

**UNIVERSIDAD  
DE ANTIOQUIA**

1 8 0 3

# Universidad de Antioquia

FACULTAD DE INGENIERÍA

PREDICCIÓN DEL CAUDAL EN EL SISTEMA MAGDALENA-CAUCA:  
ANÁLISIS DE SERIES DE TIEMPO.

*Especialización en Analítica y Ciencia de Datos.*

Autora:  
Lina María Montoya Zuluaga.

Asesor:  
Dr. Luis Alejandro Fletscher Bocanegra

Diciembre 2023



# Agradecimientos

Agradezco a mi familia por su apoyo incondicional, a mis padres y hermanos por quienes logro terminar este proceso satisfactoriamente y a quien en esta última etapa se convirtió en fuente de conocimiento para este trabajo. A mis profesores porque me encaminaron en este bello estudio de la naturaleza, especialmente a Luis Alejandro y el grupo de Investigación GITA.

Agradezco por este trabajo, porque me permitió entender que nuestros ríos están atravesando todo el planeta, fertilizando la tierra al trazar su camino, son correderos activos de vida, transportan alimentos, mercancías y son el camino para la reproducción de muchas especies. Nuestros ríos y océanos nos definen, marcan el ritmo de la vida en la tierra, conectan la dinámica de los diversos ecosistemas en el planeta y finalmente comprendí que hacemos parte de ellos y que su salud es también la nuestra.



# Resumen

El presente trabajo de tesis se enmarca en la continuidad de mi investigación durante la etapa de pregrado, donde se exploró la relación de causalidad entre los índices climáticos globales y los caudales de las principales cuencas de América del Sur. En este contexto, se estableció una conexión causal destacada entre los índices climáticos globales AMMsst, BEST, NINO3 y NINO 34 y los caudales de la estación hidrométrica Calamar, que proporciona información crucial sobre la cuenca Magdalena-Cauca.

El objetivo central de la presente investigación es avanzar en la comprensión y predicción de los caudales de la estación Calamar, utilizando enfoques innovadores. Nos proponemos emplear tres enfoques distintos, dos de ellos basados en técnicas de aprendizaje profundo (redes neuronales recurrentes LSTM y GRU), y el tercero basado en un enfoque más clásico utilizando modelos SARIMA. Estos métodos nos permitirán analizar la capacidad predictiva de cada enfoque, considerando la información de los índices climáticos mencionados junto con la serie temporal de caudales.

En la primera y segunda aproximación, se utilizarán redes neuronales recurrentes (RNN), específicamente LSTM (Long Short-Term Memory) y GRU (Gated Recurrent Unit), para incorporar la complejidad temporal de la relación entre los índices climáticos y los caudales. Estos modelos, al ser capaces de aprender patrones temporales a largo plazo, se explorarán como herramientas eficaces para la predicción hidroclimatológica.

En la tercera aproximación, se empleará un enfoque más clásico, centrándose únicamente en la serie temporal del caudal. Se aplicará un modelo SARIMA (Seasonal AutoRegressive Integrated Moving Average), una técnica bien establecida en el análisis de series temporales, para evaluar su eficacia en la predicción del caudal.

Este estudio no solo busca comparar la eficacia predictiva de diferentes enfoques, sino también proporcionar una guía valiosa para la elección de métodos en problemas hidroclimatológicos similares. Se espera que los resultados obtenidos contribuyan al avance del conocimiento en la predicción de caudales, permitiendo una mejor comprensión de la compleja relación entre los índices climáticos globales y los recursos

hídricos en la cuenca Magdalena-Cauca.

La comprensión y predicción precisas de los caudales son esenciales para la gestión sostenible de los recursos hídricos. Este estudio tiene aplicaciones prácticas significativas en la toma de decisiones relacionadas con la planificación hidroclimatológica, especialmente en regiones donde las variaciones en los patrones de precipitación y temperatura son críticas para la seguridad hídrica y la gestión de riesgos asociados.

La tesis se estructurará en secciones que abarquen desde la revisión bibliográfica hasta la presentación y discusión de los resultados obtenidos. Se prestará especial atención a la comparación entre los enfoques de aprendizaje profundo y el modelo SARIMA, analizando sus fortalezas y limitaciones en el contexto específico de la predicción de caudales en la cuenca Magdalena-Cauca.

# Contenido

<b>1. Introducción</b>	<b>13</b>
<b>2. Series de Tiempo.</b>	<b>17</b>
2.1. Modelo SARIMA (Seasonal Autoregressive Integrated Moving Average)	18
2.2. Redes Neuronales Recurrentes (RNN) . . . . .	21
2.2.1. Redes LSTM (Long Short-Term Memory) . . . . .	23
2.2.2. Redes GRU (Gated Recurrent Units) . . . . .	26
2.3. Modelos vectoriales autoregresivos (VAR) . . . . .	28
<b>3. Haciendo Predicciones en los caudales del Sistema Cauca-Magdalena.</b>	<b>31</b>
3.1. Área de estudio y datos . . . . .	32
3.2. Influencia dinámica de fenómenos globales en la hidrología del Sistema Magdalena-Cauca. . . . .	37
<b>4. Conclusiones y Perspectivas.</b>	<b>49</b>



# Índice de Figuras

2.1.	Histórico del caudal de la estación de Calamar. . . . .	18
2.2.	Las diferentes categorías del modelado de secuencias. . . . .	23
2.3.	Estructura desplegada de una celda LSTM [13]. . . . .	24
3.1.	Delimitación y Características de la zona de estudio. . . . .	33
3.2.	Esquema del Ciclo Hidrológico. . . . .	34
3.3.	Ciclo anual del caudal bimodal en la estación Calamar. . . . .	35
3.4.	Histórico con el promedio anual del caudal de la estación hidrológica de Calamar. . . . .	38
3.5.	Gráfico de Autocorrelación y Autocorrelación Parcial. . . . .	39
3.6.	Predicciones con el modelo SARIMA. . . . .	41
3.7.	Diagrama de Violín para las series de tiempo. . . . .	42
3.8.	Comparación entre las diferentes corridas para el modelo LSTM. . . .	43
3.9.	Comparación entre las diferentes corridas para las predicciones del modelo LSTM y los valores reales. . . . .	44
3.10.	Comparación entre las diferentes corridas para el modelo GRU. . . .	46
3.11.	Comparación entre las diferentes corridas para las predicciones del modelo GRU y los valores reales. . . . .	46
3.12.	Predicciones con el modelo VAR. . . . .	47



# Índice de Tablas

3.1. Descripción de los índices climáticos . . . . .	36
3.2. RMSE para las diferentes corridas de modelos LSTM. . . . .	43
3.3. RMSE para las diferentes corridas de modelos GRU. . . . .	45



# Capítulo 1

## Introducción

El agua es la representación de nuestro vínculo más estrecho con la tierra, es nuestro lazo directo con ella y con la naturaleza. El agua nos permea desde el momento mismo en que tocamos esta tierra y todas sus manifestaciones se nos hacen imprescindibles a lo largo de la vida. La supervivencia ha estado determinada por la disponibilidad de este recurso básico, además de que los ríos y mares por los que fluye son los principales ejes de las actividades comerciales que caracterizan cada región. El agua es el elemento diferenciador entre el planeta Tierra y muchos otros que hasta ahora conocemos, es el factor clave para la aparición de la vida y además, es el principal elemento integrador de sistemas oceánicos, continentales y atmosféricos, pues en su movimiento está contenido el acople dinámico que da lugar a fenómenos remotamente conectados, transportando información de un sistema a otro.

Conocer el comportamiento y la dinámica del agua resulta fundamental para poder garantizar la seguridad hídrica de algunas comunidades, da cuenta de la biodiversidad de los ecosistemas, determina la configuración espacial de los pueblos y condiciona el estilo de vida que llevan los seres que habitan ese lugar [22] [16].

A través del análisis de datos y la inteligencia artificial podemos abordar la pregunta por la dinámica de los procesos que subyacen en el movimiento del agua. Además, proporciona herramientas para explicar cómo la información que reside en los sistemas se transporta y se comparte.

La motivación subyacente en este trabajo de monografía radica en la utilización de los resultados obtenidos en una investigación previa con el objetivo de evaluar de qué manera la información derivada de los índices climáticos puede potenciar el rendimiento de los modelos empleados en la predicción del caudal. Este trabajo busca no solo aprovechar los hallazgos previos, sino también discernir la influencia específica de los índices climáticos en la capacidad predictiva de los modelos, proporcionando así una contribución significativa al conocimiento y la eficacia de los métodos empleados

en la predicción hidrológica.

En el marco de este estudio, se emprende una aproximación inicial mediante el empleo de un modelo SARIMA para la predicción del caudal del río Magdalena-Cauca, específicamente en la estación hidrométrica de Calamar. Esta metodología clásica sirve como punto de referencia para evaluar la efectividad de enfoques más avanzados. Los dos métodos subsiguientes implican el uso de modelos de redes neuronales recurrentes (RNN), específicamente un LSTM (Long Short-Term Memory) y un GRU (Gated Recurrent Unit). El propósito central de incorporar estos modelos más complejos es investigar las posibles ventajas derivadas de la inclusión de índices climáticos causalmente relacionados con el río Magdalena-Cauca en el proceso de predicción del caudal. Este enfoque no solo busca mejorar la precisión de las predicciones, sino también identificar patrones y relaciones más sofisticadas entre los factores climáticos y el comportamiento hidrológico.

Poner los ojos en la dinámica hidrológica del Sur de América significa también fijar la mirada en sistemas hídricos que son relevantes a escala mundial. Estas cuencas son determinantes en el comportamiento hidrológico y equilibrio ambiental a nivel mundial [15]. Sur América es una región con una intervención humana industrial reducida en comparación con otros continentes, cuenta con una amplia variedad de ecosistemas y está ubicada en un lugar privilegiado sobre el trópico. Nuestra zona de estudio es atravesada por la cordillera de Los Andes (la cordillera continental más larga y la segunda más alta de la tierra) que determinan significativamente la circulación de la atmósfera, por lo que son fundamentales en la reconstrucción de los patrones y la dinámica propia de los fenómenos climáticos e hidrológicos a nivel mundial.

Este proyecto fue pensado para hacer uso de las herramientas disponibles en la ciencia de datos para estudiar los sistemas de la naturaleza. Analizar las series de tiempo, entendidas como una realización misma del sistema permite determinar la estructura dinámica del fenómeno y la construcción de instrumentos de medida computacionales facilita el estudio del problema en un régimen no lineal, donde pueden expresarse completamente las características del fenómeno y detectar propiedades emergentes dentro del mismo. Aquí nos alejamos un poco de la Mecánica Clásica y nos encontramos frente a un proceso por fuera del equilibrio, que no está aislado, ni es cerrado, buscamos caracterizar un fenómeno colectivo y obtener conclusiones sobre las condiciones específicas del problema, variables de estado que pueden no ser tenidas en cuenta, pero que claramente pueden estar modificando el espacio de fase de las características tomadas.

En el segundo capítulo, se profundiza en los tres enfoques clave de modelación e inteligencia artificial que se emplearán en este estudio. Uno de estos enfoques se

centra en las redes neuronales, desglosándose en dos arquitecturas distintas del tipo RNN. Se ofrece una explicación detallada y comprensiva, tanto conceptual como matemáticamente, de los fundamentos de cada modelo, proporcionando una base sólida para su aplicación posterior.

El siguiente capítulo incluye una contextualización detallada del entorno, abordando la ubicación geográfica y las condiciones específicas de la estación hidrométrica objeto de estudio. Se explora la influencia dinámica de diversos factores ambientales y se describe cómo estos influyen en los patrones de caudal. Además, se examina la aplicación práctica de los modelos desarrollados a los conjuntos de datos disponibles.

El último capítulo, se presentan las observaciones clave derivadas de la aplicación de los modelos, proporcionando conclusiones significativas sobre la capacidad predictiva de cada enfoque. Además, se esbozan perspectivas futuras y posibles mejoras para abordar eficazmente los desafíos específicos del problema en estudio.



# Capítulo 2

## Series de Tiempo.

El estudio de las series de tiempo es fundamental en numerosos campos, desde la economía hasta la meteorología, pasando por la medicina y la ingeniería, ya que nos permite analizar y comprender cómo ciertos fenómenos evolucionan a lo largo del tiempo. En este sentido, la inteligencia artificial ha revolucionado la forma en que abordamos el análisis de series temporales. Gracias a algoritmos de aprendizaje automático y redes neuronales, la inteligencia artificial puede identificar patrones complejos en datos temporales, anticipar tendencias futuras y tomar decisiones informadas en tiempo real. Esto proporciona ventajas significativas, como la capacidad de predecir crisis económicas, gestionar eficazmente la cadena de suministro, detectar enfermedades en etapas tempranas y optimizar procesos industriales. En resumen, el estudio de las series de tiempo impulsado por la inteligencia artificial no solo es esencial para comprender el pasado, sino que también nos brinda la capacidad de anticipar y moldear el futuro de manera más precisa y efectiva.

Los estudios en hidrología y meteorología a menudo recurren a modelos de series temporales para la representación de flujos, con el fin de realizar predicciones y generar secuencias sintéticas. Estas secuencias sintéticas son esenciales para alimentar el análisis de sistemas complejos de recursos hídricos. La estación hidrométrica de Callamar se erige como un punto estratégico para monitorizar y registrar la variabilidad temporal de los flujos hídricos. Al considerar el caudal como una serie temporal, no solo se capturan las oscilaciones estacionales y anuales, sino que también se revelan patrones a largo plazo fundamentales para la gestión sostenible de recursos hídricos. La representación gráfica de la serie de tiempo del caudal en esta estación se convierte en una herramienta visual invaluable, como se ilustran en la figura (2.1). No solo proporciona una instantánea histórica de los patrones de flujo, sino que también sirve como base para predicciones futuras, crucial para la toma de decisiones informadas en el manejo de cuencas hidrográficas. La consideración de los caudales como series temporales destaca la necesidad de aplicar metodologías avanzadas, como modelos de inteligencia artificial, para desentrañar complejidades y mejorar la capacidad pre-

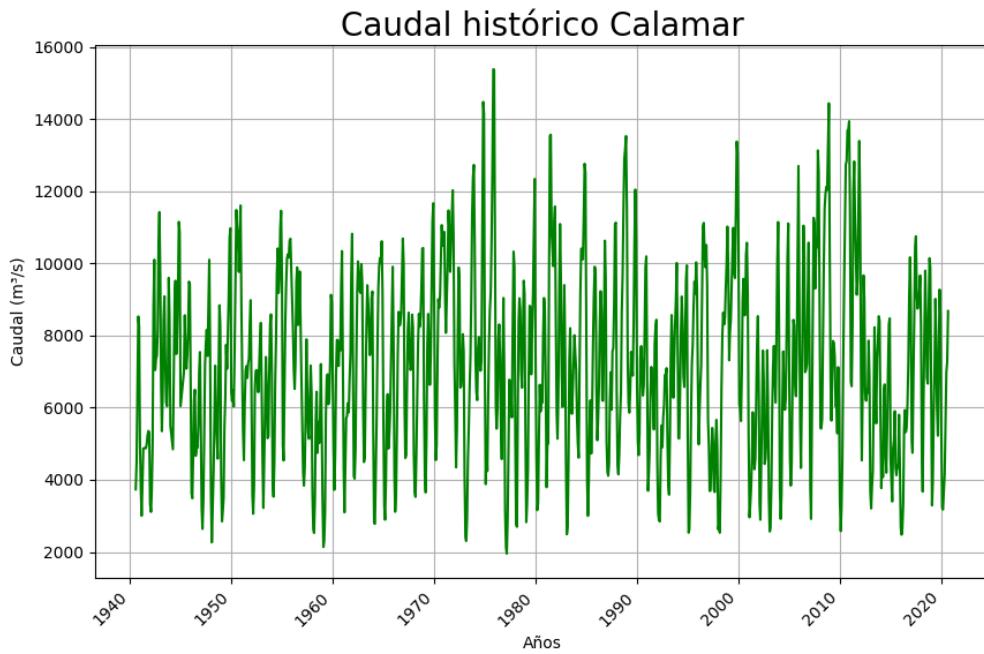


Figura 2.1: Histórico del caudal de la estación de Calamar.

dictiva, permitiendo así una gestión más eficiente y resiliente de los recursos hídricos en la región.

## 2.1. Modelo SARIMA (Seasonal Autoregressive Integrated Moving Average)

El modelo SARIMA se muestra particularmente útil en predicciones hidrológicas al considerar tanto las tendencias a largo plazo como las estacionalidades recurrentes en los datos temporales. Al profundizar primero en el modelo ARIMA (Autoregressive Integrated Moving Average), se establece una base para entender la capacidad de SARIMA para abordar patrones temporales complejos y variaciones estacionales en las series temporales hidrológicas. ARIMA se centra en la autocorrelación y la integración temporal, mientras que SARIMA amplía esta capacidad al incorporar componentes estacionales, lo que lo convierte en una herramienta valiosa para modelar y prever caudales y flujos de agua, proporcionando una visión más completa de los patrones hidrológicos a lo largo del tiempo. Explorar los conceptos de ARIMA proporciona una base sólida para comprender cómo el SARIMA aborda de manera efectiva las características estacionales presentes en las series temporales hidrológicas, mejorando así la precisión de las predicciones [12].

## 2.1. MODELO SARIMA (SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE)19

El método ARIMA es reconocido como uno de los enfoques más ampliamente utilizados en hidrología y en investigaciones asociadas a la variabilidad climática, principalmente debido a su capacidad para abordar conjuntos de datos no estacionarios[8].

ARIMA, que significa Autoregressive Integrated Moving Average, es un método estadístico utilizado en el análisis de series temporales para realizar predicciones. Este enfoque combina componentes autorregresivos (AR) y de media móvil (MA) con la diferenciación de la serie temporal para lograr la estacionarización de los datos. La estacionarización es un proceso que busca hacer que la media y la varianza de una serie temporal sean constantes a lo largo del tiempo [11].

El componente autorregresivo implica que la variable de interés se regresa a sí misma en períodos anteriores, mientras que el componente de media móvil considera la relación entre los valores observados y un término de error pasado. La diferenciación, que es la parte I.<sup>en</sup> ARIMA, se refiere al proceso de restar la observación actual de la observación anterior para obtener una serie temporal estacionaria [18].

En resumen, ARIMA es un modelo que se utiliza para prever valores futuros en una serie temporal, tomando en cuenta tanto la relación autoregresiva como la de media móvil, y aplicando la diferenciación para lograr la estacionarización de los datos. Este método es especialmente útil en la predicción de datos que exhiben tendencias y patrones estacionales.

Los modelos ARIMA están determinados por unos términos,  $p$ ,  $d$  y  $q$  que se utilizan para describir las componentes del modelo:

- $p$  (Autoregressive Order): Representa el número de términos autoregresivos en el modelo. Estos son términos que muestran la relación entre una observación y las observaciones anteriores..
- $d$  (Integrated Order): Indica el número de veces que se diferencia la serie temporal para hacerla estacionaria. La diferenciación implica restar la observación actual de la observación anterior.
- $q$  (Moving Average Order): Indica el número de términos de media móvil en el modelo. Estos términos capturan la relación entre una observación y los errores residuales de observaciones anteriores.

Es importante aclarar que los modelos ARIMA son una generalización de los modelos ARMA que se adapta específicamente a series temporales no estacionarias. Este enfoque ampliado permite modelar la serie de manera más flexible, incorporando la diferenciación necesaria para alcanzar la estacionariedad y proporcionando así una herramienta valiosa en el análisis de series temporales.

Después de establecer la necesidad de modelos ARIMA para abordar la no estacionariedad en series temporales, se introduce la definición formal de un proceso ARIMA en función de los parámetros  $p$ ,  $d$  y  $q$ :

Sea  $d$  un entero no negativo. Se dice que  $Y_t$  es un proceso  $ARIMA(p, d, q)$  si la serie  $X_t = (1 - B)^d Y_t$  es un proceso  $ARMA(p, q)$  estacionario, donde  $(1 - B)^d$  es el operador de diferenciación. Esta definición implica que la serie  $Y_t$  satisface la relación

$$\phi(B)(1 - B)^d Y_t = c + \theta(B)\epsilon_t, \quad (2.1)$$

donde  $\phi(z)$  y  $\theta(z)$  son los polinomios autorregresivo y media móvil de grado  $p$  y  $q$ , respectivamente, tales que no tienen raíces en común y  $\epsilon_t \sim N(0, \sigma^2)$  [18]. Además, debe notarse que el proceso es estacionario si y solo si  $d = 0$ , en cuyo caso se trataría de un modelo  $ARMA(p, q)$ . La forma extensa de ver la anterior ecuación es:

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

En esta ecuación,  $Y_t$  es la serie temporal,  $\alpha$  es una constante,  $\phi_1, \phi_2, \dots, \phi_p$  son coeficientes autoregresivos,  $\epsilon_t$  es un término de error en el tiempo  $t$ , y  $\theta_1, \theta_2, \dots, \theta_q$  son coeficientes de la media móvil. Este modelo implica diferenciación y combina componentes autoregresivos (AR) y de media móvil (MA) [21].

Este formalismo proporciona una herramienta poderosa para modelar la dinámica subyacente en series temporales, ofreciendo una comprensión más profunda de las relaciones entre las observaciones a lo largo del tiempo.

Teniendo claras todas las bases de un modelo ARIMA, entramos en materia con el modelo que es de nuestro interés para el presente trabajo. El modelo SARIMA amplía la definición de ARIMA para abordar patrones estacionales en series temporales. En el contexto de SARIMA, se considera una serie  $Y_t$  como un proceso  $ARIMA(p, d, q)(P, D, Q)_s$ , donde  $d$  sigue representando el grado de diferenciación temporal. El término  $(1 - B)^d$  continúa siendo el operador de diferenciación. Ahora, se introducen componentes estacionales mediante el operador  $(1 - B^s)^D$ , donde  $s$  es la periodicidad estacional y  $D$  es el grado de diferenciación estacional. La relación se expresa como:

$$\Phi(B^s)(1 - B)^d(1 - B^s)^D Y_t = c + \Theta(B^s)\epsilon_t, \quad (2.2)$$

Aquí,  $\Phi(z)$  y  $\Theta(z)$  son polinomios autorregresivos y de media móvil estacionales de grado  $P$  y  $Q$ . La estacionalidad se refleja en los términos  $B^s$ , y  $\epsilon_t \sim N(0, \sigma^2)$  sigue siendo un término de error gaussiano. La forma extendida de la ecuación SARIMA muestra la influencia de los componentes autoregresivos, de media móvil y estacionales en la serie temporal  $Y_t$ , capturando así patrones complejos y estacionales en datos hidrológicos.

## 2.2. Redes Neuronales Recurrentes (RNN)

La aplicación de Deep Learning, y en particular las Redes Neuronales Recurrentes (RNN), ha emergido como un enfoque poderoso en el estudio de sistemas hidroclimatológicos, ofreciendo una capacidad única para modelar relaciones temporales complejas y no lineales presentes en datos hidrológicos y climáticos. A diferencia de métodos tradicionales que pueden tener limitaciones para capturar patrones a largo plazo y dependencias temporales, las RNN permiten a los científicos del clima y la hidrología abordar la variabilidad temporal con mayor precisión.

Por otro lado, los modelos de redes neuronales estándar que suelen tratarse como MLP y CNN, no son capaces de manejar el orden de las muestras de entrada. De forma intuitiva, podemos decir que dichos modelos no tienen un recuerdo de las muestras vistas pasadas. Por ejemplo, las muestras pasan mediante pasos de propagación hacia atrás y de prealimentación, y los pesos se actualizan independientemente del orden en que se ha procesado la muestra. Por el contrario, las RNN están diseñadas para modelar secuencias y son capaces de recordar información pasada y procesar nuevos eventos en consecuencia[13].

Las RNN son especialmente efectivas para modelar fenómenos climáticos y hidrológicos debido a su capacidad para capturar dependencias secuenciales en los datos. Estas redes tienen la capacidad de recordar información pasada y utilizarla para prever eventos futuros, lo que resulta crucial en la comprensión de patrones climáticos estacionales, cambios en el caudal de ríos y respuestas a eventos meteorológicos extremos.

En el ámbito hidroclimatológico, las RNN se han utilizado para prever caudales, simular el ciclo hidrológico, y entender la interacción entre variables climáticas y flujos de agua. La capacidad de las RNN para adaptarse a la variabilidad temporal inherente en los datos hidroclimatológicos las convierte en herramientas valiosas para mejorar la precisión de las predicciones y comprender la compleja dinámica de los sistemas hidroclimatológicos. En resumen, el uso de Deep Learning, en particular las RNN, está transformando la manera en que abordamos y entendemos los procesos hidroclimatológicos, brindando nuevas perspectivas y oportunidades para mejorar la gestión del agua y la mitigación de riesgos asociados con eventos climáticos extremos.

Otro de los aspectos importantes a tener en cuenta se refiere a las diferentes categorías para modelar secuencias, así pues, en el contexto de series de tiempo y modelos predictivos, las expresiones "many-to-many, one-to-many, z "many-to-one" se refieren a la relación entre las series de tiempo de entrada y la variable de salida. A continuación se expone una breve explicación de cada uno de estos términos y se ilustran en la figura (2.3):

- **Many-to-Many (Muchos a Muchos):**

**Entrada:** Se refiere a tener múltiples series de tiempo como entrada, es decir, más de una variable de series temporales que se utilizan para hacer predicciones.

**Salida:** También implica generar múltiples valores en el tiempo como resultado. Por ejemplo, podrías tener varias series de tiempo de diferentes variables de entrada y estar prediciendo múltiples variables de salida en diferentes momentos.

- **One-to-Many (Uno a Muchos):**

**Entrada:** En este caso, hay una única serie de tiempo como entrada, es decir, una variable de series temporales utilizada para hacer predicciones.

**Salida:** La salida consiste en múltiples valores en el tiempo. Por ejemplo, podrías tener una serie de tiempo que representa datos históricos y predecir varios pasos en el futuro.

- **Many-to-One (Muchos a Uno):**

**Entrada:** Se refiere a tener múltiples series de tiempo como entrada, es decir, varias variables de series temporales que se utilizan para hacer predicciones.

**Salida:** La salida es un solo valor en el tiempo. En este caso, estás prediciendo una variable específica en un momento dado.

En el contexto de modelos de series temporales, estas designaciones describen cómo se estructuran las relaciones entre las variables de entrada y salida en términos de la temporalidad. Es esencial entender estas configuraciones al seleccionar y diseñar modelos para abordar problemas específicos de predicción temporal.

Además, es fundamental explorar las arquitecturas específicas de las RNN, como las LSTM (Long Short-Term Memory) y las GRU (Gated Recurrent Unit). Estas arquitecturas fueron diseñadas para abordar el problema del olvido a largo plazo y mejorar la capacidad de las RNN para capturar dependencias a largo plazo en las secuencias de datos[13].

Las LSTM introducen celdas de memoria y mecanismos de olvido selectivo, permitiendo que la red retenga información relevante durante largos períodos. Por otro lado, las GRU, aunque más simples que las LSTM, también abordan el problema del olvido a largo plazo mediante el uso de compuertas de actualización y reinicio.

La elección entre LSTM y GRU generalmente depende de la complejidad del problema y los recursos disponibles. Ambas arquitecturas han demostrado ser efectivas en el modelado de secuencias, y su implementación adecuada puede marcar la diferencia en la capacidad de la red para aprender patrones temporales complejos.

En la figura (2.3), se ilustran las diferentes categorías del modelado de secuencias, destacando cómo las LSTM y GRU pueden adaptarse a las diversas relaciones temporales presentes en los datos hidroclimatológicos.

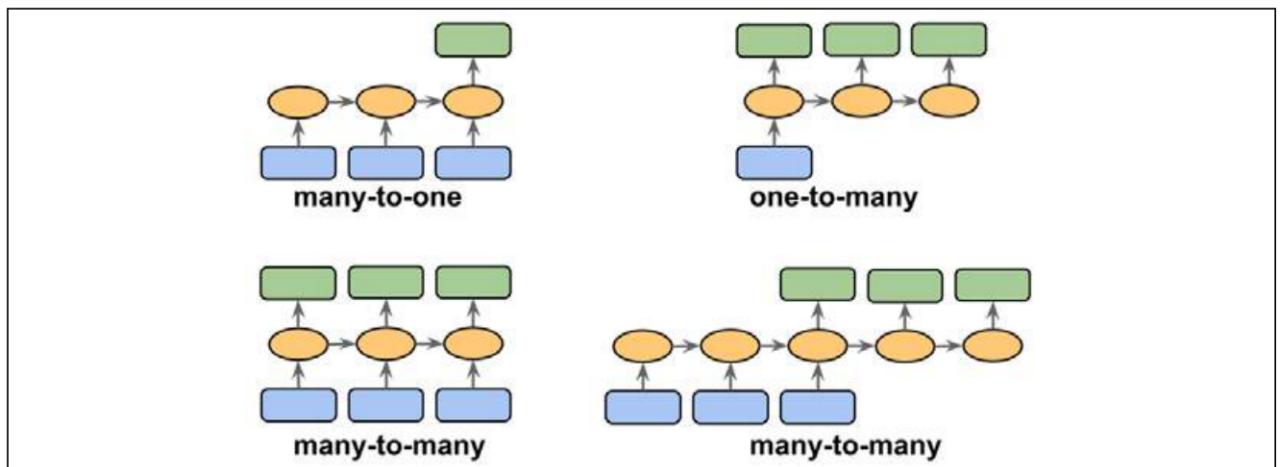


Figura 2.2: Las diferentes categorías del modelado de secuencias.

### 2.2.1. Redes LSTM (Long Short-Term Memory)

Como se mencionó antes, las LSTMs (Memorias de Corto y Largo Plazo, por sus siglas en inglés) fueron creadas para resolver un problema llamado "desvanecimiento del gradiente". Este problema ocurría en las Redes Neuronales Recurrentes estándar y dificultaba el aprendizaje de patrones a largo plazo. En una LSTM, el componente clave es una "celda de memoria", que básicamente reemplaza la capa oculta en las RNN convencionales. En cada celda de memoria, hay una conexión que ayuda a superar el problema del desvanecimiento y explosión del gradiente. La estructura desplegada de una celda LSTM moderna se puede ver en la figura (2.3).

Para entender un poco el esquema, imagina que la red neuronal tiene una especie de "memoria" que almacena información importante. Para actualizar esta memoria, usamos "puertas" que son como filtros de decisión. Estas puertas deciden qué información debe mantenerse y qué información debe olvidarse.

En la imagen,  $\odot$  significa que estamos haciendo ajustes a la información actual multiplicándola de una manera especial, y  $\oplus$  significa que estamos sumando partes espe-

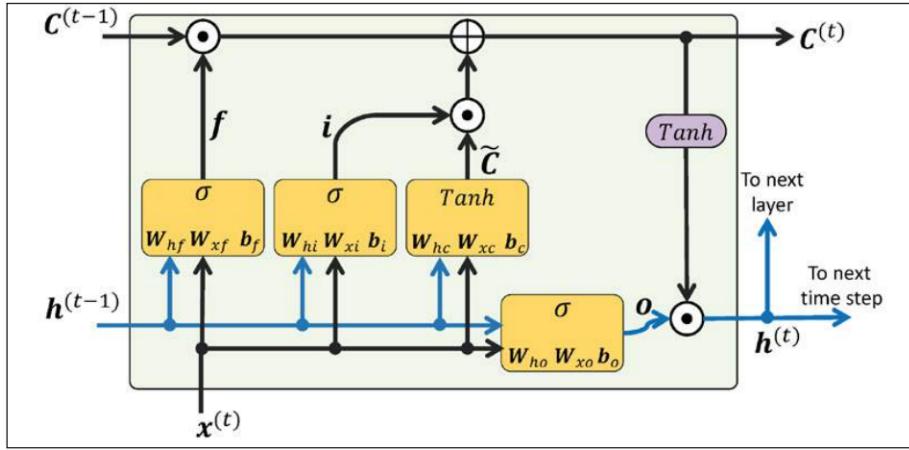


Figura 2.3: Estructura desplegada de una celda LSTM [13].

cíficas.  $x(t)$  es la nueva información que la red está observando en este momento, y  $h(t - 1)$  es la información almacenada en la memoria desde el paso anterior [13].

Estas "puertas de decisión" operaciones especiales determinan cómo la red actualiza y retiene la información clave en su "memoria". Las redes LSTM tienen tres puertas fundamentales: la puerta de olvido (forget gate), la puerta de entrada (input gate) y la puerta de salida (output gate). Cada una cumple una función específica en el manejo de la información a lo largo del tiempo en una secuencia[1]:

- **Puerta de Olvido (Forget Gate):** Permite a la red olvidar cierta información anterior que no es relevante para la tarea actual. Decide qué información retener y qué descartar, ayudando a evitar que la memoria de la red crezca indefinidamente.
- **Puerta de Entrada (Input Gate):** Controla cómo se actualiza el estado de la celda en función de la nueva información de entrada. Decide qué información es importante para agregar a la memoria actual de la red.
- **Puerta de Salida (Output Gate):** Determina cómo se actualizan las unidades ocultas y qué información se utilizará como salida en el paso de tiempo actual. Regula la información que se pasa hacia adelante en la secuencia.

En conjunto, estas puertas permiten a las LSTM modelar dependencias temporales a largo plazo en datos secuenciales al aprender qué información retener y qué descartar en cada paso de tiempo. La capacidad de gestionar la memoria a lo largo del tiempo hace que las LSTM sean especialmente efectivas en tareas que involucran secuencias, como el procesamiento del lenguaje natural y la predicción de series temporales[26].

Otro de los puntos importantes dentro de las redes neuronales son las funciones de activación, como la sigmoide y tanh mencionadas en el contexto de las "puertas" de las LSTMs, desempeñan un papel crucial en el proceso de toma de decisiones de la red neuronal [10].

- Sigmoide: La función sigmoide tiene la forma de una curva "S" que toma cualquier número y lo transforma a un valor entre 0 y 1. En el contexto de las LSTMs, se utiliza para decidir cuánta de la información actual debería ser retenida en la memoria. Piensa en ello como una especie de interruptor que controla qué partes de la información nueva deberían ser recordadas.
- tanh: La función tangente hiperbólica (tanh) también transforma los números a un rango específico, en este caso, entre -1 y 1. En el contexto de las LSTMs, se utiliza para decidir cuánta de la información almacenada anteriormente en la memoria debería ser actualizada. Funciona como otro interruptor para determinar cuánta "importancia" tiene la información antigua en comparación con la nueva.
- Lineal: el uso de una función de activación lineal en la capa de salida es adecuado y común en problemas de regresión, ya que permite a la red predecir valores continuos. Justamente el caso nuestro es un problema de predicción y es por ello una función lineal la que usamos en la última capa del modelo.

Todas estas funciones de activación son como mecanismos de control que ayudan a la red neuronal a decidir qué información retener y cuánto importa cada pieza de información. Ayudan a la red a aprender patrones y relaciones más complejas en los datos a lo largo del tiempo.

En conclusión, la arquitectura LSTM se ha revelado como una herramienta fundamental en la resolución de problemas de predicción hidroclimatológica. Su capacidad única para modelar secuencias y capturar dependencias temporales a largo plazo la convierte en una opción crucial en la comprensión de fenómenos climáticos y flujos de agua. En el ámbito hidroclimatológico, donde la variabilidad temporal es esencial, las LSTM superan las limitaciones de modelos tradicionales al permitir la retención de información relevante a lo largo del tiempo[26]. Esto se traduce en la capacidad de anticipar patrones climáticos estacionales, cambios en el caudal de ríos y respuestas a eventos meteorológicos extremos. La utilización de la arquitectura LSTM en la predicción hidroclimatológica no solo mejora la precisión de las proyecciones, sino que también ofrece una comprensión más profunda de la dinámica de los sistemas, proporcionando así herramientas valiosas para la gestión del agua y la mitigación de

riesgos asociados con eventos climáticos adversos. En resumen, la arquitectura LSTM emerge como una pieza clave en la innovación de enfoques predictivos para abordar la complejidad temporal inherente en los datos hidroclimatológicos.

### 2.2.2. Redes GRU (Gated Recurrent Units)

Una Unidad Recurrente con Compuertas (GRU), propuesta por Cho et al. [2014], es una herramienta que permite a cada unidad en una red adaptarse para capturar dependencias temporales de diferentes escalas. Similar a la unidad LSTM, la GRU utiliza compuertas para regular el flujo de información dentro de la unidad, pero a diferencia de la LSTM, no utiliza celdas de memoria separadas.

La activación  $h_j^t$  en la GRU en un momento específico  $t$  es una combinación lineal entre la activación anterior  $h_j^{t-1}$  y una nueva activación candidata  $h_j^t$ . La compuerta de actualización  $z_j^t$  determina cuánto se actualiza la activación de la unidad y se calcula usando una función sigmoide[1].

La fórmula para la activación es:

$$h_j^t = (1 - z_j^t)h_j^{t-1} + z_j^t h_j^t.$$

Aunque este proceso es similar a la LSTM, la GRU no tiene un mecanismo para controlar cuánto de su estado se expone, siempre exponiendo todo su estado. La activación candidata  $h_j^t$  se calcula de manera similar a una unidad recurrente estándar y utiliza compuertas de reinicio  $r_t$  para decidir cuánto del estado anterior debe olvidarse[1].

En resumen, la GRU adapta su estado en función de la información anterior y candidata, con compuertas que controlan las actualizaciones y reinicios. Aunque comparte similitudes con la LSTM, la GRU es más simple y no controla la exposición parcial de su estado. Estos mecanismos permiten que la GRU capture patrones en datos secuenciales de manera efectiva.

Además, las GRU son un tipo de arquitectura de RNN diseñada para abordar ciertos problemas de las RNN tradicionales, como el problema del desvanecimiento del gradiente y la incapacidad para manejar dependencias a largo plazo en secuencias de datos. Al igual que las LSTM, las GRU son una variante de las RNN que incorpora mecanismos de puertas para controlar el flujo de información.

Es importante recalcar una de las ideas previas, a diferencia de las LSTM, las GRU tienen una estructura más simplificada con solo dos puertas: una puerta de actualización y una puerta de reinicio. Estas puertas permiten a la GRU determinar qué

información retener de las entradas anteriores y cómo combinarla con la nueva información. La puerta de actualización controla cuánta información de la memoria pasada debe mantenerse, mientras que la puerta de reinicio decide cuánta información pasada se debe olvidar. Este diseño más simple en comparación con las LSTM permite que las GRU sean más eficientes computacionalmente y, en algunos casos, más fáciles de entrenar[1].

Ambas arquitecturas, LSTM y GRU, han demostrado ser efectivas en el procesamiento de secuencias temporales, y la elección entre ellas a menudo depende de la naturaleza específica del problema y los recursos computacionales disponibles. Las GRU han ganado popularidad en aplicaciones donde la complejidad computacional y la cantidad de datos disponibles son consideraciones importantes[13].

Así pues, la elección de utilizar unidades GRU en la predicción de caudales, usando también las series de tiempo de otros índices climáticos, puede ofrecer varios beneficios[29]:

- Manejo eficiente de secuencias temporales: Las GRU, al igual que las LSTM, están diseñadas para manejar eficientemente secuencias temporales y capturar dependencias a largo plazo en los datos. Esto es crucial en el caso de la predicción de caudales, donde las condiciones climáticas anteriores pueden tener un impacto significativo en los flujos de agua actuales y futuros.
- Menor complejidad computacional: Las GRU tienden a ser menos complejas computacionalmente que las LSTM debido a su estructura más simple con solo dos puertas. Esto puede resultar en un entrenamiento más rápido y un uso más eficiente de los recursos computacionales.
- Manejo eficiente de datos secuenciales: Las GRU son capaces de manejar eficientemente secuencias temporales largas y cortas. En el contexto de la predicción de caudales, donde la variabilidad temporal es esencial, esto permite capturar patrones climáticos complejos y su impacto en los flujos de agua.
- Rendimiento comparable a LSTM: Segundo investigaciones y experimentos, las GRU a menudo ofrecen un rendimiento comparable al de las LSTM en diversas tareas, incluida la predicción de secuencias temporales. Si bien cada modelo puede destacar en diferentes escenarios, la simplicidad de las GRU las hace atractivas, especialmente cuando los recursos computacionales son una consideración importante.
- Adaptabilidad a datos específicos: La elección entre GRU y otros modelos puede depender de la naturaleza específica de los datos hidroclimatológicos disponibles. Algunos conjuntos de datos pueden beneficiarse más de la capacidad de

las GRU para manejar eficientemente secuencias cortas y largas, mientras que otros pueden requerir las características específicas de las LSTM.

Es importante señalar que la efectividad de las GRU en comparación con otros modelos puede depender de la complejidad y la naturaleza específica de los datos de caudales e índices climáticos disponibles. La experimentación y la evaluación comparativa son esenciales para determinar el modelo más adecuado para una tarea específica de predicción hidroclimatológica.

### 2.3. Modelos vectoriales autoregresivos (VAR)

Un modelo VAR, o Vector Autoregression, es un tipo de modelo estadístico utilizado para analizar la relación entre múltiples variables que cambian con el tiempo. Está especialmente diseñado para trabajar con series temporales, como es nuestro caso. Aquí algunas de sus principales características[27]:

- **Vector:** La parte "Vector." en VAR indica que el modelo involucra múltiples variables. En lugar de analizar una sola variable, como en modelos univariados, un modelo VAR permite estudiar simultáneamente el comportamiento de varias variables interrelacionadas.
- **Autoregression:** La palabra ".Autoregression" se refiere a la idea de que cada variable en el modelo puede ser regresada sobre sus propios valores pasados, así como sobre los valores pasados de otras variables en el sistema. Es decir, la variable en un momento dado se explica en función de sus propios valores anteriores y de los valores anteriores de otras variables en el modelo.
- **Sistema de Ecuaciones Simultáneas:** Un modelo VAR se formula como un sistema de ecuaciones simultáneas. Cada ecuación representa una variable en términos de sus valores pasados y de los valores pasados de otras variables. Estas ecuaciones capturan las relaciones dinámicas entre las variables a lo largo del tiempo.
- **Orden del Modelo (lags):** El orden del modelo, también conocido como "lags", indica cuántos períodos pasados se incluyen en cada ecuación. Si tienes un modelo VAR(2), por ejemplo, significa que cada ecuación incluirá los valores de las variables hasta dos períodos anteriores.

Empleamos un modelo de tipo vector autoregresivo (VAR) al intentar describir las interacciones simultáneas entre un conjunto de variables. Un VAR consiste en un modelo de ecuaciones simultáneas que se compone de un sistema de ecuaciones en su forma reducida sin imponer restricciones. La reducción de la forma implica que

los valores contemporáneos de las variables en el modelo no se presentan como variables explicativas en ninguna de las ecuaciones. En cambio, el conjunto de variables explicativas para cada ecuación está compuesto por un conjunto de retardos de cada una de las variables del modelo. La falta de restricciones significa que cada ecuación contiene el mismo conjunto de variables explicativas[14].

En otras palabras, cuando usamos un modelo VAR, lo hacemos para entender cómo interactúan al mismo tiempo varias cosas diferentes, que llamamos variables. Este modelo consiste en un conjunto de ecuaciones que muestran cómo estas variables se afectan mutuamente, pero de una manera especial. En lugar de usar los valores actuales de las variables para predecir su futuro, como haríamos normalmente, el VAR utiliza versiones anteriores de esas variables. Además, todas las ecuaciones en este modelo comparten el mismo conjunto de variables, lo que significa que no hay restricciones especiales en términos de qué variables pueden influir en cuáles.

En el contexto de problemas hidroclimatológicos, los modelos VAR pueden ser herramientas valiosas para analizar las interacciones complejas entre variables climáticas y hidrológicas. Estos modelos permiten examinar cómo diferentes factores, como las precipitaciones, la temperatura y el caudal de los ríos, afectan simultáneamente unos a otros a lo largo del tiempo. Al utilizar un enfoque de ecuaciones simultáneas, los VAR capturan las relaciones dinámicas y las respuestas de las variables en distintos períodos. Esto resulta especialmente útil para comprender fenómenos como las sequías, inundaciones o cambios en los patrones climáticos, ya que facilita la identificación de conexiones temporales entre las variables. Además, al no imponer restricciones específicas en las interacciones, los modelos VAR pueden adaptarse a la complejidad inherente de los sistemas hidroclimatológicos, proporcionando comprensiones más detalladas sobre su comportamiento y evolución.

Finalmente, otra de las razones por las cuales se usa VAR como una alternativa para las predicciones de series de tiempo es que los Vectores Autorregresivos han proveído una exitosa técnica para hacer pronósticos en sistemas de variables de series de tiempo interrelacionadas, donde cada variable ayuda a pronosticar a las demás variables[2]. Así pues, dicho escenario corresponde justamente al caso que nos ocupamos de resolver en este trabajo, puesto que trabajamos con series de tiempo de diferentes sistemas, pero entre los cuales previamente hemos hallado relaciones de tipo causal entre ellas, por tanto, los resultados posteriores nos mostrarán si esta técnica puede ser adecuada para la predicción del caudal del río Magdalena, considerando algunos índices climáticos como variables propias de la dinámica del sistema y que por tanto deben hacer parte de las ecuaciones que describen el comportamiento del sistema.



# Capítulo 3

## Haciendo Predicciones en los caudales del Sistema Cauca-Magdalena.

El modelado estadísticos, el análisis de datos, la inteligencia artificial y el deep learning son herramientas útiles para estudiar los sistemas de la naturaleza. Estos sistemas son altamente complejos y generalmente describen procesos colectivos no-lineales. Las herramientas de la física y la matemática nos permiten acceder a algunas configuraciones del sistema y extraer sus relaciones dinámicas, estudiar procesos acoplados en los que cobra completo sentido preguntarse por las relaciones causales, y basados en ellas construir modelos que nos permitan hacer predicciones en estos sistemas.

En el trabajo realizado previamente, se estudió un sistema hidrometeorológico, donde se vinculan procesos locales y globales cuyas relaciones causales nos permitieron inferir los efectos de procesos climáticos a gran escala en el Sistema Magdalena-Cauca. Así pues, se seleccionan los cuatro índices climáticos con la mayor relación causal identificada con la estación hidrométrica de Calamar. Con el objetivo de realizar predicciones sobre el caudal del sistema Magdalena-Cauca, se implementan dos enfoques distintos. Por un lado, se utiliza una aproximación basada en la serie univariada del caudal y se emplea el modelo SARIMA para la predicción. Por otro lado, se recurre a modelos de redes neuronales recurrentes, como LSTM y GRU, para incorporar la información de los índices climáticos en la predicción del caudal. Finalmente, se usa también un modelo VAR para la predicción del caudal. Esta estrategia dual busca explorar la eficacia de los métodos clásicos y avanzados en la mejora de la precisión predictiva, aprovechando la información de los índices climáticos para enriquecer las estimaciones del caudal del sistema Magdalena-Cauca.

### 3.1. Área de estudio y datos

Los procesos hidrológicos y climáticos que estudiamos están enmarcados en los sistemas complejos, por lo que se toman un par de variables cuyas series de tiempo contengan la mayor información dinámica posible del fenómeno que representan. Se toman los caudales de los ríos, que son los corredores activos más importantes que tiene la naturaleza y de los que depende para mantener el equilibrio de la vida [4]. La otra variable está asociada a los océanos, donde se expresan fenómenos globales y se sostiene interacción con la dinámica atmosférica.

En general, las variables de estado tienen escalas espacio-temporales inherentes al proceso. Cada uno de los índices climáticos posee una escala de varibilidad temporal propia, algunos presentan variaciones anuales o interanuales, mientras que existen otros que lo hacen de manera multidecadal, además de que son mediciones realizadas sobre porciones específicas en el océano. Por el contrario, la hidrología presenta una fuerte componente estacional y su dinámica general la resumimos en el registro mensual del caudal.

Una serie de particularidades de orden meteorológico caracterizan la región de estudio. La ZCIT atraviesa la porción continental donde están ubicadas las cuencas, es decir que la franja en la que se encuentran los vientos alisios provenientes del hemisferio norte y el sur está sobre nuestro territorio. Habitar la ZCIT dota al lugar de características específicas en las condiciones climáticas. Pasa también la corriente de chorro ecuatorial, que se moviliza en dirección opuesta a la presentada en la corriente de chorro polar y subtropical. Fenómenos como los anteriores, sumados a otros como el monzón sudamericano, dan como resultado un comportamiento climático especial en la zona y una dinámica hidrológica en particular [22]. Todos estos procesos se constituyen como procesos de transporte de grandes masas de aire y modificaciones de los patrones de circulación a diferentes escalas, lo que termina significando una transferencia de información entre las partes del sistema.

Situemos un poco geoespacialmente nuestro problema. Desde la línea del Ecuador y avanzando hacia el sur, nos encontramos con el Sistema Magdalena-Cauca, ubicado en la zona norte y occidente de Colombia, con un área de 257400 km<sup>2</sup> que representa un 24% de la superficie continental del país y alberga aproximadamente el 80% de los colombianos, cuya vida y actividad comercial dependen directamente de esta cuenca. El Sistema Magdalena-Cauca posee una gran variedad de ecosistemas, páramos, humedales, lagos y ríos de los que depende la disponibilidad hídrica del país y la supervivencia de sus especies [5].

Las estaciones de muestreo están distribuidas como se muestra en la figura (3.1), donde se indica la cuenca a la que cada una pertenece, sus principales afluentes y la

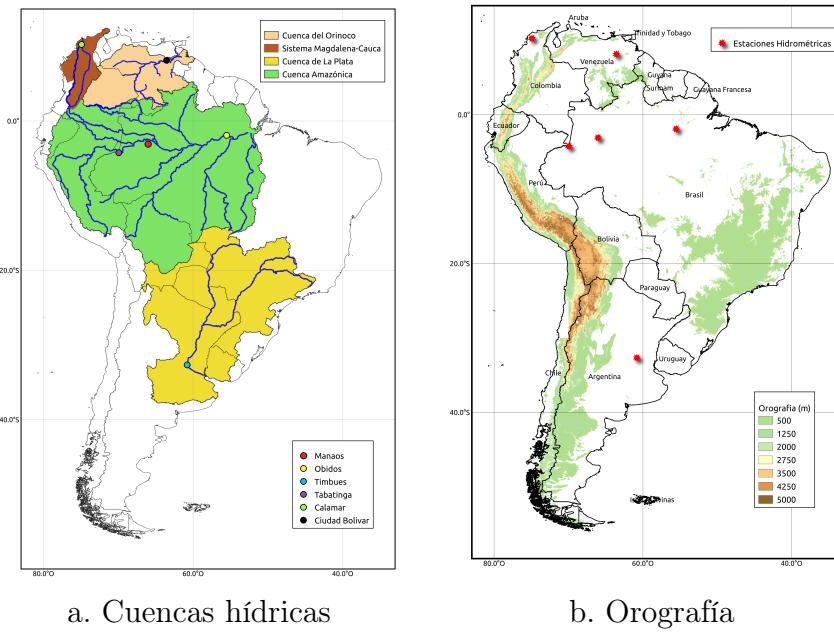


Figura 3.1: Delimitación y Características de la zona de estudio.

orografía que caracteriza el lugar. Es importante aclarar que en la imagen aparecen. La estación hidrométrica de Calamar, ubicada en el departamento del Bolívar, Colombia, guarda la información sobre uno de los principales afluentes del río Magdalena [5].

La intención no es estudiar los procesos de manera aislada, sino lograr dar cuenta de la integración del ciclo hidrológico en las cuencas. Esto se logra ubicando las estaciones hidrométricas cerca a las desembocaduras de los ríos al mar. Dicha integración es una de las bases de este trabajo, pues sugiere una relación causal entre este par de procesos que dependen del transporte hidrológico y la dinámica atmosférica. Como puede verse en la figura (3.2), los ríos funcionan como canal de comunicación entre lo que sale del océano por medio de la evapotranspiración y lo que regresa a ellos después de las precipitaciones y procesos de escorrentía. Los puntos de muestreo están ubicados justo en la interfaz, donde el agua pasa de ser parte de la dinámica fluvial a integrarse al comportamiento propio del sistema oceánico.

La integración del sistema está justamente expresada en ciclo del agua [24], compuesto por diferentes etapas, asociadas a procesos físicos y cambios en el estado de la materia del agua. Éste tiene una relación directa con los patrones de circulación atmosférica y los ciclos anuales de caudal en las diferentes cuencas. Así pues, los procesos de transporte de materia y energía por medio de la atmósfera, las precipitaciones, el desplazamiento por los ríos y su posterior llegada al mar, están determinados y acoplados por medio del ciclo hidrológico [25].

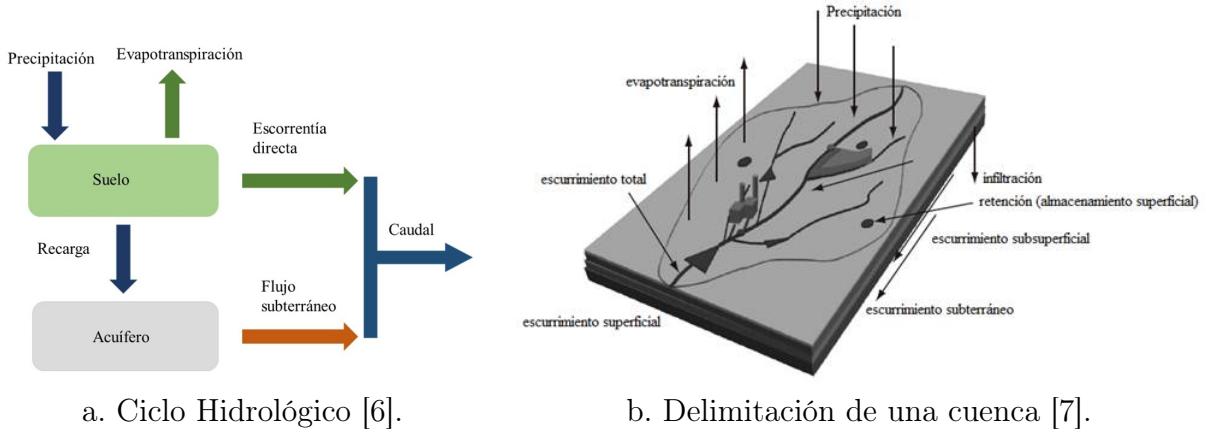


Figura 3.2: Esquema del Ciclo Hidrológico.

El caudal se toma como la variable del sistema que contiene la información hidrológica de la cuenca, la serie de tiempo se toman de la estación hidrométrica ubicada en [15] Calamar. El caudal es la cantidad de agua que circula en un canal hídrico a través de la sección transversal por unidad de tiempo. En términos matemáticos se define en la ecuación (3.1) [23]:

$$Q = A\bar{v} \quad (3.1)$$

donde  $Q$  es el caudal en  $[m^3 s^{-1}]$ ,  $A$  es el área transversal a la dirección del flujo en  $[m^2]$  y  $\bar{v}$  es la velocidad promedio que lleva el fluido en  $[m s^{-1}]$ . Esta variable contiene información explícita sobre los procesos de transporte fluvial, como canal de comunicación entre procesos.

La distribución que siga una variable en particular habla de su dinámica, su posible configuración en el espacio de fase y permite detectar eventos extremos o extraños dentro del sistema. Contamos con la dinámica hidrológica local de la zona resumida en las series de tiempo de caudal de la cuenca. Como puede verse en la figura 3.3, la estación Calamar, propia del sistema Cauca-Magdalena, se caracteriza por un ciclo anual del caudal bimodal (figura 3.3), dicha bimodalidad se debe al paso de la Zona de Convergencia Intertropical (ZCIT) sobre el territorio. La hidrología de esta cuenca depende principalmente de la precipitación.

En general, cada río puede presentar su nivel mínimo y máximo de caudal en épocas diferentes del año, condicionadas por su ubicación, el ecosistema del que hace parte, las actividades humanas e industriales que se realicen cerca a él, entre otros factores locales. Sin embargo, todos estos corredores de agua dulce se caracterizan por su ciclo anual, es decir, el comportamiento de los ríos cambia mes a mes y su dinámica vuelve a comportarse de manera análoga el mismo mes del año siguiente, cuando es posible

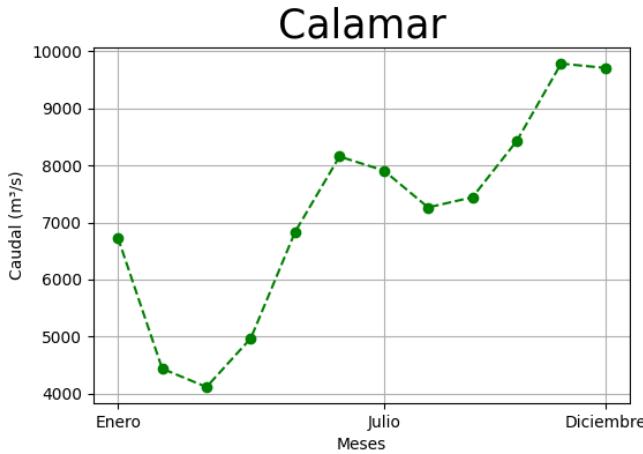


Figura 3.3: Ciclo anual del caudal bimodal en la estación Calamar.

reproducir las condiciones ambientales locales y globales; por lo que la dinámica hidrológica propia de cada afluente se puede visualizar con el ciclo anual de caudal de las cuencas.

La otra porción del sistema está compuesta por procesos de carácter meteorológico. Las bases de datos contienen los índices climáticos que dan cuenta de fenómenos y sucesos muy particulares, determinados por el estado de la atmósfera y su conexión con el océano. Las principales propiedades que determinan los indicadores climáticos están relacionados con alguna variable de tipo atmosférica, como la presión, la temperatura, la precipitación y la radiación solar o de otro tipo, con una variable de tipo oceánica como la temperatura superficial del mar (TSM) o la cobertura de hielo [28]. Los índices climáticos reportados generalmente corresponden a anomalías estandarizadas y no constituyen el fenómeno como tal, sino que funcionan para indicar su comportamiento.

Existen diversas formas de clasificar los índices climáticos, una de ellas es acorde a las siguientes características [17]:

- Teleconexiones: son relaciones significativas entre fluctuaciones simultáneas que ocurren entre variables meteorológicas de áreas geográficas separadas.
- Atmósfera: se refieren básicamente a gradientes de presión entre dos puntos de la tierra, se encarga de describir los patrones de circulación atmosférica.
- Precipitación: los registros sobre lluvias e incluso nieve o agua nieve, se convirtieron en indicadores climáticos, recolectando información sobre la intensidad y periodicidad de las precipitaciones.: (El Niño-Oscilación del Sur) es el resultado de un complejo acople entre la superficie del océano y la atmósfera.

- TSM Pacífico: temperatura superficial del mar en el Pacífico.
- TSM Atlántico: temperatura superficial del mar en el Atlántico.

En este trabajo se tomó un total de 4 indicadores climáticos, que se describen brevemente la tabla 3.1, aquellos que en la literatura aparecen como los más influyentes en esta porción Sur del continente y además los que guardan mayor relación causal con el caudal del Sistema Magdalena-Cauca [20].

Es importante tener presente la procedencia e información que contiene cada uno de los cuatro índices con los que trabajaremos

Índice Climático	Descripción
Modo Meridional del Atlántico (AMMsst)	El modo describe la variabilidad del sistema acoplado océano-atmósfera en los trópicos. [9]
Índice bivariado ENOS (BEST)	Calculado a partir de la combinación de una normalización de las series de tiempo SOI y la temperatura superficial del mar de NIÑO34. [19] [17].
El Niño y La Niña	Oscilación del sistema océano-atmósfera en el Pacífico tropical oriental. El Niño y la Niña son los eventos extremos de El Niño-Oscilación del Sur (ENOS), que es el modo dominante de variabilidad en el océano Pacífico. [22]

Tabla 3.1: Descripción de los índices climáticos

Una porción significativa de índices climáticos está determinado en el campo de los eventos teleconectados, fenómenos que tienen lugar en sitios diferentes a aquellos en los cuales las mediciones son tomadas. Esto se debe a procesos de transporte, movilización de materia, energía y momentum, que ocasionan cambios y fenómenos sobre porciones diversas del globo terráqueo, que por su complejidad no pueden incluirse con facilidad en las ecuaciones dinámicas que describen los procesos y por ello se detectan a partir de mediciones [28]. La comprensión de las implicaciones de los fenómenos teleconectados está estrechamente relacionada con el concepto de la transferencia de información, pues justamente buscamos toda esa información que las partes del sistema se están compartiendo, lo que están transportando hace parte de su descripción mecánica.

El problema dinámico subyacente a los sistemas climáticos e hidrológicos es de tipo causal, la variabilidad sobre uno de los procesos tiene consecuencia en el otro y

### **3.2. INFLUENCIA DINÁMICA DE FENÓMENOS GLOBALES EN LA HIDROLOGÍA DEL SISTEMA MAGDALENA-CAUCA**

no necesariamente de manera viceversa. La información dinámica que consideramos está contenida en las series de tiempo, que son la realización del sistema y una de sus posibles configuraciones dentro del espacio de fase. Para desencriptar toda esa dinámica causal contenida dentro del sistema en las series de tiempo se usan herramientas estadísticas y computacionales [25].

Existen factores locales y regionales que influencian el ciclo del agua, la altura del territorio, el viento, la radiación solar controlando la temperatura y la presión que influyen en la humedad del aire [3]. También, algunos factores climáticos globales tienen influencia en la dinámica y la variabilidad del régimen hidrológico. Estas características hablan de la inherente complejidad del sistema que estudiamos y la necesidad de medirlo con herramientas diseñadas para el tipo de dinámica que describe [28]. En este caso tomaremos los factores globales que tienen una influencia causal en la hidrología del Sistema Magdalena-Cauca, considerando una serie multivariada compuesta por la serie de caudal y los índices climáticos AMMsst, BEST, NIÑO3 y NIÑO34.

## **3.2. Influencia dinámica de fenómenos globales en la hidrología del Sistema Magdalena-Cauca.**

La figura (3.4) revela la presencia marcada de estacionalidad en la serie de tiempo que estamos investigando. Las fluctuaciones periódicas a lo largo del eje temporal sugieren patrones recurrentes, destacando la importancia de comprender y modelar la variabilidad estacional de los datos. Sin embargo, para respaldar de manera rigurosa nuestras observaciones gráficas, hemos empleado la prueba aumentada de Dickey-Fuller. Esta evaluación matemática añade una capa de validación cuantitativa a nuestras sospechas, permitiéndonos confirmar de manera objetiva la existencia de estacionalidad en la serie de tiempo. La combinación de análisis visual y pruebas estadísticas fortalece la robustez de nuestra comprensión de la estacionalidad presente en los datos, proporcionando así una base sólida para interpretar y modelar adecuadamente la serie temporal en cuestión.

Finalmente, con un  $p - value$  de 0,000047 garantizamos que los datos del caudal de Calamar con los que trabajamos son estacionarios.

Además, en nuestro análisis de la serie de tiempo del caudal en el Sistema Magdalena-Cauca, no solo nos limitamos a verificar la estacionalidad visualmente, sino que también empleamos herramientas analíticas más profundas para comprender la dependencia temporal en los datos. Realizamos gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF), todo en la figura (3.5), con el objetivo de obtener información valiosa sobre los hiperparámetros del primer modelo que empleamos en

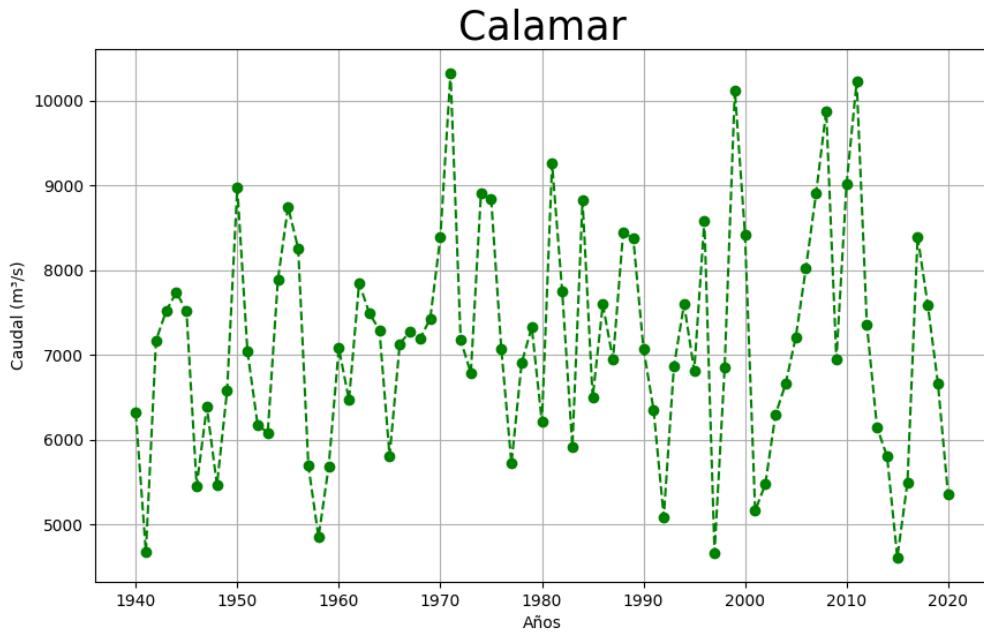


Figura 3.4: Histórico con el promedio anual del caudal de la estación hidrológica de Calamar.

nuestras predicciones, el SARIMAX. Estos gráficos son fundamentales en el análisis de series temporales y proporcionan una visión detallada de la correlación entre las observaciones a lo largo del tiempo.

El ACF nos revela la correlación entre la serie temporal y sus valores rezagados en distintos intervalos de tiempo. Esta herramienta nos ayuda a identificar patrones repetitivos, ciclos y estacionalidades en los datos. Por otro lado, el PACF se centra en la correlación directa entre observaciones a una distancia específica, excluyendo las contribuciones de los rezagos intermedios. Este gráfico es esencial para entender la estructura de dependencia temporal y nos proporciona información clave para determinar el orden apropiado del modelo autorregresivo (AR) en nuestro enfoque SARIMAX.

En el contexto del análisis de series temporales, estos gráficos desempeñan un papel crucial en la identificación de patrones, la selección de modelos y la mejora de la precisión en las predicciones. Al comprender la correlación entre observaciones en diferentes momentos, podemos ajustar adecuadamente los hiperparámetros de nuestros modelos, permitiéndonos realizar predicciones más precisas y contextualmente informadas sobre el caudal en el sistema hidrográfico estudiado.

Mientras que los gráficos de autocorrelación y autocorrelación parcial nos pro-

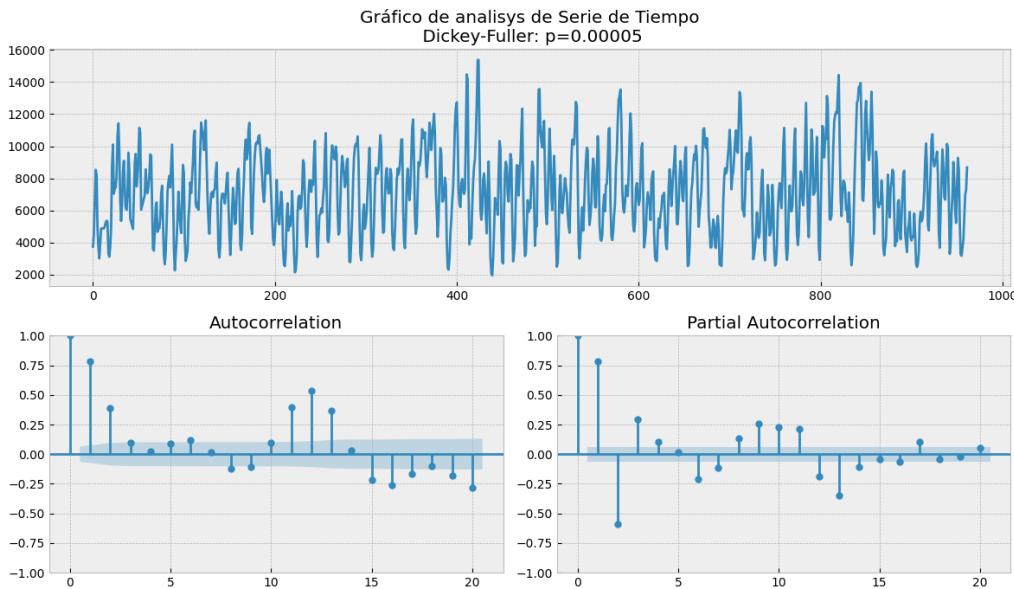


Figura 3.5: Gráfico de Autocorrelación y Autocorrelación Parcial.

porcionaron valiosa información sobre la dependencia temporal en la serie de tiempo del caudal en el Sistema Magdalena-Cauca, ahora nos enfocaremos en otro aspecto crucial del análisis de series temporales: la partición adecuada de los datos. En el contexto de las series temporales, conservar el orden temporal de las observaciones es esencial, ya que este orden guarda información vital sobre la dinámica del problema. La correcta partición de los datos nos permite entrenar nuestros modelos con información histórica y evaluar su desempeño en intervalos futuros, replicando de manera más precisa las condiciones del mundo real y mejorando así la capacidad predictiva de nuestros modelos. Así, la transición de explorar patrones temporales a abordar la partición estratégica de los datos se presenta como un paso fundamental en nuestro análisis, contribuyendo a la robustez y eficacia de nuestras predicciones en el contexto hidroclimatológico que investigamos.

Así pues, teniendo en cuenta todo lo mencionado anteriormente, el primer modelo que usamos para hacer las predicciones es el SARIMAX, la diferencia entre las dos corridas que se realizaron con este método fue la elección de usar o no variables exógenas. El uso de un modelo con variables exógenas frente a un modelo basado únicamente en la serie temporal de caudal tiene varias implicaciones:

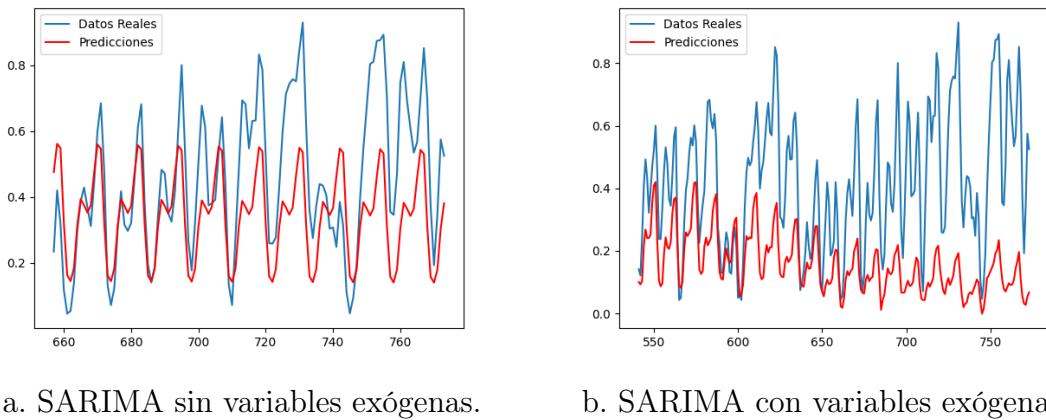
- **Información Adicional:** Al incluir variables exógenas, el modelo tiene acceso a información adicional que puede ayudar a capturar patrones y tendencias en los datos. Estas variables podrían representar factores externos que afectan el caudal y que no están inherentemente reflejados en la serie temporal.

- Mejora de la Precisión: Si las variables exógenas están relacionadas de alguna manera con el caudal, la inclusión de estas puede mejorar la precisión de las predicciones. Por ejemplo, si las variables exógenas representan condiciones climáticas, la temperatura del agua, o características geográficas, podrían influir en el caudal y ser útiles para hacer predicciones más precisas.
- Complejidad del Modelo: La inclusión de variables exógenas puede aumentar la complejidad del modelo. Aunque esto puede ser beneficioso para capturar relaciones más sofisticadas en los datos, también puede aumentar el riesgo de sobreajuste, especialmente si hay demasiadas variables exógenas en comparación con la cantidad de datos disponibles.
- Requisitos de Datos: El uso de variables exógenas implica que debes tener datos disponibles para esas variables en el período de predicción. Si las variables exógenas no están disponibles para el futuro, el modelo no podrá utilizar esa información para hacer predicciones.
- Interpretación: Modelos con variables exógenas pueden ser más difíciles de interpretar, ya que ahora estás considerando múltiples variables en lugar de solo la serie temporal de caudal. La interpretación de cómo cada variable afecta las predicciones puede requerir un análisis más detenido.

En última instancia, la elección entre usar un modelo con variables exógenas o basado únicamente en la serie temporal depende de la naturaleza de los datos, la disponibilidad de información adicional relevante y los objetivos específicos de la predicción. Se recomienda realizar experimentos y evaluaciones comparativas para determinar cuál enfoque funciona mejor para tu conjunto de datos y contexto particular.

Adicional a lo que logra verse en la figura (3.6), sobre las mejores predicciones del modelo sin las variables exógenas, es decir, utilizando únicamente la serie de tiempo del caudal, la métrica del Root Mean Squared Error (RMSE) para el conjunto de prueba muestra mejores resultados para el SARIMA sin variables exógenas. Los valores para el MSE del modelo SARIMA sin variables exógenas y con variables exógenas, respectivamente son, 0,202 y 0,321, teniendo en cuenta que los valores menores del MSE indican errores menos significativos y por tanto, mejores aproximaciones.

Por otro lado, para proporcionar una visión completa de la distribución y variabilidad de estos datos a lo largo del tiempo, se emplearon gráficos de violín (figura 3.7), una herramienta visual poderosa que permite apreciar la forma y dispersión de las series temporales. Estos violines fueron generados para cada serie temporal, tanto para las covariables climáticas como para la variable a predecir, en los conjuntos de entrenamiento, validación y prueba.



a. SARIMA sin variables exógenas.      b. SARIMA con variables exógenas.

Figura 3.6: Predicciones con el modelo SARIMA.

El diagrama de violín no solo destaca la distribución de los valores en cada serie temporal, sino que también revela posibles patrones y variaciones estacionales a lo largo del tiempo. Este análisis visual es esencial para comprender la complejidad inherente de las series temporales hidroclimatológicas, ya que proporciona insights sobre la estacionalidad, la presencia de outliers y la variabilidad de las variables en diferentes períodos.

Particularmente en el contexto de las redes LSTM, este enfoque visual es valioso. Las LSTM, como modelos de redes neuronales recurrentes, son altamente sensibles a patrones temporales y dependencias a largo plazo. Al visualizar la distribución de las series temporales, especialmente en relación con la variable objetivo (caudal), se pueden identificar características clave que las LSTM podrían aprender y aprovechar para mejorar la precisión de las predicciones. Este análisis visual, complementado con técnicas de preprocesamiento y ajuste de hiperparámetros, contribuye a un enfoque integral para el modelado y la predicción de series temporales en el ámbito hidroclimatológico.

El empleo de modelos de redes neuronales recurrentes, como el LSTM, para la predicción de caudales es una estrategia prometedora que ha demostrado su eficacia en el ámbito hidroclimatológico. En este contexto, se ha seleccionado un conjunto de índices climáticos que han revelado tener una conexión significativa con la cuenca Magdalena-Cauca, según el trabajo realizado previamente en el pregrado. Estos índices, que incluyen variables como la temperatura superficial del mar y patrones climáticos relevantes, son incorporados como covariables en el modelo LSTM para capturar mejor las complejidades de las relaciones climáticas que influyen en los caudales.

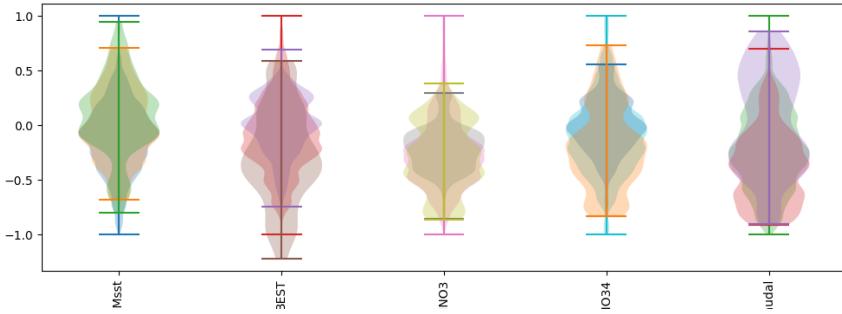


Figura 3.7: Diagrama de Violín para las series de tiempo.

En una primera iteración del modelo, se realizaron experimentos ajustando hiperparámetros clave y decidiendo que se dejarán fijas la cantidad de neuronas en la capa LSTM. Esto busca aprovechar la capacidad del modelo para aprender patrones temporales y adaptarse a las relaciones no lineales presentes en los datos. Los resultados de esta configuración inicial proporcionan una base para la comparación con las corridas posteriores.

En una segunda corrida del modelo LSTM, se optó por ajustar la tasa de aprendizaje, tomando el mismo optimizador pero una tasa de aprendizaje más pequeña, incrementar el número de épocas y aumentar el batch size. Estos ajustes buscan refinamientos adicionales en la capacidad del modelo para realizar predicciones precisas en series temporales hidrológicas. La comparación de resultados entre la primera y segunda corrida permitirá evaluar el impacto de estos cambios en la calidad de las predicciones.

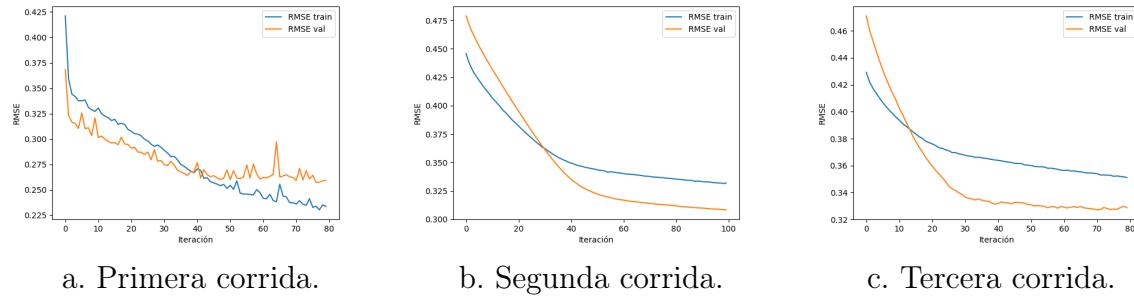
En un enfoque más simplificado, la tercera corrida del modelo LSTM se llevó a cabo utilizando únicamente la serie de tiempo del caudal como entrada. Esto representa un modelo LSTM univariado con predicción de un solo paso adelante, donde la red neural se entrena exclusivamente con la información del caudal pasado para realizar pronósticos futuros. Este enfoque busca evaluar la capacidad intrínseca del modelo para aprender y predecir el comportamiento del caudal sin la influencia de variables climáticas adicionales.

Estas distintas configuraciones y experimentos con el modelo LSTM ofrecen una visión detallada de cómo factores como la complejidad del modelo, la influencia de índices climáticos y la variación en los hiperparámetros afectan la capacidad del modelo para realizar predicciones precisas en el contexto específico de la cuenca Magdalena-Cauca.

Como puede notarse en la tabla 3.2, la mejor métrica para los datos de prueba se obtiene en la primera corrida, pero el histórico para entrenamiento y validación,

LSTM	RMSE train	RMSE val	RMSE test
Corrida 1	0.233	0.259	0.309
Corrida 2	0.327	0.305	0.395
Corrida 3	0.348	0.328	0.471

Tabla 3.2: RMSE para las diferentes corridas de modelos LSTM.



a. Primera corrida.

b. Segunda corrida.

c. Tercera corrida.

Figura 3.8: Comparación entre las diferentes corridas para el modelo LSTM.

aunque tiende a disminuir, presenta fluctuaciones y dificultad para estabilizarse. El historial del RMSE que muestra una tendencia a disminuir con muchas fluctuaciones puede sugerir la presencia de overfitting, una situación en la cual el modelo no solo aprende patrones generales en los datos, sino también el ruido específico presente en el conjunto de entrenamiento. Estas fluctuaciones, o picos y valles en la métrica de rendimiento, indican que el modelo está capturando detalles particulares de los datos de entrenamiento que no son generalizables a nuevas observaciones. En otras palabras, el modelo podría estar adaptándose demasiado a las peculiaridades del conjunto de entrenamiento en lugar de aprender patrones más amplios y aplicables.

Por otro lado, las fluctuaciones en el historial del RMSE también podrían sugerir dificultades en la convergencia del modelo. Esto puede ser el resultado de una tasa de aprendizaje inapropiada o de la complejidad inherente de los datos. Si el modelo tiene problemas para llegar a una solución estable durante el proceso de entrenamiento, es posible que veamos variaciones significativas en la métrica de rendimiento a lo largo del tiempo. En este caso, ajustar la tasa de aprendizaje, la inicialización de pesos u otros hiperparámetros relevantes podría ayudar a estabilizar el proceso de entrenamiento y mejorar la convergencia hacia una solución más consistente y generalizable.

Así mismo, en la figura 3.9, la cual muestra la comparación entre los diferentes

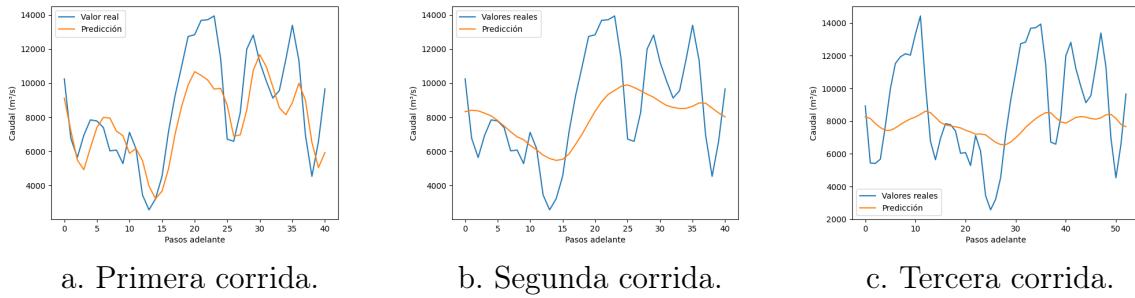


Figura 3.9: Comparación entre las diferentes corridas para las predicciones del modelo LSTM y los valores reales.

modelos LSTM y los valores reales del caudal para el Sistema Magdalena-Cauca, es de notar que la refinación de los hiperparámetros y la simplificación del modelo al usar únicamente la serie de tiempo del caudal, no logran capturar la dinámica real de la hidrología en la cuenca, lo que es verdaderamente importante a la hora de hacer predicciones y más al considerar que para las cuestiones hidrológicas los máximos y los mínimos en el caudal son eventos extremos que es fundamental anticipar, ya que representan acontecimientos que inciden notablemente en la vida del ecosistema y de la comunidad que habita en las cercanías de la misma. Por otro lado, la primera corrida de estas redes LSTM sugiere una mejor aproximación, acercándose de manera más adecuada a los eventos extremos del sistema, sin embargo, es importante es futuros trabajos realizar otro tipo de pruebas que nos permitan descartar sobreajuste, puesto que las fluctuaciones en el histórico del RMSE lo sugieren y dado que la disponibilidad de datos no es muy alta, es difícil identificar si el modelo está logrando hacer generalizaciones sobre el problema o si en realidad está memorizando los datos, cosa que no es nada conveniente en modelos predictivos y menos al tratarse de series de tiempo.

En la segunda parte de la modelación hacemos uso de otro tipo de arquitecturas tipo RNN, para evaluar qué clase de modelos permite acercarse mejor al problema. El uso de modelos GRU en la predicción de caudales, especialmente cuando se trata de series de tiempo multivariadas, puede ser crucial debido a su capacidad para manejar dependencias temporales en datos complejos. Aunque los GRU son una versión simplificada de las LSTM, su estructura con dos puertas (una de actualización y una de reinicio) les permite capturar patrones temporales de manera efectiva. Esta simplicidad a menudo se traduce en una computación más eficiente y, en algunos casos, en un mejor rendimiento en comparación con modelos más complejos.

En el contexto de la predicción de caudales, donde múltiples variables climáticas pueden afectar la dinámica del flujo de agua, los GRU ofrecen la posibilidad de adaptarse eficazmente a la variabilidad de los datos. Su capacidad para retener y olvidar

información pasada mediante las puertas de actualización y reinicio permite modelar la influencia de diversas variables exógenas en el caudal, proporcionando así una representación más precisa de la interacción compleja entre los factores climáticos y los flujos hídricos.

Aunque los GRU son menos complejos que las LSTM, su rendimiento comparable y, en ocasiones, superior, destaca su utilidad en problemas específicos como la predicción de caudales. Además, su estructura más simple facilita la interpretación de los resultados y puede ser preferible cuando se cuenta con recursos computacionales limitados. La elección entre modelos más complejos y modelos GRU dependerá en última instancia de la naturaleza específica del problema y de los datos disponibles, siendo los GRU una opción valiosa para la predicción precisa de caudales en entornos hidroclimatológicos.

GRU	RMSE train	RMSE val	RMSE test
Corrida 1	0.063	0.060	0.083
Corrida 2	0.139	0.128	0.254
Corrida 3	0.137	0.123	0.252

Tabla 3.3: RMSE para las diferentes corridas de modelos GRU.

En estas corridas para los modelos GRU, el que presenta una mayor diferencia en sus hiperparámetros es la primera corrida, la cual tiene *tanh* como su función de activación, un regularizador *L2* de 0,01, no tiene capas de dropout y trabaja con un optimizador Adam con tasa de aprendizaje 0,001. La arquitectura de los otros dos modelos GRU conserva la misma arquitectura y la misma cantidad de neuronas por capa, solo que trabaja con función de activación *relu*, tasas de aprendizaje más finas, optimizador Nesterov Adam y la diferencia es que el segundo tiene dos capas dropout y la tercera tiene tan solo una capa de ellas.

La tabla (3.3), además de la figura (3.10), muestran mejores resultados para los hiperparámetros iniciales en el modelo GRU, la curva del histórico del RMSE es más estable, mientras que las otras presentan fluctuaciones mientras decaen, principalmente en el entrenamiento, lo que da cuenta de dificultades para capturar y generalizar patrones, lo único es que sí tiene un efecto agregar capas de dropout, pero la mejoría puede no ser significativa y los hiperparámetros más representativos tienen que ver más con las funciones de activación, la regularización y el optimizador que se usa.

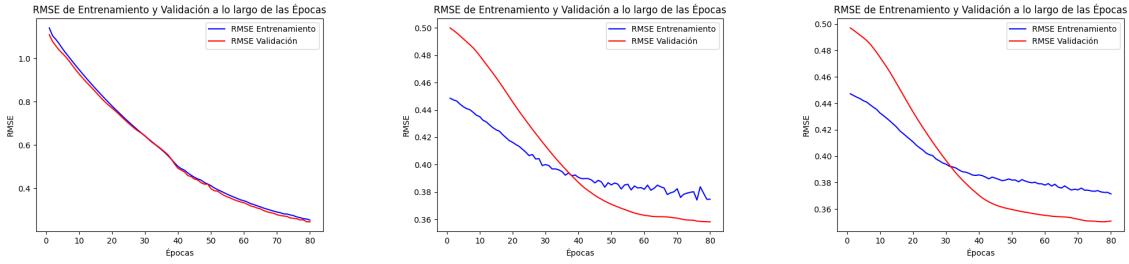


Figura 3.10: Comparación entre las diferentes corridas para el modelo GRU.

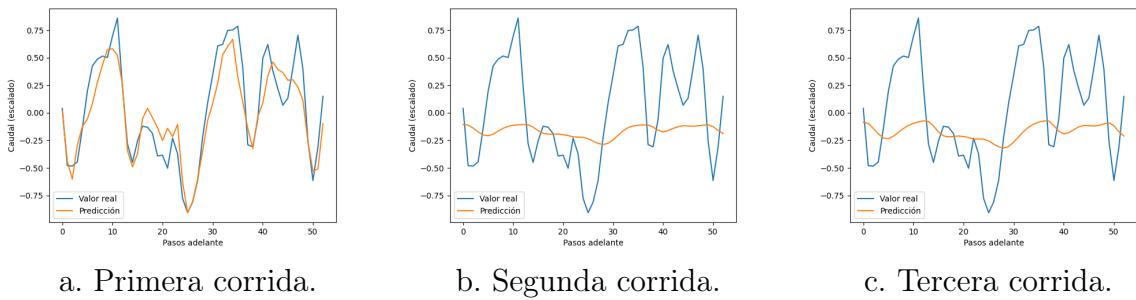
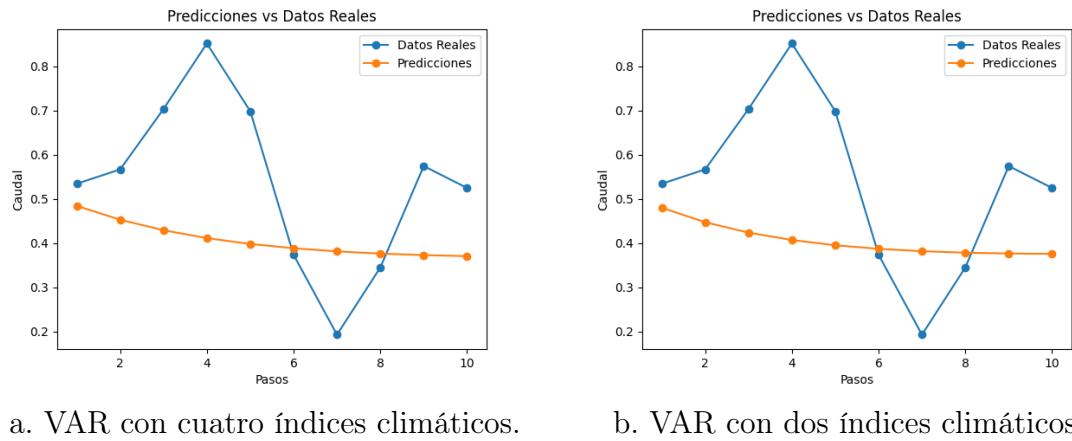


Figura 3.11: Comparación entre las diferentes corridas para las predicciones del modelo GRU y los valores reales.

Teniendo en mente las diferencias entre las iteraciones, concluimos que la predicción de caudales mediante la implementación de la arquitectura GRU, junto con la inclusión de cuatro índices climáticos como variables exógenas, demuestra una capacidad efectiva para adaptarse a los patrones generales que rigen la dinámica del sistema hidroclimatológico (figura 3.11). Este enfoque logra no solo capturar las tendencias generales de la serie temporal, sino también realizar aproximaciones precisas de los valores extremos, aspecto crucial en la predicción de eventos hidrológicos significativos.

En contraste, los otros dos casos de predicción presentan limitaciones notables. Dichas predicciones muestran una tendencia a producir valores extremadamente estables, cercanos a un valor central. Esta estabilidad constante se aleja significativamente de la realidad observada en los datos reales de caudales, donde se espera una variabilidad más pronunciada y respuestas a condiciones climáticas diversas.

En resumen, la implementación de GRU con variables exógenas emerge como una estrategia robusta que logra abordar de manera efectiva la complejidad inherente a



a. VAR con cuatro índices climáticos.

b. VAR con dos índices climáticos.

Figura 3.12: Predic平ones con el modelo VAR.

la predicción de caudales, ofreciendo resultados que reflejan con mayor fidelidad las fluctuaciones y patrones característicos del sistema hidroclimatológico estudiado.

Para concluir, la última aproximación empleó un modelo de tipo VAR (figura 3.12). La distinción clave entre las dos iteraciones radicó en que la primera incorporó cuatro variables climáticas junto con el caudal, mientras que la segunda utilizó únicamente dos índices climáticos además del caudal. A pesar de esta diferencia en la configuración, las disparidades en el rendimiento son mínimas. Esto podría sugerir que la dinámica climática, al menos con las variables consideradas, no aporta significativamente a la predicción de los caudales. Curiosamente, las predicciones tampoco logran ajustarse de manera precisa al comportamiento general del sistema, lo que plantea interrogantes adicionales sobre la complejidad y las interrelaciones en el sistema hidroclimatológico estudiado. Estos hallazgos resaltan la importancia de una evaluación continua y detallada de distintos enfoques y modelos en la búsqueda de comprender y prever de manera efectiva los caudales en la cuenca Magdalena-Cauca.

Es importante señalar que, aunque los modelos Vector AutoRegressive (VAR) han demostrado ser eficaces en una variedad de contextos, su rendimiento puede verse afectado en problemas hidroclimatológicos complejos, como la predicción de caudales en la cuenca Magdalena-Cauca. La aparente falta de mejora sustancial al reducir el número de variables climáticas sugiere que la dinámica subyacente del sistema puede no estar totalmente capturada por las variables seleccionadas. Además, la variabilidad extrema y la dependencia temporal pueden introducir desafíos que los modelos VAR, que asumen estacionariedad y homocedasticidad, podrían no abordar de manera efectiva. Es crucial reconocer que la elección del modelo adecuado depende en gran medida de la naturaleza específica del sistema y de la complejidad de las interacciones entre las variables. Estos resultados destacan la necesidad continua de

explorar y adaptar enfoques de modelado para abordar las particularidades únicas de la predicción de caudales en la región de la cuenca Magdalena-Cauca.

# Capítulo 4

## Conclusiones y Perspectivas.

De acuerdo con los resultados de este trabajo, los fenómenos climáticos globales tienen influencia dinámica en la variabilidad de la hidrología del Sistema Magdalena-Cauca, sin embargo, no todos los modelos logran capturar dichas relaciones. El estudio lo realizamos por medio de los análisis de las series de tiempo, puesto que ellas contienen las propiedades físico-estadísticas y las relaciones dinámicas causales que guardan los procesos, junto con el patrón de variabilidad espacio-temporal propio para cada fenómeno.

El Sistema Magdalena-Cauca está acoplado principalmente a los eventos climáticos asociados con el ENOS, lo que explica que la dinámica de esta cuenca esté condicionada por los patrones de precipitación. Las variables de estado son incluidas en el sistema por medio de las series de tiempo, las cuales guardan las propiedades físico-estadísticas y relaciones causales del proceso.

Los modelos empleados en este estudio revelan diferentes perspectivas y desafíos en la predicción de caudales en la cuenca Magdalena-Cauca. La tendencia general sugiere que los modelos basados en la periodicidad, como SARIMA, logran capturar los ciclos recurrentes del sistema hidroclimático, aunque muestran limitaciones en la predicción de eventos extremos. Por otro lado, LSTM, un modelo robusto en el aprendizaje de patrones temporales, exhibe cierta propensión al sobreajuste con los hiperparámetros utilizados, indicando una memorización de datos en lugar de una comprensión profunda. Este riesgo de sobreajuste subraya la importancia crítica de la selección cuidadosa de hiperparámetros y la necesidad de abordar el equilibrio entre complejidad y generalización.

Los modelos VAR, a pesar de su utilidad en diversos contextos, parecen no ser la elección óptima para la predicción de caudales en esta cuenca. Su tendencia a sobre-simplificar la dinámica del sistema, asumiendo estacionariedad y homocedasticidad, puede ser una limitación significativa en la representación de las complejas interac-

ciones entre las variables climáticas y el caudal. Por último, la arquitectura GRU destaca como la más efectiva, especialmente al incluir índices climáticos como variables exógenas. Este enfoque demuestra una capacidad sólida para generalizar resultados y manejar eventos extremos, lo que subraya la importancia de la flexibilidad y la adaptabilidad de los modelos en la predicción hidroclimatológica. En perspectiva, se puede hacer un estudio más extenso de este mismo enfoque sobre la zona de estudio. Lo primero es tomar varios puntos de medición sobre la cuenca, es decir, hacer un mapeo con diferentes estaciones hidrométricas sobre el Sistema Magdalena-Cauca para determinar si la influencia climática es homogénea o si pueden existir factores puntuales que modifiquen la dinámica de una estación a otra, aunque hagan parte de la misma cuenca. Es importante incluir otros índices climáticos para conocer con mayor exactitud la gama de eventos que están condicionando la hidrología de la región.

Para nutrir la dinámica del sistema y la coherencia de los resultados, es importante estudiar también otras factores como la precipitación y la humedad con la cuenca del Magdalena-Cauca, esto permite determinar si la dinámica hidrológica está asociada al comportamiento de los índices climáticos o existe otro fenómeno que esté predominando en la evolución que exhibe la cuenca y por tanto pueda producir mejores resultados a la hora de hacer predicciones en el caudal de la estación Calamar.

Finalmente, usar bases de datos más grandes puede funcionar muy bien a la hora de entrenar con mayor precisión los modelos y aumentar la capacidad de generalización y predicción para el caudal de la estación Calamar.

# Bibliografía

- [1] Junyoung Chung. Caglar Gulcehre. KyungHyun Cho. Yoshua Bengio. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.* Université de Montréal., 2014.
- [2] Gustavo Herminio Trujillo Calagua. *La Metodología del Vector Autorregresivo: Presentación y Algunas Aplicaciones.* Universidad Científica del Sur., 2010.
- [3] Ma. del Carmen Jiménez Quiroz. *INDICADORES CLIMATICOS. Una manera para identificar la variabilidad climática a escala global.* Anexo del Informe Técnico:Elaboración de un boletín con información hidroclimática de los mares de México.
- [4] Patricio Rubio Carolina Martínez Alfonso Fernández. *Caudales y variabilidad climática en una cuenca de latitudes medias en Sudamérica: río Aconcagua, Chile Central.* Boletín de la Asociación de Geógrafos Españoles., 2012.
- [5] Acuerdo IDEAM - Cormagdalena. *Estudio Ambiental de la Cuenca Magdalena - Cauca y elementos para su Ordenamiento Territorial.* 2001.
- [6] Jonathan Romero Cuellar; Andres Buitrago Vargas; Tatiana Quintero Ruiz y Félix Francés. *Simulación hidrológica de los impactos potenciales del cambio climático en la cuenca hidrográfica del río Aipe, en Huila, Colombia.* Revista Iberoamericana del Agua., 2018.
- [7] Yaset Martínez Valdés Víctor Michel Villalejo García. *Ecohidrología-Ecohídrica: claves para la gestión integrada de los recursos hídricos.* Ingeniería Hidráulica y Ambiental., 2019.
- [8] Thomas Guerrero. Gloria Amaris Humberto Ávila. *Aplicación de modelo ARI-MA para el análisis de series de volúmenes anuales en el río Magdalena.* Universidad Distrital Francisco José de Caldas., 2017.
- [9] Sofía Prado González. *Caracterización de la Circulación Meridional Atlántica en 26,5°N.* Universidad de Vigo, 2017.
- [10] Damián Jorge Matich. *Redes Neuronales: Conceptos Básicos y Aplicaciones.* Universidad Tecnológica Nacional de Argentina., 2001.
- [11] José Alberto Mauricio. *Introducción al Análisis de Series Temporales.* Universidad Complutense de Madrid., 2007.

- [12] María Paula Llano. Melanie Meis. *Modelado Estadístico del Caudal Mensual en la Baja Cuenca del Plata*. CONICET, 2017.
- [13] Sebastian Raschka. Vahid Mirjalili. *Python Machine Learning*. Marcombo, 2019.
- [14] Alfonso Novales. *Modelos vectoriales autoregresivos (VAR)*. Universidad Complutense de Madrid, 2017.
- [15] Thomas T. Veblen; Kenneth R. Young Antony R. Orme. *The Physical Geography of South America*. Oxford University., 2007.
- [16] Humberto Peña. *Desafíos de la seguridad hídrica en América Latina y el Caribe*. Recursos Naturales e Infraestructura-CEPAL., 2016.
- [17] Carolina Ramírez C.; Jorge J. Vélez U.; Andrés J. Peña Q. *Analizando índices climáticos para predecir la lluvia mensual en una región agrícola de los andes del norte (Caldas, Colombia)*. Investigación Geográfica de Chile., 2018.
- [18] David Gutiérrez Ramos. *Análisis de Series Temporales: Estudio de una Caso Práctico*. Universidad Politécnica de Madrid., 2022.
- [19] Luis Alfonso Naranjo Sarmiento; José Enrique García Ramos. *Influencia de la Oscilación del Sur - El Niño y La Niña sobre la temperatura y la velocidad del viento en la Subcuenca de los ríos Blancos y del Sector Cordón del Plata - Argentina*. Universidad Internacional de Andalucia., 2016.
- [20] I. Hoyos; J. Cañón Barriga; T. Arenas Suárez; F. Dominguez; B. A. Rodríguez. *Variability of regional atmospheric moisture over Northern South America: patterns and underlying phenomena*. Springer Nature, 2018.
- [21] Jorge Guerra Rodríguez. *Fundamentos y variantes de los modelos ARIMA para el análisis de series temporales. Aplicación a la estadística universitaria*. Universidad de La Laguna., 2022.
- [22] Mauricio Bedoya; Claudia Contreras; Franklin Ruiz. *Alteraciones del Régimen Hidrológico y de la Oferta Hídrica por Variabilidad y Cambio Climático*. Estudio Nacional del Agua, 2010.
- [23] Alonso Sepúlveda S. *Hidrodinámica*. Universidad de Antioquia, 2013.
- [24] F Javier Sánchez San Román. Dpto Geología. Universidad de Salamanca (España). *Ciclo Hidrológico*. url : <http://hidrologia.usal.es>.
- [25] María Fernanda Reynoso Savio. *Índices q De Tsallis En Un Sistema De No Equilibrio: La Capa De Ozono Estratosférico*. Universidad Nacional de la Pampa., 2012.
- [26] Darwin Giusseppe Marín Vilca. Ian Augusto Pineda Torres. *Modelo predictivo Machine Learning aplicado a análisis de datos Hidrometeorológicos para un SAT en Represas*. Universidad Tecnológica del Perú., 2019.

- [27] Pablo Alvarez-De-Toledo. Adolfo Crespo. Fernando Núñez. Carlos Usabiaga. *Introducción de elementos autorregresivos en modelos de dinámica de sistemas.* Revista de Dinámica de Sistemas, 2006.
- [28] Daniel S Wilks. *Statistical Methods in the Atmospheric Sciences. Second edition.* Department of Earth y Atmospheric Sciences Cornell University, 2006.
- [29] Shudong Yang. Xueying Yu. Ying Zhou. *LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example.* International Workshop on Electronic Communication y Artificial Intelligence., 2020.