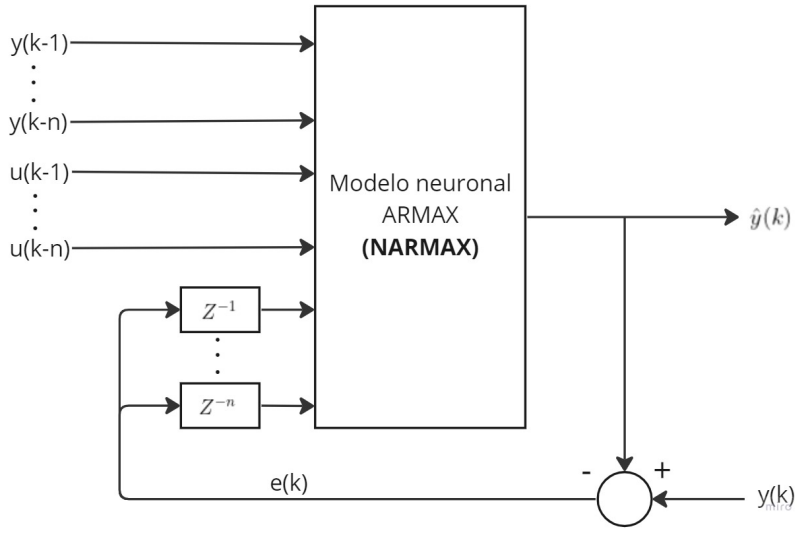


Figura 2.5: Modelo NNARMAX.

La función de transferencia del modelo NNARMAX viene dada por:

$$\begin{aligned} y(k+1) = fW[\, & y(k), y(k-1),\ \dots, y(k-n_y), \\ & u_1(k), u_1(k-1),\ \dots, u_1(k-n_{u1}),\ \dots, \\ & u_2(k), u_2(k-1),\ \dots, u_2(k-n_{u2}),\ \dots, \\ & u_r(k), u_r(k-1),\ \dots, u_r(k-n_{ur}), \\ & e_k(k), e(k-1), ..., e(k-n_e)\ ] \end{aligned} \tag{2.6}$$

Donde, $y(k+1)$ es la salida; n_y son los rezagos de la salida; $u_x(k)$ son las variables de entrada (términos exógenos) con $x \in [1, r]$; n_{ux} son los rezagos de los términos exógenos; k es el tiempo de muestreo; $e(k)$ es una secuencia de ruido; $n_u \geq 0$ es el máximo de rezagos; y fW es una función no lineal. $e(k) = y(k) - \hat{y}(k)$, donde $\hat{y}(k)$ es el valor pronosticado en el instante de tiempo k, y n_e son los rezagos de las respectivas diferencias de $e(k)$.

El modelo NNARMAX puede ser implementado usando un perceptrón multicapa (MLP) y regularización bayesiana. Las funciones de activación pueden ser tangente hiperbólica y lineal para las capas ocultas y de salida, respectivamente. En la fase de entrenamiento de la red, se puede entrenar en bucle abierto, mientras que la fase de validación se realiza en bucle cerrado, asumiendo una media móvil igual a cero y desviación estándar igual a uno.

Capítulo 3

Fuente de datos y Metodología

En la solución propuesta se definió un diseño e implementación modular, esto con el fin de segmentar las diferentes etapas desarrolladas a lo largo de la investigación. En este capítulo se desarrolla el bloque de Fuente de datos y Metodología (Ver Fig. 1.1). Inicialmente se define la zona de estudio, posteriormente se presenta el bloque de fuente de datos, en donde se realiza un análisis exploratorio con los registros de lluvias de estaciones *in situ*. Luego, se propone usar datos satelitales como alternativa para superar la dificultad de tener información no homogénea de las estaciones *in situ*, para ello se realiza previamente la validación de la base de datos satelital seleccionada. Con los datos satelitales como punto de partida se continua al bloque de Metodología, en este bloque se abordan los diferentes análisis y técnicas utilizadas para determinar las mejores variables predictoras y posteriormente usarlas en el modelo de pronóstico.

3.1. Zona de estudio

La zona de estudio corresponde al departamento de Nariño, se ubica en el suroeste de Colombia entre los $0°21'$ y $2°40'$ de latitud norte y los $76°\ 50'$ y $79°02'$ de longitud oeste [60]. Limita al norte con el departamento del Cauca, al sur con la República de Ecuador, al oriente con el departamento del Putumayo y al occidente con el Océano Pacífico (Fig. 3.1). Nariño tiene una extensión total de $33,268 km^2$ [61], equivalente al 2.9 % de la superficie del país [60].

El departamento es una zona de diversidad desde el punto de vista climático, de ecosistemas y de especies. Respecto al componente climático, Nariño cuenta con 15 tipos de climas según la clasificación Caldas-Lang, esta considera el factor térmico y factor de humedad [62]. Respecto a las especies, Nariño cuenta con el 10 % de los recursos de flora, el 11 % de las especies de anfibios, el 17 % de las especies de reptiles, el 56.20 % de las especies en aves y el 38 % de las especies de mamíferos respecto al total del país [61].

Topográficamente el departamento va desde los 0 metros sobre el nivel del mar (msnm) hasta los 4764 msnm correspondiente a la cima del volcán Cumbal (el volcán más alto del departamento y el octavo pico más alto de Colombia). Nariño se encuentra atravesado por la cordillera de los Andes de sur a norte, lo que lo convierte en un departamento bastante accidentado topográficamente como se observa en la Fig. 3.1. La cordillera es atravesada por

Figura 3.1: Ubicación geográfica del departamento de Nariño y modelo de elevación digital.

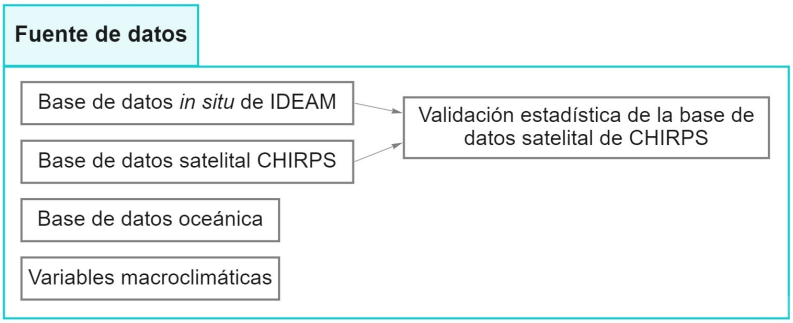
el río Patía que irrumpe en el piedemonte al norte del departamento, dándose paso a la llanura de Pacífico, una región plana de cara al Océano Pacífico con poca variación altitudinal. Por su parte, al sureste del departamento las altitudes son menores en la vertiente del Amazonas, donde el río Putumayo es el principal afluente al este de la cordillera de los Andes nariñenses.

3.2. Fuente de datos

En la La Fig. 3.2 se presenta los datos usados como insumo principal para los diferentes cálculos y análisis realizados a lo largo de está investigación. La información base para evaluar la sequía meteorológica son los datos de lluvia. Los registros usados inicialmente fueron datos de lluvia mensual registrada por las estaciones *in situ* dentro del límite político del departamento de Nariño y administradas por el Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia (IDEAM). De la localización espacial de las estaciones de medición se evidenció una distribución inequitativa de las estaciones meteorológicas y escasez de las mismas en la región Occidental del área de estudio, por lo cuál se buscó alternativas de información satelital que permitieran superar la falta de información y garantizar una distribución de datos uniforme. Para ello, se tomaron diferentes productos satelitales y se seleccionó el de mejor características en términos de cobertura, resolución temporal y

espacial. Posteriormente se validaron los datos satelitales respecto a los datos de lluvia *in situ*, para verificar si el producto satelital representa los datos *in situ*.

Figura 3.2: Solución propuesta - Fuente de datos.



3.2.1. Base de datos *in situ* de IDEAM

Los datos *in situ* para el departamento de Nariño son proporcionados por IDEAM, para el análisis se contaron con 47 estaciones meteorológicas. Sin embargo, se descartaron estaciones que no cumplieran con los siguientes criterios:

- Periodo de registro mayor o igual a 30 años.
- Datos faltantes de lluvia mensual superior al 20 %.
- Vigencia de operación actual.

Teniendo en cuenta los criterios mencionados anteriormente, el resultado final fue una red meteorológica de 46 estaciones con un periodo de registro de 38 años (1983-2021), estaciones localizadas de manera no uniforme, como se puede observar en la Fig. 3.3. A los datos de lluvia mensual de estas estaciones se les realizó un análisis preliminar y exploratorio de datos donde se priorizó conocer los valores de lluvia anual, la variabilidad de los datos y el porcentaje de datos faltantes (ver Tabla 3.1).

Tabla 3.1: Análisis descriptivo de las estaciones de precipitación mensual en Nariño (1983-2021).

Estación	**Sigla**	**Precipitación Anual (mm)**	**Desviación estándar (mm/mes)**	**Coeficiente de variación mensual**	**Datos faltantes mensual (%)**
Barbacoas	BAR	6740	260	0.46	3.6
Mira	MIR	2980	158	0.64	1.8
El Charco	CHA	3663	159	0.52	7.2
Mosquera	MOS	3660	203	0.67	4.1
Remolino	REM	2919	188	0.77	11.5
José Tapaje	JOS	4872	234	0.58	2.5
Salahonda	SAL	4959	253	0.61	3.2

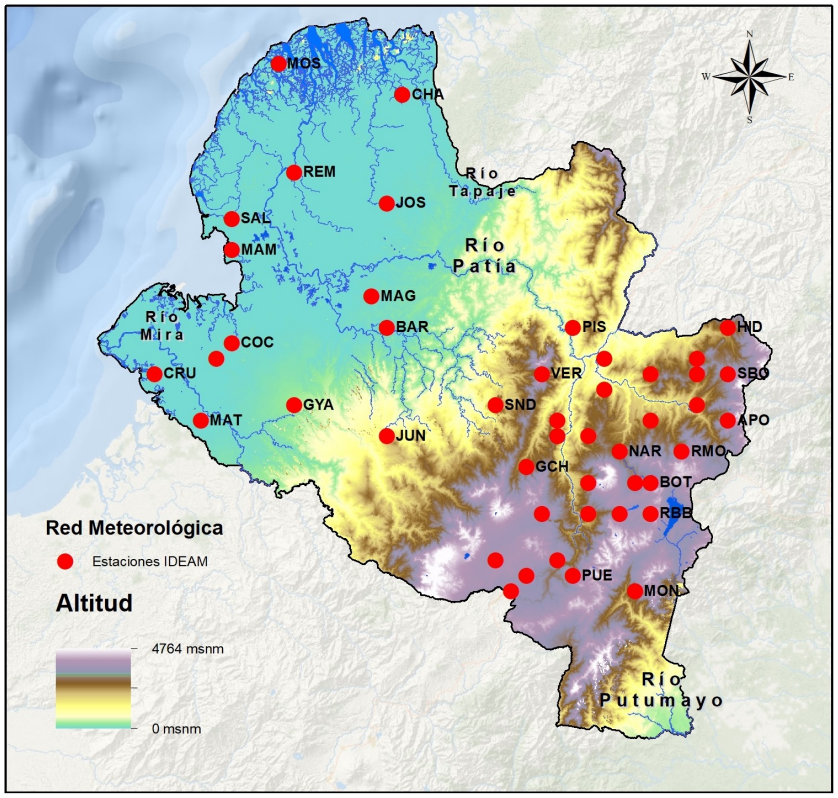
Continua en la siguiente página

Tabla 3.1 – Continua de la página anterior

Estación	Sigla	Precipitación Anual (mm)	Desviación estándar (mm/mes)	Coeficiente de variación mensual	Datos faltantes mensual (%)
Magüí	MAG	4847	242	0.60	5.9
Guayacana	GYA	6129	242	0.47	6.6
Coco	COC	2664	182	0.82	2.9
Mataje	MAT	3458	195	0.68	9.0
Berruecos	BER	1729	111	0.77	3.8
Bomboná	BOM	1043	62	0.71	1.4
Botana	BOT	929	40	0.52	3.6
Buesaco	BUE	1275	91	0.86	1.8
Chiles	CHI	1094	61	0.67	1.8
Cumbal	CUM	896	48	0.64	0.2
Paraiso	PAR	997	53	0.64	0.7
Peñol	PEÑ	1108	66	0.71	0.2
Guachavez	GCH	1654	93	0.67	0.9
Gualmatán	GMT	937	52	0.67	2.0
Hidromayo	HID	1346	90	0.80	1.6
Imués	IMU	1021	63	0.74	1.1
Junín	JUN	8697	272	0.38	2.7
La Unión	UNI	1980	115	0.70	1.1
Mamaconde	MAM	1336	95	0.86	1.8
Nariño	NAR	2007	127	0.76	1.4
Obonuco	OBO	812	51	0.75	7.2
Pisanda	PIS	1265	79	0.75	0.0
Puerres	PUE	1025	49	0.57	0.2
Rio Bobo	RBB	1094	53	0.59	0.7
Rosal Monte	RMO	1357	90	0.79	0.9
Samaniego	SAM	1440	96	0.80	0.2
San Bernardo	SBO	2023	118	0.70	1.6
Sandoná	SAN	1144	78	0.82	1.6
Taminango	TAM	1686	96	0.68	0.5
Tanamá	TAN	1351	80	0.71	0.5
Tangua	TGA	1001	61	0.73	0.5
La Cruz	CRU	1348	91	0.81	0.5
Monopamba	MON	3189	124	0.47	3.4
Apt. San Luis	ASL	878	44	0.60	0.2
Apt. A. Nariño	AAN	1193	73	0.73	0.7
Aponte	APO	1552	115	0.89	0.7
Vergel	VER	2549	148	0.70	3.4
Guasca	GCA	598	47	0.95	1.1
Sande	SND	4510	236	0.63	1.3

Del análisis exploratorio inicial de los datos, se destaca la variación en los valores extremos de la lluvia anual, con un valor mínimo de 598 mm al año en la estación Guasca hasta 8697 mm al año en la estación Junín; la desviación estándar y el coeficiente de variación de la lluvia mensual resalta la alta variabilidad de la lluvia a escala mensual. El porcentaje de datos

Figura 3.3: Localización de estaciones *in situ* seleccionadas.



faltantes de las estaciones seleccionadas oscila desde 0 % en Pisanda hasta un valor máximo de la estación Remolino con 11.5 %. En la Fig. 3.3 se observa la inequitativa distribución espacial de las estaciones *in situ* seleccionadas, donde el 76 % de las estaciones pluviométricas se encuentran en la región Andina del departamento, con una densidad de una estación cada 470 km^2, cubriendo el 48 % del área total del departamento, y el 24 % de las estaciones pluviométricas en la región del Pacífico con una densidad de una estación cada 1442 km^2. La región Pacífico es las más lluviosa del departamento y la que tiene menos estaciones pluviométricas *in situ*. Por está razón se recurrió a información satelital como una fuente alterna con mejor resolución espacial y equitativa distribución de los datos para la zona de estudio.

Se analizó la resolución espacial, temporal, dominio temporal y cobertura geográfica de las principales fuentes de información pluviométrica por teledetección definidas:

- Global Precipitation Climatology Centre (GPCC) [63].
- ERA5 [64].
- ERA-Interim [65].
- Tropical Rainfall Measuring Mission (TRMM) [66].
- Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks series (PERSIANN) [67].
- Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) [68].

El resumen de las características de las fuentes de información satelital se presenta en la Tabla 3.2.

Tabla 3.2: Principales características del conjunto de datos de precipitación satelital

Producto	**Resolución espacial**	**Resolución temporal**	**Dominio temporal**	**Cobertura**	**Ref**
GPCC	2.5°	Mensual	1951 - 2000	Global $60°N-S$	[63]
GPCP	0.5°	Diario/Mensual	1979 - 2004	$60°N-S$	[69]
PERSIANN	0.25°	6-horario/Diario	1983 - presente	Global $60°N-S$	[67]
TRMM	0.25°	3-horario/Diario	1998 - presente	Global $50°N-S$	[66]
ERA Interim	0.75°	6-horario	1950 - presente	Global $60°N-S$	[65]
ERA5	0.25°	horario	1979 - presente	Global $60°N-S$	[64]
CHIRPS	0.05°	Diario a Anual	1981 - presente	$50°N-S$	[68]

El principal criterio de selección para la base de datos satelitales fue la resolución espacial, de las bases de datos mostradas en la Tabla 3.2. CHIRPS presenta la resolución espacial más precisa de 0.05°. Esta resolución se puede interpretar como una "estación artificial" con distribución espacial uniforme que registra datos de lluvia mensual cada 5.3 km. Adicionalmente, la resolución temporal de CHIRPS abarca datos mensuales desde el año 1983 hasta la actualidad, lo cual concuerda con los datos *in situ* proporcionados por IDEAM. La cobertura espacial abarca todo el departamento de Nariño. Por estas razones y las características planteadas a continuación, se escogió CHIRPS como fuente de información pluviométrica alternativa. La validacion de CHIRPS se realizó en el articulo de coautoría *"A spatiotemporal assessment of the high-resolution CHIRPS rainfall dataset in southwestern Colombia using combined principal component analysis"* [70] y se complementa en la sección 3.2.4.

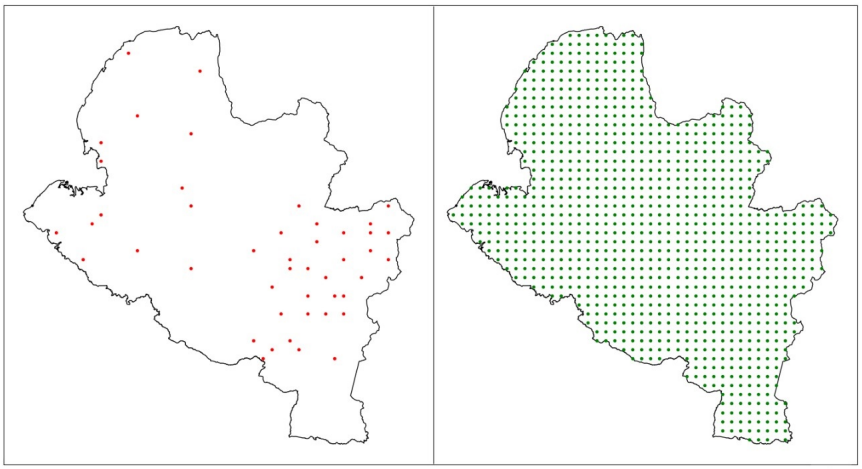
3.2.2. Base de datos satelital CHIRPS

El conjunto de datos Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) tiene registros de resolución temporal de las precipitaciones diarias, pentadales, decadales y mensuales con una resolución espacial de 0.05° ($\sim 5.3km$) y una cobertura casi global (dentro de los $50°S$, $50°N$, $180°O$ y $180°E$) desde 1983 hasta la actualidad. Estos datos son usados para estudiar diferentes problemáticas asociadas a la disponibilidad del recurso hídrico, por ejemplo la sequía [71]. Los datos de precipitación son estimados de manera remota por satélites administrados por el USGS (Servicio Geológico de Estados Unidos, por sus siglas en inglés) y la Universidad de California, Santa Bárbara. El departamento de Nariño esta cubierto por 1016 píxeles (estaciones) con una distribución espacial uniforme como se muestra en la Fig. 3.4.

3.2.3. Base de datos oceánica e índices macroclimáticos

Adicionalmente se usaron datos mensuales de TSM del conjunto de datos ERA5 [64]. El ERA5 proporciona estimaciones horarias de un gran número de variables climáticas atmosféricas, terrestres y oceánicas. Los datos cubren la Tierra en una cuadrícula de 30 km y analizan la atmósfera utilizando 137 niveles desde la superficie hasta una altura de

Figura 3.4: Distribución espacial de estaciones (a) IDEAM y (b) CHIRPS.



80 km. El ERA5 incluye información sobre las incertidumbres de todas las variables con resoluciones espaciales y temporales reducidas.

ERA5 tiene una resolución temporal mensual (recomendada) desde 1959 hasta el presente [72], estos datos se publican en un plazo de 3 meses en tiempo real. El ERA5 combina grandes cantidades de observaciones históricas en estimaciones globales utilizando sistemas avanzados de modelización y asimilación de datos.

Adicionalmente se utilizaron variables macroclimáticas proporcionadas por la Administración Nacional Oceánica y Atmosférica de estados unidos (NOAA[1], por sus siglas en inglés). El uso de las VM y su explicación se detalla en la sección 3.3.7.

3.2.4. Validación estadística de la base de datos satelital de CHIRPS

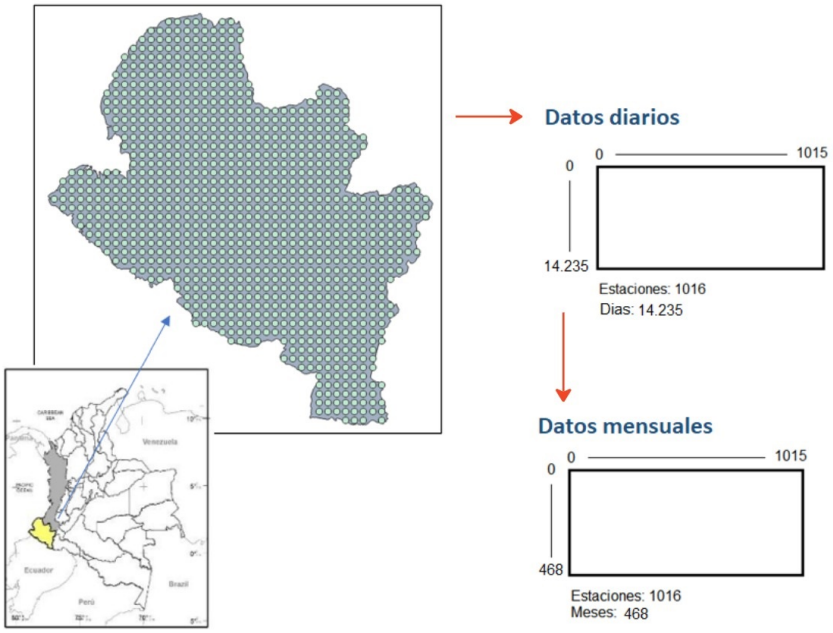
Inicialmente se extrajeron los datos de lluvias a partir de las imágenes satelitales con resolución temporal diaria. En total fueron 14.235 imágenes diarias procesadas, para obtener las series de tiempo de lluvias diarias en el periodo de estudio (1983-2021). Posteriormente se calcularon las series de tiempo mensuales para un total de 468 meses (Ver Fig. 3.5).

La validación de la información satelital consistió en una comparación entre la base de datos de lluvias de IDEAM y la base de datos de lluvias de CHIRPS. Para el análisis se construyó una matriz punto-píxel, en el que el punto representa el dato de lluvia estimado por la estación pluviométrica *in situ* o variable observada del IDEAM y el píxel representa la cuadrícula del conjunto de datos estimado de CHIRPS sobre la ubicación de la estación *in situ*. De los 1016 pixeles de CHIRPS que cubren la zona de estudio se escogieron 46 pixeles para construir la matriz de comparación punto-píxel. La selección de estos 46 pixeles coincide con el mismo espacio geográfico que ocupa cada estación de IDEAM, resaltando que ninguna celda de la cuadrícula contenía más de una estación.

El proceso de validacion se realizó aplicando métricas de desempeño para la comparación

[1] `https://www.esrl.noaa.gov/psd/data/climateindices/list/`

Figura 3.5: Extracción de datos de lluvia a partir de CHIRPS.



entre series de tiempo. Se estimaron cuatro métricas de desempeño, las cuales se describen a continuación, especificando las respectivas ecuaciones en la Tabla 3.3.

1. **Sesgo**, valores superiores a cero representa una sobre estimación de la información satelital, respecto a la información *in situ* [73].
2. **El Error Cuadrático Medio (ECM)** mide la diferencia media absoluta entre ambas bases de datos, donde cero es el valor ideal [74].
3. **Coeficiente de correlación de Pearson (CCP)** evalúa la independencia de la fuente de los datos a través de las relaciones existentes entre las bases de datos [75].
4. **Correlación Media Ponderada (CMP)** es una medida de similitud entre bases de datos basado en la mediana, en lugar de la media, por lo tanto, es menos sensible a valores atípicos [76].

Tabla 3.3: Métricas de desempeño para la comparación entre series de tiempo.

Nombre	**Ecuación**	**Unidades**	**Valor ideal**
Sesgo	$BIAS = \sum_{i=1}^{n} \frac{P_i - P_o}{N}$ P_i = lluvia de CHIRPS durante el periodo i. P_o = lluvia IDEAM durante el periodo i. n = número total de observaciones N.	mm	0
Error cuadrático medio (ECM)	$ECM = \sqrt{\sum_{i=1}^{n} \frac{(P_i - P_o)^2}{N}}$ P_i = lluvia de CHIRPS durante el periodo i. P_o = lluvia IDEAM durante el periodo i. n = número total de observaciones N.	mm	0

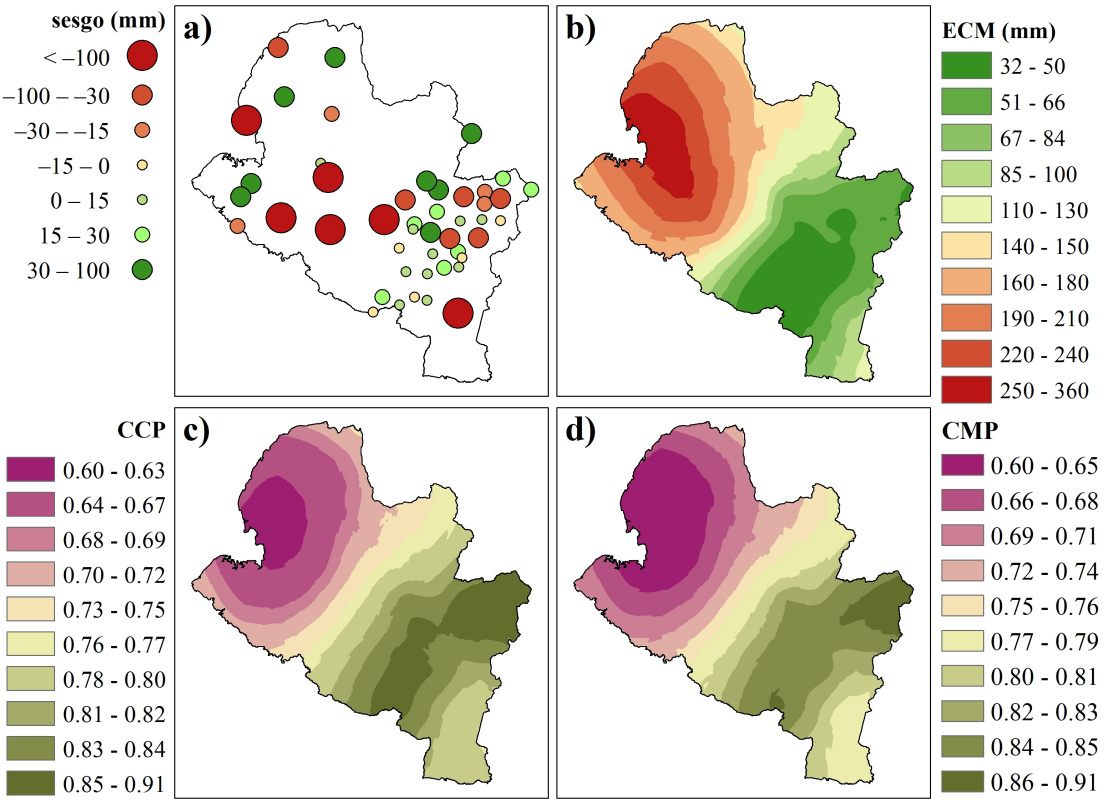
Continua en la siguiente página

Tabla 3.3 – Continua de la página anterior

Nombre	Ecuación	Unidades	Valor ideal
Coeficiente de correlación de Pearson (CCP)	$CCP = \frac{\frac{1}{n}\sum_{i=1}^{n}(P_i-\hat{P}_i)(P_o-\hat{P}_o)}{\sqrt{\sum_{i=1}^{n}(P_i-\hat{P}_i)^2}\sqrt{\sum_{i=1}^{n}(P_o-\hat{P}_o)^2}}$ P_i = lluvia de CHIRPS durante el periodo i. P_o = lluvia IDEAM durante el periodo i. n = número total de observaciones N.	-	1
Correlación media ponderada (CMP)	$CMP = \frac{\zeta xy}{\sqrt{\zeta xx\ \zeta yy}}$ ζxx = media ponderada de xx. ζyy = media ponderada de yy. ζxy = media ponderada de xx y yy.	-	1

La validación consiste en la comparación estadística entre una base de datos asumida como real o de menor incertidumbre (datos de lluvia mensual de IDEAM) con una base de datos estimada (datos de lluvia mensual de CHIRPS). Las dos fuentes de información están midiendo la misma variable (lluvia); por lo tanto, al demostrar estadísticamente la similitud entre ambas bases de datos es factible utilizar la información de CHIRPS para distintos análisis climáticos. Los resultados de la aplicación de las cuatro métricas de desempeño se resumen espacialmente en la Fig. 3.6.

Figura 3.6: Espacialización de las métricas de validación usadas entre CHIRPS e IDEAM. (a) Sesgo, (b) ECM, (c) CCP y (d) CMP.



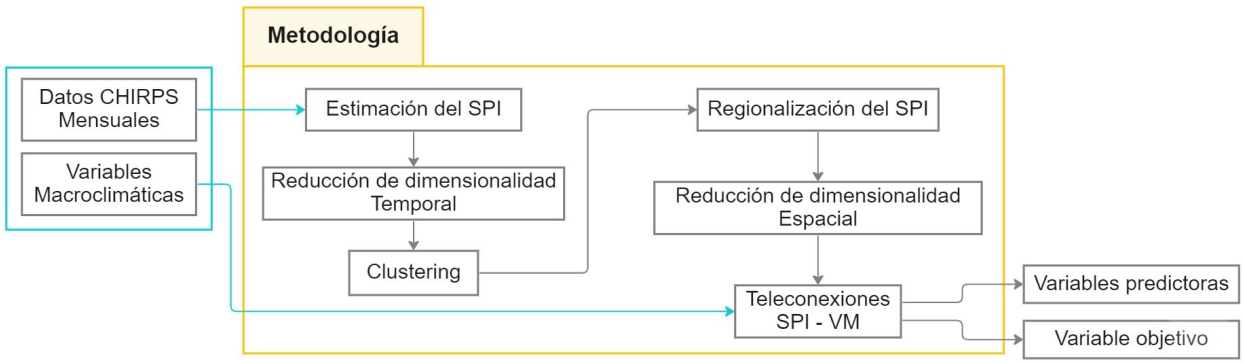
Las metricas de Sesgo y ECM son sensibles a la magnitud de la lluvia, por lo tanto, los errores mayores deben contrastarse con las isoyetas anuales de lluvia (Fig. 3.6). El Sesgo

muestra errores significativos de subestimación de -309 mm en la estación Junín con lluvia anual de 8700 mm/año, y Sesgo de sobreestimación de hasta 64 mm en regiones con lluvias de 600 mm/año a 2000 mm/año. Según el ECM se evidencian errores bajos ($32mm < ECM < 84mm$) en la región Andina; mientras que se tienen errores significativos ($131mm < ECM < 358mm$) en la región lluviosa del Pacífico; en el piedemonte o zona de transición, se producen errores intermedios ($85mm < ECM < 130mm$). Las metricas de CCP y CMP muestran resultados similares; se presentan altas correlaciones en la región Andina ($0.76 < CCP < 0.91$ y $0.77 < CMP < 0.91$), y una menor correlación en la región Pacífico ($0.60 < CCP < 0.75$, $0.62 < CMP < 0.76$), sin embargo, se considera una correlación aceptable entre las dos series analizadas. En términos generales los resultados obtenidos del proceso de validación, usando las diferentes métricas evidenciaron que la base de datos CHIRPS tiene una buena representación de los datos de lluvia *in situ* en la mayor parte de la zona de estudio, principalmente en las regiones con menor magnitud de lluvia, respecto a las regiones más lluviosas.

3.3. Metodología

Teniendo los datos mensuales de CHIRPS y las variables macroclimáticas como insumo principal y entrada al bloque de Metodología, se realizan una serie de etapas de procesamiento de los datos, como se puede observar en la Fig. 3.7. Inicialmente se realizó la estimación del SPI (en diferentes escalas temporales) con los datos de lluvia; luego con el SPI calculado, se redujo la dimensionalidad temporal, para emplear métodos de *clustering* y agrupar los datos con similar comportamiento. Con la agrupación de datos, se segmento espacialmente regiones SPI para hacer una reducción de dimensionalidad espacial, luego se utilizó la memoria climática almacenada en distintas variables oceánicas y atmosféricas (VM) para analizar las teleconexiones existentes con el objetivo de seleccionar las mejores VM que permiten explicar o relacionar el comportamiento del SPI en las regiones segmentadas. Las VM seleccionadas son los predictores del modelo de pronóstico.

Figura 3.7: Solución propuesta - Metodología.



3.3.1. Estimación del SPI

El índice de precipitación estandarizado (SPI), es un índice ideado para cuantificar el déficit de lluvia para varias escalas temporales, las cuales reflejan el impacto de la sequía en la disponibilidad de los diferentes recursos hídricos. Este índice ha sido recomendado para medir la intensidad, duración y extensión espacial de la sequía [77]. El SPI se calcula a partir de datos de precipitación mensuales con más de 30 años de registro [42] para agrupaciones temporales usualmente de 1, 3, 6, 12 meses; que permite analizar diferentes tipos de sequía debido las condiciones meteorológicas y de humedad del suelo que responden a anomalías de precipitación en escalas temporales relativamente cortas (entre 1 y 6 meses), mientras que los caudales fluviales, el almacenamiento en reservorios y las aguas subterráneas responden a anomalías de precipitación a más largo plazo (6 meses y hasta 24 meses o más) [42].

En esta investigación, se calculó el SPI para cuatro escalas temporales: 1 mes (SPI1), 3 meses (SPI3), 6 meses (SPI6) y 12 meses (SPI12). El SPI se convierte en la variable a analizar y la interpretación del índice puede detallarse en la Tabla 2.2. Los datos SPI fueron calculados usando una función de distribución de probabilidad Gamma, recomendada por [78].

Para el calculó del SPI se implementó el siguiente algoritmo:

```
Inicio
Inicializar variables
Iterar mes en (1, 3, 6 y 12 meses):
    Agrupar los datos de lluvias en el mes.
    Ajustar los parámetros de la función Gamma.
    Calcular la función de densidad de probabilidad Gamma.
    Calcular la probabilidad acumulada de precipitación.
    Calcular los parámetros de escala alpha y beta.
    Calcular el factor de corrección
    Calcular la probabilidad acumulada total.
    Obtener la función de densidad de probabilidad acumulativa.
    Convertir a una distribución normal estándar (SPI).
    Guardar el SPI procesado
Fin
```

En general los valores positivos del SPI muestran exceso de precipitación y valores negativos del SPI denotan déficit de precipitación. Por lo tanto, los siguientes análisis se centran en los valores negativos del SPI para el estudio de la sequía meteorológica.

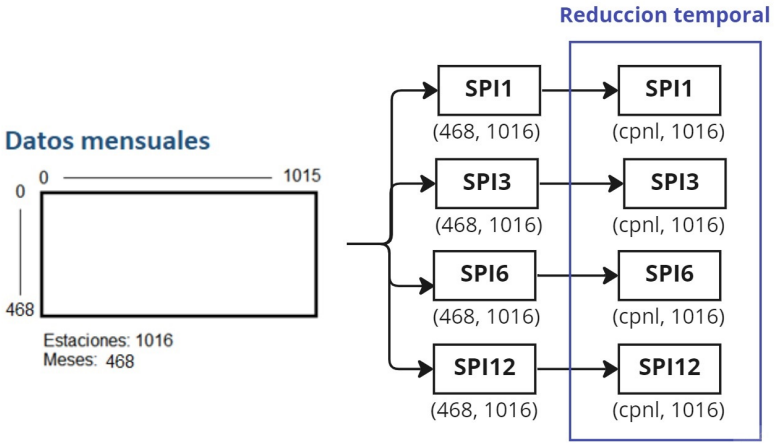
3.3.2. Reducción de dimensionalidad temporal

Con los datos SPI calculados en diferentes escalas temporales, se realizó la reducción de dimensionaliad temporal para realzar la distribución espacial de las estaciones, esta reducción se explica en la Fig. 3.8. En donde, cada dataset SPI contiene 468 filas (meses) y 1016 columnas (estaciones), aplicando la reducción de dimensionalidad se puede obtener un número menor de filas (CPNL) permitiendo utilizar métodos de *clustering* con mayor efectividad, para identificar zonas espaciales de similar comportamiento.

Para el cálculo de las CPNLs se utilizó el conjunto de funcionalidades (toolbox) de Matlab[2], que permite la implementación de ACPNL utilizando el enfoque de redes

[2] `http://www.nlpca.org/matlab.html`

Figura 3.8: Reducción de dimensionalidad temporal



neuronales. La forma de evaluar el desempeño del proceso de entrenamiento de este método es observando la varianza explicada de cada componente no lineal para diferentes topologías establecidas. La selección del número de CPNLs (capa latente), se basó en ACP, evaluando el número de valores propios superiores a uno según [79, 80]. Se realizó un conjunto de entrenamientos con datos del SPI para diferentes escalas de tiempo, con el objetivo de evaluar el desempeño y comparar con los resultados obtenidos a través del método lineal de análisis de componentes principales (ACP) y el método no lineal (ACPNL).

La Tabla 3.4 muestra los resultados del ACP, el primer componente principal (CP1) explica una varianza superior al 50 % para todos los casos, el segundo componente (CP2) explica una varianza superior al 10 % y la suma de la varianza explicada de las cinco primeras componentes principales es superior al 80 %.

Tabla 3.4: Cálculo de ACP para SPI en diferentes escalas temporales.

Datos	**Varianza explicada (%)**				
	CP1	**CP2**	**CP3**	**CP4**	**CP5**
SPI1	57.176	19.835	6.448	3.131	2.219
SPI3	54.044	21.334	7.567	3.419	2.577
SPI6	52.559	23.215	8.044	4.197	2.504
SPI12	53.717	22.724	8.601	4.958	2.136

Los entrenamientos realizados para el cálculo de los ACPNL fueron en total 120, donde se realizaron variaciones en la topología de la red, haciendo 10 entrenamientos por variación de topología para cada SPI (SPI1, SPI3, SPI6 y SPI12). La Tabla 3.5 consolida los mejores resultados obtenidos para variaciones de topologías en las diferentes escalas de SPI.

Tabla 3.5: Cálculo de ACPNL para SPI en diferentes escalas de temporales.

Datos	**Topología**	**Varianza explicada (%)**				
		CPNL1	**CPNL2**	**CPNL3**	**CPNL4**	**CPNL5**
SPI1	468-200-20-5-20-200-468	76.188	12.982	2.045	0.160	0.043
	468-150-15-5-15-150-468	66.216	10.984	2.658	0.892	0.068
	468-100-10-5-10-100-468	61.159	12.529	2.748	0.256	0.098
SPI3	468-200-20-5-20-200-468	74.476	13.485	1.218	0.645	0.341
	468-150-15-5-15-150-468	59.153	15.154	2.265	0.894	0.064
	468-100-10-5-10-100-468	60.253	11.784	2.189	0.659	0.084
SPI6	468-200-20-5-20-200-468	77.867	13.371	1.437	0.834	0.283
	468-150-15-5-15-150-468	70.892	10.453	2.658	0.822	0.198
	468-100-10-5-10-100-468	70.982	11.448	2.595	0.869	0.255
SPI12	468-200-20-5-20-200-468	78.128	19.000	1.484	0.543	0.351
	468-150-15-5-15-150-468	72.565	15.145	0.974	0.445	0.256
	468-100-10-5-10-100-468	70.496	14.156	0.785	0.156	0.257

La topología 468-200-20-5-20-200-468 fue la de mejores resultados puesto que las dos primeras componentes (CPNL1 y CPNL2) tiene una varianza explicadas para SPI1 de 89.17 %, SPI3 de 87.96 %, SPI6 de 91.23 % y SPI12 de 97.12 %. De los resultados se puede evidenciar que ACPNL tiene una varianza explicada más significativa que las obtenidas mediante el ACP. El primer componente lineal (CP1) en su mejor varianza explicada es de 57.176 %; la varianza explicada es menor que la varianza del CPNL1 para todos los casos. Además, la suma de la varianza explicada de los cinco primeras componentes principales para el método lineal y no lineal da como resultado una varianza explicada de 80 % y 88 %, respectivamente, resultado que confirma que ACPNL tiene una varianza explicada más significativa que ACP para los conjuntos de datos procesados.

3.3.3. Clustering en función del SPI

Con los datos reducidos temporalmente, se utilizaron cinco métodos de clustering (SOM, HC_{div}, HC_{aglo}, $Fuzzy$ y $Kmeans$) y se evaluó el número óptimo de agrupamiento $\widehat{K}$ haciendo variaciones del número de agrupaciones ($k = 2, 3, 4, 5$) con los diferentes métodos. Los métodos de *clustering* empleados se resumen en la Tabla 3.6.

Para determinar las regiones homogéneas a partir del SPI, se utilizaron los diferentes conjuntos de datos SPI reducidos temporalmente y se aplicaron los métodos de agrupamiento descritos en la Tabla 3.6. Las regiones obtenidas de la aplicación de estos métodos se evaluaron con métricas/índices de validación interna de *clustering*, para verificar las correlaciones intra-grupos e inter-grupos. Existen variedad de índices que permiten evaluar el desempeño de las regiones resultantes, pero los más usados según [81], y [82] son los mostrados en la Tabla 3.7.

Tabla 3.6: Descripción de los métodos de clustering utilizados.

Método	**Descripción**	**Referencia**
Mapas Auto-organizados (SOM)	Es un tipo de RNA entrenada mediante el aprendizaje no supervisado para producir una representación discretizada de baja dimensión (generalmente bidimensional) del espacio de entrada de las muestras de entrenamiento. Los mapas auto-organizados se diferencian de otras RNA en que aplican el aprendizaje competitivo en oposición al aprendizaje con corrección de errores (como la propagación hacia atrás con descenso de gradiente) y en el sentido de que utilizan una función de vecindad para preservar las propiedades topológicas del espacio de entrada. La principal idea de funcionamiento de SOM, es que cada dato del conjunto de datos se reconoce a sí mismo compitiendo por la representación en el arreglo bidimensional de neuronas.	[83]
Agrupamiento difuso ($Fuzzy$)	Este tipo de algoritmo se aparta de la forma exclusiva, que un dato pertenezca a un único clúster, para este caso se introduce un concepto llamado función de pertenencia, esta función permite conocer si un punto del conjunto de entrada, puede pertenecer a un conjunto o pertenecer a múltiples conjuntos. La salida de este algoritmo es un número que indica la probabilidad de que el punto de entrada pertenezca a un conjunto en particular.	[84]
Cluster Jerárquico aglomerativo (HC_{aglo})	En este método cada punto de datos se considera una región individual, en cada iteración lo grupos similares se fusionan con otros grupos, esto hasta que se forma un solo grupo. De esta forma, no se muestra un agrupamiento sino las relaciones de proximidad que existen entre los elementos.	[85]
Cluster Jerárquico Divisivo (HC_{div})	En este método se considera todos los puntos de datos como una única región y en cada iteración, separamos los puntos de datos de la región que no son similares. Cada punto de datos que se separa se considera como una región individual. Al final, se obtienen n regiones.	[85]
Agrupamiento por la media ($Kmeans$)	Es un algoritmo de regionalización no supervisado que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o región. Se suele usar la distancia euclidiana como forma de medir la similitud entre el centroide de un conjunto de datos y el dato mismo.	[86]

Tabla 3.7: Descripción de metricas/índices de evaluación de conglomerados.

Nombre	Descripción	Ecuación	Interpretación	Ref
Índice Silhouette	Es un índice de interpretación y validación de la coherencia dentro de grupos de datos. El valor de la Silueta mide el grado de similitud de un objeto a su propio grupo (cohesión) en comparación con otros grupos (separación).	$Silhouette = \frac{b-a}{max(a,b))}$ a: La distancia media entre un punto y todos los demás puntos del mismo clúster. b: La distancia media entre un punto y todos los demás puntos del siguiente clúster más cercano.	La mayor cohesión en la agrupación se representa con 1, y la menor cohesión con el valor de -1. Los valores cercanos a 0 indican agrupaciones superpuestas.	[82]
Índice Davies Bouldin	Se define como la medida de similitud promedio de cada región con su par más similar, donde la similitud es la relación entre las distancias intra-clúster y las distancias inter-clúster. Por lo tanto, las regiones que estén más separadas y menos dispersas darán como resultado una mejor puntuación.	$R_{ij} = \frac{s_i+s_j}{d_{ij}}$ $DB = \frac{1}{k}\sum_{i=1}^{k} max_{i\neq j}\, R_{ij}$ R_{ij}: Medida de similitud. s_i: Distancia promedio entre cada punto del clúster y el centroide del clúster. d_{ij}: Distancia entre los centroides del clúster i y j.	La puntuación mínima es 0, y los valores más bajos indican una mejor agrupación.	[82]
Índice Calinski Harabasz	Conocido como criterio de relación de varianza, es una medida de qué tan similar es un objeto a su propio grupo (cohesión) en comparación con otros grupos (separación). Aquí la cohesión se estima en función de las distancias desde los puntos de datos en un grupo a su centroide de grupo y la separación se basa en la distancia de los centroides de grupo desde el centroide global.	$CH = \frac{tr(B_k)}{tr(W_k)}\ x\ \frac{n_E-k}{k-1}$ $tr(B_k)$: Trazo de la matriz de dispersión entre clusters. $tr(W_k)$: Trazo de la matriz de dispersión entre clusters.	El valor óptimo es de mayor magnitud. La puntuación es más alta cuando las regiones son densas y están bien separadas.	[82]

Continua en la siguiente página

Tabla 3.7 – Continua de la página anterior

Nombre	Descripción	Ecuación	Interpretación	Ref
Índice CS	Método que determina la distancia euclidiana entre las regiones y relacionándolo con el número de regiones identificados. Funciona bien para regiones con diferentes densidades y/o tamaños de regiones.	$CS = \frac{\sum_{i=1}^{c}\left\{\frac{1}{\lvert A_i\rvert}\sum max\{d(x_i,x_k)\}\right\}}{\sum_{i=1}^{c}\{min\{d(v_i,v_j)\}\}}$ c: Número de clusters identificados. d): Función distancia, usualmente se usa distancia euclidiana.	El valor óptimo es de menor magnitud, el cual indica una buena selección de agrupamientos.	[82]

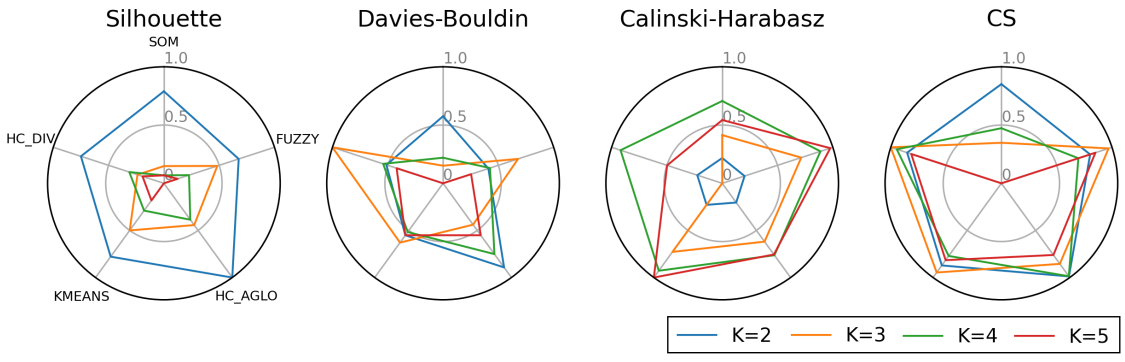
El resultado de los métricas de validación de *clustering* determina el número óptimo de agrupamiento $\widehat{K}$ así como también permite definir cual de los métodos de *clustering* empleados registra el mejor desempeño para el conjunto de datos evaluados.

Cada método de *clustering* (5 métodos) fue usado en 4 posibles grupos (k): 2, 3, 4 y 5 grupos; por lo tanto, se obtuvieron 20 posibles agrupaciones. Estas posibles agrupaciones se evaluaron mediante las 4 métricas de desempeño descritas en la Tabla 3.7: índice de Silhouette, Davies-Bouldin, Calinski-Harabasz y CS.

La interpretación de los resultados de cada índice difiere por la magnitud, por lo tanto, se homogeneizaron los resultados al normalizarlos para facilitar su interpretación. Resultado de la normalización, el valor de 1 es considerado el óptimo para el método de *clustering* en los cuatro índices utilizados, el resultado gráfico se muestra en diagramas de radar, donde cada vértice representa un método de *clustering* utilizado y la línea representa la magnitud de cada métrica. Se muestra el gráfico de radar del número óptimo de regiones para el SPI1 en la Fig. 3.9, SPI3 en la Fig. 3.10, SPI6 en la Fig. 3.11 y SPI12 en la Fig. 3.12.

El número óptimo de agrupamiento para el SPI1 no es concluyente, como se observa en la Fig. 3.9, cada índice muestra un número óptimo de agrupamiento distinto. Según el índice Silhouette $\widehat{K} = 2$, según el índice de Davies-Bouldin $\widehat{K} = 3$, el índice Calinski-Harabasz $\widehat{K} = 5$ y el índice CS índica que $\widehat{K} = 3$.

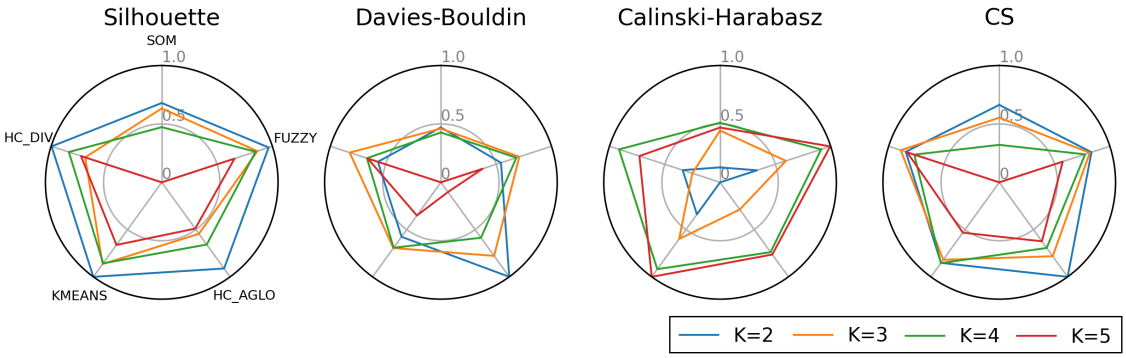
Figura 3.9: Métricas de desempeño para clustering, SPI1



En el caso de SPI3 el número óptimo de agrupamiento no es concluyente, en la Fig. 3.10 cada índice muestra un número óptimo de agrupamiento distinto. El índice Silhouette $\widehat{K} = 2$,

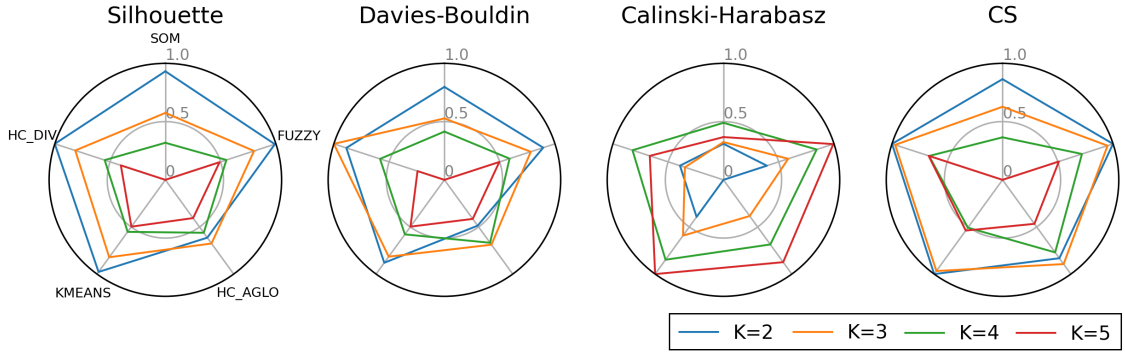
según el índice de Davies-Bouldin $\widehat{K} = 3$, el índice Calinski-Harabasz $\widehat{K} = 5$ y el índice CS índica que $\widehat{K} = 2$.

Figura 3.10: Métricas de desempeño para clustering, SPI3



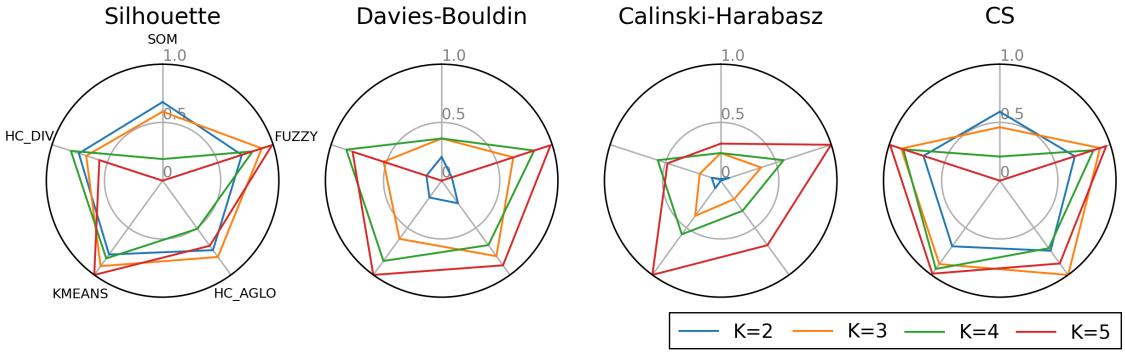
Para el SPI6, los índices de Silhouette, Davis-Bouldin y CS indican que el agrupamiento óptimo es 2 (ver Fig. 3.11 donde $\widehat{K} = 2$).

Figura 3.11: Métricas de desempeño para clustering, SPI6



Para el SPI12, los cuatro índices usados para evaluar la regionalización muestran que el agrupamiento óptimo es 5 (ver Fig. 3.12 donde $\widehat{K} = 5$).

Figura 3.12: Métricas de desempeño para clustering, SPI12



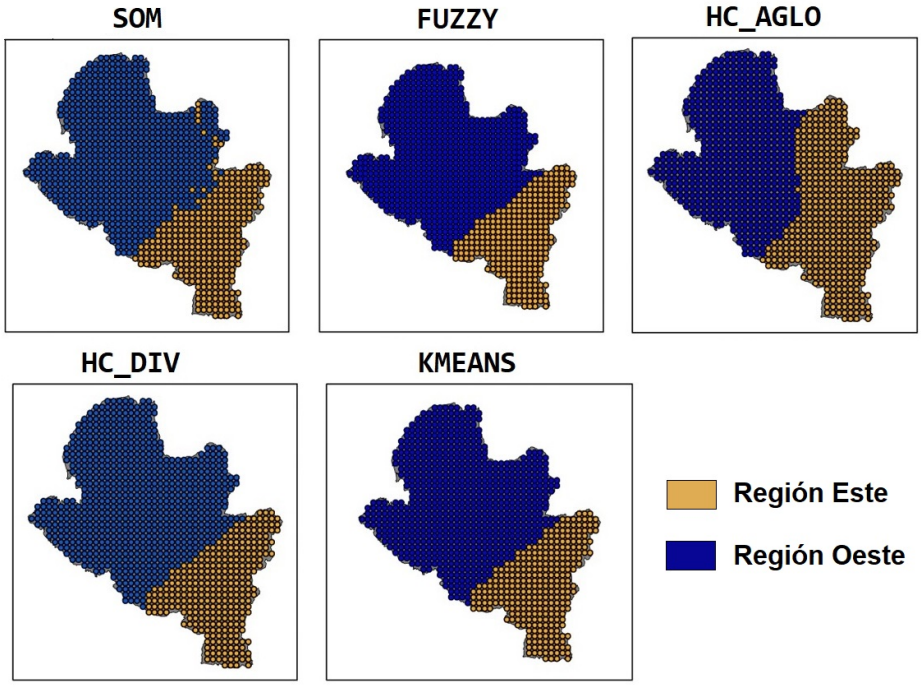
De los gráficos de radar se puede evidenciar que los resultados son concluyentes para los datos SPI6 y SPI12 con $\widehat{K} = 2$ y $\widehat{K} = 5$ respectivamente. Además, el método de *clustering* que mejor logró agrupar los datos fue *Kmeans* para SPI6 y SPI12.

3.3.4. Regionalización del SPI

Teniendo en cuenta el resultado obtenido del proceso de *clustering* para el SPI en las diferentes escalas temporales, se encontraron resultados consistentes para SPI6 y SPI12 con $\widehat{K} = 2$ y $\widehat{K} = 5$ regiones. Con los datos agrupados se realizó una espacialización dando como resultado un mapa de la región de estudio con regiones espaciales homogéneas por cada método de *clustering*, donde los datos tienen un comportamiento similar. Las regiones obtenidas se describen a continuación. Los resultados al espacializar el SPI6, son dos regiones claramente definidas (ver Fig. 3.13) que se denominaron:

- "Región Este", delimitada en amarillo, se ubica al oriente del área de estudio y la cruza de sur a norte. Esta región coincide con la Cordillera de los Andes y la vertiente del Amazonas.
- "Región Oeste", delimitada en azul, está ubicada en la parte occidental del área de estudio, cubriendo la mayor parte de la zona, incluida la llanura del Pacífico colombiano y las estribaciones de la vertiente del Pacífico de los Andes.

Figura 3.13: Espacialización de las regiones identificadas, SPI6.

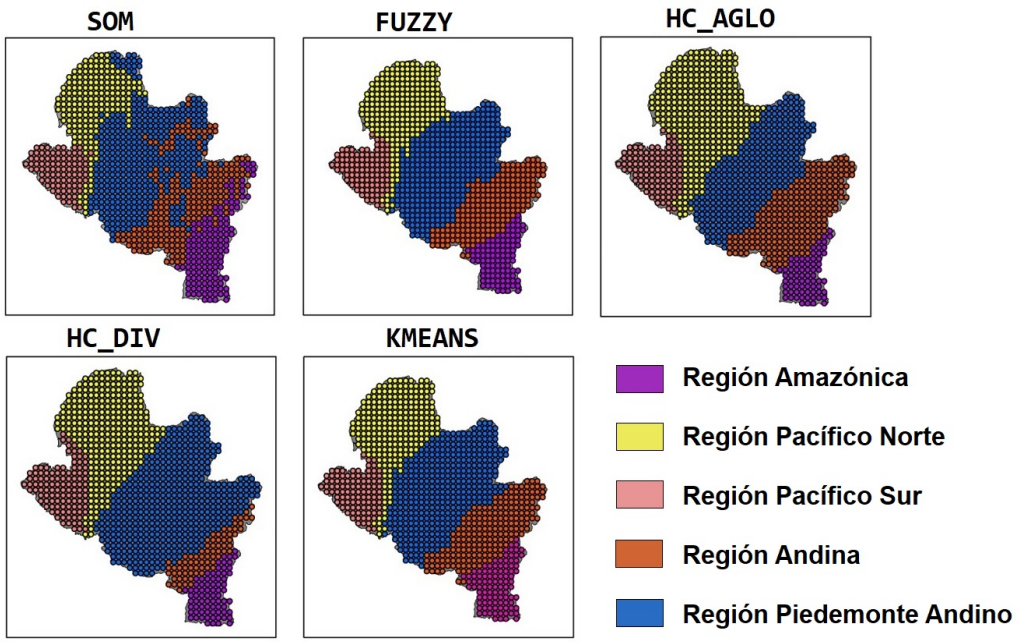


Para el SPI12 como se muestra en la Fig. 3.14. La espacialización de las regiones obtenida se describen en cinco regiones:

- "Región Amazónica", delimitada en violeta. está ubicada en el suroriente del área de estudio y coincide con el flanco oriental de los Andes en el flanco del Amazonas.
- "Región Pacífico Norte", delimitada en amarillo, está ubicada en el noroccidente de la zona de estudio con una delgada línea que conecta el norte y el sur. La región está ubicada en la zona plana del flanco del Pacífico colombiano.

- "Región Pacífico Sur",delimitada en rosa, está ubicada en el suroriente de la zona de estudio, limitando con el país vecino de Ecuador, el Océano Pacífico y la región del Pacífico Norte.
- "Región Andina", delimitada en naranja, está ubicada al este de la zona de estudio, una franja que cubre la parte más alta de la cordillera de los Andes y cruza la zona de estudio de sur a norte.
- "Región Piedemonte Andino", delimitada en azul, se ubica en el centro del área de estudio y la cruza de sur a norte como una franja paralela a la Región Andina. Esta región coincide con la transición topográfica entre la parte más baja de la zona de estudio (Cuenca del Pacífico - Región Sur y Pacífico Norte) a la parte más alta de la zona de estudio (cordillera de los Andes - Región Andina).

Figura 3.14: Espacialización de las regiones identificadas, SPI12.

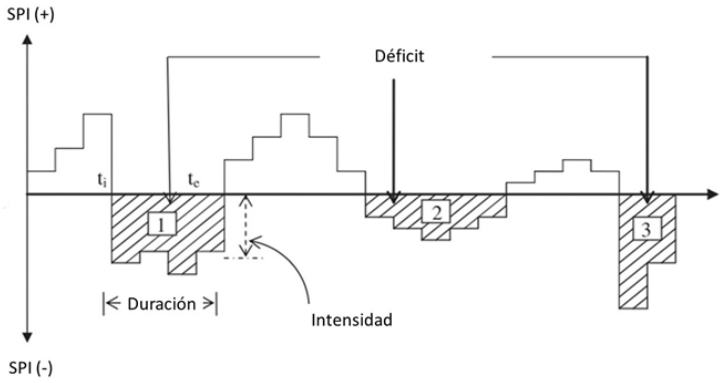


3.3.5. Evaluación y caracterización de las regiones SPI

Teniendo en cuenta la regionalización obtenida por el método *Kmeans* que según las métricas de validación fue el método con mejores resultados al agrupar los datos, se realizó la caracterización de las regiones para datos SPI6 (Fig. 3.13) y SPI12 (Fig. 3.14). La caracterización de cada región obtenida en función del SPI, se realizó graficando el promedio del SPI para cada región identificada (regiones SPI6 y regiones SPI12). Al graficar el SPI es posible determinar características propias de los eventos extremos de lluvias, como la intensidad, duración y severidad. Gráficamente, la duración de un evento se puede identificar desde el momento t cuando se presentan valores menores a -0.5 de SPI hasta que cambia a valores mayores a $+0.5$ de SPI; la intensidad está determinada por la magnitud que alcanza el valor del SPI; la severidad es la combinación de la duración e intensidad del evento (está representada por el área bajo la curva) como se puede observar en la Fig. 3.15. Como ejemplo, en la Fig. 3.15 el evento 2 es el de mayor duración, mientras que el evento 3 es de mayor

intensidad. El criterio del área bajo la curva integra ambas características de un evento extremo, el criterio de área bajo la curva es de mayor interés y sirve para listar en orden de importancia los eventos extremos más severos. En la Fig. 3.15, el evento 1 (evento de sequía) es el de mayor severidad dado que integra mayor área sombreada [4].

Figura 3.15: Características numéricas de los eventos SPI. Adaptado de [4].



3.3.5.1. Caracterización regiones SPI 6

Los estadísticos obtenidos de la Región Oeste (R.OE) que representa el 70 % del área de estudio y la Región Este (R.E) que representa el 30 %, se presentan en la Tabla 3.8.

Tabla 3.8: Estadísticos por región, SPI6.

Estadístico	**R.E**	**R.OE**
Media	0.00	-0.00
Minimo	-3.91	-3.35
Maximo	3.48	3.97
Percentil5	-1.64	-1.59
Percentil25	-0.66	-0.67
Percentil50	-0.04	-0.04
Percentil75	0.69	0.65
Percentil95	1.66	1.71

De acuerdo a los resultados, la sequía es más marcada en la R.E en comparación con la R.OE, la R.E tiene un SPI de -3.91, mientras que el valor más intenso de SPI en la R.OE correspondió a -3.35. Además, el percentil 5 (P5) fue -1.64 en la R.E y -1.59 en la R.OE; esto indica que el 5 % de los valores de SPI6 más negativos tienen una magnitud mayor en la R.E que en la R.OE. Los eventos de humedad son más notorios en la R.OE con un SPI de 3.97. El percentil 95 (P95) indica un valor de 1.71 en la R.OE y 1.66 en la R.E, caracterizando la humedad extrema más comúnmente presente en la R.OE. El valor promedio es cero en ambas regiones, mientras que el percentil 50 (P50) muestra una ligera asimetría de ambos conjuntos de datos hacia la izquierda, los percentiles: 25 (P25), 75 (P75) indica un comportamiento muy similar entre regiones.

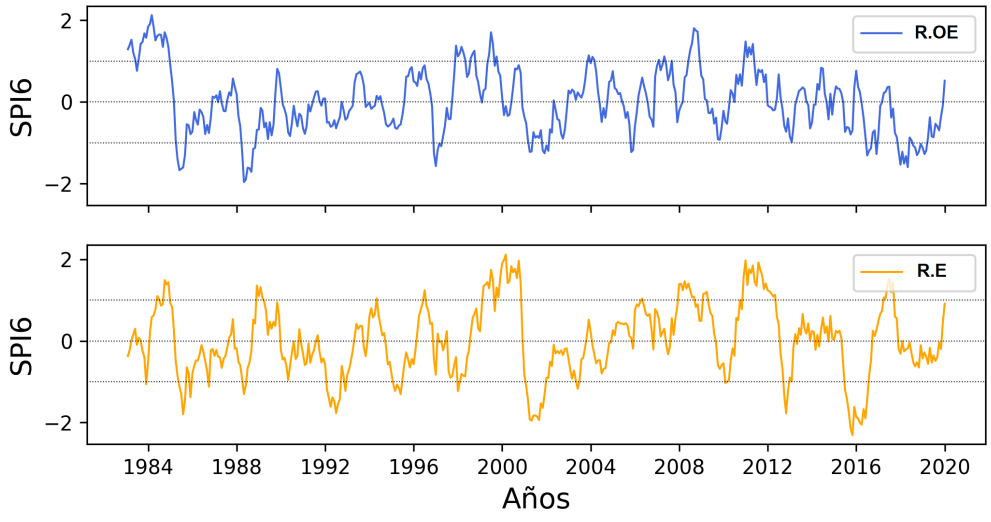
La Tabla 3.9 presenta las características de los eventos de sequía y humedad extrema, acordes con el promedio del SPI6 para cada región (Fig. 3.16). Sólo se registran los eventos con una duración superior a 3 meses, puesto que los cultivos de secano, comienzan a verse seriamente afectados después de 3 meses ante la ocurrencia de un evento extremo (estos registros superiores a 3 meses también se aplican para la Tabla 3.11).

Se puede observar que la ocurrencia de eventos extremos es mayor en la R.E con 34 eventos, en comparación con la R.OE con 24 eventos. En términos generales, la R.E se caracteriza por el predominio de la sequía, considerando el área bajo la curva para valores negativos de SPI6 de 146, mientras que el área bajo la curva para la R.OE es de 122.

Tabla 3.9: Características de los eventos por región, SPI6.

Características	**Signo**	**R.E**	**R.OE**
Número de eventos	SPI +	11	18
	SPI -	13	16
Mayor Duración (meses)	SPI +	29	26
	SPI -	34	28
Mayor magnitud	SPI +	2	2.1
	SPI -	2.2	1.7
Área bajo la curva	SPI +	127	111
	SPI -	146	122

Figura 3.16: Promedio por regiones, SPI6. De arriba a abajo: R.OE y R.E



3.3.5.2. Caracterización regiones SPI 12

Las regiones SPI12 identificadas (Fig. 3.14) fueron cinco. El área que ocupa cada región es: Amazónica (R.AM) - 18 %, Pacífico Norte (R.PN) - 23 %, Pacífico Sur (R.PS) - 14 %, Andina (R.A) - 20 % y Piedemonte andino (R.PA) - 25 %. Los estadísticos básicos se presentan en la Tabla 3.10.

Tabla 3.10: Estadísticos por región, SPI12.

Características	**R.AM**	**R.PN**	**R.PS**	**R.A**	**R.PA**
Media	3e-6	-165e-6	-689e-6	-98e-6	-118e-6
Minimo	-4.03	-3.03	-2.68	-3.37	-3.13
Maximo	3.27	3.32	3.77	3.17	3.61
Percentil5	-1.53	-1.60	-1.50	-1.56	-1.61
Percentil25	-0.63	-0.67	-0.71	-0.68	-0.70
Percentil50	-0.06	-0.05	-0.05	-0.10	-0.01
Percentil75	0.67	0.65	0.52	0.70	0.65
Percentil95	1.68	1.75	1.87	1.75	1.69

Los estadísticos muestran que la R.A y la R.AM presentan los valores más bajos de SPI12 con $-3,37$ y $-4,03$, respectivamente. La R.A también en su percentil 5 (P5) tiene un valor bajo de $-1,56$; destacando esta región con características de sequía asociadas a ser la región con menor precipitación en el área de estudio (ver la distribución espacial de la precipitación en la Fig. 3.1). La R.AM presenta valores de SPI negativos asociados a la desviación de los valores medios de precipitación, dado que esta región tiene una precipitación considerablemente alta.

La R.PS presenta los eventos de humedad extrema más altos y un valor máximo de humedad de 3.77. Además, el percentil 95 (P95) es 1.87, el más alto en comparación con las demás regiones. La región del Pacífico (R.PS y R.PN) está dividida por características opuestas de los eventos extremos analizados; El SPI12 negativo es predominante en R.PN y lo contrario en R.PS. La división del Pacífico en sur y norte también se presenta en [80] para la variable de precipitación en 4 de los 5 métodos de regionalización analizados. Los valores promedio de SPI12 son cercanos a cero y predominan en todas las regiones, los percentiles 25, 50 y 75 muestran un comportamiento estándar sin diferenciación importante de una región a otra.

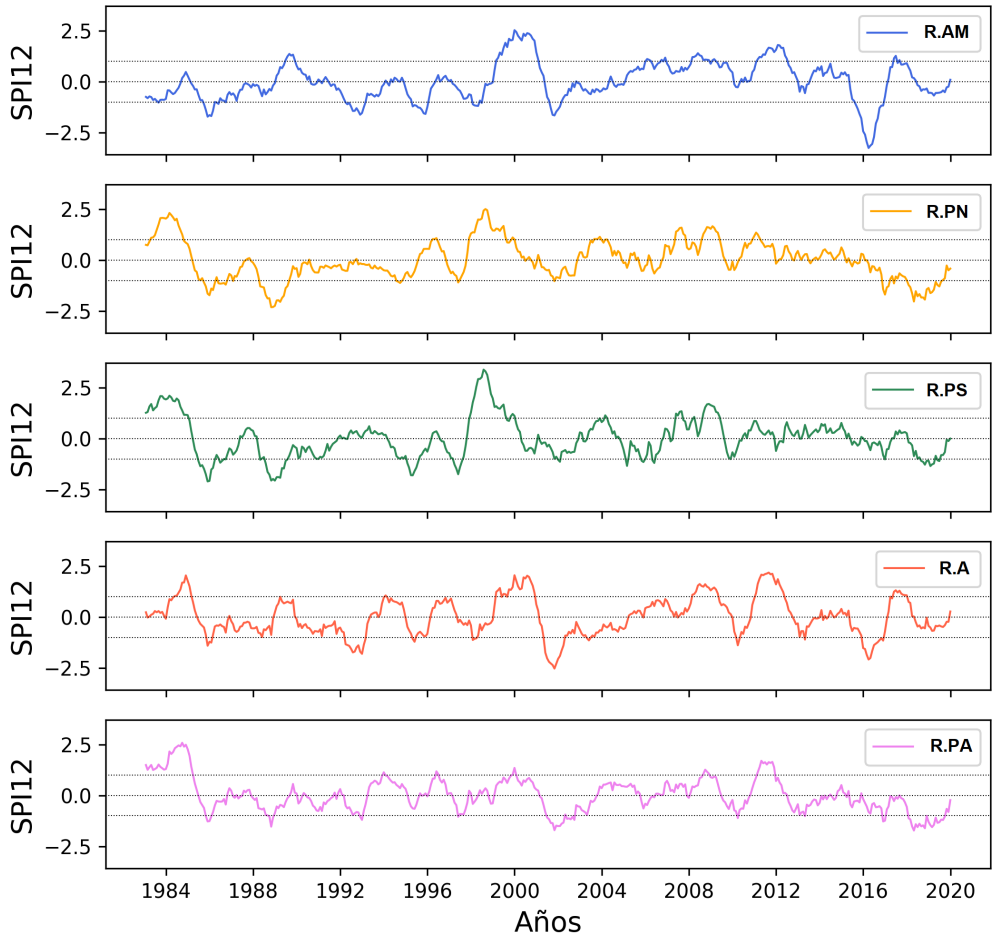
La magnitud de los eventos en las cinco regiones a través del área bajo la curva de la Fig. 3.17, muestra que la R.A presenta características de sequía más marcadas en toda el área de estudio con un valor de 184 (SPI negativo), siendo la mayor en magnitud en comparación con las demás regiones.

La R.PS tiene el evento positivo más extremo de 3.4. En cambio, La R.A tiene el evento negativo más extremo de -2,5. Por otro lado, la R.PA y R.AM presentan eventos con mayor frecuencia y menor magnitud; la R.PA registró 24 eventos en total, el más alto entre todas las regiones; esta frecuencia es compensada por la baja magnitud de los eventos.

Tabla 3.11: Características de los eventos por región, SPI12.

Características	Signo	R.AM	R.PN	R.PS	R.A	R.PA
Número de	SPI +	8	9	10	8	12
Eventos	SPI -	10	9	10	9	12
Mayor Duración	SPI +	58	35	37	55	30
(meses)	SPI -	45	92	31	47	55
Mayor magnitud	SPI +	2.5	2.5	3.4	2.1	2.5
	SPI -	1.6	2.3	2.1	2.5	1.7
Área bajo la	SPI +	166.1	165.9	166.3	168.5	141.8
curva	SPI -	165.8	166.2	167.2	168.8	142.5

Figura 3.17: Promedio por regiones, SPI12. De arriba a abajo: R.AM, R.PN, R.PS, R.A y R.PA



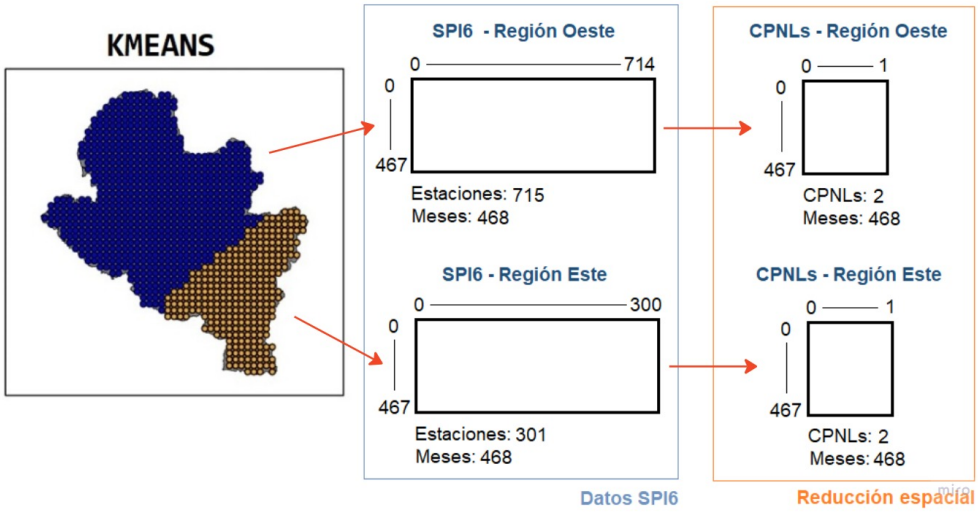
Nota: Considerando que el SPI6 es el índice recomendado para estudiar las condiciones meteorológicas y de humedad del suelo en la agricultura[3] [87, 47]. En adelante, se utiliza el SPI6 para el desarrollo del modelo de pronóstico y los análisis correspondientes.

[3] La OMM, recomienda en su manual de usuario del SPI la escala semestral, para estudiar las condiciones meteorológicas y de humedad del suelo en la agricultura.

3.3.6. Reducción de dimensionalidad espacial

Con las regiones SPI6 identificadas y caracterizadas, se realiza una reducción de dimensionalidad espacial donde las estaciones son reducidas a CPNLs para cada región SPI6 identificada, en la Fig. 3.18 se muestra el proceso de reducción de dimensionalidad, inicialmente se parte de la regionalización, para las dos regiones identificadas se toman las series de tiempo, la R.E tiene un tamaño de 715 estaciones y 468 meses; la R.OE tiene 301 estaciones y 468 meses. Luego se utiliza ACPNL para la reducción de dimensionalidad quedando las series reducidas a 2 CPNLs y 468 meses.

Figura 3.18: Reducción de dimensionalidad espacial - SPI6



Al igual que la reducción de dimensionalidad temporal, la forma de evaluar el desempeño del proceso de entrenamiento es observando la varianza explicada de cada componente no lineal para diferentes topologías establecidas. La selección del número de CPNLs (capa latente), se basó en ACP, se seleccionó el número de valores propios superiores a uno según [79, 80]. Se realizó un total de 150 entrenamientos con datos SPI6 de las regiones homogéneas identificadas. La Tabla 3.12 consolida los mejores resultados obtenidos y la topología de red usada.

Tabla 3.12: Cálculo de ACPNL para SPI6 en la R.E y la R.OE.

Región	**Topología**	**Varianza explicada (%)**	
		CPNL1	**CPNL2**
	468-200-20-2-20-200-468	82.12	9.43
R.E	468-200-10-2-10-200-468	70.36	8.43
	468-200-15-2-15-200-468	75.59	8.59
	468-250-50-2-50-250-468	71.89	19.33
R.OE	468-200-20-2-20-200-468	68.15	15.87
	468-200-15-2-15-200-468	70.68	18.43

Para la R.E la varianza explicada total fue de 91.55 % y para la R.OE la varianza explicada total fue de 91.22 %. Las varianzas explicadas están por encima del 90 % para las dos regiones, lo que puede considerarse una buena varianza explicada en dos CPNLs.

3.3.7. Teleconexiones del SPI6 con variables macroclimáticas

Teniendo la reducción de dimensionalidad espacial de cada región identificada previamente, se utilizó la memoria climática almacenada en distintas variables oceánicas y atmosféricas. Se recopilaron correlaciones entre la lluvia y VM con estudios previos realizados en Colombia, Ecuador y Nariño; las VM utilizadas en esta investigación provienen de la NOAA[4]. En total son 22 VM preseleccionadas en la Tabla 3.13 se presenta en nombre y la descripción de cada una.

Tabla 3.13: Variables macroclimáticas seleccionadas.

Variable macroclimática	**Descripción**	**Tipo**	**Ubicación**
*MEI: Índice Multivariado del ENSO	Serie de tiempo de la Función Ortogonal Empírica (EOF) que combina cinco variables diferentes: presión, TSM, viento, radiación de onda larga (componente zonal y meridional).	Oceánico y Atmosférico	Pacífico
*NAO: Oscilación del Atlántico Norte	Anomalía de un dipolo en la TSM del Norte del Atlántico.	Oceánico	Atlántico
*NINO12: TSM en la región Niño 1+2	Anomalía de la TSM en el Extremo Este del Pacifico Tropical (0-10S, 90W-80W).	Oceánico	Pacífico tropical
*NINO3: TSM en la región Niño 3	Anomalía de la TSM en el Pacifico Tropical Central (5N-5S,150W-90W).	Oceánico	Pacífico tropical
*NINO4: TSM en la región Niño 3.4	Anomalía de la TSM entre la región Niño 3 y Niño 4 (-5S-5N, 170W-120W).	Oceánico	Pacífico tropical
*NINO4: TSM en la región Niño 4	Anomalía de la TSM en el Pacifico Oriental (5N-5S, 160E-150W).	Oceánico	Pacífico tropical
*ONI: Índice del Niño Oceánico	Tres meses de media de la anomalía de TSM en la región de El Niño 3.4 basado en el cambio de periodo base de 30 años de registro.	Oceánico	Pacífico
*SOI: Índice de Oscilación del Sur	Diferencia de presión atmosférica entre Tahití y Darwin.	Atmosférico	Pacífico
*TNI: Índice del Niño Trans	Es el T-Nino 1+2 estandarizado menos el T-Niño 4 con una media corrida de 5 meses aplicada que luego se estandariza usando el período 1950-1979.	Oceánico	Pacífico

Continua en la siguiente página

[4] https://www.esrl.noaa.gov/psd/data/climateindices/list/

Tabla 3.13 – Continua de la página anterior

Variable macroclimática	Descripción	Tipo	Ubicación
*TNA: Índice del Atlántico Norte Tropical	Anomalía el promedio mensual del TSM de 5.5N a 23.5N y 15W a 57.5W.	Oceánico	Atlántico
*ENSO: Índice de Precipitación del ENSO	Serie temporal que utiliza los datos de las precipitaciones en el Pacífico Tropical para describir los eventos del ENSO.	Atmosférico	Pacífico
NTA: Índice de la TSM del Atlántico Norte Tropical	Anomalía de TSM en 60W a 20W, 6N a 18N y de 20W a 10W, 6N a 10N. Las anomalías están suavizadas por el procedimiento de media corrida de tres meses y proyectadas sobre 20 FOEs principales	Oceánico	Atlántico
CAR: Índice del Caribe	Anomalía de TSM sobre el Caribe	Oceánico	Atlántico
*WP: Índice del Pacífico Occidental	Variabilidad de baja frecuencia sobre el Pacifico norte	Oceánico	Pacífico
*QBO: Oscilación Quiasibienal	Media zonal del viento a 30 mb en el Ecuador (estratosfera tropical)	Atmosférico	Pacífico
*AMO: Oscilación del Atlántico Norte	TSM promediada en 0°-60°N, 0°-80°W menos la TSM promediada en 60°S-60°N	Oceánico	Atlántico
NP: Patrón del Pacifico Norte	Es la presión al nivel del mar ponderada por área en la región 30N-65N, 160E-140W	Atmosférico	Pacífico
*AMM: Modo Meridional del Atlántico	Aplicación del análisis de covarianza máxima (MCA) a la TSM y el campo de viento de 10 m	Oceánico y Atmosférico	Atlántico
*TSA: Índice del Atlántico Sur Tropical	Anomalía el promedio mensual del TSM de Eq-20S y 10E-30W	Oceánico	Atlántico
*BEST: Serie de tiempo bivariada del ENSO	Resultado de la combinación estandarizada del SOI y la TSM del Niño 3.4	Oceánico y Atmosférico	Pacífico
*PDO: Oscilación Decadal del Pacífico	Anomalía mensual de TSM sobre el Océano Pacífico Norte, desde 20N hacia el polo	Oceánico	Pacífico
EMI: Índice del Niño Modoki	Posee 3 términos derivados de un promedio de área de la anomalía de TSM en las siguientes regiones A (165°E-140°W, 10°S-10°N), B (110°W-70°W, 15°S-5°N9	Oceánico	Pacífico

En la Tabla 3.13 se presentan las VM reportadas en la literatura, pero algunas no están disponibles o están desactualizadas, las variables que están disponibles y actualizadas están marcadas con (*). Con las 18 VM disponibles, se realizó un análisis donde el objetivo es seleccionar las mejores VM que permiten explicar o relacionar el comportamiento del SPI6 en las dos regiones (R.OE y R.E). Para seleccionar las VM que tienen mayor relación con el

SPI6, se estimó la correlación de Pearson cruzada sincrónica y con rezago entre las CPNLs de cada región del SPI6 y las VM disponibles.

Para considerar una significancia estadística, se estimó un limite mínimo de correlación positiva y negativa, esta estimación se realizo considerando los siguientes pasos metodológicos [54].

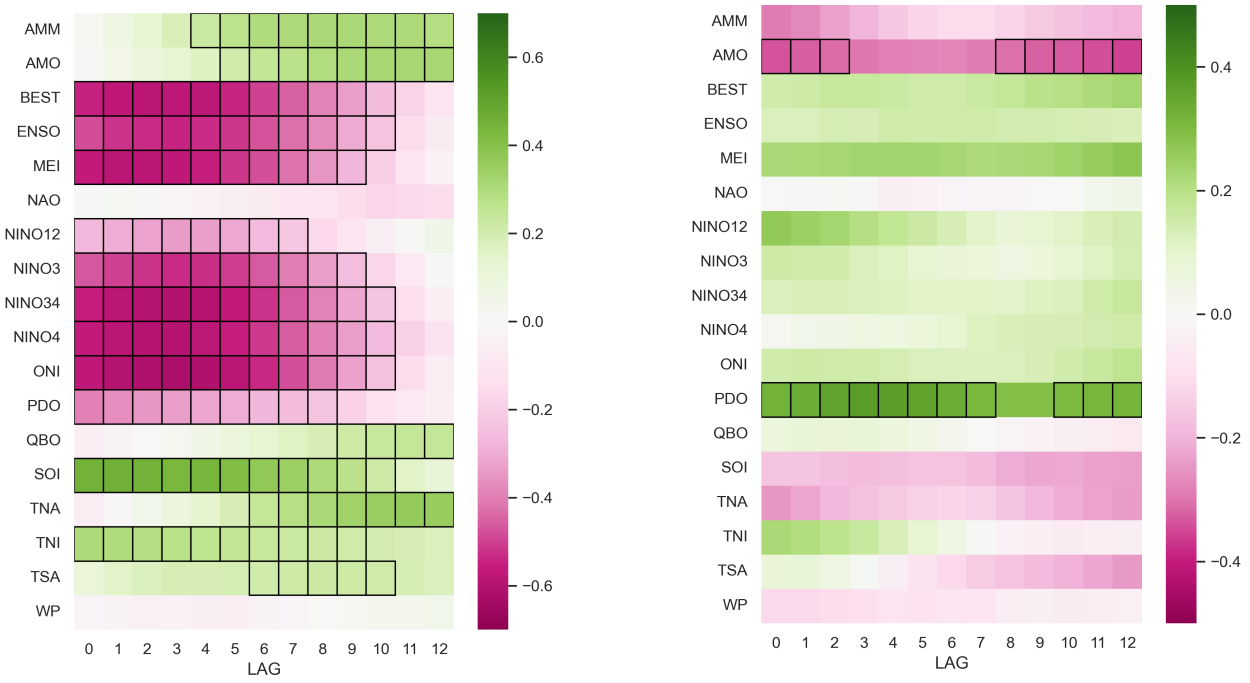
```
- Se estimó la autocorrelación de cada CPNL.
- Se observo el grado de memoria para cada serie de tiempo.
  Esto se realizó analizando el número de meses que supera
  los limites de confianza de la función de autocorrelación.
- Se calcularon los grados de libertad de cada CPNL
  realizando el cociente entre la cantidad de meses y el grado de memoria de cada CPNL.
- Con los grados de libertad se calculó el intervalo de confianza al 95% para cada CPNL.
- Los intervalos calculados son el limite de correlación positiva y negativa
  usado para determinar que VM tienen una correlación con significancia estadística.
```

Los limites de correlación para cada CPNL, son los siguientes:

- CPNL1 - R.E: significancia estadística con una correlación mayor o igual a $|0.257|$.
- CPNL2 - R.E: significancia estadística con una correlación mayor o igual a $|0.377|$.
- CPNL1 - R.OE: significancia estadística con una correlación mayor o igual a $|0.257|$.
- CPNL2 - R.OE: significancia estadística con una correlación mayor o igual a $|0.222|$.

En la Fig. 3.19 y la Fig. 3.20 se analizan las correlaciones entre las VM y las dos CPNL de la R.E y la R.OE, respectivamente. Estas correlaciones se describen en términos de la amplitud de memoria de las variables representadas en la cantidad de meses de rezago, también en la relación directa o inversa que mantienen con las CPNLs y la magnitud de la correlación presentada.

3.3.7.1. Teleconexiones del SPI6 - R.E



(a) Correlación R.E - CPNL1.

(b) Correlación R.E - CPNL2.

Figura 3.19: Correlación entre las CPNL de R.E y las VM, las correlaciones significativas están marcadas con un recuadro negro.

La CPNL1 de la R.E presenta una correlación directa, significativa y de largo rezago (entre 0 y 9 meses) para SOI, TNI; a su vez, presenta una correlación inversa, significativa y de rezago largo con BEST, ENSO, MEI, NINO12, NINO3, NINO34, NINO4, ONI y PDO y un rezago directo amplio (mas de 6 meses) con AMM, AMO, QBO, TNA y TSA. Las correlaciones cruzadas del SPI con VM son equivalentes con las correlaciones de las anomalías de precipitación con las VM. Canchala et al.[24] muestran correlaciones negativas (positivas) con el ONI, MEI, NINO3, NINO3.4, NINO4 (SOI) demostrando una relación inversa (directa) con índices oceánicos (atmosféricos), resultados equivalentes a las correlaciones encontradas con el SPI6. La relación inversa entre las VM del Niño como NINO3, NINO4, NINO34, ONI, ENSO y BEST indica que en la medida que la anomalía es positiva (negativa), los valores del SPI disminuye (aumenta) en la R.E. Esto coincide con el comportamiento de la fase cálida (fría) de ENOS [88] y la disminución (aumento) de precipitaciones en la región Andina de Colombia [11].

Las correlaciones encontradas para la CPNL1 de la R.E, muestra el alto grado de influencia de las anomalías en la cuenca del Pacífico sobre la variabilidad explicada por esta componente. Las correlaciones con largo rezago fueron TNI y SOI. Esta última es una variable macroclimática de tipo atmosférica que mide la diferencia de presión entre Tahití y Darwin (ver Tabla 3.13), la atmósfera suele tener una memoria de corto plazo que las anomalías de temperatura en el océano. En contraste, el TNI se basa en una relación de la TSM de regiones específicas del Pacífico, la región de NINO12 se encuentra contigua al departamento de Nariño y limita con la parte continental (ver Fig. 3.1), por lo tanto se

espera una respuesta en las precipitaciones de corto rezago.

La CPNL1 de la R.E presenta una correlación directa con amplio rezago con TNA y TSA, estas dos VM tienen un alto potencial de pronóstico a largo plazo. Aunque la mayor parte de las correlaciones significativas con la CPNL1 de la R.E corresponden al Pacífico, las VM asociadas al Atlántico son las que presentan el mayor rezago. Esto indica que la macrocuenca del Atlántico tiene influencia sobre el SPI en Nariño, pero la respuesta al cambio de la TSM en el Atlántico se refleja hasta un año después. Este amplio rezago de NTA y TSA con la CPNL1 de la R.E se explica por la distancia física entre el Atlántico y esta región. Adicionalmente Poveda et al.[89] explica que la vertiente Este de los Andes es fuertemente influenciada por la humedad proveniente del Amazonas, esta humedad viene del proceso de evaporación del océano Atlántico. Teóricamente, este proceso tarda varios meses dado que la selva Amazónica tiene un desarrollo cíclico de evapotranspiración-precipitación de hasta 6 meses, esto explica el amplio rezago entre la anomalía de la TSM en el océano (NTA y TSA) y las variaciones del SPI en la R.E de Nariño.

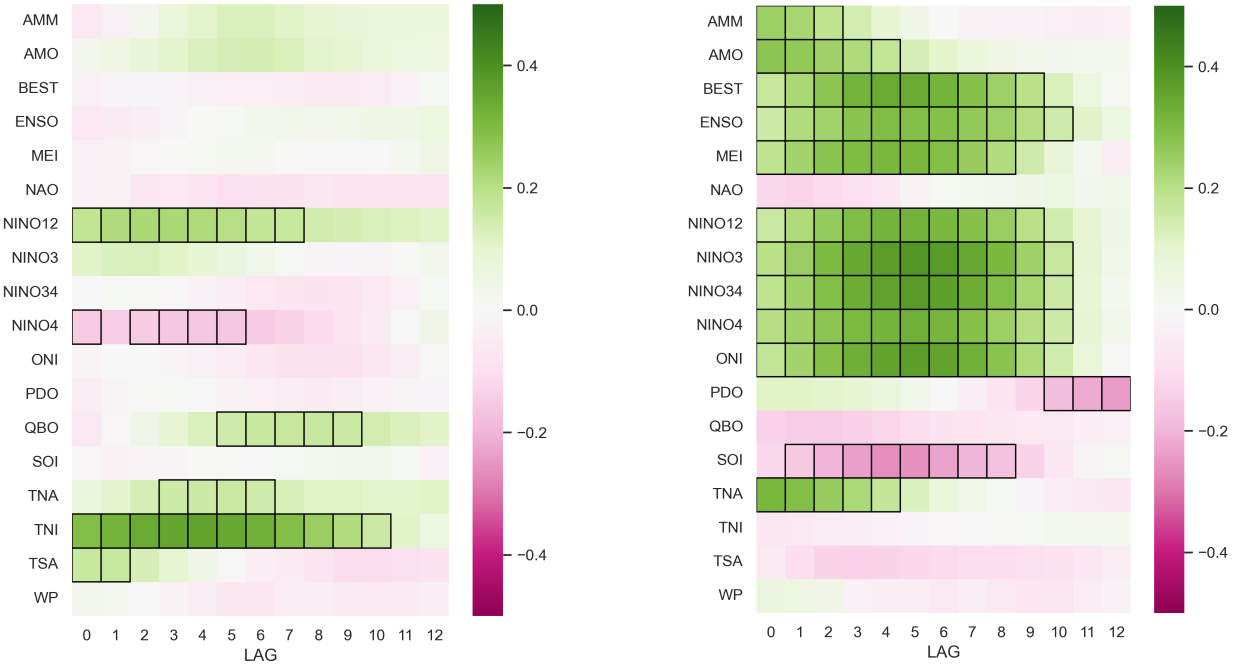
Comparando las correlaciones encontradas con estudios antecedentes, existe una relación de la R.E del SPI6 con la región la región Andina colombiana, y la región Andina ecuatoriana. Por ejemplo, Moran-Tejeda et al.[77] reportan una fuerte relación inversa (directa) entre la precipitación de los Andes ecuatorianos y NINO34 (SOI), esto concuerda con las correlación inversa (directa) entre NINO34 (SOI) y la CPNL1 de la R.E; además, Vicente-Serrano et al.[90] reporta una relación entre la sequía en los Andes ecuatorianos con NINO34. En Colombia, Canchala et al.[10] estudiaron las teleconexiones con la precipitación en Nariño de la región Andina y Pacífico, ellos encontraron una correlación positiva con el SOI y negativa con el ENSO, coherente con las correlaciones encontradas para el SPI de la R.E. Adicionalmente, Montealegre et al.[91] reportó una reducción (incremento) de las precipitaciones mensuales, equivalente a una reducción (incremento) del SPI, con una relación directa (inversa) con los índices asociados al ENSO. En el estudio de Cerón et al.[13] se muestra una correlación negativa entre el SPI en Cali (región Andina de Colombia, equivalente a la R.E en Nariño) con las VM del ONI y MEI, y una correlación positiva con el SOI, coherente el sentido de las correlaciones encontradas con la CPNL1 de la R.E.

La CPNL2 de la R.E presenta una correlación significativa de una manera sincrónica (rezago 0) y a lo largo de todos los rezagos (entre 1 y 12 meses) con AMO y PDO. Ambas VM están asociadas a la explicación de fenómenos macroclimáticos de baja frecuencia como la OAN y la ODP; Morán-Tejada et al.[77] reporta que la PDO es mejor modulador de la precipitación en los Andes Ecuatorianos que el ENSO, coherente con lo mostrado en esta investigación.

La relación directa entre las VM del Atlántico (AMO) y el CPNL2 indica que ante el calentamiento del Atlántico se incrementa los valores del SPI en la R.E, pero a su vez tiene una relación inversa con las VM del Pacífico (PDO), lo que indica que ante el calentamiento del Pacífico se disminuye el SPI.

3.3.7.2. Teleconexiones del SPI6 - R.OE

La Fig. 3.20 muestra las correlaciones de R.OE y la CPNL1 de esta región, se presenta una alta correlación significativa y directa con NINO12, QBO, TNA, TNI, TSA. Por otro lado, la CPNL1 de la R.OE muestra una correlación significativa negativa con NINO4.



(a) Correlación CPNL1.

(b) Correlación CPNL2.

Figura 3.20: Correlación entre las CPNL de R.OE y las VM, las correlaciones significativas están marcadas con un recuadro negro.

Canchala et al. [10] reportaron correlaciones positivas de la región Pacífico nariñense con NINO12 y NINO3, y correlación negativa con NINO34; esto es congruente por los reportado en las correlaciones de la CPNL1 de la R.OE y las VM usadas. Adicionalmente, en ambos estudios no se encontró significancia estadística en las correlaciones, con unas débiles teleconexiones con la región Pacífico. Vicente-Serrano et al.[90] relaciona la sequía del Pacífico ecuatoriano con la VM NINO12, similar a lo reportado en las correlaciones calculadas. Por otro lado, de Guenni et al.[92] reportó una correlación positiva de NINO3 con la precipitación de la costa ecuatoriana y se relaciona NINO12 con las lluvias [93] y caudales [94] de la costa ecuatoriana, similar a lo reportado en las correlaciones de la R.OE calculadas.

La VM TNI, es la diferencia entre la TSM en la región del Niño 1+2 y la TSM de la región del Niño 4, por lo tanto, valores negativos de TNI indican un mayor calentamiento de la región Niño 1+2 (más cerca del continente) respecto a la región del Niño 4 (más alejado del continente). La relación directa entre el TNI y la CPNL1 de la R.OE indica que mientras la TSM en la región del Niño 1+2 sea mayor que la TSM en la región del Niño 4 se tendera a valores negativos del SPI en la R.OE.

La CPNL2 de la R.OE presenta una correlación significativa y directa con las VM AMO, AMM, TNA, TSA, como se observa en la Fig. 3.20. La magnitud de estas variables supera

el límite de confianza de 0.222 en los rezagos menores a 10 meses, por lo tanto se consideran potenciales variables exógenas para el pronóstico. Estas relaciones directas indican que en la medida que el Océano Atlántico se calienta, la R.OE tiende a presentar valores negativos del SPI. El efecto que presenta el calentamiento del Atlántico sobre el departamento es contrario entre las dos regiones, un calentamiento anómalo del Atlántico conlleva a un incremento del SPI en la R.E, mientras que provoca valores negativos de SPI en la R.OE.

Las correlaciones encontradas con las CPNL de ambas regiones, pone en evidencia la clara influencia del ENOS sobre la sequía en el departamento. Las correlaciones significativas de NINO12, NINO3, NINO34. NINO4, ENOS, TNI entre otras VM asociadas al ENOS.

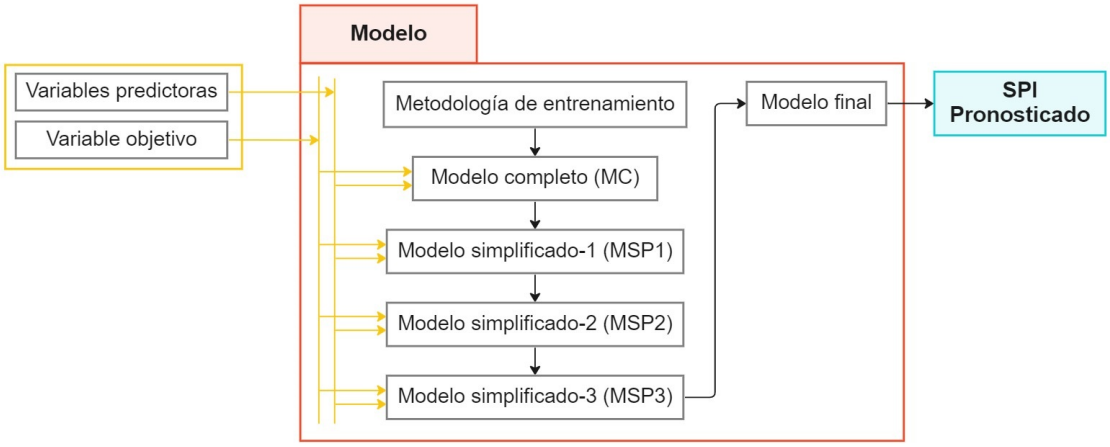
Capítulo 4

Modelo y SPI pronosticado

Machine learning (ML) es una forma de la IA que permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita. El proceso de entrenamiento de un modelo de ML consiste en proporcionar datos de entrenamiento de los cuales aprende un algoritmo de ML, también llamado algoritmo de aprendizaje. El término modelo, se refiere al artefacto de modelo resultante del proceso de entrenamiento. Para un modelo, los datos de entrenamiento deben contener la respuesta correcta, que se conoce como destino o variable objetivo. El algoritmo de aprendizaje encuentra patrones en los datos de entrenamiento que asignan los atributos de los datos de entrada al destino (la respuesta que desea predecir) y genera un modelo de ML que captura dichos patrones.

Los pasos metodológicos utilizados para desarrollar el modelo final de pronóstico se presentan en la Fig. 4.1. Inicialmente se presenta la metodología de entrenamiento que es basada en agentes, una forma centralizada y distribuida usada en la actualidad para entrenar modelos de ML. Seguido se presenta el desarrollo de los modelos completos como base inicial para realizar iteraciones de simplificación y llegar a modelos simplificados que cumplan criterios establecidos. Finalmente con el modelo simplificado se detalla el proceso para calcular el SPI pronosticado.

Figura 4.1: Esquema general del desarrollo metodológico del bloque Modelo.



El objetivo del modelo final es pronosticar el SPI6 de la R.E y la R.OE, para llegar a ese modelo se agruparon cuatro modelos NNARMAX que tienen como objetivo la CPNL1-R.E,

CPNL2-R.E, CPNL1-R.OE y la CPNL2-R.OE; luego se aplica la transformación inversa de ACPNL y se obtiene el SPI6 de las dos regiones. Para el entrenamiento de los modelos se utilizaron las VM con una correlación estadísticamente significativa, las correlaciones se resumen en la Tabla 4.1 estas VM provienen del análisis y selección de los mejores predictores, proceso detallado en el capitulo anterior.

Tabla 4.1: Principales VM en correlación con el SPI6 en la R.E y R.OE.

CPNL - Región	**VM**			
CPNL1 - R.E	AMM	AMO	BEST	ENSO
	MEI	NINO12	NINO3	NINO34
	NINO4	ONI	PDO	QBO
	SOI	TNA	TNI	TSA
CPNL2 - R.E	AMO	PDO		
CPNL1 - R.OE	NINO12	QBO	TNI	NINO4
	TNA	TSA		
CPNL1 - R.OE	AMM	MEI	NINO4	AMO
	BEST	ENSO	NINO1	NINO3
	NINO34	ONI	SOI	TNA

4.1. Metodología de entrenamiento

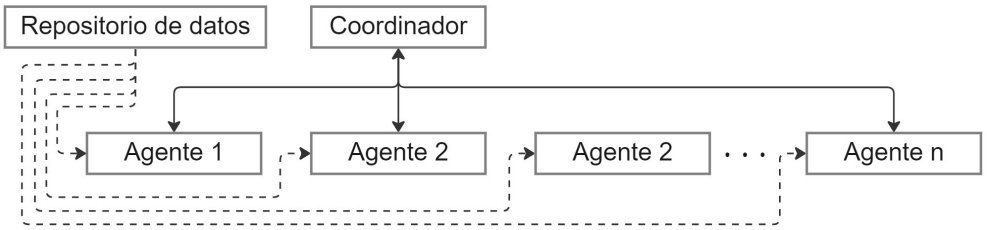
En la Fig. 4.2 se presenta el esquema general de la metodología de entrenamiento de los modelos, esta metodología es basada en agentes. Un agente es una unidad de procesamiento que esta conectada a un coordinador, el coordinador es el encargado de distribuir los parámetros e hiperparámetros de entrenamiento a cada agente conectado; cada agente entrenan una cantidad determinada de modelos y los retorna al coordinador para ser almacenado remotamente. Para el entrenamiento de los modelos, los datos deben también estar centralizados y disponibles para ser consultados o descargados por los agentes. Las ventajas de tener una metodología de entrenamiento basada en agentes son:

- Información y artefactos disponibles de forma centralizada, esto evita errores locales y la carga manual de datos.
- El coordinador tiene la capacidad de realizar diferentes barridos de parámetros e hiperparámetros para encontrar el mejor modelo.
- Computación distribuida, los agentes permiten tener diferentes máquinas entrenando modelos de forma paralela y reportando sus resultados al coordinador.

La herramienta usada para configurar el coordinador y los agentes es WandB[1], esta herramienta permite configurar barridos (*sweeps*) de parámetros e hiperparámetros, esta

[1] WandB, es un panel de control central que permite realizar un seguimiento de los hiperparámetros, las métricas del sistema y las predicciones para comparar modelos de ML en tiempo real y compartir los resultados. `https://wandb.ai/site`

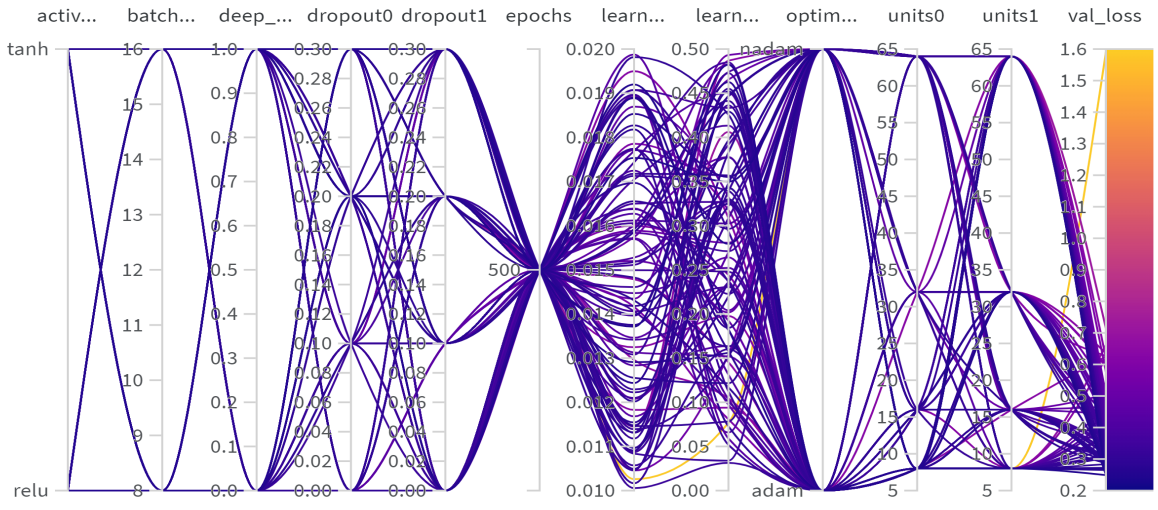
Figura 4.2: Esquema general de la metodología de entrenamiento.



configuración se realiza por rangos y de forma aleatoria. El coordinador va asignando un grupo de parámetros que el agente recibe, con el grupo de parámetros recibido el agente entrena los modelos y retorna los artefactos y logs (historial) al coordinador.

De forma general los diferentes sweeps de cada modelo, fueron configurados para realizar variaciones de diferentes hiperparámetros y posteriormente los agentes asociados a los sweeps realizan el entrenamiento de la red neuronal. Las variaciones fueron en el tamaño de datos por batch, tasa de aprendizaje, algoritmo de aprendizaje, cantidad de capas, neuronas por capa, *dropout* por capa y funciones de activación. Estos barridos fueron definidos tomando un modelo inicial (modelo completo CPNL1-R.E) y haciendo variaciones con rangos grandes, luego al observar la distribución de los parámetros e hiperparámetros (ver Fig. 4.3) y determinar la convergencia de estos, se definieron los rangos de variación de los sweeps posteriores.

Figura 4.3: Distribución de parámetros e hiperparámetros.



La configuración de los sweeps para los modelos se definió de la siguiente forma.

```
sweep_config = {
    "method" : "random",
    "metric":{
        "goal":"minimize",
        "name":"val_root_mean_squared_error"
    },
    "parameters" : {
        "epochs" : { "values" : [500] },
        "batch_size":{ "values": [8, 16] },
        "learning_rate_sgd" :{ "min": 0.01, "max": 0.02 },
        "learning_rate_adam" :{ "min": 0.01, "max": 0.02 },
        "learning_rate_nadam" :{ "min": 0.01, "max": 0.02 },
        "optimizer":{ "values": ["SDG", "adam", "nadam"] },
        "deep_layers":{ "values": [0, 1] },
        "units0":{ "values": [8, 16, 32, 64] },
        "units1":{ "values": [8, 16, 32, 64] },
        "dropout0":{ "values": [0.0, 0.1, 0.2, 0.3] },
        "dropout1":{ "values": [0.0, 0.1, 0.2, 0.3] },
        "activation":{ "values": ["relu", "tanh"] }
    }
  }
}
```

La selección del mejor modelo se realizó analizando el ECM en los datos de entrenamiento (loss) y los datos de validación (val_loss), para cada modelo se calcularon y graficaron en un diagrama de dispersión los errores, donde la finalidad es encontrar el modelo con los errores más bajos, modelos más cercanos al origen de coordenadas.

4.2. Modelo completo (MC)

Como punto inicial de comparación, se entrenaron los cuatro modelos completos NNARMAX que tienen como objetivo las CPNLs de la R.E y la R.OE. Estos modelos son entrenados con todas las VM (ver Tabla 4.1) y sus rezagos significativos para cada componente (ver Fig. 3.19 y Fig. 3.20). Los modelos completos son el modelo "ideal" puesto que capturan todos los predictores y la información estadísticamente significativa, sin tener en cuenta el horizonte de pronóstico. Las funciones que definen cada modelo se presentan a continuación.

$$
\begin{aligned}
y(k+1) = f[\ & y(k), y(k-1), y(k-2), x_1(k-3), x_1(k-4), x_1(k-5), x_1(k-6), x_1(k-7), x_1(k-8), x_1(k-9), \\
& x_1(k-10), x_1(k-11), x_2(k-4), x_2(k-5), x_2(k-6), x_2(k-7), x_2(k-8), x_2(k-9), x_2(k-10), x_2(k-11), \\
& x_3(k), x_3(k-1), x_3(k-2), x_3(k-3), x_3(k-4), x_3(k-5), x_3(k-6), x_3(k-7), x_3(k-8), x_3(k-9), \\
& x_4(k), x_4(k-1), x_4(k-2), x_4(k-3), x_4(k-4), x_4(k-5), x_4(k-6), x_4(k-7), x_4(k-8), x_5(k-6), x_5(k-7), \\
& x_4(k-9), x_5(k), x_5(k-1), x_5(k-2), x_5(k-3), x_5(k-4), x_5(k-5), x_5(k-8), x_6(k), x_6(k-1), x_6(k-2), \\
& x_6(k-3), x_6(k-4), x_6(k-5), x_6(k-6), x_7(k), x_7(k-1), x_7(k-2), x_7(k-3), x_7(k-4), x_7(k-5), x_7(k-6), \\
& x_7(k-7), x_7(k-8), x_8(k), x_8(k-1), x_8(k-2), x_8(k-3), x_8(k-4), x_8(k-5), x_8(k-6), x_8(k-7), x_8(k-8), \\
& x_8(k-9), x_9(k), x_9(k-1), x_9(k-2), x_9(k-3), x_9(k-4), x_9(k-5), x_9(k-6), x_9(k-7), x_9(k-8), x_9(k-9), \\
& x_{10}(k), x_{10}(k-1), x_{10}(k-2), x_{10}(k-3), x_{10}(k-4), x_{10}(k-5), x_{10}(k-6), x_{10}(k-7), x_{10}(k-8), x_{10}(k-9), \\
& x_{11}(k), x_{11}(k-1), x_{11}(k-2), x_{11}(k-3), x_{11}(k-4), x_{11}(k-5), x_{11}(k-6), x_{11}(k-7), x_{12}(k-8), x_{12}(k-9), \\
& x_{12}(k-10), x_{12}(k-11), x_{12}(k), x_{12}(k-1), x_{12}(k-2), x_{12}(k-3), x_{12}(k-4), x_{12}(k-5), x_{12}(k-6), x_{12}(k-7), \\
& x_{12}(k-8), x_{12}(k-9), x_{14}(k-5), x_{14}(k-6), x_{14}(k-7), x_{14}(k-8), x_{14}(k-9), x_{14}(k-10), x_{14}(k-11), x_{15}(k), \\
& x_{15}(k-1), x_{15}(k-2), x_{15}(k-3), x_{15}(k-4), x_{15}(k-5), x_{15}(k-6), x_{15}(k-7), x_{15}(k-8), x_{16}(k-5), \\
& x_{16}(k-6), x_{16}(k-7), x_{16}(k-8), x_{16}(k-9), e(k), e(k-1), e(k-2)\ ]
\end{aligned}
\tag{4.1}
$$

Donde: $y = CPNL1_{RE},\ x_1 = AMM,\ x_2 = AMO,\ x_3 = BEST,\ x_4 = ENSO,\ x_5 = MEI, x_6 = NINO12,$
$x_7 = NINO3,\ x_8 = NINO34,\ x_9 = NINO4,\ x_{10} = ONI,\ x_{11} = PDO, x_{12} = QBO,\ x_{12} = SOI,\ x_{14} = TNA,$
$x_{15} = TNI,\ x_{16} = TSA$

$$
\begin{aligned}
y(k+1) = f[\ & y(k), y(k-1), y(k-2), x_1(k), x_2(k), x_2(k-1), x_2(k-2), x_2(k-3), x_2(k-4), x_2(k-5), x_2(k-6), \\
& x_2(k-7), x_2(k-8), x_2(k-9), x_2(k-10), x_2(k-11), x_3(k-9), x_3(k-10), x_3(k-11), x_4(k), x_4(k-1), \\
& x_4(k-2), x_4(k-3), x_4(k-4), x_4(k-5), x_4(k-6), x_4(k-7), x_4(k-8), x_4(k-9), x_4(k-10), x_4(k-11), \\
& e(k), e(k-1), e(k-2)\ ]
\end{aligned} \tag{4.2}
$$

Donde: $y = CPNL2_{RE},\ x_1 = AMM,\ x_2 = AMO,\ x_3 = MEI,\ x_4 = PDO$

$$
\begin{aligned}
y(k+1) = f[\ & y(k), y(k-1), y(k-2), x_1(k), x_1(k-1), x_1(k-2), x_1(k-3), x_1(k-4), x_1(k-5), x_1(k-6), x_2(k), \\
& x_2(k-1), x_2(k-2), x_2(k-3), x_2(k-4), x_3(k-4), x_3(k-5), x_3(k-6), x_3(k-7), x_3(k-8), x_4(k-2), \\
& x_4(k-3), x_4(k-4), x_4(k-5), x_5(k), x_5(k-1), x_5(k-2), x_5(k-3), x_5(k-4), x_5(k-5), x_5(k-6), \\
& x_5(k-7), x_5(k-8), x_5(k-9), x_5(k-10), x_5(k-11), x_6(k), e(k), e(k-1), e(k-2)\ ]
\end{aligned} \tag{4.3}
$$

Donde: $y = CPNL1_{ROE},\ x_1 = NINO12,\ x_2 = NINO4,\ x_3 = QBO,\ x_4 = TNA,\ x_5 = TNI,\ x_6 = TSA$

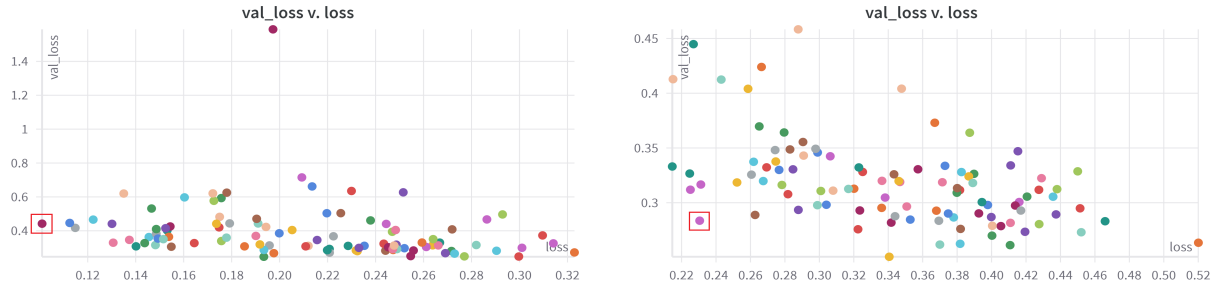
$$
\begin{aligned}
y(k+1) = f[\ & y(k), y(k-1), y(k-2), x_1(k), x_1(k-1), x_2(k), x_2(k-1), x_2(k-2), x_2(k-3), x_3(k-2), x_3(k-3), \\
& x_3(k-4), x_3(k-5), x_3(k-6), x_3(k-7), x_3(k-8), x_4(k-1), x_4(k-2), x_4(k-3), x_4(k-4), x_4(k-5), \\
& x_4(k-6), x_4(k-7), x_4(k-8), x_4(k-9), x_5(k-1), x_5(k-2), x_5(k-3), x_5(k-4), x_5(k-5), x_5(k-6), \\
& x_5(k-7), x_6(k), x_6(k-1), x_6(k-2), x_6(k-3), x_6(k-4), x_6(k-5), x_6(k-6), x_6(k-7), x_6(k-8), \\
& x_7(k), x_7(k-1), x_7(k-2), x_7(k-3), x_7(k-4), x_7(k-5), x_7(k-6), x_7(k-7), x_7(k-8), x_7(k-9), \\
& x_8(k), x_8(k-1), x_8(k-2), x_8(k-3), x_8(k-4), x_8(k-5), x_8(k-6), x_8(k-7), x_8(k-8), x_8(k-9), \\
& x_9(k), x_9(k-1), x_9(k-2), x_9(k-3), x_9(k-4), x_9(k-5), x_9(k-6), x_9(k-7), x_9(k-8), x_9(k-9), \\
& x_{10}(k), x_{10}(k-1), x_{10}(k-2), x_{10}(k-3), x_{10}(k-4), x_{10}(k-5), x_{10}(k-6), x_{10}(k-7), x_{10}(k-8), \\
& x_{11}(k-9), x_{11}(k-10), x_{11}(k-11)x_{12}(k), x_{12}(k-1), x_{12}(k-2), x_{12}(k-3), x_{12}(k-4), x_{12}(k-5), \\
& x_{12}(k-6), x_{12}(k-7), x_{13}(k), x_{13}(k-1), x_{13}(k-2), x_{13}(k-3), e(k), e(k-1), e(k-2)\ ]
\end{aligned} \tag{4.4}
$$

Donde: $y = CPNL2_{ROE},\ x_1 = AMM,\ x_2 = AMO,\ x_3 = BEST,\ x_4 = ENSO,\ x_5 = MEI,\ x_6 = NINO12,$
$x_7 = NINO3,\ x_8 = NINO34,\ x_9 = NINO4,\ x_{10} = ONI,\ x_{11} = PDO,\ x_{12} = SOI,\ x_{13} = TNA$

Por cada ecuación descrita, se entrenaron 100 modelos, usando el sweep definido anteriormente (sección 4.1). Los mejores modelos fueron seleccionados teniendo en cuenta el ECM en los datos de validación. Los datos de test fueron los últimos 36 meses del periodo de estudio (1983-2019), es decir el periodo (2019-2021), se debe recordar que los datos de test fueron separados de los datos de entrenamiento de los modelos para probar la capacidad de inferencia de cada modelo entrenado. En la Fig. 4.4 se muestran los errores obtenidos para los modelos entrenados de la R.E (CPNL1 y CPNL2).

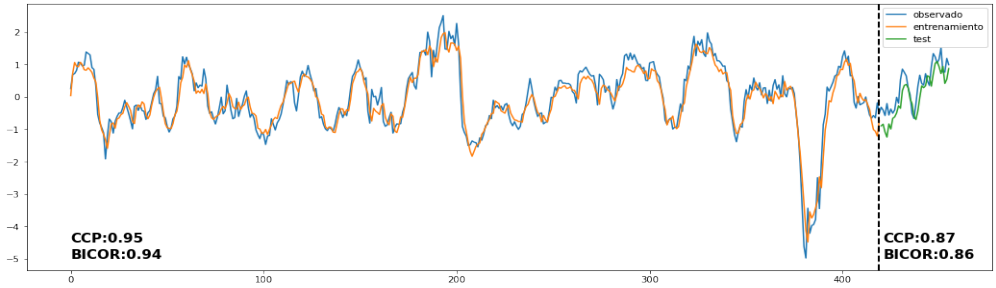
El mejor modelo para la CPNL1 obtuvo un ECM en la validación de 0.44, para la CPNL2 se obtuvo un ECM de 0.28. Para evaluar los modelos completos de la R.E se calculó el CCP y BICOR[2] como metricas de similitud entre las CPNLs observadas y el pronóstico del modelo. En la Fig. 4.5 se muestran las series de tiempo resultantes para las CPNLs observadas y el resultado de los modelos para los datos de entrenamiento y test.

[2] La correlación media biponderada (también llamada BICOR) es una medida de similitud entre muestras. Se basa en la mediana, en lugar de en la media, por lo que es menos sensible a los valores atípicos, y puede ser una alternativa robusta a otras métricas de similitud, como el CCP [76].

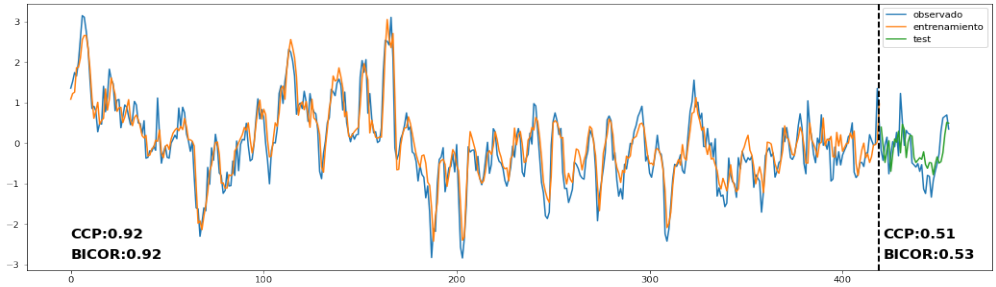


(a) CPNL1-R.E (b) CPNL2-R.E

Figura 4.4: Entrenamientos del modelo completo R.E. El mejor modelo se marca con un recuadro rojo.



(a) CPNL1-R.E

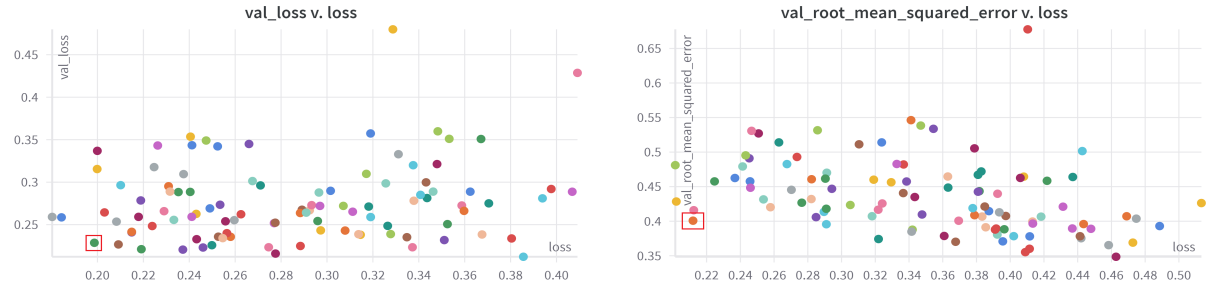


(b) CPNL2-R.E

Figura 4.5: MC - Inferencia sobre los datos de entrenamiento y test para la CPNLs de R.E.

De los resultados se puede observar similitud en el comportamiento y magnitud de las dos series de tiempo de las CPNLs observadas y pronosticadas. Esto se comprueba con un CCP de 0.95 y 0.92 para la primera y segunda CPNL (ver Fig. 4.5), incluso en los valores del test representado en los últimos 36 meses de la serie, se observa una similitud considerable para la CPNL1 con BICOR de 0.86 y una similitud moderada para la CPNL2 con BICOR de 0.53.

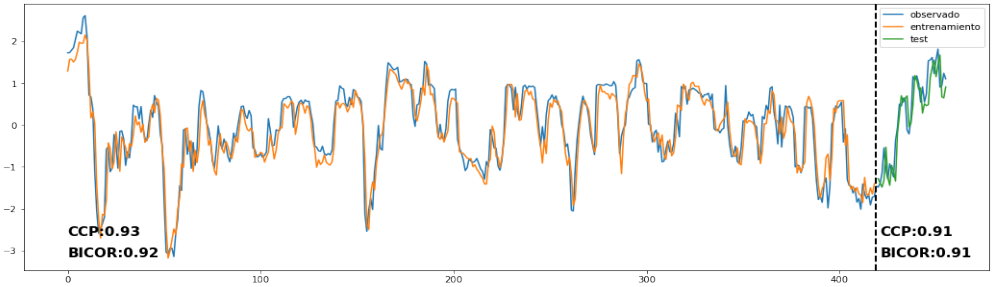
En la R.OE los diferentes entrenamientos se muestran en la Fig. 4.4. El mejor modelo para la CPNL1 obtuvo un ECM en la validación de 0.22, para la CPNL2 se obtuvo un ECM de 0.38.



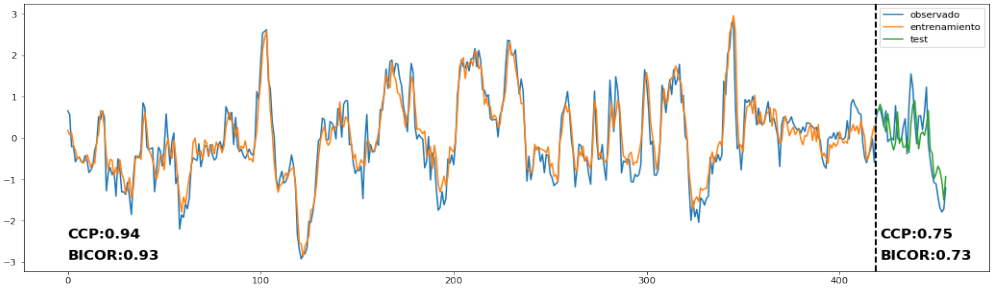
(a) CPNL1-R.OE (b) CPNL2-R.OE

Figura 4.6: Entrenamientos del modelo completo R.OE. El mejor modelo se marca con un recuadro rojo.

Evaluando los mejores modelos de la R.OE (ver Fig. 4.7), existe una similitud en el comportamiento y magnitud en las dos series de tiempo de las CPNLs observadas y pronosticadas. La CPNL1 obtuvo un CPP de 0.93 y BICOR de 0.92 en los datos de entrenamiento, en los datos de test un CPP de 0.91 y BICOR de 0.91. Estos errores son bajos y levemente mejores que la CPN1 de la R.E. Para la CPNL2 se obtuvo un CPP de 0.94 y BICOR de 0.93 en los datos de entrenamiento, en los datos de test un CPP de 0.72 y BICOR de 0.73.



(a) CPNL1-R.OE



(b) CPNL2-R.OE

Figura 4.7: MC - Inferencia sobre los datos de entrenamiento y test para la CPNLs de R.OE.

Los modelos completos muestran que las CPNL2 tienen un error más alto a comparación de las CPNL1, esto se puede atribuir a multicolinealidad de las VM y rezagos que no aportan a la buena inferencia de los modelos, estos modelos pueden mejorar en métricas de test utilizando criterios para reducir el número de VM. Comprendiendo como idea principal de un modelo de pronóstico el simplificar los predictores y tener un horizonte de pronóstico lejano.

El modelo completo es el punto de partida para entrenar los modelos simplificados, los criterios de simplificación son.

Criterio 1 - Correlación con rezago mayor: La finalidad de un regresor es tener un horizonte de pronóstico lejano, esto se traduce en la capacidad de predecir la variable objetivo lo mas alejado posible del presente. Rezagos superiores o iguales a 6 meses es un tiempo considerable para preparar los sistemas productivos agrícolas, típicamente de secano en la región, en mitigación y adaptación al evento de sequía venidero con medidas agronómicas y de gestión integrada del recurso hídrico [47].

Criterio 2 - Correlación con significancia estadística: Previamente se definió el valor mínimo desde el cual la correlación es considerada estadísticamente significativa para cada CPNL (ver sección 3.3.7). Las VM que tienen una correlación superior a 0.257 (0.377) para la CPNL1 (CPNL2) de la R.E y las VM con una correlación superior a 0.257 (0.222) para la CPNL1 (CPNL2) de la R.OE, son consideradas como posibles predictores de los modelos simplificados.

Criterio 3 - Multicolinealidad: Las VM de la R.E y la R.OE pueden presentar correlaciones altas entre ellas, esto significa que algunas VM están siendo representadas por otras. Se debe tener en cuenta que una ventaja de los modelos de regresión con RNA es que no requieren del criterio de no multicolinealidad [24], común en variables de tipo climático. Sin embargo, se decidió omitir las VM con alta relación con otras que se usan en el pronóstico, buscando simplificar el número de predictores para cada modelo.

Los criterios anteriores permiten llegar a un modelo simplificado con un horizonte de pronóstico de utilidad para el contexto de la investigación, pero se debe tener una métrica o criterio para decidir que tan desviado está un modelo simplificado del modelo completo. El criterio estadístico usado es el análisis de la varianza (ANOVA por sus siglas en inglés, *Analysis of Variance*) este test permite conocer si un modelo es similar a un modelo base (modelo completo). ANOVA[3] Se utilizó calculando el residual de los datos de test para cada modelo y se validaron los supuestos de normalidad y homogeneidad de varianzas requeridos. A continuación, se presentan los modelos simplificados, en donde se aplican los criterios mencionados.

Nota: La metodología para entrenar el modelo completo y los criterios de selección de predictores, se utilizan en adelante para el entrenamiento de todos los modelos.

4.3. Modelo simplificado-1 (MSP1)

La primera iteración para llegar a un modelo simplificado final, se realizó buscando ampliar el horizonte de pronóstico. En este caso, se entrenaron los modelos para tener un horizonte de tres meses, manteniendo la mayoría de las VM usadas en el modelo completo, pero conforme se va aumentando el horizonte algunas VM van siendo eliminadas puesto que sus rezagos significativos son de corto plazo, para el caso de CPNL2-R.E la VM AMO fue

[3] ANOVA es una técnica estadística que se utiliza para comparar las medias de diferentes grupos independientes (modelos), considerando la variabilidad de los datos al interior de cada grupo.

eliminada, para la CPNL1-R.OE la VM TSA y CPNL2-R.OE la VM AMM también fueron eliminadas. Las funciones que definen cada modelo simplificado se presentan a continuación.

$$
\begin{aligned}
y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-3), x_1(k-4), x_1(k-5), x_1(k-6), x_1(k-7), x_1(k-8), x_1(k-9), \\
& x_1(k-10), x_1(k-11), x_2(k-4), x_2(k-5), x_2(k-6), x_2(k-7), x_2(k-8), x_2(k-9), x_2(k-10), x_2(k-11), \\
& x_3(k-2), x_3(k-3), x_3(k-4), x_3(k-5), x_3(k-6), x_3(k-7), x_3(k-8), x_3(k-9), x_4(k-2), x_4(k-3), \\
& x_4(k-4), x_4(k-5), x_4(k-6), x_4(k-7), x_4(k-8), x_4(k-9), x_5(k-2), x_5(k-3), x_5(k-4), x_5(k-5), \\
& x_5(k-6), x_5(k-7), x_5(k-8), x_6(k-2), x_6(k-3), x_6(k-4), x_6(k-5), x_6(k-6), x_7(k-2), x_7(k-3), \\
& x_7(k-4), x_7(k-5), x_7(k-6), x_7(k-7), x_7(k-8), x_8(k-2), x_8(k-3), x_8(k-4), x_8(k-5), x_8(k-6), \\
& x_8(k-7), x_8(k-8), x_8(k-9), x_9(k-2), x_9(k-3), x_9(k-4), x_9(k-5), x_9(k-6), x_9(k-7), x_9(k-8), \\
& x_9(k-9), x_{10}(k-2), x_{10}(k-3), x_{10}(k-4), x_{10}(k-5), x_{10}(k-6), x_{10}(k-7), x_{10}(k-8), x_{10}(k-9), \\
& x_{11}(k-2), x_{11}(k-3), x_{11}(k-4), x_{11}(k-5), x_{11}(k-6), x_{11}(k-7), x_{12}(k-8), x_{12}(k-9), x_{12}(k-10), \\
& x_{12}(k-11), x_{13}(k-2), x_{13}(k-3), x_{13}(k-4), x_{13}(k-5), x_{13}(k-6), x_{13}(k-7), x_{13}(k-8), x_{13}(k-9), \\
& x_{14}(k-5), x_{14}(k-6), x_{14}(k-7), x_{14}(k-8), x_{14}(k-9), x_{14}(k-10), x_{14}(k-11), x_{15}(k-2), x_{15}(k-3), \\
& x_{15}(k-4), x_{15}(k-5), x_{15}(k-6), x_{15}(k-7), x_{15}(k-8), x_{16}(k-5), x_{16}(k-6), x_{16}(k-7), x_{16}(k-8), \\
& x_{16}(k-9), e(k), e(k-1), e(k-2)\,]
\end{aligned}
\tag{4.5}
$$

Donde: $y = CPNL1_{RE},\ x_1 = AMM,\ x_2 = AMO,\ x_3 = BEST,\ x_4 = ENSO,\ x_5 = MEI,\ x_6 = NINO12,$ $x_7 = NINO3,\ x_8 = NINO34,\ x_9 = NINO4,\ x_{10} = ONI,\ x_{11} = PDO,\ x_{12} = QBO,\ x_{13} = SOI,$ $x_{14} = TNA,\ x_{15} = TNI,\ x_{16} = TSA$

$$
\begin{aligned}
y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-2), x_1(k-3), x_1(k-4), x_1(k-5), x_1(k-6), x_1(k-7), x_1(k-8), \\
& x_1(k-9), x_1(k-10), x_1(k-11), x_2(k-9), x_2(k-10), x_2(k-11), x_3(k-2), x_3(k-3), x_3(k-4), x_3(k-5), \\
& x_3(k-6), x_3(k-7), x_3(k-8), x_3(k-9), x_3(k-10), x_3(k-11), e(k), e(k-1), e(k-2)\,]
\end{aligned}
\tag{4.6}
$$

Donde: $y = CPNL2_{RE},\ x_1 = AMO,\ x_2 = MEI,\ x_3 = PDO$

$$
\begin{aligned}
y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-2), x_1(k-3), x_1(k-4), x_1(k-5), x_1(k-6), x_2(k-2), x_2(k-3), \\
& x_2(k-4), x_3(k-4), x_3(k-5), x_3(k-6), x_3(k-7), x_3(k-8), x_4(k-2), x_4(k-3), x_4(k-4), x_4(k-5), \\
& x_5(k-2), x_5(k-3), x_5(k-4), x_5(k-5), x_5(k-6), x_5(k-7), x_5(k-8), x_5(k-9), x_5(k-10), x_5(k-11), \\
& e(k), e(k-1), e(k-2)\,]
\end{aligned}
\tag{4.7}
$$

Donde: $y = CPNL1_{ROE},\ x_1 = NINO12,\ x_2 = NINO4,\ x_3 = QBO,\ x_4 = TNA,\ x_5 = TNI$

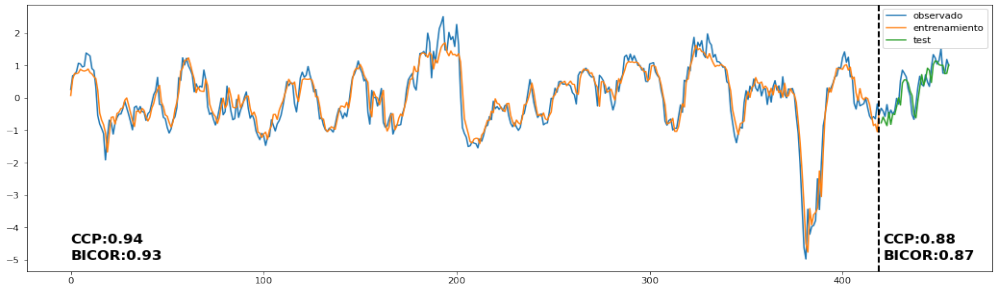
$$
\begin{aligned}
y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-2), x_1(k-3), x_2(k-2), x_2(k-3), x_2(k-4), x_2(k-5), x_2(k-6), \\
& x_2(k-7), x_2(k-8), x_3(k-2), x_3(k-3), x_3(k-4), x_3(k-5), x_3(k-6), x_3(k-7), x_3(k-8), x_3(k-9), \\
& x_4(k-2), x_4(k-3), x_4(k-4), x_4(k-5), x_4(k-6), x_4(k-7), x_5(k-2), x_5(k-3), x_5(k-4), x_5(k-5), \\
& x_5(k-6), x_5(k-7), x_5(k-8), x_6(k-2), x_6(k-3), x_6(k-4), x_6(k-5), x_6(k-6), x_6(k-7), x_6(k-8), \\
& x_6(k-9), x_7(k-2), x_7(k-3), x_7(k-4), x_7(k-5), x_7(k-6), x_7(k-7), x_7(k-8), x_7(k-9), x_8(k-2), \\
& x_8(k-3), x_8(k-4), x_8(k-5), x_8(k-6), x_8(k-7), x_8(k-8), x_8(k-9), x_9(k-2), x_9(k-3), x_9(k-4), \\
& x_9(k-5), x_9(k-6), x_9(k-7), x_9(k-8), x_{10}(k-9), x_{10}(k-10), x_{10}(k-11), x_{11}(k-2), x_{11}(k-3), \\
& x_{11}(k-4), x_{11}(k-5), x_{11}(k-6), x_{11}(k-7), x_{12}(k-2), x_{12}(k-3), e(k), e(k-1), e(k-2)\,]
\end{aligned}
\tag{4.8}
$$

Donde: $y = CPNL2_{ROE},\ x_1 = AMO,\ x_2 = BEST,\ x_3 = ENSO,\ x_4 = MEI,\ x_5 = NINO12,$ $x_6 = NINO3,\ x_7 = NINO34,\ x_8 = NINO4,\ x_9 = ONI,\ x_{10} = PDO,\ x_{11} = SOI,\ x_{12} = TNA$

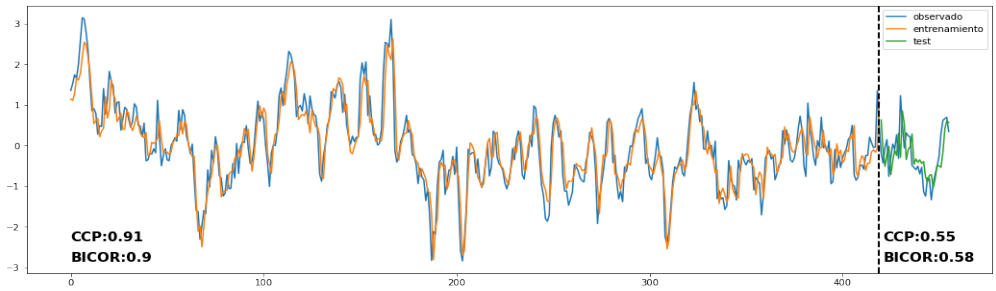
Los mejores modelos fueron seleccionados a partir de 100 entrenamientos por cada ecuación descrita, para visualizar los sweeps ejecutados se puede consultar el informe[4] generado para los modelos simplificados. El mejor modelo para CPNL1-R.E obtuvo un ECM en la validación de 0.25, para CPNL2-R.E obtuvo un ECM de 0.26. Adicionalmente, el mejor

[4] `https://wandb.ai/cefernal/Regressor/reports/Untitled-Report--VmlldzozMDgzODE5?accessToken=rtbo39nkwv1til1fqbwr78gpess2ws7z2o21kstxdybg4tss6c0n9gng6yf4ylxu`

modelo para la CPNL1-R.OE obtuvo un ECM de 0.3 y para CPNL2-R.OE un ECM de 0.18. Las series de tiempo resultantes para las CPNLs observadas y el resultado de los modelos para los datos de entrenamiento y test se muestran en la Fig. 4.8 para la R.E y la R.OE en la Fig. 4.9.

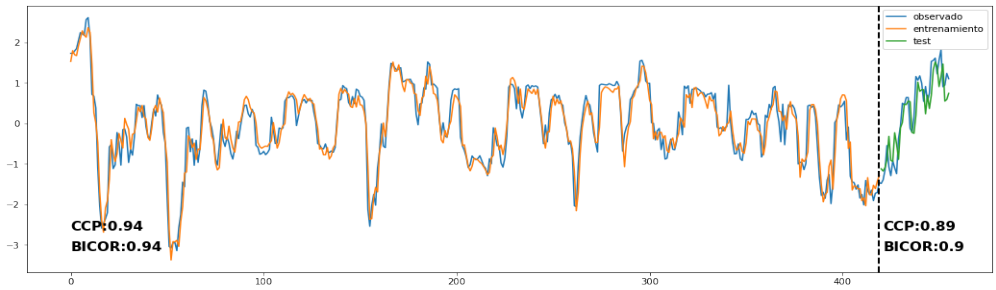


(a) CPNL1-R.E

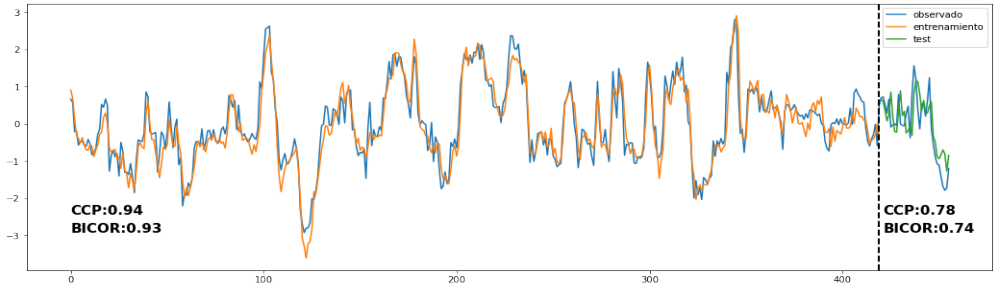


(b) CPNL2-R.E

Figura 4.8: MSP1 - Inferencia sobre los datos de entrenamiento y test para la CPNLs de R.E.



(a) CPNL1-R.OE



(b) CPNL2-R.OE

Figura 4.9: MSP1 - Inferencia sobre los datos de entrenamiento y test para la CPNLs de R.OE.

Los resultados obtenidos para la CPNL1-R.E indican un CPP de 0.93 y BICOR de 0.92 en los datos de entrenamiento, los datos de test con un CPP de 0.87 y BICOR de 0.86. Para CPNL2-R.E se calculó un CPP de 0.92 y BICOR de 0.92 en los datos de entrenamiento, para el test un CPP de 0.55 y BICOR de 0.58. En comparación con el modelo completo para la CPNL1-R.E el modelo simplificado-1 tiene mejores métricas pero con una diferencia en el test usando BICOR de 0.01; es decir, los dos modelos son prácticamente iguales. Lo mismo ocurre con la CPNL2-R.E tiene mejores métricas pero con una diferencia en el test usando BICOR de 0.02, lo cual no es una diferencia significativa.

En la R.OE los resultados obtenidos son similares a la R.E. Para la CPNL1-R.OE se obtuvo un CPP de 0.92 y BICOR de 0.92 en los datos de entrenamiento, para el test un CPP de 0.55 y BICOR de 0.58. La CPNL2-R.E obtuvo un CPP de 0.91 y BICOR de 0.9 en los datos de entrenamiento, para el test un CPP de 0.55 y BICOR de 0.58. Al igual que la R.E existe una mejora en las métricas, pero no es significativa.

4.3.1. ANOVA del MSP1

Para determinar si el modelo simplificado es estadísticamente igual al modelo completo. Es decir, la media del residual es similar entre el modelo simplificado y el modelo completo, se calculó ANOVA para los 8 modelos entrenados (4 modelos completos y 4 modelos simplificados). La interpretación de la prueba se hace evaluando el *p-value* y determinando cual de las hipótesis se cumple.

- Hipótesis nula (H_0): la media del residual es la misma en los diferentes modelos, los modelos son similares.
- Hipótesis alternativa (H_a): al menos dos medias difieren de forma significativa, algún modelo no es similar.

Para un *p-value* menor o igual al nivel de significancia de 0.05 ($p \leq 0.05$) se debe considerar la H_a y para un *p-value* mayor al nivel de significancia de 0.05 ($p > 0.05$) se debe considerar la H_0. La finalidad de entrenar modelos simplificados es tener similitud con el modelo completo, por lo tanto es ideal que siempre se considere la H_0. El cálculo de ANOVA se realizó para cada región, para los modelos de la R.E, ANOVA arrojó un *p-value* de 0.182 lo que indica que se debe considerar la H_0. En la R.OE, ANOVA arrojó un *p-value* de 0.115 indicando que también se debe considerar la H_0. Como conclusión los modelos simplificados-1 son estadísticamente iguales a los modelos completos.

4.4. Modelo simplificado-2 (MSP2)

La segunda iteración para llegar al modelo simplificado final, se realizó con el fin de ampliar el horizonte de pronóstico para cumplir el criterio 1 y criterio 2, detallados anteriormente. En esta simplificación, el horizonte de pronóstico se amplió hasta seis, debido al incremento en el modelo CPNL1-R.OE las VM NINO4 y TNA fueron eliminadas, para el

modelo CPNL2-R.OE las VM AMO y TNA también se eliminaron. Los modelos entrenados en esta iteración son expresados en las siguientes ecuaciones.

$$
\begin{aligned}
y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-5), x_1(k-6), x_1(k-7), x_1(k-8), x_1(k-9), x_1(k-10), x_1(k-11), \\
& x_2(k-5), x_2(k-6), x_2(k-7), x_2(k-8), x_2(k-9), x_2(k-10), x_2(k-11), x_3(k-5), x_3(k-6), x_3(k-7), \\
& x_3(k-8), x_3(k-9), x_4(k-5), x_4(k-6), x_4(k-7), x_4(k-8), x_4(k-9), x_5(k-5), x_5(k-6), x_5(k-7), \\
& x_5(k-8), x_6(k-5), x_6(k-6), x_7(k-5), x_7(k-6), x_7(k-7), x_7(k-8), x_8(k-5), x_8(k-6), x_8(k-7), \\
& x_8(k-8), x_8(k-9), x_9(k-5), x_9(k-6), x_9(k-7), x_9(k-8), x_9(k-9), x_{10}(k-5), x_{10}(k-6), x_{10}(k-7), \\
& x_{10}(k-8), x_{10}(k-9), x_{11}(k-5), x_{11}(k-6), x_{11}(k-7), x_{12}(k-8), x_{12}(k-9), x_{12}(k-10), x_{12}(k-11), \\
& x_{13}(k-5), x_{13}(k-6), x_{13}(k-7), x_{13}(k-8), x_{13}(k-9), x_{14}(k-5), x_{14}(k-6), x_{14}(k-7), x_{14}(k-8), \\
& x_{14}(k-9), x_{14}(k-10), x_{14}(k-11), x_{15}(k-5), x_{15}(k-6), x_{15}(k-7), x_{15}(k-8), x_{16}(k-5), x_{16}(k-6), \\
& x_{16}(k-7), x_{16}(k-8), x_{16}(k-9), e(k), e(k-1), e(k-2)\,]
\end{aligned}
\tag{4.9}
$$

Donde: $y = CPNL1_{RE},\ x_1 = AMM,\ x_2 = AMO,\ x_3 = BEST,\ x_4 = ENSO,\ x_5 = MEI,\ x_6 = NINO12,\ x_7 = NINO3,\ x_8 = NINO34,\ x_9 = NINO4,\ x_{10} = ONI,\ x_{11} = PDO,\ x_{12} = QBO,\ x_{13} = SOI,\ x_{14} = TNA,\ x_{15} = TNI,\ x_{16} = TSA$

$$
\begin{aligned}
y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-5), x_1(k-6), x_1(k-7), x_1(k-8), x_1(k-9), x_1(k-10), x_1(k-11), \\
& x_2(k-9), x_2(k-10), x_2(k-11), x_3(k-5), x_3(k-6), x_3(k-7), x_3(k-8), x_3(k-9), x_3(k-10), \\
& x_3(k-11), e(k), e(k-1), e(k-2)\,]
\end{aligned}
\tag{4.10}
$$

Donde: $y = CPNL2_{RE},\ x_1 = AMO,\ x_2 = MEI,\ x_3 = PDO$

$$
\begin{aligned}
y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-5), x_1(k-6), x_2(k-5), x_2(k-6), x_2(k-7), x_2(k-8), x_3(k-5), \\
& x_3(k-6), x_3(k-7), x_3(k-8), x_3(k-9), x_3(k-10), x_3(k-11), e(k), e(k-1), e(k-2)\,]
\end{aligned}
\tag{4.11}
$$

Donde: $y = CPNL1_{ROE},\ x_1 = NINO12,\ x_2 = QBO,\ x_3 = TNI$

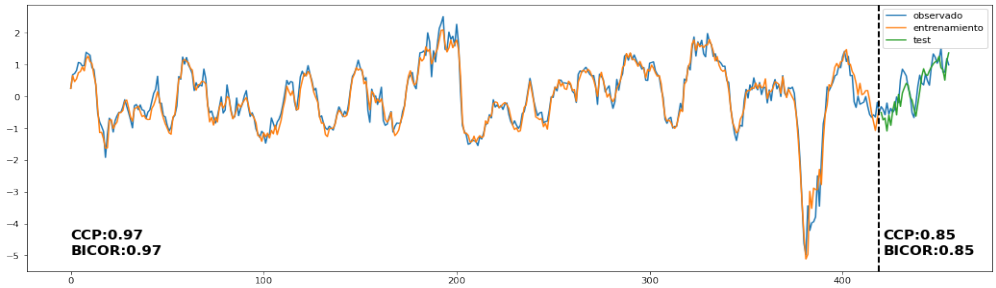
$$
\begin{aligned}
y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-5), x_1(k-6), x_1(k-7), x_1(k-8), x_2(k-5), x_2(k-6), x_2(k-7), \\
& x_2(k-8), x_2(k-9), x_3(k-5), x_3(k-6), x_3(k-7), x_4(k-5), x_4(k-6), x_4(k-7), x_4(k-8), x_5(k-5), \\
& x_5(k-6), x_5(k-7), x_5(k-8), x_5(k-9), x_6(k-5), x_6(k-6), x_6(k-7), x_6(k-8), x_6(k-9), x_7(k-5), \\
& x_7(k-6), x_7(k-7), x_7(k-8), x_7(k-9), x_8(k-5), x_8(k-6), x_8(k-7), x_8(k-8), x_9(k-9), x_9(k-10), \\
& x_9(k-11), x_{10}(k-5), x_{10}(k-6), x_{10}(k-7), e(k), e(k-1), e(k-2)\,]
\end{aligned}
\tag{4.12}
$$

Donde: $y = CPNL2_{ROE},\ x_1 = BEST,\ x_2 = ENSO,\ x_3 = MEI,\ x_4 = NINO12,\ x_5 = NINO3,\ x_6 = NINO34,\ x_7 = NINO4,\ x_8 = ONI,\ x_9 = PDO,\ x_{10} = SOI$

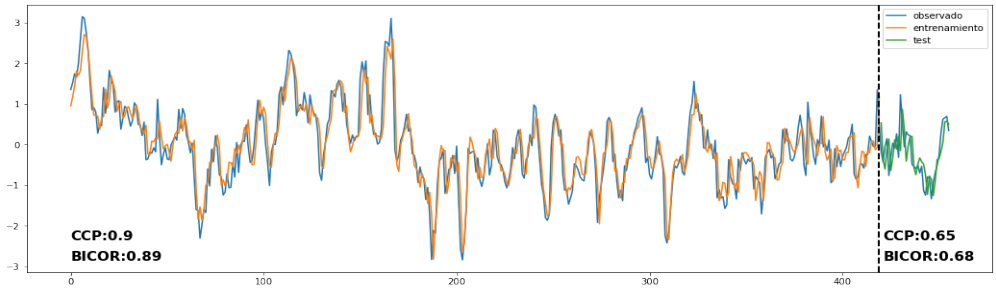
De los 100 entrenamientos realizados por cada modelo descrito, se seleccionaron los cuatro mejores correspondientes a las dos regiones y las dos CPNLs, los sweeps definidos y los modelos se pueden observar en el informe generado[5]. Los mejores modelos fueron: CPNL1-R.E con ECM de 0.34, CPNL2-R.E con ECM de 0.27, CPNL1-R.OE con ECM de 0.27 y el CPNL2-R.OE con 0.18. Las series de tiempo resultantes para las CPNLs observadas y el resultado de los modelos para los datos de entrenamiento y test se muestran en la Fig. 4.10 para la R.E y la R.OE en la Fig. 4.11.

Los resultados obtenidos para la R.E y la R.OE, muestran un grado de similitud en comportamiento y magnitud de las series de tiempo observadas y pronosticadas. Los resultados obtenidos para la CPNL1-R.E indican un CPP de 0.97 y BICOR de 0.97 en los datos de entrenamiento, los datos de test obtuvieron un CPP de 0.9 y BICOR de 0.89. La

[5] https://wandb.ai/cefernal/Regressor/reports/Untitled-Report--VmlldzozMDgzODE5?accessToken=rtbo39nkwv1til1fqbwr78gpess2ws7z2o21kstxdybg4tss6c0n9gng6yf4ylxu

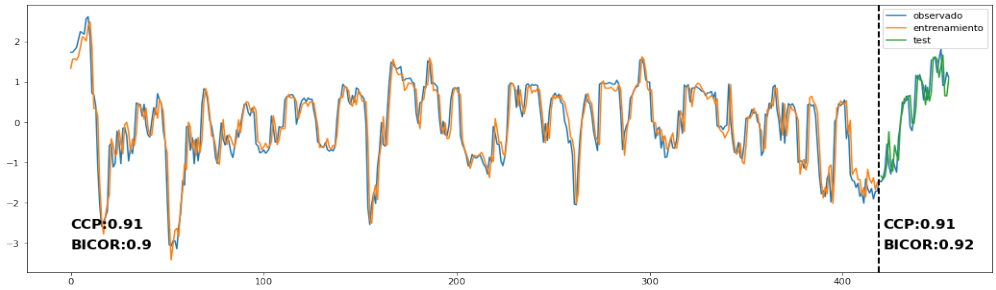


(a) CPNL1-R.E

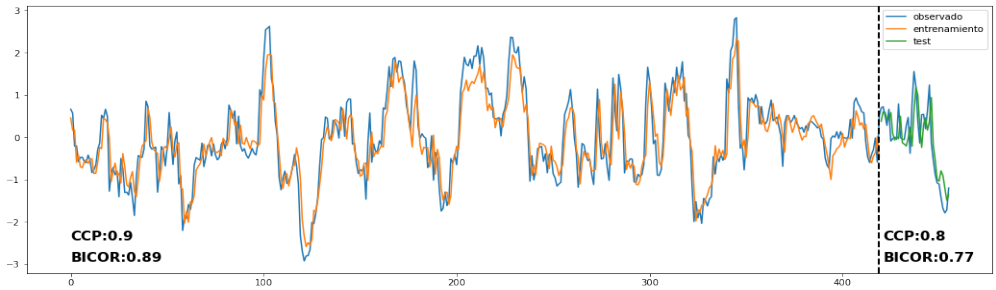


(b) CPNL2-R.E

Figura 4.10: MSP2 - Inferencia sobre los datos de entrenamiento y test para la CPNLs de R.E.



(a) CPNL1-R.OE



(b) CPNL2-R.OE

Figura 4.11: MSP2 - Inferencia sobre los datos de entrenamiento y test para la CPNLs de R.OE.

CPNL2-R.E obtuvo un CPP de 0.9 y BICOR de 0.89 en los datos de entrenamiento, para el test un CPP de 0.65 y BICOR de 0.68. En comparación con el modelo completo la mejora significativa se obtuvo en la CPNL2-R.E, en el modelo completo se obtuvo en BICOR de 0.53 y en modelo simplificado un BICOR de 0.65, es notable el incremento en la precisión y mejora para el modelo.

En la R.OE los modelos obtenidos fueron mejores respecto al modelo completo, pero no

con aumento significativo, en la mayoría de casos el incremento en los datos de test teniendo en cuenta el BICOR fue de 0.02. Los resultados en la CPNL1-R.OE indican un CPP de 0.91 y BICOR de 0.9 en los datos de entrenamiento, los datos de test obtuvieron un CPP de 0.91 y BICOR de 0.92. En la CPNL2-R.OE se obtuvo un CPP de 0.9 y BICOR de 0.89 en los datos de entrenamiento, en los datos de test un CPP de 0.8 y BICOR de 0.77.

Comparando la primera iteración (MSP1) y la segunda iteración (MSP2) los modelos resultantes en MSP2 tienen mejores resultados, lo destacable es el modelo CPNL2-R.E que tuvo un incremento comparando el BICOR del modelo completo de 0.15 en los datos de test. Con este incremento se espera que ANOVA arroje como conclusión que los modelos simplificados sean iguales estadísticamente al modelo completo.

4.4.1. ANOVA del MSP2

Para determinar si el modelo simplificado es estadísticamente igual al modelo completo; es decir, la media del residual es similar entre el modelo simplificado y el modelo completo, al igual que MSP1 se calculó ANOVA comparando el MC con MSP2. En a R.E, ANOVA arrojó un *p-value* de 0.538 lo que indica que se debe considerar la H_0. En la R.OE, ANOVA arrojó un *p-value* de 0.066 indicando que se debe considerar también la H_0. Como conclusión los modelos simplificados-2 son estadísticamente iguales a los modelos completos.

4.5. Modelo simplificado-3 (MSP3)

Teniendo presente que el MSP2 es equivalente al MC, se puede seguir iterando para conseguir un modelo simplificado más compacto y cumplir con el criterio 3 (multicolinealidad de VM). Se analizaron las VM usadas en el MSP2 para determinar si están siendo representadas por otras VM dentro del modelo. El análisis consistió en calcular el factor de inflación de varianza[6] (VIF, por sus siglas en inglés) que es una medida de la cantidad de multicolinealidad en un análisis de regresión. La multicolinealidad crea un problema en el modelo de regresión porque todas las entradas del modelo (predictores) influyen entre sí. Por lo tanto, en realidad no son independientes y es difícil probar cuánto afecta la combinación los predictores (VMs) a la variable predecida (CPNLs), dentro del modelo de regresión. Para determinar el grado de relación y eliminar algunas VM, se tuvieron en cuenta los siguientes criterios [95].

- VIF igual a 1 = las variables no están correlacionadas.
- VIF entre 1 y 5 = las variables están moderadamente correlacionadas.
- VIF mayor que 5 = las variables están altamente correlacionadas.

En la Tabla 4.2, se calculó VIF el conjunto de VM de cada modelo. Las VM que se tuvieron en cuenta para el MSP3 fueron aquellas con valor entre 1 y 5.

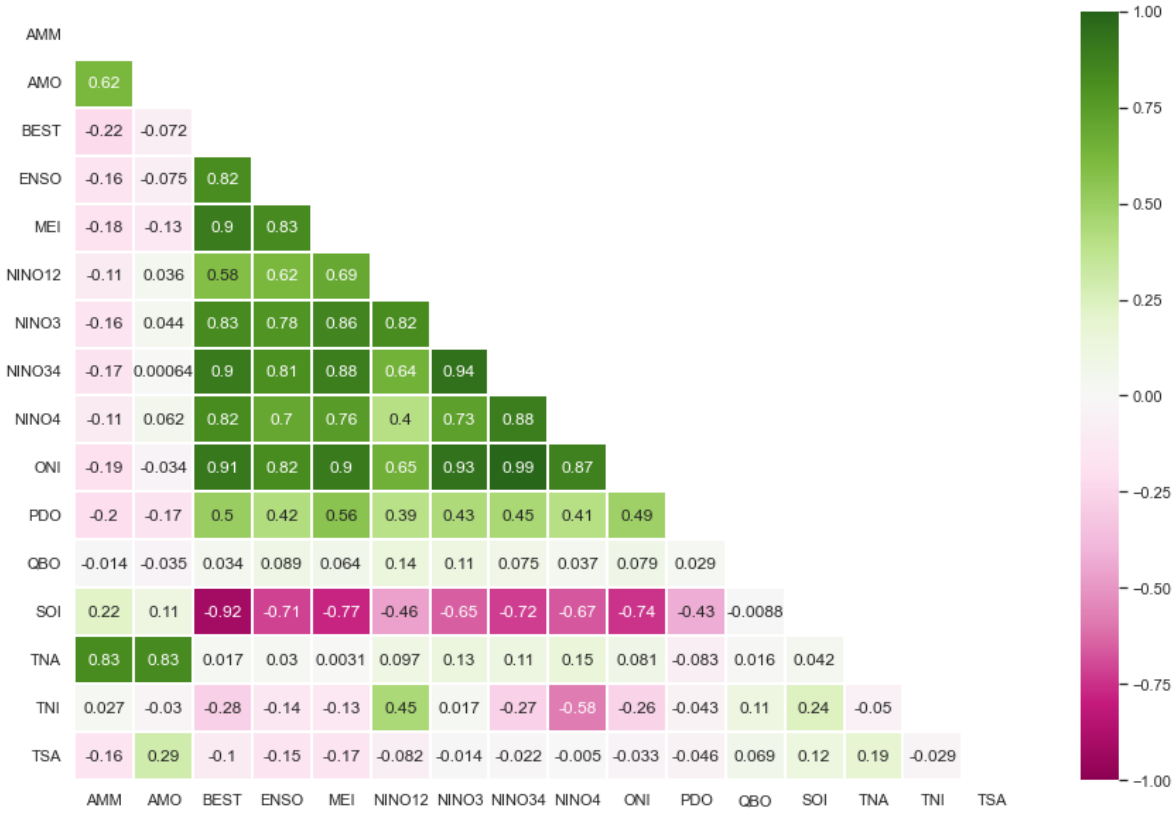
[6] La implementación y analisis se desarrolló con base en `https://www.reneshbedre.com/blog/variance-inflation-factor.html`

Para contrastar los resultados arrojados por VIF se calculó la matriz de correlación de las VM y se comprobó que tuvieran una correlación baja. En la Fig. 4.12, se presenta la matriz de correlación calculada.

Tabla 4.2: Factor de inflación de la varianza.

CPNL1-R.E		**CPNL2-R.E**		**CPNL1-R.OE**		**CPNL2-R.OE**	
VM	**VIF**	**VM**	**VIF**	**VM**	**VIF**	**VM**	**VIF**
AMM	8.81	AMO	1.54	NINO12	1.26	BEST	28.31
AMO	4.40	MEI	1.45	QBO	1.02	ENSO	3.73
BEST	29.25	PDO	1.47	TNI	1.25	MEI	8.71
ENSO	3.78					NINO12	7.65
MEI	9.43					NINO3	49.80
NINO12	13.29					NINO34	124.25
NINO3	55.00					NINO4	9.02
NINO34	128.82					ONI	74.25
NINO4	17.23					PDO	1.61
ONI	77.65					SOI	9.96
PDO	1.65						
QBO	1.05						
SOI	10.07						
TNA	13.15						
TNI	8.95						
TSA	2.19						

Figura 4.12: Correlación sincrónica entre las variables macroclimáticas.



Como resultado final se obtuvo un número bajo de predictores con poca colinealidad, en comparación con el modelo completo. La cantidad de predictores para el modelo CPNL1-R.E fue de 5, para los modelos CPNL2-R.E, CPNL1-R.OE y CPNL2-R.OE fueron 3 predictores. Una cantidad baja puesto que inicialmente se consideraron 18 VM. Los predictores seleccionados para el MSP3 se listan a continuación.

- CPNL1-R.E: AMO, ENSO, PDO, QBO y TSA.
- CPNL2-R.E: AMO, MEI y PDO.
- CPNL1-R.OE: NINO12, QBO y TNI.
- CPNL1-R.OE: ENSO, MEI y PDO.

Los modelos entrenados en esta iteración son expresados en las siguientes ecuaciones.

$$\begin{aligned} y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-5), x_1(k-6), x_1(k-7), x_1(k-8), x_1(k-9), x_1(k-10), x_1(k-11), \\ & x_2(k-5), x_2(k-6), x_2(k-7), x_2(k-8), x_2(k-9), x_3(k-5), x_3(k-6), x_3(k-7), x_4(k-8), x_4(k-9), \\ & x_4(k-10), x_4(k-11), x_5(k-5), x_5(k-6), x_5(k-7), x_5(k-8), x_5(k-9), e(k), e(k-1), e(k-2)\,] \\ & \textbf{Donde: } y = CPNL1_{RE},\ x_1 = AMO,\ x_2 = ENSO,\ x_3 = PDO,\ x_4 = QBO,\ x_5 = TSA \end{aligned} \tag{4.13}$$

$$\begin{aligned} y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-5), x_1(k-6), x_1(k-7), x_1(k-8), x_1(k-9), x_1(k-10), x_1(k-11), \\ & x_2(k-9), x_2(k-10), x_2(k-11), x_3(k-5), x_3(k-6), x_3(k-7), x_3(k-8), x_3(k-9), x_3(k-10), \\ & x_3(k-11), e(k), e(k-1), e(k-2)\,] \\ & \textbf{Donde: } y = CPNL2_{RE},\ x_1 = AMO,\ x_2 = MEI,\ x_3 = PDO \end{aligned} \tag{4.14}$$

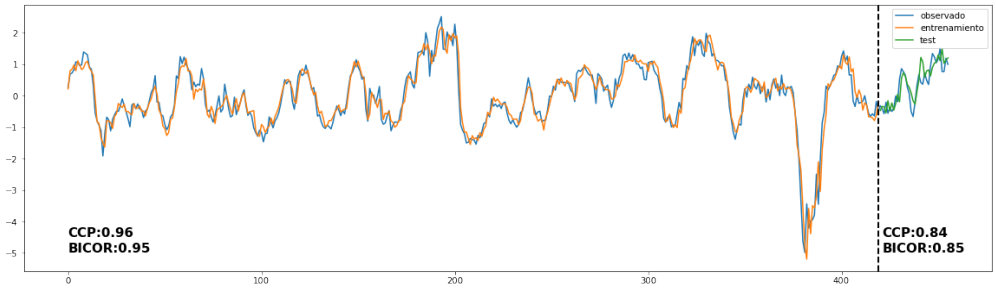
$$\begin{aligned} y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-5), x_1(k-6), x_2(k-5), x_2(k-6), x_2(k-7), x_2(k-8), x_3(k-5), \\ & x_3(k-6), x_3(k-7), x_3(k-8), x_3(k-9), x_3(k-10), x_3(k-11), e(k), e(k-1), e(k-2)\,] \\ & \textbf{Donde: } y = CPNL1_{ROE},\ x_1 = NINO12,\ x_2 = QBO,\ x_3 = TNI \end{aligned} \tag{4.15}$$

$$\begin{aligned} y(k+1) = f[\, & y(k), y(k-1), y(k-2), x_1(k-5), x_1(k-6), x_1(k-7), x_1(k-8), x_1(k-9), x_2(k-5), x_2(k-6), \\ & x_2(k-7), x_3(k-9), x_3(k-10), x_3(k-11), e(k), e(k-1), e(k-2)\,] \\ & \textbf{Donde: } y = CPNL2_{ROE},\ x_1 = ENSO,\ x_2 = MEI,\ x_3 = PDO \end{aligned} \tag{4.16}$$

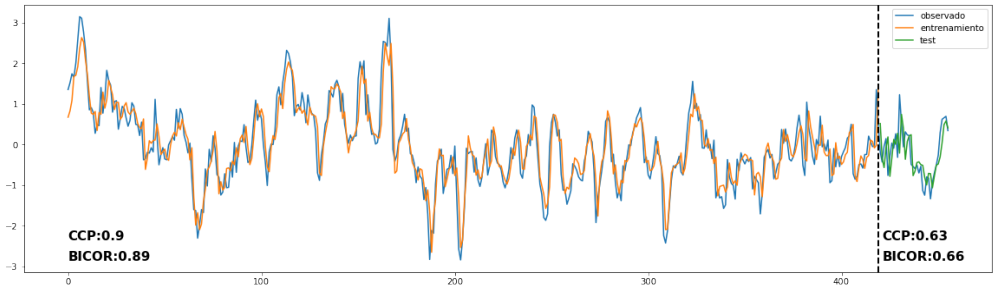
Al igual que las anteriores iteraciones de 100 entrenamientos realizados por cada modelo descrito, se seleccionaron los cuatro mejores correspondientes a las dos regiones y las dos CPNLs, los sweeps definidos y los modelos se pueden observar en el informe generado[7]. Los modelos resultantes fueron: CPNL1-R.E con ECM de 0.31, CPNL2-R.E con ECM de 0.25, CPNL1-R.OE con ECM de 0.26 y el CPNL2-R.OE con 0.11. Las series de tiempo resultantes para las CPNLs observadas y el resultado de los modelos para los datos de entrenamiento y test se muestran en la Fig. 4.10 para la R.E y la R.OE en la Fig. 4.11.

Los resultados obtenidos para la R.E y la R.OE, muestran similitud en comportamiento y magnitud de las series de tiempo observadas y pronosticadas. la CPNL1-R.E indican un CPP de 0.96 y BICOR de 0.95 en los datos de entrenamiento, los datos de test obtuvieron un CPP de 0.84 y BICOR de 0.85. La CPNL2-R.E obtuvo un CPP de 0.9 y BICOR de 0.89 en los datos de entrenamiento, en el test un CPP de 0.63 y BICOR de 0.66. En comparación

[7] `https://wandb.ai/cefernal/Regressor/reports/Untitled-Report--VmlldzozMDgzODE5?accessToken=rtbo39nkwv1til1fqbwr78gpess2ws7z2o21kstxdybg4tss6c0n9gng6yf4ylxu`

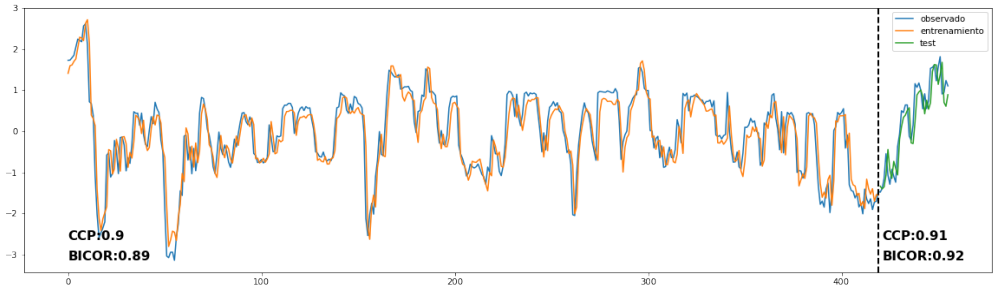


(a) CPNL1-R.E

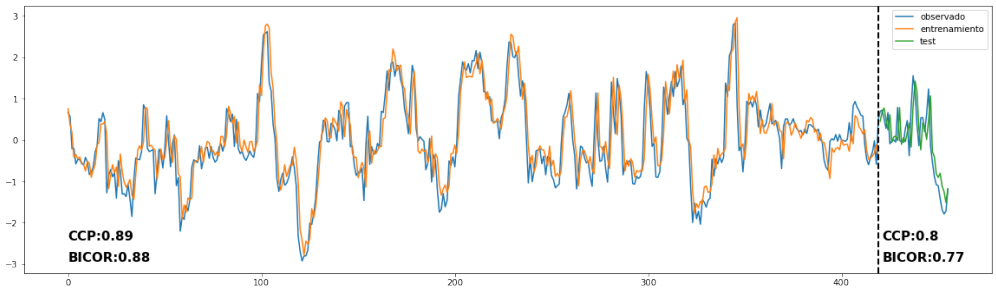


(b) CPNL2-R.E

Figura 4.13: MSP3 - Inferencia sobre los datos de entrenamiento y test para la CPNLs de R.E.



(a) CPNL1-R.OE



(b) CPNL2-R.OE

Figura 4.14: MSP3 - Inferencia sobre los datos de entrenamiento y test para la CPNLs de R.OE.

con el MC los MSP3 de la R.E, tienen una leve desviación no significativa de 0.01 (BICOR) que se puede atribuir a la reducción de predictores realizada.

En la R.OE las series obtenidas determinaron para la CPNL1-R.OE un CPP de 0.96 y BICOR de 0.95 en los datos de entrenamiento, para el test un CPP de 0.84 y BICOR de 0.85. La CPNL2-R.E obtuvo un CPP de 0.9 y BICOR de 0.89 en los datos de entrenamiento, en el test un CPP de 0.63 y BICOR de 0.66.

Comparando la segunda y tercera iteración, los modelos resultantes en MSP3 tienen un desempeño igual al MSP2, lo destacable se resume en la disminución de los predictores (VM) usados. Esto se traduce en la menor cantidad de información exógena requerida para realizar un pronóstico.

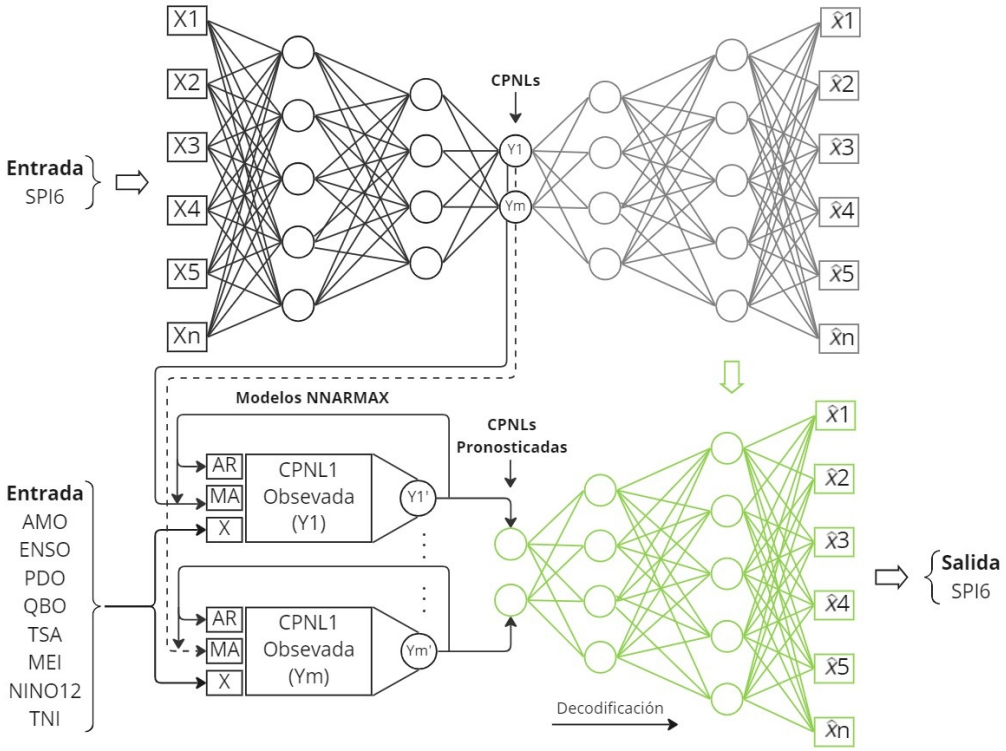
4.5.1. ANOVA del MSP3

Para determinar la similitud los MSP3 y los MC, se calculó ANOVA entre los modelos resultantes y los modelos completos. En a R.E, ANOVA arrojó un *p-value* de 0.692 que indica la similitud entre los modelos. En la R.OE, ANOVA arrojó un *p-value* de 0.150 indicando que se debe considerar también la H_0. Como conclusión los MSP3 son estadísticamente iguales a los MC.

4.6. Modelo final

El modelo final es el resultado de la unión de la arquitectura de ACPNL y los regresores de cada CPNL. Es decir, los cuatro modelos NNARMAX que realizan el pronóstico de cada CPNL de la R.E y R.OE. De las tres iteraciones previas, los MSP3 cumple los criterios de simplificación definidos previamente (ver sección 4.2). Por lo tanto los MSP3 se definen como los regresores que conforman el modelo final de pronóstico del SPI.

Figura 4.15: Modelo final de pronóstico del SPI6.



El modelo final se detalla en la Fig. 4.15 en donde la entrada inicial es el SPI6, luego el SPI6 es reducido en dimensionalidad espacial (ver sección 3.3.6); las CPNLs y VM son utilizadas como entradas en los modelos NNARMAX, como salida de los modelos se

obtiene las CPNLs pronosticadas; con las CPNLs se utiliza la sección de decodificación de la arquitectura ACPNL, para realizar la transformación inversa y obtener el SPI6 pronosticado para la R.E o la R.OE.

4.6.1. Evaluación del pronóstico

La evaluación del pronóstico se realizó al comparar gráficamente las series de tiempo pronosticadas y observadas de la CPNL de la R.E (Fig. 4.13) y la CPNL de la R.OE (Fig. 4.14), se complementó la comparación con los resultados de CCP y BICOR, en la sección 4.5 se presentó el análisis para el pronóstico de las CPNLs en las dos regiones. Partiendo de ese análisis anterior, se aplicó el ACPNL inverso para obtener las series de tiempo SPI6 pronosticadas en la distribución espacial de la R.E y la R.OE.

Las series SPI6 observadas y pronosticadas, se promediaron y compararon gráficamente. También, se calcularon metricas de similitud y error (CPP, BICOR y ECM); aunque el SPI es adimensional, es posible interpretar la magnitud del ECM del pronóstico. En la Fig. 4.16 se muestran las series SPI6 de la R.E y en la Fig. 4.17 las series SPI6 de la R.OE.

Figura 4.16: Comparación series de tiempo SPI6 reales y estimadas - R.E.

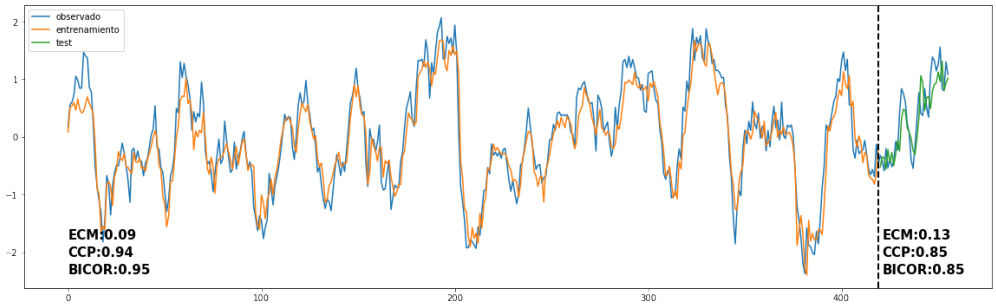
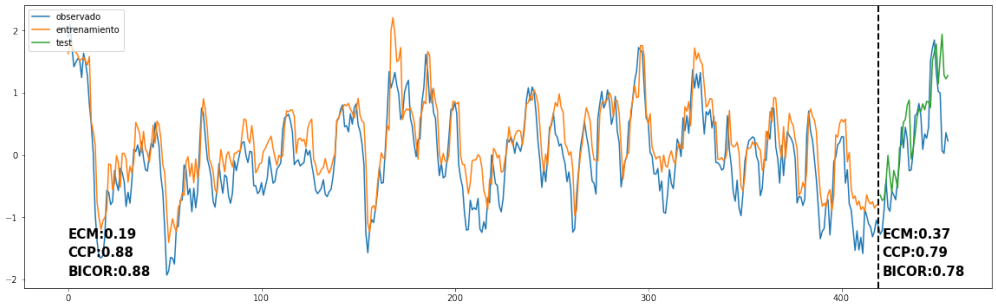


Figura 4.17: Comparación series de tiempo SPI6 reales y estimadas - R.OE.



De los resultados mostrados, se presenta un mejor desempeño temporal del pronóstico en la R.E, respecto a la R.OE.

- La R.E presenta en los datos de entrenamiento un ECM de 0.09, CPP de 0.94 y BICOR de 0.95, en el test un ECM de 0.13, CPP de 0.85 y BICOR de 0.85.
- La R.OE presenta en los datos de entrenamiento un ECM de 0.19, CPP de 0.88 y BICOR de 0.88, en el test un ECM de 0.37, CPP de 0.79 y BICOR de 0.78.

La diferencia en el desempeño de pronóstico entre ambas regiones se evidencia en los últimos 36 meses del test. Gráficamente se ven algunas diferencias importantes en los últimos meses de la serie de tiempo en la R.OE, pero se muestra una tendencia visible que trata de seguir los datos observados. La R.OE presenta un menor desempeño pero las metricas no son malas, están por encima de 0.7 para BICOR y CPP.

Finalmente, se evaluó el desempeño del pronóstico en términos espaciales según calculando el ECM (ver Fig. 4.18), el CCP (ver Fig. 4.19) y BICOR (ver Fig. 4.20), para el periodo de tiempo de los datos de entrenamiento (1983-2019) y para los datos de test (2019-2021).

Figura 4.18: Distribución espacial del ECM en el pronóstico del SPI6. Izquierda serie de tiempo entrenamiento, derecha la serie de tiempo test.

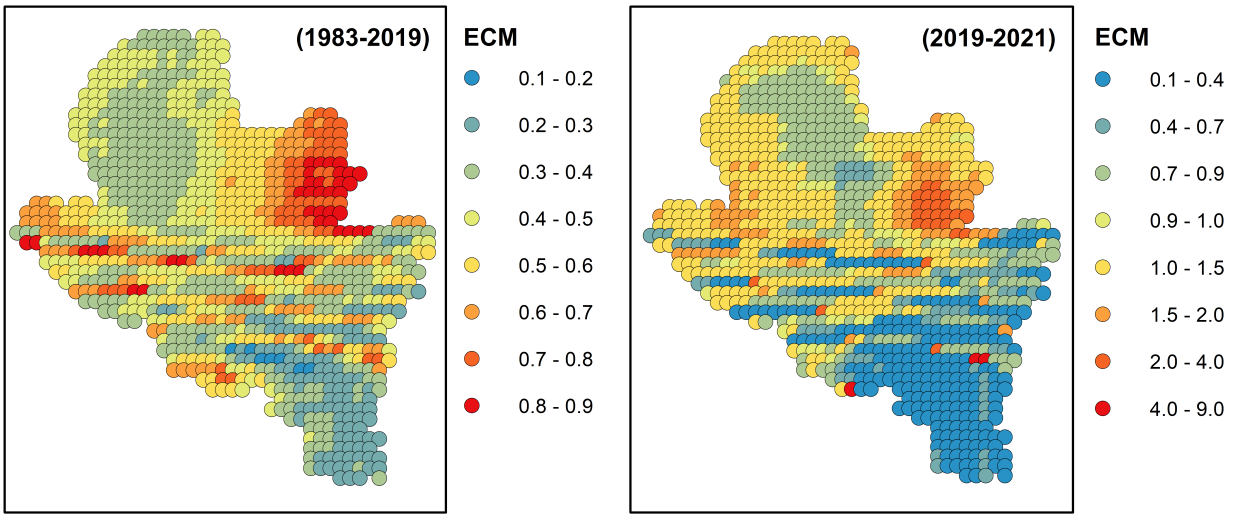


Figura 4.19: Distribución espacial del CCP en el pronóstico del SPI6. Izquierda serie de tiempo entrenamiento, derecha la serie de tiempo test.

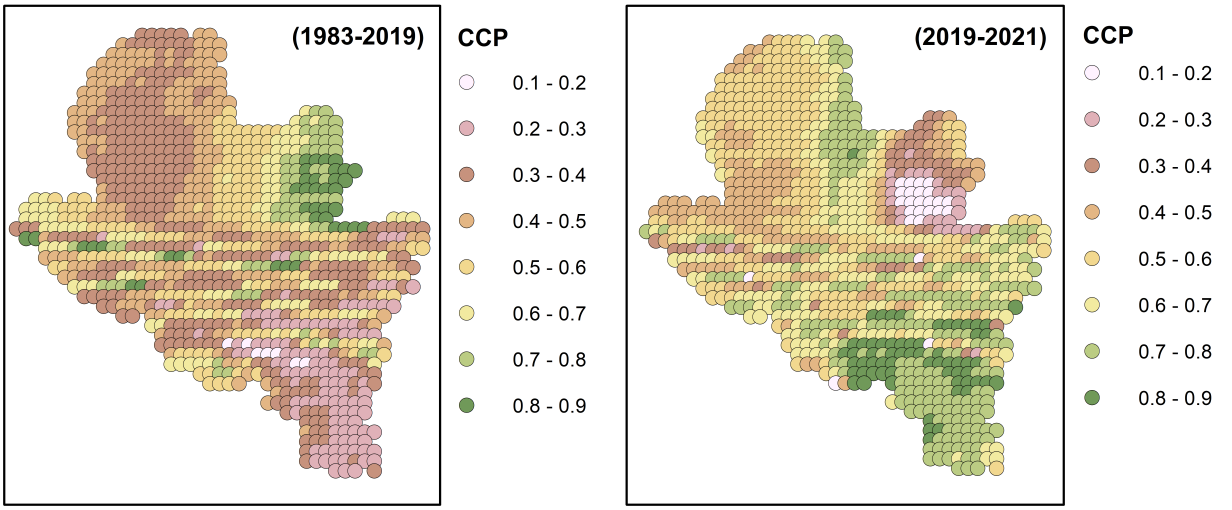
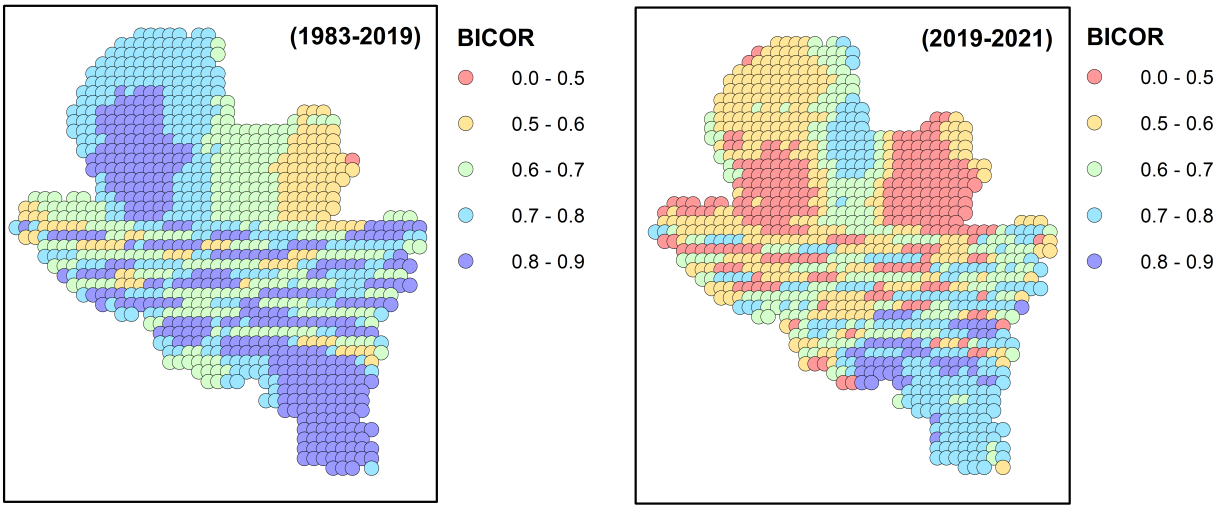


Figura 4.20: Distribución espacial del BICOR en el pronóstico del SPI6. Izquierda serie de tiempo entrenamiento, derecha la serie de tiempo test.



Los desempeños del pronóstico principalmente se validaron con los datos de test. El mejor desempeño se presenta en la R.E. Observando gráficamente en la mayoría de la región; el ECM presenta valores de 0.1 a 0.4, en CCP se presentan valores de 0.7 a 0.9 y BICOR presenta valores de 0.7 a 0.9 denotando la limitación de la R.E respecto a la R.OE. Este resultado tiene coherencia puesto que en el pronóstico de las CPNLs la R.E obtuvo mejores indicadores que la R.OE.

La región con menor desempeño de pronóstico se encuentra en noreste del departamento, con ECM entre 1.0 a 2.0, CPP entre 0.1 a 0.4 y BICOR de 0.0 a 0.6. Esto coincide con la región montañosa limite de la R.E y la R.OE. Por otro lado, la R.E presenta una mayor uniformidad en las magnitudes del ECM y el CCP con pocas variaciones espaciales, contrario a la R.OE donde la distribución es heterogénea y hay cambios fuertes entre un pixel y otro colindante. Este último aspecto se puede asociar al efecto de las diferencias altitudinales en la ocurrencia de las lluvias y su posible relación con variables macroclimáticas. A pesar de ello, se considera un desempeño aceptable en el pronóstico de la sequía para esta región.

Capítulo 5

Disposiciones finales

5.0.1. Conclusiones

En este trabajo de investigación, se aprendieron y plasmaron conceptos sobre la climatología en Nariño en comparación a lo conocido en el periodo de tiempo de los últimos 3 años, se investigó la sequía meteorológica mediante el SPI, brindando herramientas para reducir la vulnerabilidad del departamento a los eventos de sequía venideros cada vez más intensos y duraderos por el cambio climático. Finalmente, se construyó una herramienta de pronóstico del SPI que permite conocer con 6 meses de anticipación los eventos de sequía futuros. A continuación, se discuten algunas conclusiones de esta investigación, que se complementan con lo tratado en los capítulos anteriores.

La utilización de información satelital es una buena herramienta al momento de sortear dificultades de distribución espacial de estaciones meteorólogas *in situ*, esta información es valiosa al momento de caracterizar la lluvia en regiones con mucha variabilidad espacio-temporal. Los datos de CHIRPS en el departamento de Nariño son representativos de la lluvia mensual en departamento. La validación estadística realizada confirma el uso de la base de datos para futuros estudios climáticos. Se debe tener en cuenta algunas advertencias en cuanto al desempeño de CHIRPS sobre el departamento: (i) el mejor desempeño se presenta en la región Andina en comparación a la región del Pacífico, (ii) se presenta una ligera sobrestimación de la lluvia sobre las zonas menos lluviosas y una subestimación en las regiones más lluviosas. Sin embargo el uso de CHIRPS es apropiado para los estudios realizados.

El fenómeno ENOS es el de mayor influencia sobre la variabilidad interanual del clima y del recurso hídrico en Colombia [11]. Sin embargo, la lluvia en Nariño esta modulada por fenómenos a escala planetaria de variabilidad interanual como PDO y NAO. Estas oscilaciones tienen una influencia directa sobre los eventos extremos de lluvias, principalmente el ENOS. El Niño (La Niña) tiene un efecto de reducción (aumento) de las lluvias en la región Andina nariñense, similar a la región Andina colombiana y ecuatoriana; por otro lado, El Niño o la Niña presentan un efecto fluctuante de reducción o aumento de las lluvias en el Pacífico nariñense, adoptando el efecto climático del ENOS para Colombia. Detectar y almacenar estos eventos puede contribuir a mejorar el pronóstico de eventos futuros de sequía en el departamento.

La reducción de dimensionalidad espacial y temporal es un aporte valioso para analizar datos de alta dimensionalidad, como fueron las series de SPI calculadas a partir de CHIRPS. Con la reducción de dimensionalidad se logró encontrar regiones homogéneas SPI y segmentar los datos para entrenar modelos de pronóstico más precisos.

El cálculo y análisis de las teleconexiones de las variables macroclimáticas y su influencia en las sequías en Nariño permiten construir un modelo de pronóstico con errores relativamente bajos de 0.13 (0.37) en el ECM, 0.85 (0.79) en el CPP y de 0.85 (0.78) en BICOR para la R.E (R.OE). Adicionalmente el modelo presenta un horizonte de pronóstico de 6 meses, suficiente para la preparación de los sistemas productivos agrícolas ante un posible evento de sequía. Los riesgos asociados con la sequía y el cambio climático, ponen en frente la necesidad de comprender sobre los procesos de prevención de desastres climáticos y los impactos negativos de la sequía en el ambiente y las sociedades. La región con mayores eventos de sequía coincide con la zona montañosa, de bajas precipitaciones, más poblada y con amplitud de cultivos de secano; lo cual empeora el posible efecto de la sequía en la sociedad y la economía. La identificación y caracterización de estas regiones permite realizar una planificación y mitigación del riesgo más adecuada ante los efectos adversos de futuros eventos de sequía.

Los modelos NNARMAX entrenados bajo la metodología basada en agentes es una buena estrategia para tener procesamiento en paralelo y realizar entrenamientos masivos en corto tiempo. En esta investigación se implementó la metodología y se reportaron los modelos entrenados, usando herramientas online accesibles y centralizadas. Además, se entrenaron en total 1200 regresores que bajo un enfoque metodológico de simplificación se llegaron a 4 modelos que permitieron realizar el pronóstico del SPI6 en el departamento.

5.0.2. Trabajos futuros

La metodología de regionalización implementada, permitió encontrar regiones SPI homogéneas, segmentarlas y contrastarlas con estudios previos que coincidieron con lo hallado. Con base en la regionalización del SPI6 se construyó el modelo de pronóstico, pero una posibilidad está en ampliar el horizonte de pronóstico para el SPI6. También, se puede emplear el SPI12 y construir modelos para el pronóstico a esa escala temporal.

Las variables macroclimáticas usadas están disponibles para su consulta, una mejora a los modelos de pronóstico es emplear técnicas para descargar la información de forma periódica y guardarla usando algún recurso de almacenamiento, que sirva para estar en constante monitoreo y mejora de los modelos existentes. Al día de hoy existe una metodología llamada MLOPS, que es un conjunto de prácticas que tienen como objetivo implementar y mantener modelos de aprendizaje automático actualizados en ciclos de entrenamiento y validación. Es posible implemenar MLOPS a los resultados obtenidos en esta investigación.

El modelo general de pronóstico se puede liberar en forma de API (como un servicio) e integrar un sistema de alerta temprana de los posibles eventos de sequía pronosticados. Este tipo de sistemas anidados a estrategias de adaptación y mitigación ante la sequía reduciría notablemente la vulnerabilidad del territorio, puesto que podrían ser usados masivamente.

Bibliografía

[1] B. Lloyd-Hughes and M. A. Saunders, "The relationship of drought frequency and duration to time sacales," *International Journal of Climatology*, vol. 22, no. 13, pp. 1571–1592, 2002.

[2] CAZALAC, *Atlas de sequías Latinoamerica de América Latina y el Caribe*. 2017.

[3] M. Scholz, M. Fraunholz, and J. Selbig, "Nonlinear principal component analysis: Neural network models and applications," *Lecture Notes in Computational Science and Engineering*, vol. 58, pp. 44–67, 2008.

[4] A. K. Mishra and V. P. Singh, "A review of drought concepts," *Journal of Hydrology*, vol. 391, no. 1-2, pp. 202–216, 2010.

[5] IDEAM, "La variabilidad climática y el cambio climático en Colombia," p. 28, 2018.

[6] N. Hoyos, J. Escobar, J. C. Restrepo, A. M. Arango, and J. C. Ortiz, "Impact of the 2010-2011 La Niña phenomenon in Colombia, South America: The human toll of an extreme weather event," *Applied Geography*, vol. 39, pp. 16–25, 2013.

[7] Y. Carvajal, "Efectos de la Variabilidad Climática (vc) y el Cambio Climático (cc) en los Recursos Hídricos de Colombia," *Entre Ciencia e Ingeniería*, no. 9, pp. 33–61, 2011.

[8] A. Estupiñan, "Estudio de la variabilidad espacio temporal de la precipitación en Colombia. Doctoral Dissertation," p. 118, 2016.

[9] M. Bedoya, C. Contreras, and F. Ruiz, "Alteraciones del régimen hidrológico y de la oferta hídrica por variabilidad y cambio climático. Estudio Nacional del agua 2010.," *Estudio nacional del agua 2010*, p. 421, 2010.

[10] T. Canchala, W. Alfonso-Morales, W. L. Cerón, Y. Carvajal-Escobar, and E. Caicedo-Bravo, "Teleconnections between monthly rainfall variability and large-scale climate indices in Southwestern Colombia," *Water*, pp. 1–20, 2020.

[11] G. Poveda, "La Hidroclimatología de Colombia: Una Síntesis Desde la Escala Inter-Decadal Hasta la Escala Diura por Ciencias de La Tierra," *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales. Vol. 28, no. 107*, pp. 201–222, 2004.

[12] Á. J. Ávila Díaz, Y. Carvajal Escobar, and S. E. Gutiérrez Serna, "Análisis de la influencia de El Niño y La Niña en la oferta hídrica mensual de la cuenca del río Cali," *Revista Tecnura*, vol. 18, no. 41, p. 120, 2014.

[13] W. L. Cerón, Y. Carvajal-Escobar, R. V. A. De Souza, M. T. Kayano, and N. G. López, "Spatio-temporal analysis of the droughts in Cali, Colombia and their primary relationships with the El Nino-Southern Oscillation (ENSO) between 1971 and 2011," *Atmosfera*, vol. 33, no. 1, pp. 51–69, 2020.

[14] UNGRD, *Fenómeno El Niño análisis comparativo 1997-1998 // 2014-2016*. 2016.

[15] M. G. Jacox, E. L. Hazen, K. D. Zaba, D. L. Rudnick, C. A. Edwards, A. M. Moore, and S. J. Bograd, "Impacts of the 2015–2016 El Niño on the California Current System: Early assessment and comparison to past events," *Geophysical Research Letters*, vol. 43, no. 13, pp. 7072–7080, 2016.

[16] G. Poveda, W. Rojas, M. L. Quinones, I. D. Velez, R. I. Mantilla, D. Ruiz, J. S. Zuluaga, and G. L. Rua, "Coupling between Annual and ENSO Timescales in the Malaria: Climate Association in Colombia," *Environmental Health Perspectives*, vol. 109, no. 5, p. 489, 2001.

[17] FAO, "El agua y la agricultura," 2002.

[18] Y. Zhang, X. Liu, W. Jiao, X. Zeng, X. Xing, L. Zhang, J. Yan, and Y. Hong, "Drought monitoring based on a new combined remote sensing index across the transitional area between humid and arid regions in China," *Atmospheric Research*, vol. 264, no. September, p. 105850, 2021.

[19] Q. Liu, G. Zhang, S. Ali, X. Wang, G. Wang, Z. Pan, and J. Zhang, "SPI-based drought simulation and prediction using ARMA-GARCH model," *Applied Mathematics and Computation*, vol. 355, pp. 96–107, 2019.

[20] Asociación Nacional de Industriales (ANDI), "Colombia: Balance 2017 y Perspectiva 2018," *Andi*, pp. 1–68, 2018.

[21] Corponariño and WWF, *Plan Territorial de Adaptación Climática del departamento de Nariño*. 2016.

[22] R. Giuseppe and V. Teodoro, *Methods and tools for drought analysis and management*. 2007.

[23] Organización Meteorológica Mundial (OMM) and Asociación Mundial para el Agua, *Manual de indicadores e índices de sequía*. 2016.

[24] T. Canchala, W. Alfonso-Morales, Y. Carvajal-Escobar, W. L. Cerón, and E. Caicedo-Bravo, "Monthly Rainfall Anomalies Forecasting for Southwestern Colombia Using Artificial Neural," *water*, vol. 12, no. 9, p. 2628, 2020.

[25] T. B. McKee, N. J. Doesken, and J. Kleist, "The relationship od drought frecuency and duration to time scales," *International Journal of Climatology*, vol. 22, no. 13, pp. 1571–1592, 1993.

[26] IPCC, "Cambio Climático 2013 Bases físicas," 2013.

[27] A. K. Mishra, V. R. Desai, and V. P. Singh, "Drought forecasting using a hybrid stochastic and neural network model," *Journal of Hydrologic Engineering*, vol. 12, no. 6, pp. 626–638, 2007.

[28] X. Pan and J. Wu, "Bayesian neural network ensemble model based on partial least squares regression and its application in rainfall forecasting," *Proceedings of the 2009 International Joint Conference on Computational Sciences and Optimization, CSO 2009*, vol. 2, pp. 49–52, 2009.

[29] K. Fernandes, W. Baethgen, S. Bernardes, R. Defries, D. G. Dewitt, L. Goddard, W. Lavado, D. E. Lee, C. Padoch, M. Pinedo-Vasquez, and M. Uriarte, "North Tropical Atlantic influence on western Amazon fire season variability," *Geophysical Research Letters*, vol. 38, no. 12, pp. 1–5, 2011.

[30] M. Rezaeianzadeh and H. Tabari, "MLP based drought forecasting in different climatic regions," *Theoretical and Applied Climatology*, vol. 109, no. 3-4, pp. 407–414, 2012.

[31] J. Abbot and J. Marohasy, "Application of artificial neural networks to rainfall forecasting in Queensland, Australia," *Advances in Atmospheric Sciences*, vol. 29, no. 4, pp. 717–730, 2012.

[32] X. Gu and N. Li, "Study of droughts and floods predicting system based on Spatial-temporal Data Mining," *Proceedings - 2012 6th International Conference on New Trends in Information Science, Service Science and Data Mining (NISS, ICMIA and NASNIT), ISSDM 2012*, pp. 248–253, 2012.

[33] A. Belayneh, J. Adamowski, B. Khalil, and B. Ozga-Zielinski, "Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural networks and wavelet support vector regression models," *Journal of Hydrology*, vol. 508, pp. 418–429, 2014.

[34] M. Khadr, "Forecasting of meteorological drought using Hidden Markov Model (case study: The upper Blue Nile river basin, Ethiopia)," *Ain Shams Engineering Journal*, vol. 7, no. 1, pp. 47–56, 2016.

[35] D. Hong and K. A. Hong, "Drought forecasting using MLP neural networks," *Proceedings - 8th International Conference on u- and e-Service, Science and Technology, UNESST 2015*, pp. 62–65, 2016.

[36] G. G. Netto, A. C. Barbosa, M. N. Coelho, A. R. Miranda, V. N. Coelho, M. J. Souza, F. G. Guimarães, and A. J. Reis, "A hybrid evolutionary probabilistic forecasting

model applied for rainfall and wind power forecast," *Proceedings of the 2016 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS 2016*, pp. 73–78, 2016.

[37] O. Kisi, A. Docheshmeh Gorgij, M. Zounemat-Kermani, A. Mahdavi-Meymand, and S. Kim, "Drought forecasting using novel heuristic methods in a semi-arid environment," *Journal of Hydrology*, vol. 578, no. August, p. 124053, 2019.

[38] S. Poornima and M. Pushpalatha, "Drought prediction based on SPI and SPEI with varying timescales using LSTM recurrent neural network," *Soft Computing*, vol. 1, 2019.

[39] S. Qureshi, J. Koohpayma, M. K. Firozjaei, and A. A. Kakroodi, "Evaluation of Seasonal and Drought Conditions Effect on satellite-based Precipitation Data Accuracy over different Climatic Conditions," no. December, 2020.

[40] S. C. Caress, A. A. Belen, I. J. Esguerra, H. D. Wacan, F. D. Poso, and M. B. Solomon, "Rainfall and Meteorological Drought Forecasting in Albay, Philippines Using Artificial Neural Network," *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2021*, pp. 1–6, 2021.

[41] N. Rachpibool and S. Kajornkasirat, "Assessment of Drought Forecasting Model for Northeastern, Thailand," *ITC-CSCC 2022 - 37th International Technical Conference on Circuits/Systems, Computers and Communications*, vol. 2015, no. 2046735, pp. 116–119, 2022.

[42] OMM, *Manual de indicadores e índices de sequía*. 2016.

[43] Priyamvada and R. Wadhvani, "Review on various models for time series forecasting," *Proceedings of the International Conference on Inventive Computing and Informatics, ICICI 2017*, no. Icici, pp. 405–410, 2018.

[44] "Balances de glaciares y clima en Bolivia y Peru: impacto de los eventos ENSO," *Bulletin - Institut Francais d'Etudes Andines*, vol. 24, no. 3, pp. 661–670, 1995.

[45] K. E. Trenberth and D. P. Stepaniak, "Indices of El Niño evolution," *Journal of Climate*, vol. 14, no. 8, pp. 1697–1701, 2001.

[46] T. Canchala-Nastar and Y. Carvajal-Escobar, "Identificación de escalas dominantes de variabilidad de precipitacion mensual en el departamento de nariño asociadas con fenomenos macroclimáticos," vol. 1, 2018.

[47] W. Loaiza, "Evaluación de sequías meteorológicas y procesos de adaptación de las comunidades agrícolas de la cuenca del río Dagua – Valle del Cauca. Caso de estudio: microcuenca la Centella.," p. 128, 2014.

[48] A. Zargar, R. Sadiq, B. Naser, and F. I. Khan, "A review of drought indices," *Environmental Reviews*, vol. 19, no. 1, pp. 333–349, 2011.

[49] W. C. Palmer, "Meteorological Drought," 1965.

[50] W. Gibbs and J. Majer, "Rainfall Deciles as Drought Indicators," *Bureau of Meteorology Bull*, vol. 48, 1967.

[51] W. C. Palmer, "Keeping track of crop moisture conditions, nationwide: The new crop moisture index," *Weatherwise*, vol. 21, no. 4, pp. 156–161, 1968.

[52] "Large-scale droughts/ floods and monsoon circulation.," 1980.

[53] B. A. Shafer and L. E. Dezman, "Development of a Surface Water Supply Index (SWSI) to assess the severity of drought conditions in snowpack runoff areas (Colorado)," *Proceedings: Eastern Snow Conference, 39th annual meeting*, pp. 164–175, 1982.

[54] L. W. Ceron, R. Andreoli, K. M. Toshie, d. S. R. Ferreria, N. T. Canchala, and Y. Carvajal-Escobar, "Comparison of spatial interpolation methods for annual and seasonal rainfall in two hotspots of biodiversity in South America," *Anais de Academia Brasileira de Ciencias*, 2020.

[55] J. J. Miró, V. Caselles, and M. J. Estrela, "Multiple imputation of rainfall missing data in the Iberian Mediterranean context," *Atmospheric Research*, vol. 197, no. July, pp. 313–330, 2017.

[56] W. W. Hsieh, "Nonlinear principal component analysis by neural networks," *Tellus, Series A: Dynamic Meteorology and Oceanography*, vol. 53, no. 5, pp. 599–615, 2001.

[57] B. Martín del Brío and C. Serrano Cinca, "Fundamentos de redes neuronales artificiales: hardware y software," *Scire: representación y organización del conocimiento*, no. June, pp. 103–125, 1995.

[58] N. Khalili, S. R. Khodashenas, K. Davary, M. M. Baygi, and F. Karimaldini, "Prediction of rainfall using artificial neural networks for synoptic station of Mashhad: a case study," *Arabian Journal of Geosciences*, vol. 9, no. 13, 2016.

[59] R. Oppenheim, "Forecasting via the Box-Jenkins method," *Journal of the Academy of Marketing Science*, vol. 6, no. 3, pp. 206–221, 1978.

[60] IGAC, "Estudio general de suelos y zoonificación de tierras. Departamento de Nariño," 2004.

[61] Gobernacion de Nariño, "Plan de Desarrollo Departamental Gobernación de Nariño," *Plan de Desarrollo Departamental. Gobernación de Nariño.*, p. 255, 2016.

[62] IDEAM, "Clsificacion de climas," 2011.

[63] U. Schneider, A. Becker, P. Finger, A. Meyer-Christoffer, M. Ziese, and B. Rudolf, "GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle," *Theoretical and Applied Climatology*, vol. 115, no. 1-2, pp. 15–40, 2014.

[64] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J. N. Thépaut, "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[65] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. Mcnally, B. M. Monge-Sanz, J. J. Morcrette, B. K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J. N. Thépaut, and F. Vitart, "The ERA-Interim reanalysis: Configuration and performance of the data assimilation system," *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 656, pp. 553–597, 2011.

[66] C. Kummerow, W. Barnes, T. Kozu, J. Shiue, and J. Simpson, "The Tropical Rainfall Measuring Mission (TRMM) sensor package," *Journal of Atmospheric and Oceanic Technology*, vol. 15, no. 3, pp. 809–817, 1998.

[67] K. L. Hsu, X. Gao, S. Sorooshian, and H. V. Gupta, "Precipitation estimation from remotely sensed information using artificial neural networks," *Journal of Applied Meteorology*, vol. 36, no. 9, pp. 1176–1190, 1997.

[68] C. Funk, P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell, and J. Michaelsen, "The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes," *Scientific Data*, vol. 2, pp. 1–21, 2015.

[69] G. J. Huffman, R. F. Adler, P. Arkin, A. Chang, R. Ferraro, A. Gruber, J. Janowiak, A. McNab, B. Rudolf, and U. Schneider, "The Global Precipitation Climatology Project (GPCP) Combined Precipitation Dataset," *Bulletin of the American Meteorological Society*, vol. 78, no. 1, pp. 5–20, 1997.

[70] C. Ocampo-Marulanda, C. Fernández-Álvarez, W. L. Cerón, T. Canchala, Y. Carvajal-Escobar, and W. Alfonso-Morales, "A spatiotemporal assessment of the high-resolution CHIRPS rainfall dataset in southwestern Colombia using combined principal component analysis," *Ain Shams Engineering Journal*, vol. 13, no. 5, p. 101739, 2022.

[71] C. C. Funk, P. J. Peterson, M. F. Landsfeld, D. H. Pedreros, J. P. Verdin, J. D. Rowland, B. E. Romero, G. J. Husak, J. C. Michaelsen, and A. P. Verdin, "A

Quasi-Global Precipitation Time Series for Drought Monitoring," *U.S. Geological Survey Data Series*, vol. 832, p. 4, 2014.

[72] ECMWF, "Centro Europeo de Previsiones Meteorológicas a Plazo Medio (ECMWF) - ERA5," 2021.

[73] G. T. Ayehu, T. Tadesse, B. Gessesse, and T. Dinku, "Validation of new satellite rainfall products over the Upper Blue Nile Basin, Ethiopia," *Atmospheric Measurement Techniques*, vol. 11, no. 4, pp. 1921–1936, 2018.

[74] H. E. Beck, N. Vergopolan, M. Pan, V. Levizzani, A. I. van Dijk, G. P. Weedon, L. Brocca, F. Pappenberger, G. J. Huffman, and E. F. Wood, "Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling," *Advances in Global Change Research*, vol. 69, pp. 625–653, 2020.

[75] G. N. Caroletti, R. Coscarelli, and T. Caloiero, "Validation of satellite, reanalysis and RCM data of monthly rainfall in Calabria (Southern Italy)," *Remote Sensing*, vol. 11, no. 13, 2019.

[76] M. Karamouz, S. Nazif, and M. Fallahi, "Rainfall downscaling using statistical downscaling model and canonical correlation analysis: A case study," *World Environmental and Water Resources Congress 2010: Challenges of Change - Proceedings of the World Environmental and Water Resources Congress 2010*, pp. 4579–4587, 2010.

[77] E. Morán-Tejeda, J. Bazo, J. I. López-Moreno, E. Aguilar, C. Azorín-Molina, A. Sanchez-Lorenzo, R. Martínez, J. J. Nieto, R. Mejía, N. Martín-Hernández, and S. M. Vicente-Serrano, "Climate trends and variability in Ecuador (1966–2011)," *International Journal of Climatology*, vol. 36, no. 11, pp. 3839–3855, 2016.

[78] M. Dehghani, B. Saghafian, F. Nasiri Saleh, A. Farokhnia, and R. Noori, "Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation," *International Journal of Climatology*, vol. 34, no. 4, pp. 1169–1180, 2014.

[79] R. W. Preisendorfer, "Principal component analysis in meteorology and oceanography," 1988.

[80] N. T. Canchala, C. Ocampo-Marulanda, W. Alfonso-Morales, Y. Carvajal-Escobar, W. Ceron, and E. Caicedo-Bravo, "Comparing methods to regionalizarion of monthly rainfall in soutwestern Colombia," *Annals of the Brazilian Academy of Sciences*, p. In print, 2020.

[81] E. León Guzmán, "Métricas para la validación de Clustering Contenido," 2016.

[82] B. Desgraupes, "Clustering Indices," *CRAN Package*, no. April, pp. 1–10, 2013.

[83] J. Marin, "Los mapas auto-organizados de Kohonen (SOM) Introducción," pp. 1–13, 1982.

[84] M. A. Coimbra de Araujo, Jerry Uribe-Opazo and E. C. Camargo, "Classification of areas associated with soybean yield and agrometeorological variables through fuzzy clustering," *Ciencia e investigación agraria*, vol. 40, no. 3, pp. 617–627, 2013.

[85] F. Berzal, "Clustering jerárquico - Métodos de agrupamiento," 2017.

[86] "A Clustering Method Based on K-Means Algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 2012.

[87] OMM, "Índice normalizado de precipitación. Guía del usuario," *Revista Climatologica*, pp. 97–2, 1997.

[88] "The El Niño with a difference," *Nature*, vol. 461, no. 7263, pp. 481–484, 2009.

[89] G. Poveda, "El papel de la Amazonia en el clima global continental: Impactos del Cambio Climático y la deforestación.," in *Amazonia colombiana: imaginarios y realidades*, pp. 145–156, Escuela de Geociencias y Medio Ambiente. Universidad Nacional de Colombia., 2011.

[90] S. M. Vicente-Serrano, E. Aguilar, R. Martínez, N. Martín-Hernández, C. Azorin-Molina, A. Sanchez-Lorenzo, A. El Kenawy, M. Tomás-Burguera, E. Moran-Tejeda, J. I. López-Moreno, J. Revuelto, S. Beguería, J. J. Nieto, A. Drumond, L. Gimeno, and R. Nieto, "The complex influence of ENSO on droughts in Ecuador," *Climate Dynamics*, vol. 48, no. 1-2, pp. 405–427, 2017.

[91] J. E. Montealegre Bocanegra, "Estudio de la variabilidad climática de la precipitación en Colombia asociada a procesos oceánicos y atmosféricos de meso y gran escala," *Ideam*, p. 54, 2009.

[92] L. B. de Guenni, M. García, Á. G. Muñoz, J. L. Santos, A. Cedeño, C. Perugachi, and J. Castillo, "Predicting monthly precipitation along coastal Ecuador: ENSO and transfer function models," *Theoretical and Applied Climatology*, vol. 129, no. 3-4, pp. 1059–1073, 2017.

[93] *A Bayesian Network Approach to Identity Climate Teleconnections Within Homogeneous Precipitation Regions in Ecuador*, vol. 1099. Springer International Publishing, 2020.

[94] C. Quishpe-Vásquez, S. R. Gámiz-Fortis, M. García-Valdecasas-Ojeda, Y. Castro-Díez, and M. J. Esteban-Parra, "Tropical Pacific sea surface temperature influence on seasonal streamflow variability in Ecuador," *International Journal of Climatology*, vol. 39, no. 10, pp. 3895–3914, 2019.

[95] J. Neter, W. Wasserman, and M.-H. Kutner, "Applied linear regression models," 1997.

[96] M. D. Mahecha, F. Gans, S. Sippel, J. F. Donges, T. Kaminski, S. Metzger, M. Migliavacca, D. Papale, A. Rammig, and J. Zscheischler, "Detecting impacts of extreme events with ecological in situ monitoring networks," *Biogeosciences*, vol. 14, no. 18, pp. 4255–4277, 2017.

[97] WMO, *Weather extremes in a Changing Climate: Hindsight on Foresight*. No. 1075, 2012.

[98] Q. Zhang, J. Li, V. P. Singh, and Y. Bai, "SPI-based evaluation of drought events in Xinjiang, China," *Natural Hazards*, vol. 64, no. 1, pp. 481–492, 2012.

[99] P. N. Lal, T. Mitchell, P. Aldunce, H. Auld, R. Mechler, A. Miyan, L. E. Romano, S. Zakaria, A. Dlugolecki, T. Masumoto, N. Ash, S. Hochrainer, R. Hodgson, T. U. Islam, S. Mc Cormick, C. Neri, R. Pulwarty, A. Rahman, B. Ramalingam, K. Sudmeier-Reiux, E. Tompkins, J. Twigg, and R. Wilby, *National systems for managing the risks from climate extremes and disasters*, vol. 9781107025. 2012.

[100] A. S. A. da Silva, R. S. C. Menezes, L. Telesca, B. Stosic, and T. Stosic, "Fisher Shannon analysis of drought/wetness episodes along a rainfall gradient in Northeast Brazil," *International Journal of Climatology*, no. July 2020, pp. 2097–2110, 2020.

[101] L. A. Espinosa, M. M. Portela, J. D. Pontes Filho, T. M. d. C. Studart, J. F. Santos, and R. Rodrigues, "Jointly modeling drought characteristics with smoothed regionalized SPI series for a small island," *Water (Switzerland)*, vol. 11, no. 12, pp. 1–27, 2019.

[102] W. Wang, M. W. Ertsen, M. D. Svoboda, and M. Hafeez, "Propagation of drought: From meteorological drought to agricultural and hydrological drought," *Advances in Meteorology*, vol. 2016, no. 4, 2016.

[103] J. F. Santos, M. M. Portela, and I. Pulido-Calvo, "Regional Frequency Analysis of Droughts in Portugal," *Water Resources Management*, vol. 25, no. 14, pp. 3537–3558, 2011.

[104] P. A. Baeriswyl and M. Rebetez, "Regionalization of precipitation in Switzerland by means of principal component analysis," *Theoretical and Applied Climatology*, vol. 58, no. 1-2, pp. 31–41, 1997.

[105] University of Wyoming, "Department of Atmospheric Science. Weather - Soundings," no. 634, 2016.

[106] N. B. Guttman, "Comparing the palmer drought index and the standardized precipitation index," *Journal of the American Water Resources Association*, vol. 34, no. 1, pp. 113–121, 1998.

[107] N. B. Guttman, "Accepting the standardized precipitation index: A calculation algorithm," *Journal of the American Water Resources Association*, vol. 35, no. 2, pp. 311–322, 1999.

[108] M. J. Hayes, M. D. Svoboda, D. A. Wilhite, and O. V. Vanyarkho, "Monitoring the 1996 Drought Using the Standardized Precipitation Index," *Bulletin of the American Meteorological Society*, vol. 80, no. 3, pp. 429–438, 1999.

[109] C. Funk, P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell, and J. Michaelsen, "The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes," *Scientific Data*, vol. 2, pp. 1–21, 2015.

[110] W. Wu, Y. Li, X. Luo, Y. Zhang, X. Ji, and X. Li, "Performance evaluation of the CHIRPS precipitation dataset and its utility in drought monitoring over Yunnan Province, China," *Geomatics, Natural Hazards and Risk*, vol. 10, no. 1, pp. 2145–2162, 2019.

[111] C. Ocampo-Marulanda, C. Fernández-Álvarez, W. Cerón, T. Canchala, Y. Carvajal-Escobar, and W. Alfonso-Morales, "Spatio-temporal assessment of the hight-resolution CHIRPS rainfall dataset for Southwestern Colombia 1983-2019," *Journal of Pure and Applied Geophysics*, vol. In evaluat, 2020.

[112] L. Pereira, I. Cordery, and I. Iacovides, *Coping with Water Scarcity: Addressing the Challenges*. 03 2009.

[113] "SPI Modes of Drought Spatial and Temporal Variability in Portugal: Comparing Observations, PT02 and GPCC Gridded Datasets," *Water Resources Management*, vol. 29, no. 2, pp. 487–504, 2015.

[114] J. Nuñez Cobo and K. Verbist, *Atlas de sequías de América Latina y el Caribe*. UNESCO Publishing, 2018.

[115] J. V. De la Hoz, "Economía del departamento de Nariño: ruralidad y aislamiento geográfico," *Centro De Estudios Economicos Regionales (Ceer)*, 2007.

[116] V. Urrea, A. Ochoa, and O. Mesa, "Seasonality of Rainfall in Colombia," *Water Resources Research*, vol. 55, no. 5, pp. 4149–4162, 2019.

[117] C. H. Chou, M. C. Su, and E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Analysis and Applications*, vol. 7, no. 2, pp. 205–220, 2004.

[118] G. T. Ayehu, T. Tadesse, B. Gessesse, and T. Dinku, "Validation of new satellite rainfall products over the upper blue nile basin, ethiopia," *Atmospheric Measurement Techniques*, vol. 11, no. 4, pp. 1921–1936, 2018.

[119] G. Tsakiris, A. Loukas, D. Pangalou, H. Vangelis, D. Tigkas, G. Rossi, and A. Cancelliere, "Drought characterization [Part 1. Components of drought planning. 1.3. Methodological component]," *Drought management guidelines technical annex*, vol. 58, no. 58, pp. 85–102, 2007.

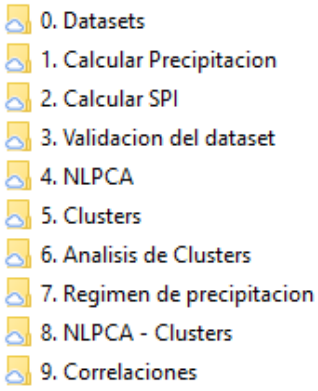
Capítulo 6

Anexos

Distribución del código fuente

El código fuente de esta investigación esta liberado para el publico en general, las carpetas están enumeradas y distribuidas con los pasos metodológicos empleados a lo largo de esta investigación, como se muestra en la Fig. 6.1.

Figura 6.1: Distribución de las carpetas, código fuente.



`https://github.com/lsqrtl`

En el repositorio llamado **Forecasting of SPI using NNARMAX**. Se encontrará el código fuente disponible, para alguna duda o comentario escribir.

- cristhian.fernandez@correounivalle.edu.co
- cefernal@gmail.com