EVALUACIÓN DE METODOLOGÍAS PARA EL PRONÓSTICO DE SERIES DE PRECIPITACIÓN EN LA CUENCA DEL RIO SOGAMOSO EN EL DEPARTAMENTO DE SANTANDER COLOMBIA

LAURA STEPHANY CAMARGO RUEDA
ALEJANDRA PAEZ CANABAL

UNIVERSIDAD SANTO TOMAS
FACULTAD DE INGENIERIA AMBIENTAL
BOGOTA D.C., COLOMBIA
2019

EVALUACIÓN DE METODOLOGÍAS PARA EL PRONÓSTICO DE SERIES DE PRECIPITACIÓN EN LA CUENCA DEL RIO SOGAMOSO EN EL DEPARTAMENTO DE SANTANDER COLOMBIA

LAURA STEPHANY CAMARGO RUEDA
ALEJANDRA PAEZ CANABAL

PROYECTO DE GRADO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERA AMBIENTAL

DIRECTOR:
MIGUEL ANGEL CAÑON RAMOS
INGENIERO AMBIENTAL
MSC(C) EN HIDROSISTEMAS

CODIRECTOR:
DARWIN MENA RENTERIA
INGENIERO AMBIENTAL Y SANITARIO
ESPECIALISTA EN GESTIÓN AMBIENTAL
MAGISTER EVALUACIÓN DE RECURSOS HÍDRICOS

UNIVERSIDAD SANTO TOMAS
INGENIERIA AMBIENTAL
BOGOTA D.C., COLOMBIA
2019

PÁGINA DE ACEPTACIÓN

_____

_____

_____

NOTA DE ACEPTACIÓN

_____

Presidente del Jurado

_____

Jurado

_____

Jurado

_____

Ciudad y fecha de presentación

DEDICATORIA Y AGRADECIMIENTOS

A nuestras familias por la confianza y el apoyo que nos brindaron durante todo el proceso.

A nuestros profesores por guiarnos en cada paso y decisión tomada.

A Dios por la fortaleza moral y emocional, por ser guía incondicional de nuestro proceso de formación.

A nuestras familias por el apoyo incondicional en todos los aspectos, por los valores y la ética que nos inculcaron durante nuestro crecimiento personal y profesional, por los consejos y las enseñanzas dadas.

A los ingenieros Miguel Ángel Cañón Ramos y Darwin Mena Rentería acompañantes durante todo el recorrido y elaboración de este trabajo, por ser nuestros guías, consejeros, y también por su tiempo, paciencia y dedicación.

Finalmente, a todas las personas que depositaron su confianza y fe en nosotras desde el comienzo de nuestra carrera.

TABLA DE CONTENIDO

CONTENIDO DE FIGURAS

CONTENIDO DE TABLAS

CONTENIDO DE ECUACIONES

GLOSARIO

Algoritmo: método y notación de las distintas formas de cálculo. Ciencia del cálculo aritmético y algebraico; teoría de los números.

Amenaza: potencial ocurrencia de un hecho que pueda manifestarse en un lugar específico, con una duración e intensidad determinada. Se considera la materialización del riesgo.

Análisis: examen detallado de una cosa para conocer sus características o cualidades o su estado, y extraer conclusiones que se realiza separando o considerando por separado las partes que las constituyen.

Anomalía: cambio o desviación respecto a lo que es normal, regular, natural o previsible.

Climatología: ciencia que estudia el clima, sus variedades, cambios y las causas de estos. Conjunto de condiciones atmosféricas propias de un determinado clima.

Conglomeración: unión fuerte de partículas o fragmentos de una o varias sustancias formando una masa compacta.

Convección: forma de transferencia de calor por medio de un fluido que transporta el calor entre zonas con diferentes temperaturas.

Convergencia: propiedad que poseen dos o más cosas que confluyen en un mismo punto.

Cuenca: territorio cuyas aguas afluyen todas a un mismo rio, lago o mar.

Código: es el texto escrito en un lenguaje de programación que ha de ser compilado o interpretado para ejecutarse en una computadora.

Desertificación: proceso de degradación ecológica en el que el suelo fértil pierde total o parcialmente el potencial de producción.

Fenómeno: conjunto de manifestaciones que caracterizan un proceso.

Galpón: construcción relativamente grande que se destina al depósito de mercancía o maquinaria. Son construcciones con una sola puerta.

Hidrología: parte de las ciencias naturales que trata de las aguas, tanto superficiales como subterráneas.

Impacto ambiental: efecto que produce una actividad humana sobre el medio ambiente, puede extenderse a los efectos de un fenómeno natural catastrófico.

Metodología: conjunto de métodos que se siguen en una investigación o en una exposición doctrinal.

Orográfico: elevaciones que pueden existir en una zona en particular.

Precipitación: lluvia, nieve o granizo que se deposita en la tierra.

Probabilidad: cálculo matemático de las posibilidades que existen de que una cosa se cumpla o suceda al azar.

Pronóstico: calendario en que se anuncia uno o varios fenómenos meteorológicos.

Temperatura: estado de un cuerpo, grado o nivel térmico de un cuerpo o la atmosfera.

Variabilidad climática: según el Centro Internacional para la Investigación del fenómeno de El Niño, se refiere a variabilidad climática a una medida del rango en que los elementos climáticos, como temperatura o lluvia, varían de un año a otro.

RESUMEN

Se presenta la evaluación de diferentes modelos de pronóstico de precipitación en la cuenca del rio Sogamoso en Santander, Colombia. Se partió del estudio de cada una de las metodologías seleccionadas, se realizó una recopilación de los datos de las estaciones meteorológicas de la zona de estudio, para la identificación de las variables que se utilizaron en cada uno de los modelos seleccionados. Se desarrollaron de acuerdo a los datos encontrados por cada uno de los modelos, el análisis de consistencia de la información en donde se realizan los respectivos estudios del comportamiento que tuvo cada modelo de acuerdo a sus condiciones iniciales e ideales.

Para el desarrollo del proyecto se establecieron por modelos combinaciones de estaciones meteorológicas y los datos correspondientes a las variables (precipitación, humedad relativa, evapotranspiración, temperatura media y brillo solar) que se analizaron de acuerdo a las necesidades de cada uno de los modelos, con estas combinaciones y los resultados obtenidos según cada modelo, se dio respuesta a cuál es el más apto para la zona de estudio. Los modelos que se analizaron en este proyecto fueron regresión lineal simple, regresión lineal múltiple, modelos lineales generalizados, análisis de componentes principales y filtro de Kalman, cada uno requirió datos diferentes, ya que las características de cada uno son distintas.

Con las combinaciones de cada uno de los modelos se realizó el proceso de análisis de cumplimiento de los criterios a evaluar, y para esto se utilizaron el 70% del total de datos de cada una de las estaciones para la simulación y con el 30% restante la validación de cada uno. Los cinco modelos planteados se desarrollaron a partir de algoritmos y series de las variables dependiendo de los criterios de cada uno, con base en los resultados de los pronósticos desarrollados en los modelos se realizó la evaluación de las metodologías aplicando las métricas de desempeño.

Con base al desempeño de cada uno de los modelos, sus resultados, las métricas de desempeño y otras características, se determinó cual modelo se adecua a los datos de las estaciones y los requerimientos del modelo en el que se ejecutó, se identificaron cuales modelos presentan más errores y son descartables para la zona de la cuenca. Como resultado, se encontró que, de acuerdo con los modelos realizados y analizados, junto con la familia Gamma, el modelo cuyos mejores resultados se obtuvieron son los modelos lineales generalizados (MLG).

PALABRAS CLAVE: PRONÓSTICO, PRECIPITACIÓN, CÓDIGO, HIDROSOGAMOSO, MÉTRICAS, VARIABLES.

ABSTRACT

The evaluation of different precipitation forecasting models is presented in the Sogamoso River in Santander, Colombia. From the study of each of the selected methodologies, a compilation of the data of the meteorological stations of the study area was carried out, for the identification of the variables that were used in each of the selected models. They were developed according to the data found by each of the models, the consistency analysis of the information in which the respective studies of the behavior of each model were carried out according to their initial and ideal conditions.

For the development of the project, combinations of meteorological stations and data corresponding to the variables (precipitation, relative humidity, evapotranspiration, mean temperature and solar brightness) were established, which were analyzed according to the needs of each of the models, with these combinations and the results obtained according to each model, was answered which is the most suitable for the study area. The models that were analyzed in this project were simple linear regression, multiple linear regression, generalized linear models, principal component analysis and Kalman filter, each one required different data, since the characteristics of each are different.

With the combinations of each of the models, the process of analyzing compliance with the criteria to be evaluated was performed, and for this, 70% of the total data from each of the stations was used for the simulation and with the remaining 30% the forecast of each one. The five proposed models were developed from algorithms and series of variables depending on the criteria of each, based on the results of the forecasts developed in the models, the evaluation of the methodologies was carried out applying the performance metrics.

Based on the performance of each of the models, their results, performance metrics and other characteristics, it was determined which model fits the data of the stations and the requirements of the model in which it was executed, which models were identified. more errors and they are disposable for the area of the basin. As a result, it was found that, according to the models made and analyzed, together with the Gamma family, the model whose best results were obtained are the generalized linear models (MLG).

1. MARCO CONCEPTUAL

1.1. Introducción

Los pronósticos de precipitación han sido una herramienta que durante muchos años ha proporcionado información necesaria en la ocurrencia de eventos. Con los avances tecnológicos, la incertidumbre que se generaba en estos análisis ha disminuido, sin embargo, el constante cambio climático ha generado variaciones significativas en la dinámica de la precipitación, lo que dificulta la elaboración de un pronóstico [1].

El rio Sogamoso cuyo nacimiento yace en la confluencia del rio Suarez, queda ubicado en el departamento de Santander, este abastece a varios municipios del departamento y es el principal afluente de la hidroeléctrica Hidro Sogamoso. Este rio es representativo para la cuenca, no solo en temas ambientales sino también económicos, esto se debe a que el uso de su recurso logra abastecer hasta el 10% de la población colombiana en cuanto a temas energéticos [2]. Es por esto que se seleccionó la cuenca del rio Sogamoso, para realizar pronósticos de precipitación.

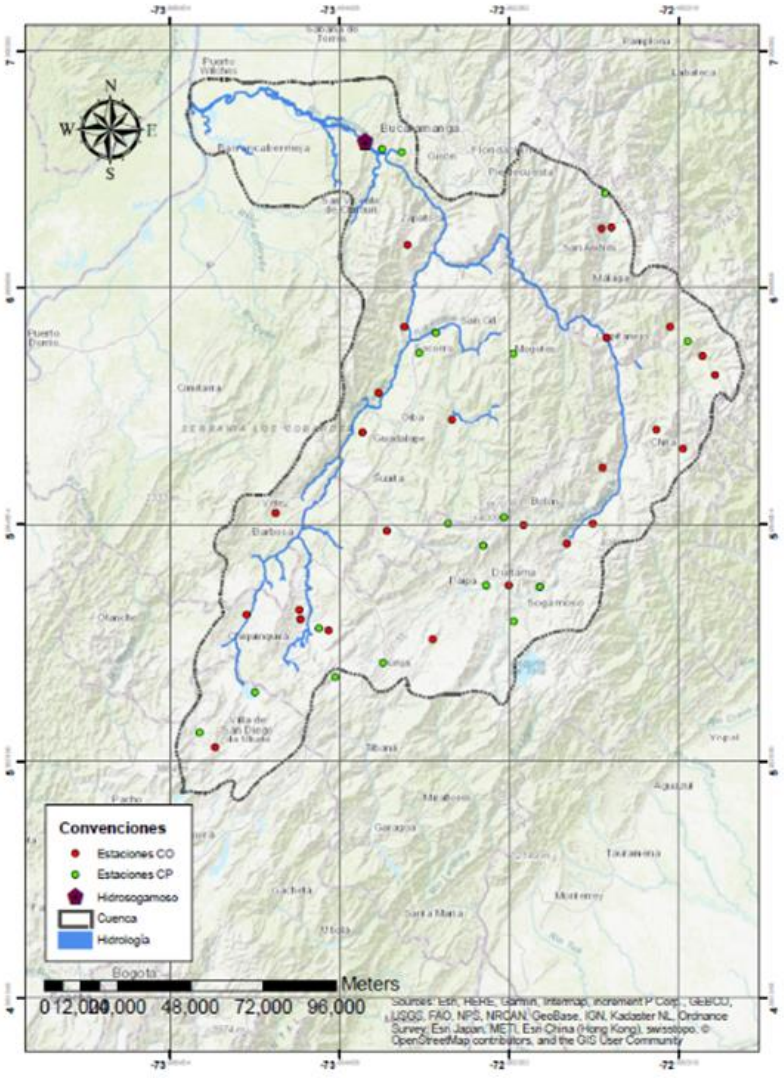
Primero se realizó una revisión bibliográfica para la selección de los modelos de pronóstico a desarrollar en la zona del rio Sogamoso fue fundamental, ya que por medio de esta se encontraron los modelos adecuados a la zona y sus características. En Colombia se presenta variabilidad climática a causa de la barrera orográfica que es constituida por tres ramificaciones las cuales son las cordilleras de los Andes, las que permiten la generación de diferentes microclimas en una región, ya que recibe la influencia del Caribe, el Pacifico y la circulación atmosférica de la cuenca del Amazonas [3]. Desde el campo de la ingeniería ambiental, el estudio del comportamiento de fenómenos naturales y los posibles pronósticos que se les pueden realizar es fundamental, ya que se pueden dar soluciones a problemáticas de interés no solo para Colombia sino para el mundo y su constante cambio diario. Utilizarlos en zonas vulnerables a constantes precipitaciones y cuyo fenómeno genera problemáticas ha hecho que sea posible verificar su funcionalidad con respecto a los datos que se tienen de las estaciones, es una solución viable ya que permite la identificación de zonas de riesgo donde se pueden presentar precipitaciones y a su vez es una herramienta que permite reconocer zonas de generación de energía. Los modelos de pronósticos, son modelos existentes e implementados en diferentes campos, sin embargo, fue imperativo adaptarlos a las necesidades, datos y variables que se utilizaron en este trabajo**.**

1.2. Descripción del área de estudio

- **Cuenca del rio Sogamoso**

Es un rio que nace en la confluencia del rio Suarez y el rio Chicamocha en el departamento de Santander, es uno de los principales afluentes que constituye el rio Magdalena. Es fuente hídrica principal de municipios como Los Santos, Zapatoca, Betulia, San juan de Girón, Chucurí y Barrancabermeja. Nace en el municipio de Puerto Wilches a 370 metros de altura, tiene una longitud de 135 kilómetros [4]**.**

Ilustración 1 Mapa cuenca rio Sogamoso



Fuente: autores

2. OBJETIVOS

2.1. Objetivo general

Determinar el modelo de pronóstico de precipitaciones más apropiado para la cuenca baja del rio Sogamoso ubicada en el departamento de Santander en Colombia tomando los datos climatológicos correspondientes.

2.2. Objetivos específicos

- Desarrollar una evaluación de los fundamentos matemáticos de las ecuaciones seleccionadas para la elaboración de pronósticos de precipitación.

- Construir los algoritmos teniendo en cuenta las estaciones seleccionadas para el desarrollo de los pronósticos de precipitación.

- Generar los pronósticos con respecto a los datos correspondientes a las estaciones seleccionadas.

- Identificar el modelo de pronóstico óptimo para la zona de estudio, implementando las métricas de desempeño seleccionadas.

3. MARCO TEORICO

3.1. Pronóstico de precipitación

Es el proceso de estimación que se realiza ante situaciones de incertidumbre, se refiere a la estimación de series temporales o datos obtenidos de forma instantánea. Se considera el pronóstico como un proceso crítico y continuo cuyos resultados son obtenidos mediante un proceso de planificación estadístico [5].

3.2. Linealidad

La linealidad examina qué tan exactas son las mediciones en todo el rango esperado de mediciones. La linealidad indica si el sistema de medición tiene la misma exactitud para todos los valores de referencia. La linealidad se sigue analizando mediante métodos de consistencia gráfica, sin embargo, el análisis numérico siempre es necesario cuando se requiere de una evaluación cuantitativa. En este escenario el cálculo de curvas de ajuste mediante el método de mínimos cuadrados se hace imprescindible [6]. Se entiende por linealidad la capacidad de un método analítico de obtener resultados proporcionales a la concentración de analito en la muestra dentro de un intervalo determinado [7].

3.3. Colinealidad

La colinealidad es la propiedad según la cual un conjunto de puntos está situado sobre la misma línea recta. Que una variable X_1 sea combinación lineal de otra X_2, significa que ambas están relacionadas por la expresión $X_1 = b_1 + b_2X_2$, siendo b_1 y b_2 constantes, por lo tanto, el coeficiente de correlación entre ambas variables será 1. Otro modo, por tanto, de definir la colinealidad es decir que existe colinealidad cuando alguno de los coeficientes de correlación simple o múltiple entre algunas de las variables independientes es 1, es decir, cuando algunas variables independientes están correlacionadas entre sí [8].

3.4. Regresión simple ajustada con mínimos cuadrados (RSL)

Es un modelo óptimo para circunstancias en donde se encuentran patrones de demanda con tendencia, ya sean crecientes o decrecientes, es decir, patrones que presentan una relación entre demanda y tiempo, para el tema de precipitación se

utiliza en la interacción que se tiene entre la cantidad de lluvia con el tiempo en que sucede la misma. El objetivo de la regresión lineal es determinar la relación que existe entre una variable dependiente y una independiente [1]. Cuando se hablan acerca de variables independientes, se deben utilizar relaciones lineales. El análisis se determina por medio de la intensidad entre variables a través de coeficientes de correlación y determinación [9].

Las ecuaciones que se utilizan en el modelo de regresión simple son las siguientes:

$$Y(x) = aX + b \quad (1)$$

En donde

$$Y(x) = \text{Variable a predecir.}$$
$$\text{b} = \text{Ordenada.}$$
$$\text{a} = \text{Pendiente}$$

El método más efectivo para determinar los parámetros a y b se conoce como mínimos cuadrados Para poder determinar la variable ***b*** se debe tener en cuenta la siguiente ecuación:

$$b = \frac{\Delta Y}{\Delta X} \quad (2)$$

Este es un método explicativo, desarrollar este tipo de modelo facilita una mejor comprensión de la situación y permite la experimentación con combinaciones diferentes de insumos para estudiar sus efectos en los pronósticos [10].

3.5. Regresión lineal múltiple

La metodología busca ajustar modelos lineales o linealizables entre una variable dependiente y más de una variable explicativa. En la regresión lineal múltiple vamos a utilizar más de una variable explicativa, lo cual ofrece la ventaja de utilizar más información en la construcción del modelo y realizar estimaciones más precisas. Se considera que los valores de la variable dependiente y han sido generados por una combinación lineal de los valores de una o más variables explicativas y un término aleatorio [11]:

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_i x_i + u \quad (3)$$

Los coeficientes son seleccionados de forma que la suma de cuadrados entre los valores observados y los pronosticados sea mínima, es decir que se va a minimizar la varianza residual [11].Este modelos permite extender a las “k” variables de la regresión lineal simple, en esta línea la variable y puede expresar mediante una función lineal de las variables ($x_1, x_2, x_3 \dots x_k$), para esto se dispone de un modelo donde se fijan los valores de las variables regresoras x_{ki} y obtiene al azar los correspondientes valores de y_i [8].

3.6. Modelos lineales generalizados (MLG)

En este modelo se tienen variables de respuesta asociadas a covariables, y permite incluir respuestas no normales como binomial, Poisson, o Gamma y en la teoría clásica la estimación se realiza mediante el método de máxima verosimilitud [13].

Se plantea que el MLG posee tres componentes:

Componente aleatoria: la variable de respuesta Y tiene una distribución:

Componente Sistemática: el vector de covariables x’= (x1, x2, xp) que da origen al predictor lineal

Función de enlace: relaciona las dos componentes μ y ŋ

$$g(\mu) = \eta \ (4)$$

Los modelos lineales generalizados permiten extensiones [13]:

- **Enlaces canónicos**

En el caso normal mostramos que si $Y \sim N(\mu, \theta^2)$ el parametro canonico es ɵ = μ.

En el caso binomial Y $\sim$ Bin (n,p) en el que se considera $\frac{Y}{n}$ vimos que el canónico es $\theta = logit(\pi)$. Estos son los link canónico o natural. Es conveniente usar el link $\eta = \theta$ el modelo tiene el link canónico o natural. Es conveniente usar el link natural, ya que algunas cosas se simplifican, pero la posibilidad de usarlo dependerá de los datos con los que estemos trabajando [13].

- Normal: ŋμ
- Poisson: ŋ = log μ
- Gamma: ŋ = μ^{-1}

3.7. Regresión lineal con análisis de componentes principales (PCA)

El análisis de componentes principales es una técnica de reducción de dimensión, que permite pasar de una gran cantidad de variables interrelacionadas a componentes principales. El método consiste en buscar combinaciones lineales de las variables originales que representan lo mejor posible a la variedad presente en los datos. A partir de las combinaciones lineales que son las componentes principales se comprende la información de una manera más sencilla. Al mismo tiempo la forma en la que se construyen las componentes y su relación con una y otras variables originales sirven para entender la estructura de correlación inherente a los datos y estas componentes son utilizadas para análisis estadísticos [14].

Descomposición de un vector aleatorio en sus componentes principales. Las componentes principales se obtienen a partir de la matriz de covarianza de un vector aleatorio a un conjunto de datos a el cual se le aplica una distribución de probabilidad discreta equiprobable pobre los vectores observados [15]. De modo que multiplicando por una constante podemos hacer la varianza tan grande o tan pequeña como se necesite, por esta razón es necesario normalizar, pidiendo que [14]

$$\left||v|\right| = \sqrt{v'_1 v} = 1 \ (5)$$

La primera componente principal de X adopta la forma

$$Z_1 = v'_1 \ (6)$$

La segunda componente principal de X adopta la forma

$$Z_2 = v'_2 X \ (7)$$

La componente principal es la variable aleatoria unidimensional $Z_1 = v_1 X$. Al proyectar el vector X sobre la recta que pasa por el origen con la dirección v_1, se obtiene el vector aleatorio [15].

$$Z_1^m = v_1 Z_1 = v_1 v'_1 X \ (8)$$

Que coincide con la representación de la primera componente Z_1 en $\mathbb{R}^p$ a lo largo de la direccion v_1. Con una matriz de covarianza:

$$Cov(Z_1^m, Z_1^m) = Cov(v_1 Z_1, v_1 Z_1) = Var(Z_1) v_1 v'_1 = \lambda v_1 v'_1 \ (9)$$

La cual es una matriz de rango y no es más que el primer sumando de la descomposición espectral de$\sum$.Centramos el vector Z_1^m en el vector de medias de X, $\mu = E(X)$. Este vector aleatorio Z_1^m es el resultado de proyectar el vector X sobre la recta que pasa por el vector de medias μ y tiene dirección v_1 [15].

3.8. Filtro de Kalman (FK)

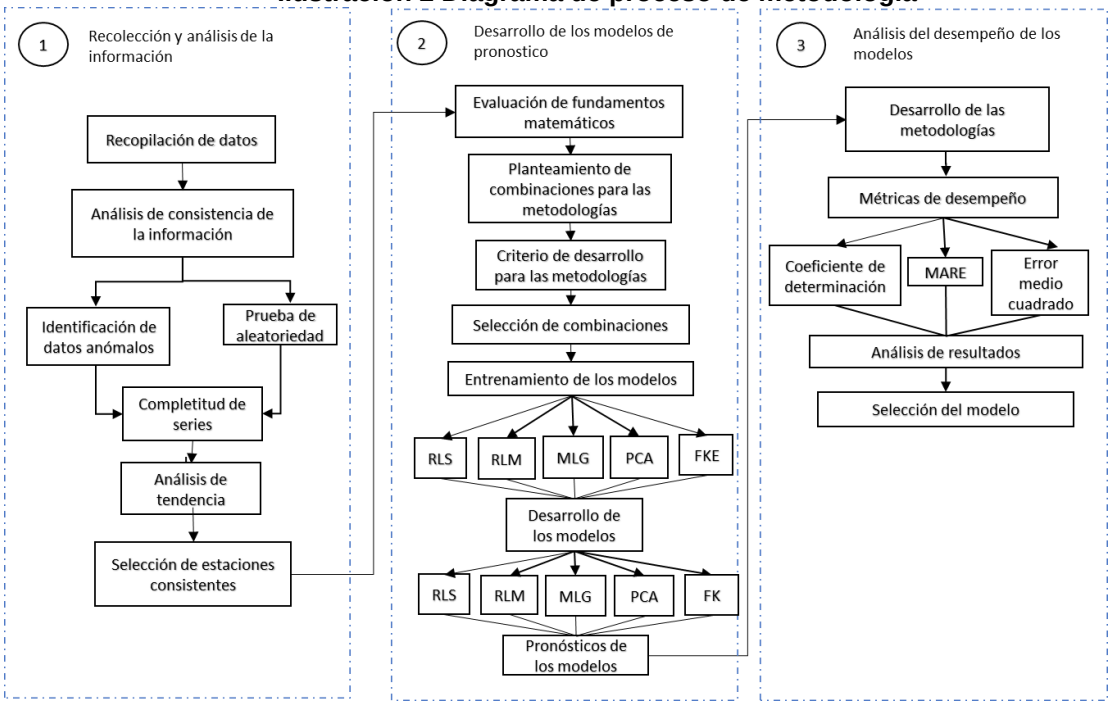
El filtro de Kalman es un método que permite estimar variables de estado no observables a partir de variables observables que pueden contener algún error de medición. Es un algoritmo que requiere dos tipos de ecuaciones: las que relacionan las variables de estado con las variables observables (ecuaciones principales) y las que determinan la estructura temporal de las variables de estado (ecuaciones de estado). Para utilizar la metodología del filtro de Kalman se deben tener en cuenta estimar las variables de estado utilizando su propia dinámica (etapa de predicción) y Mejorar esa primera estimación utilizando la información de las variables observables (etapa de corrección).[16]:

Una vez que el algoritmo pronostica el nuevo estado en el momento t, añade un término de corrección y el nuevo estado “corregido” sirve como condición inicial en la siguiente etapa, t+1. De esta forma, la estimación de las variables de estado utiliza toda la información disponible hasta ese momento y no sólo la información hasta la etapa anterior al momento en el cual se realiza la estimación [17].

4. METODOLOGÍA

En el desarrollo de este proyecto se llevaron a cabo tres fases, en la primera fase se realizó la respectiva recolección de los datos de las estaciones meteorológicas, la selección de las series a utilizar en la zona de estudio, filtro, tratamiento y análisis de consistencia de la información para dar inicio a la segunda fase, donde se seleccionó la información a utilizar en los modelos según el análisis de cumplimiento de los criterios obligatorios en las metodologías del proyecto, a partir del proceso anterior se desarrollaron los modelos con las series resultantes, en sus diferentes escenarios según la metodologías propuestas y en la tercera fase se realizó el análisis del rendimiento de los modelos utilizando las métricas de desempeño y los respectivos análisis de resultados obtenidos de cada uno de los modelos.

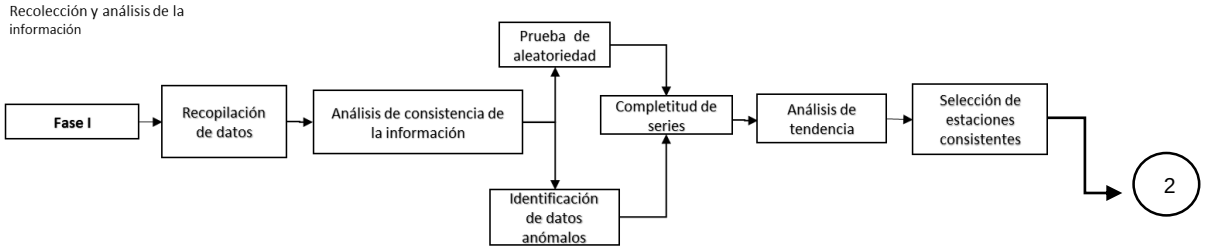
Ilustración 2 Diagrama de proceso de metodología



Fuente: autores

4.1. Fase I: recolección y análisis de la información

Ilustración 3. Diagrama de proceso fase 1



Fuente: autores

En esta fase se realizó el proceso para la obtención de la información necesaria para alimentar los modelos, haciendo una solicitud de catálogos a la entidad encargada, estas series corresponden a las variables meteorológicas de cada una de las estaciones ubicadas en la cuenca del rio Sogamoso.

Una vez se obtuvo la información, se realizó la selección de las estaciones teniendo como principal objetivo la cercanía de las mismas a la cuenca, al drenaje principal y su respectiva cercanía. Para dar explicación detallada de las estaciones seleccionadas, se utilizó la herramienta QGis el cual sirvió para georreferenciar la zona de estudio y se elaboró un mapa que permite la visualización general de su ubicación. Una vez se seleccionaron, se realizó el análisis de consistencia de la información donde se identificaron los datos anómalos, adicional a esto se realizó una prueba aleatoria para el completado de datos de las variables seleccionadas previamente, con el fin de desarrollar los modelos con las series que cumpla los criterios.

- **Recopilación de datos**

A partir del área de estudio seleccionada se identificaron las estaciones meteorológicas que recolectan los datos de las variables necesarias para el desarrollo de los modelos, esta recopilación se centró en los reportes anuales del IDEAM (Instituto de hidrología, Meteorología y estudios ambientales) los cuales constan de datos mensuales del año 1970 al 2016 de las variables de precipitación, brillo solar y temperatura media. Por medio del sistema de información geográfico se realizó la selección de los datos meteorológicos y estaciones seleccionadas aledañas a la hidroeléctrica hidro Sogamoso las cuales permitieron utilizar datos representativos de esta cuenca.

- **Análisis de la consistencia de la información**

Inicialmente se describió la cantidad de estaciones que presentaron datos de las variables seleccionadas, a partir de la Guía de Prácticas Hidrológicas de la OMM (Organización Mundial de Meteorología) y los estándares de calidad aplicables para Colombia, se seleccionaron las series de las variables en un periodo de 30 años. El análisis de consistencia de la información permite el primer filtro para la selección de las estaciones que cumplen con los criterios planteados. Se realizaron logaritmos los cuales generan una serie de pasos para la identificación y desarrollo de los datos e interpretación para utilizarlos en un pronóstico.

4.1..1. Identificación de datos anómalos

Los datos anómalos son observaciones que se encuentran alejadas de las demás observaciones, la prueba de evaluación de datos anómalos (Prueba Grubbs) asume que estos siguen una distribución normal, a partir de su identificación se realiza el retiro de estos en las series de tiempo. A partir de la identificación de los datos anómalos se procedió a realizar el análisis de los datos los cuales deben cumplir con dos criterios, el periodo de registro de completitud de los datos en un periodo de 30 años y presentar un porcentaje de completitud del 70%. Donde se considera la longitud esperada y observada de la serie [17]. La detección de datos atípicos es fundamental ya que esos son observaciones que no parecen corresponder con las demás observaciones del conjunto de datos de estudio y pueden alterar el resultado de los pronósticos de precipitación y su desarrollo.

4.1..2. Prueba de aleatoriedad

La prueba de aleatoriedad realizada fue la de rachas, las cual consiste en la identificación de parámetros y características que describen el comportamiento de los conjuntos estadísticos de manera aleatoria. El número de rachas en una muestra que genera un indicio si hay o no hay aleatoriedad, ya que supone que las series están en secuencia de datos [17]. Para aprobar el análisis de consistencia debe cumplir con los criterios mencionados anteriormente de manera simultánea, a partir del paso anterior con las series que presentan cumplimiento se procede a realizar completado de estas.

- **Completitud de series**

Para realizar la complementación de los datos, se utilizó el test de Rachas el cual consistió en verificar si las sucesivas observaciones son independientes y de esta forma se identifica el número de rachas que se presentan en las series seleccionadas, las rachas identificadas representan los valores consecutivos iguales o superiores a la media muestral. Este proceso se desarrolló con el fin de identificar las series que no cumplen con el 70 % de aleatoriedad de las variables seleccionadas para los diferentes modelos las cuales son precipitación, brillo solar y temperatura.

- **Análisis de tendencia**

El análisis de tendencia se desarrolla para aislar el componente que determina el comportamiento a largo plazo de las series en este caso se utiliza el ajuste por mínimos cuadrados ordinarios, al identificar que la tendencia de las series es lineal, esto se logra a partir del cálculo de las medias anuales y la tendencia anual, a partir de la tendencia anual se obtiene la tendencia para los distintos subperiodos.

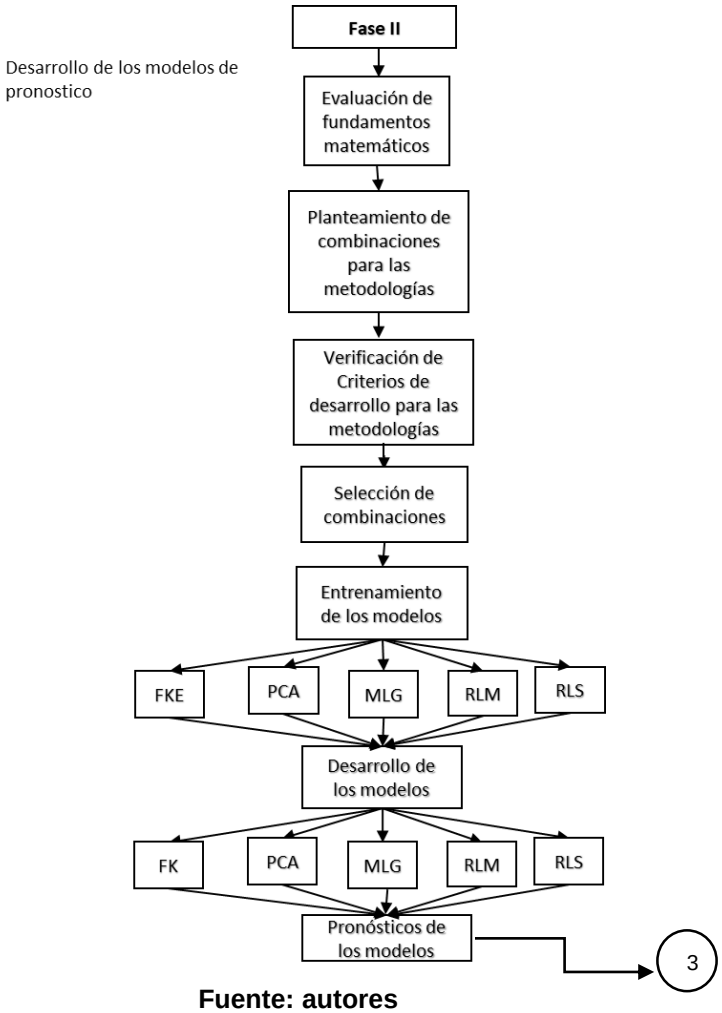
- **Selección de estaciones consistentes**

Las estaciones seleccionadas fueron las que cumplieron con los criterios anteriormente planteados y posterior a esto se les realizo un completado de datos para poder ser utilizados en los modelos planteados.

4.2. Fase II: desarrollo de los modelos de pronostico

En esta fase se tuvo en cuenta para alimentar los modelos, las series que cumplieron con la etapa de tratamiento de datos para plantear las combinaciones correspondientes a cada uno de los cinco modelos a ejecutar, a estas combinaciones se les realizó una evaluación de fundamentos matemáticos y verificación de criterios para comprobar el cumplimiento de los supuestos obligatorios de las series para cada uno de los modelos propuestos y partir de del paso anterior seleccionar las combinaciones aptas para entrenamiento y posterior pronóstico de cada uno de los modelos.

Ilustración 4. Diagrama proceso fase 2



Fuente: autores

4.3. Evaluación de fundamentos matemáticos

Los análisis de los modelos se llevaron a cabo teniendo en cuenta los tipos de datos seleccionados, es por esto, que se realizó un cuadro en donde se observan las ventajas y desventajas que tiene cada uno de los modelos, las variables, parámetros y características. Esto se realiza para la identificación de la aptitud del modelo y su desarrollo.

4.4. Planteamiento de combinaciones para las metodologías

Se plantean las combinaciones para cada uno de los modelos a partir de los criterios obligatorios en cada uno de estos (linealidad, normalidad, colinealidad, homocedasticidad, independencia de los residuos) las series seleccionadas se agruparon a partir de los criterios necesarios para analizar la influencia de las distancias entre estaciones en los pronósticos.

4.5. Verificación de criterios de desarrollo para las metodologías

Se realizó previamente una revisión bibliográfica de los criterios obligatorios de cada uno de los métodos a desarrollar a continuación se describen los requerimientos para cada uno de los modelos planteados y seguidos de esto la metodología de identificación del cumplimiento de estos.

Tabla 1. Criterios obligatorios de cada modelo

	Criterios				
Modelo	**Linealidad**	**Normalidad**	**Independencia de residuos**	**Homocedasticidad**	**Colinealidad**
RLS	X	X	x	x	
RLM	X	X	x	x	x
GLM		X		x	x
PCA	X	X			x
FK	X	X			

Fuente: autores

Linealidad

Para verificar el supuesto de linealidad se analizó la tabla de resumen del modelo donde se muestra el valor de coeficiente de correlación para identificar la relación lineal dependiendo de su cercanía a 1, así garantizar una relación fuerte entre los datos. Para poder identificar el cumplimiento de los criterios de la regresión se realizó el análisis iniciando con el desarrollo del ANOVA, el cual analiza la varianza donde

se prueba la hipótesis que la variable independiente es igual o diferente de 0 como se muestra a continuación:

$$H_0 = B_1 = 0 \quad El\ modelo\ no\ es\ significativo$$
$$H_0 = B_1 \neq 0 \quad El\ modelo\ es\ significativo$$

A partir del resultado del p-valor existe suficiente evidencia para aceptar o rechazar las hipótesis anteriormente planteadas pues esta muestra si tiene significancia o no. Al identificar los valores de B_0 B_1 plantea nuevamente la hipótesis de para los coeficientes

$$H_0 = B_0 = 0 \qquad H_0 = B_1 = 0 \Rightarrow Los\ coeficientes\ no\ tienen\ significancia$$

$$H_1 = B_0 \neq 0 \qquad H_1 = B_1 = 0 \Rightarrow Los\ coeficientes\ tienen\ significancia$$

A partir de estas hipótesis, se identificó si los valores de los coeficientes que presentan son significativos en la regresión de cada una de las combinaciones planteadas. Si el p-valor es > 0,05 se plantea que el coeficiente no es significativo y si es < 0,05 no existe suficiente evidencia estadística para rechazar la hipótesis nula por consiguiente el coeficiente será significativo Se utilizo un nivel de confianza del 95%.

Normalidad

Se realizó el análisis visual del histograma del residuo estandarizado y del grafico Q-Q Plot identificando la distribución de los datos en estos gráficos. Para el caso del histograma este debe cumplir con la centralidad, dispersión y asimétrica y el grafico de Q-Q Plot los datos debes estar distribuidos lo más cercano a la línea de regresión. Adicional a esto se realizó prueba Kolmogórov-Smirnov ya que la serie es mayor a 50 datos y con el análisis de los dos pasos anteriores no queda clara la normalidad de los datos en algunos casos. Para estos se realizó la prueba para identificar el cumplimiento del criterio de normalidad, analizando el p-valor de los residuos y de esta forma se identificó si su distribución es normal, a partir de las siguientes hipótesis:

$$H_0 = Los\ residuos\ no\ se\ distribuyen\ como\ una\ distribucion\ normal$$
$$H_1 = Los\ residuos\ se\ distribuyen\ como\ una\ distribucion\ normal$$

El p-valor al ser > 0,05 afirma el H_1 y si por el contrario es < se 0,05 acepta la hipótesis H_0 obteniendo como resultado la hipótesis H_1.

Independencia de los residuos

Para identificar la independencia de los datos se utilizó la prueba de Durbin-Watson, ya que son utilizados datos que tienen secuencias de tiempo, esta prueba busco identificar la autocorrelación de los residuos. Si el DW es un valor cercano a 2 tendremos cumplimiento del H_0 y si el DW es muy pequeño la independencia se califica como inexistente. Si el p-valor es grande significa que siendo la hipótesis nula el valor observado del estadístico D era esperable, de esta forma no se rechaza dicha hipótesis. Así mismo si el p-valor fuese pequeño indicaría que siendo la hipótesis nula es muy difícil que se produjera el valor D.

$$Si\ p-valor\ es\ cercano\ a\ 2 \quad \rightarrow \quad Aceptar\ H_0 = Hay\ independencia$$
$$Si\ p-valor\ es\ lejano\ a\ 2 \quad \rightarrow \quad Rechazar\ H_0 = No\ hay\ independencia$$
$$0 \leq DW \leq 4$$

Homocedasticidad

Para comprobar el supuesto de Homocedasticidad, se analizó gráficamente los valores pronosticados versus los residuos estandarizados de cada una de las combinaciones en el gráfico de dispersión, los valores de las gráficas deben encontrarse distribuidos entre -2 a 2 a lo largo del grafico para afirmar el cumplimiento de este criterio. Posterior a esto se plantearon tres tipos de escenarios en cada una de las combinaciones que cumplieron con los criterios anteriormente analizados y a partir de esto se realizó el análisis del comportamiento de los escenarios, inicialmente se plantean las ecuaciones de la recta y de esta forma hallar los valores pronosticados de las diferentes estaciones.

No colinealidad

Al evaluar la existencia de la colinealidad en las series utilizadas es preciso tener en cuenta el grado máximo de relación permisible entre las variables dependientes, al identificar este criterio se tuvo en cuenta el estadístico F ya que este evaluó el ajuste general de la ecuación, los coeficientes de relación parcial (1 a -1) y los coeficientes de correlación parcial.

- **Aplicación de criterios**

Se realiza la aplicación de los criterios matemáticos obligatorios a las series de datos a utilizar en cada uno de los modelos verificando el cumplimiento de cada uno de estos, las series o combinaciones que no cumplen con los criterios serán descartadas para el desarrollo del entrenamiento y pronostico

4.6. Entrenamiento de los modelos

Para el desarrollo del entrenamiento de los modelos se utiliza el 70% de las series de cada una de las variables correspondientes ya que a partir de esto se entrena el modelo y se identifica el comportamiento y a partir de esto se desarrolla el pronóstico, este entrenamiento se realizó como se muestra a continuación:

- **Regresión lineal simple (RLS)**

Como primer paso para el desarrollo de este modelo, a cada una de las combinaciones que se realizaron se les plantearon dos tipos de escenarios, cada uno con una situación distinta para observar los posibles comportamientos que se tenía en los datos. Inicialmente se plantean las ecuaciones de la recta y de esta forma hallar los valores pronosticados de las diferentes estaciones.
A continuación, se muestra un cuadro con las situaciones en cada uno de los escenarios:

Tabla 2. Escenarios RLS

Escenario 1	Escenario 2
Se utilizó el 70% de los datos de las estaciones la combinación utilizada consta de dos estaciones, para hallar el pronóstico de precipitación se toma una estación como dependiente y otra como independiente y a partir de esto se obtiene la ecuación de la recta y se reemplaza los datos de la variable independiente en la recta, de esta forma se obtiene el Y simulado.	Se realizaron dos gráficas, en la primera los datos de la primera estación correspondían a X y la segunda a Y. Por el contrario, en la segunda grafica se toman los datos de la primera estación como Y y el de la segunda como X. de esta manera la ecuación de la recta fue encontrada para ambas situaciones y así determinar cuál estación era dependiente y cual la independiente.

Fuente: autores

- **Regresión lineal múltiple**

A partir de las combinaciones propuestas anteriormente se identificaron las combinaciones que cumplen con las especificaciones necesarias, se seleccionaron como datos de entrada el 70% de las series aprobadas anteriormente y esto se procedió a desarrollar el modelo. En este método se utilizaron varias variables explicativas, esto presenta una mayor ventaja con respecto al modelo anterior ya que se encontraron datos más precisos, los coeficientes que se utilizaron en el modelo se eligieron al realizar la verificación del cumplimiento de la suma de los cuadrados entre valores simulados y observados sea mínima, todo esto para garantizar una mínima varianza residual.

En este proceso se identifican los coeficientes b los cuales indican el incremento en de Y o sea la precipitación por el incremento unitario de la correspondiente variable explicativa. A partir de la generación de la ecuación del modelo se identifica el coeficiente de regresión múltiple y el coeficiente de regresión múltiple al cuadrado y seguido de esto se realiza la validación del modelo la cual consistió en analizar si la variable de la variable criterio atribuida a la regresión, en este caso al efecto del conjunto de variables predictoras es lo suficientemente grande con respecto a las variabilidad no explicada o residual, esta validación se efectúa con el índice F el cual es una prueba estadística que evaluó esta relación. También se identificó el error típico de la estimación.

- **Modelos lineales generalizados (MLG)**

A partir del análisis de cumplimiento de criterios específicos de este modelo se seleccionaron las combinaciones aptas para el correcto desarrollo del modelos, a partir de esta selección el modelo identifica las variables exógenas (brillo solar, evaporación, temperatura) y la variable endógena (precipitación), como datos de entrada se utilizó el 70% de las series de cada una de las variables, estos modelos están constituidos por tres componentes básicos: Componente aleatoria que identifico la variable de respuesta y su distribución de probabilidad, la componente sistemática que especifica las variables explicativas y la función link que es la función del valor esperado como combinación lineal de las variables predictoras.

A partir de la entrada de datos en forma matricial se realizó la validación del modelo, ya que los datos utilizados son continuos, se utilizan las familias Gauss, Poisson, Gamma y Gauss inversa pues estos permitieron incluir respuestas no normales. El modelo a partir de las series utilizadas selecciono cual es la familia y función de vinculo, estos datos son llevados a una escala a partir de una transformación

adecuada dependiendo del tipo de dato que se utilizó mediante la función de vinculo, esta función está relacionada con el predictor lineal, para devolver a la escala original de medida el valor ajustado es la función inversa de la transformación, el modelo procede a evaluar el predictor lineal para cada valor de la variable dependiente y luego comparar el valor predicho con la transformación de las series, de esta forma determina el ajuste del modelo ya que este utilizo la función de vinculo dependiendo de la familia, valorando la adecuación del modelo a los datos o la desvianza residual, identificando la que mejor se adapta para realizar el entrenamiento.

- **Regresión lineal por análisis componentes principales**

Se tomaron parejas de estaciones en donde en cada pareja estaba conformada por una estación de precipitación, esta estación fue relacionada con las otras variables (evapotranspiración, humedad relativa, brillo solar y temperatura media). Para el entrenamiento de los datos se utilizaron el 70% de los datos de cada una de las estaciones. Para el desarrollo de este modelo, se utilizó Python, el cual realiza la transformación lineal de los datos originales de las estaciones y por medio de una matriz de covarianza genera un nuevo sistema de coordenadas el cual tiene las varianzas que pasaron a convertirse en los componentes principales.

Para determinar la prioridad de cada uno de los componentes, se verifico cual varianza era la que tenía mayor relevancia, esta se tomó como el componente principal, luego se tomó el segundo dato más cercano, y así se encontraron cada uno de los componentes para el modelo. Para tener resultados del análisis de PCA concretos, es necesario retener los factores que tienen valor propio superior a 1. El componente principal es una variable que se obtiene del resultado de una combinación lineal de todas las variables que se consideraron al principio y cada una de estas variables da forma a la componente.

En PCA se utilizó regresión logística, la cual es un tipo de análisis de regresión que se utiliza para la predicción del resultado de una variable categórica (puede adoptar un sin número de categorías) en una función de las variables independientes. Este tipo de regresión es útil para modelar la probabilidad de un evento que ocurre como función de otro, esta regresión también es utilizada comúnmente en MLG [14].

- **Filtro de Kalman**

Para el desarrollo del entrenamiento de este modelo se utiliza el 70% de las series de precipitación seleccionadas a partir del cumplimiento de criterios para este método, este filtro requirió la obtención de una distribución del estado inicial para esto fue necesario hallar la media y la desvianza para dar inicio a su desarrollo, este modelo necesito identificar el tipo de distribución, lo cual obtuvo a partir de la derivación de la distribución inicial a partir de información externa del modelo considerando la estacionariedad del estado y la varianza se determina con la ecuación de transición resolviéndola iterativamente hasta alcanzar la convergencia.

Este modelo se desarrolló en dos fases, la primera fase se realizó la actualización de tiempo o también llamada predicción la cual inicia con la proyección del estado a partir de las matrices dadas por el modelo, seguido de esto se proyectó la covarianza del error donde está representada la proyección y la varianza del ruido del proceso. En la segunda fase del modelo se realizó la actualización de medidas o corrección del modelo, esta corrección se desarrolló en tres pasos: el primero calculo la ganancia de Kalman dado por los coeficientes de señal de entrada y la covarianza del error medio, el segundo paso fue la actualización del estado con mediciones a partir del estado a priori y el estado que se predice y finalmente se realiza la actualización de la covarianza del error, a partir de lo anterior se obtuvo el pronóstico de precipitación.

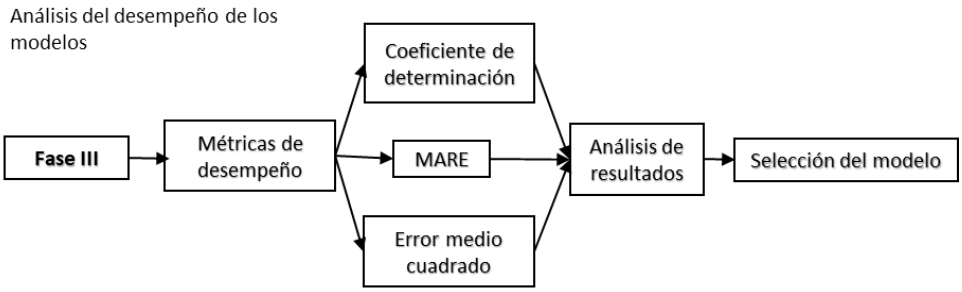
4.7. Desarrollo de los modelos

En del desarrollo de los modelos se utiliza el 30% de las series de cada una de las variables utilizadas y se realiza el mismo proceso planteado como metodología en el ítem de entrenamiento de los modelos.

4.8. Fase III: análisis de desempeño de los modelos

En esta fase se desarrolló la aplicación de métricas de desempeño las cuales identifican que modelo desarrollado presenta menor error para este proceso se encuentran gran variedad de métricas de desempeño las cuales facilitan la evaluación y comparación de las salidas y funcionamiento de modelos climatológicos. En este proyecto se seleccionaron a partir de revisión bibliográfica una métrica absoluta, relativa y dimensional para el respectivo análisis de los modelos.

Ilustración 5. Diagrama de proceso fase 3



Fuente: autores

- **Coeficiente de determinación**

Esta métrica adimensional muestra la proporción de la varianza estadística de los datos observados que se utilizan en el modelo. Maneja rangos entre cero y uno donde los valores cercanos a cero muestran modelos pobres y por el contrario los valores cercanos a uno indican modelos perfectos [18]. Esta métrica presenta alta sensibilidad a valores atípicos y se expresa en la siguiente formula:

$$r^2 = \left(\frac{\sum_{i=1}^{n} (\hat{x}_i - \hat{x})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (\hat{x}_i - \hat{x})^2}} \right) (10)$$

$$\hat{x} = Media\ valores\ pronosticos$$
$$x_i = Valor\ observado$$
$$\hat{x}_i = Valores\ pronosticados$$
$$\bar{x} = Media\ de\ valores\ observados$$

- **Error cuadrático medio**

También llamado ECM, esta métrica corresponde a las absolutas y mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo

que se estima. El ECM es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática. La diferencia se produce debido a la aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa [18].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \hat{Q}_i)^2}{n}} \quad (11)$$

- **Error medio absoluto relativo**

Esta métrica es relativa y comprende la media del error absoluto convertida en la relativa para el registro observado, no tiene un límite superior y para modelos perfectos el valor es cero [18].

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - \hat{x}_1|}{x_i} \quad (12)$$

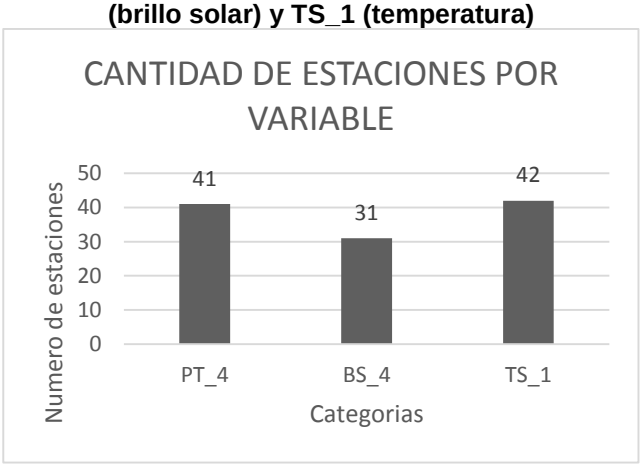
5. RESULTADOS

5.1. Fase I: recolección y análisis de la información

- **Recopilación de datos**

Para el análisis de la consistencia de la información se utilizaron los datos mensuales del año 1970 al año 2016. En la selección de las estaciones se tomaron las más cercanas a la hidroeléctrica de Sogamoso ubicada en el río Sogamoso en el departamento de Santander. Inicialmente se presenta la siguiente cantidad de estaciones por variables seleccionadas

Gráfico 1. Cantidad de estaciones por variables meteorológicas PT_4 (precipitación), BS_4 (brillo solar) y TS_1 (temperatura)



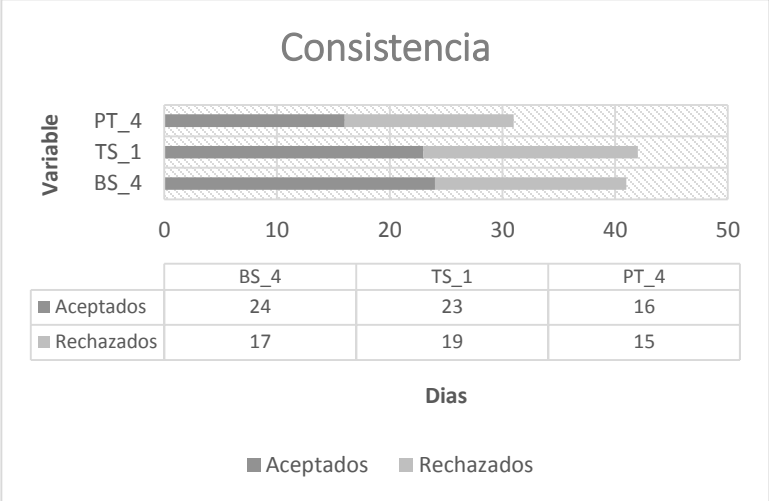
Fuente: autores

- **Análisis de la consistencia de la información**

5.1..1. Identificación de datos anómalos

Se realiza la identificación de datos anómalos seleccionando las estaciones que cumplen los criterios basados en la prueba de Grubbs como se muestra a continuación:

Gráfico 2. Cantidad de estaciones rechazadas/aceptadas en el análisis de completitud



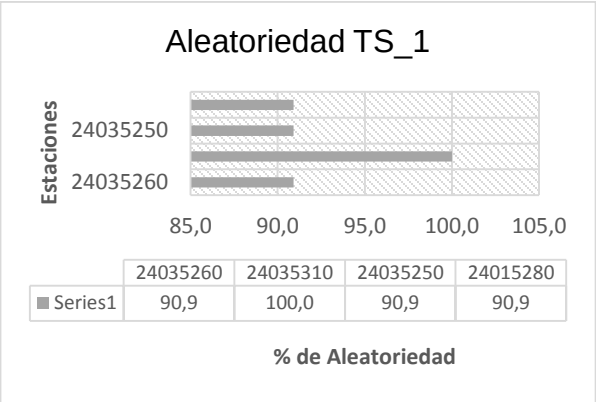
Fuente: autores

Para la variable de precipitación se aceptaron 24 estaciones, para Temperatura media 23 estaciones y para brillo solar 16 estaciones las cuales fueron utilizadas como base para el siguiente paso.

5.1..2. Prueba de aleatoriedad (prueba de rachas)

Se identificaron las series que cumplían con el 70% de aleatoriedad obteniendo como resultado que, en las variables de precipitación y brillo solar, todas las estaciones de entrada cumplen con el porcentaje de aleatoriedad mínimo como se observa a continuación:

Gráfico 3. Porcentaje de cumplimiento de aleatoriedad de las series de precipitación



Fuentes: autores

De las 16 estaciones de entrada solo cumplieron con el criterio de aleatoriedad mayor al 70% son las estaciones mencionadas en el anterior gráfico.

- **Completitud de las series**

A partir de datos de las series que se muestran en las siguientes tablas que han sido seleccionados a partir del cumplimiento de los criterios de análisis de datos se realizó la completitud de las series resultantes. Las estaciones de las variables precipitación, brillo solar y temperatura son utilizadas para completar las series, la cantidad de datos de entrada son los siguientes.

En la variable de precipitación se obtuvieron 11 estaciones

Tabla 3. Estaciones finales con periodos de tiempo de precipitación

	Estaciones	Inicio	Final	Datos	Años
1	24035240	1970-01-01	2016-12-01	470	39.17
2	24035260	1970-01-01	2016-12-01	469	39.08
3	24035310	1970-01-01	2016-12-01	459	38.25
4	24035250	1970-01-01	2016-12-01	478	39.83
5	24035320	1970-01-01	2016-12-01	465	38.75
6	24015260	1970-01-01	2016-12-01	472	39.33
7	24015280	1970-01-01	2016-12-01	447	37.25
8	24055030	1970-01-01	2016-12-01	468	39.00
9	24015250	1970-01-01	2016-12-01	481	40.08
10	24015270	1970-01-01	2016-12-01	459	38.25
11	24025050	1970-01-01	2016-12-01	459	38.25

Fuente: autores

En la variable de temperatura media se obtuvieron 11 estaciones

Tabla 4.Estaciones finales con periodos de tiempo de temperatura media

	Estaciones	Inicio	Final	Datos	Años
1	24035240	1976-01-01	2016-12-01	387	32.25
2	24035260	1976-01-01	2016-12-01	432	36.00
3	24035310	1976-01-01	2016-12-01	425	35.42
4	24035250	1976-01-01	2016-12-01	380	31.67
5	24035320	1976-01-01	2016-12-01	414	34.50
6	24015260	1976-01-01	2016-12-01	414	34.50
7	24015280	1976-01-01	2016-12-01	418	34.83
8	24055030	1976-01-01	2016-12-01	441	36.75
9	24015250	1976-01-01	2016-12-01	423	35.25
10	24015270	1976-01-01	2016-12-01	391	32.58
11	24025050	1976-01-01	2016-12-01	436	36.33

Fuente: autores

Y en la variable de brillo solar se obtuvieron 2 estaciones

Tabla 5.Estaciones finales con periodos de tiempo de brillo solar

	Estaciones	Inicio	Final	Datos	Años
1	24035260	1976-01-01	2013-12-01	379	31.58
2	24035310	1976-01-01	2013-12-01	393	32.75

Fuente: autores

Las series anteriormente completadas son utilizadas para analizar la tendencia como se muestra en el siguiente ítem.

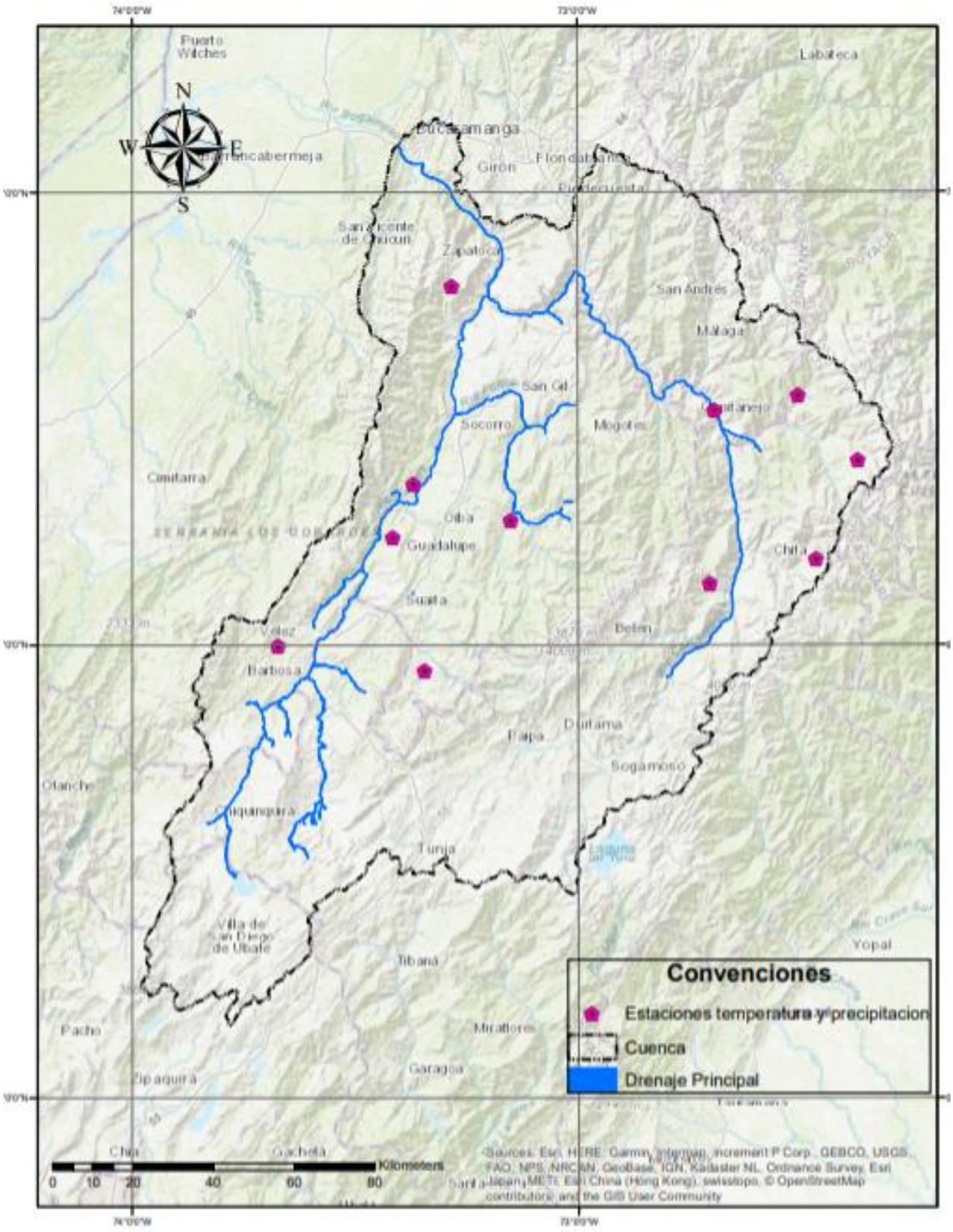
- **Análisis de tendencia**

Ya que las estaciones utilizadas disponen de una larga serie de observaciones permite el correcto desarrollo del análisis de tendencia a partir del método de mínimos cuadrados, este método representa la tendencia a partir de los valores de la variable dependiente Y en el tiempo, se calculan las medias anuales, la tendencia anual y a partir de esta la tendencia para los subperiodos.

- **Selección de estaciones consistentes**

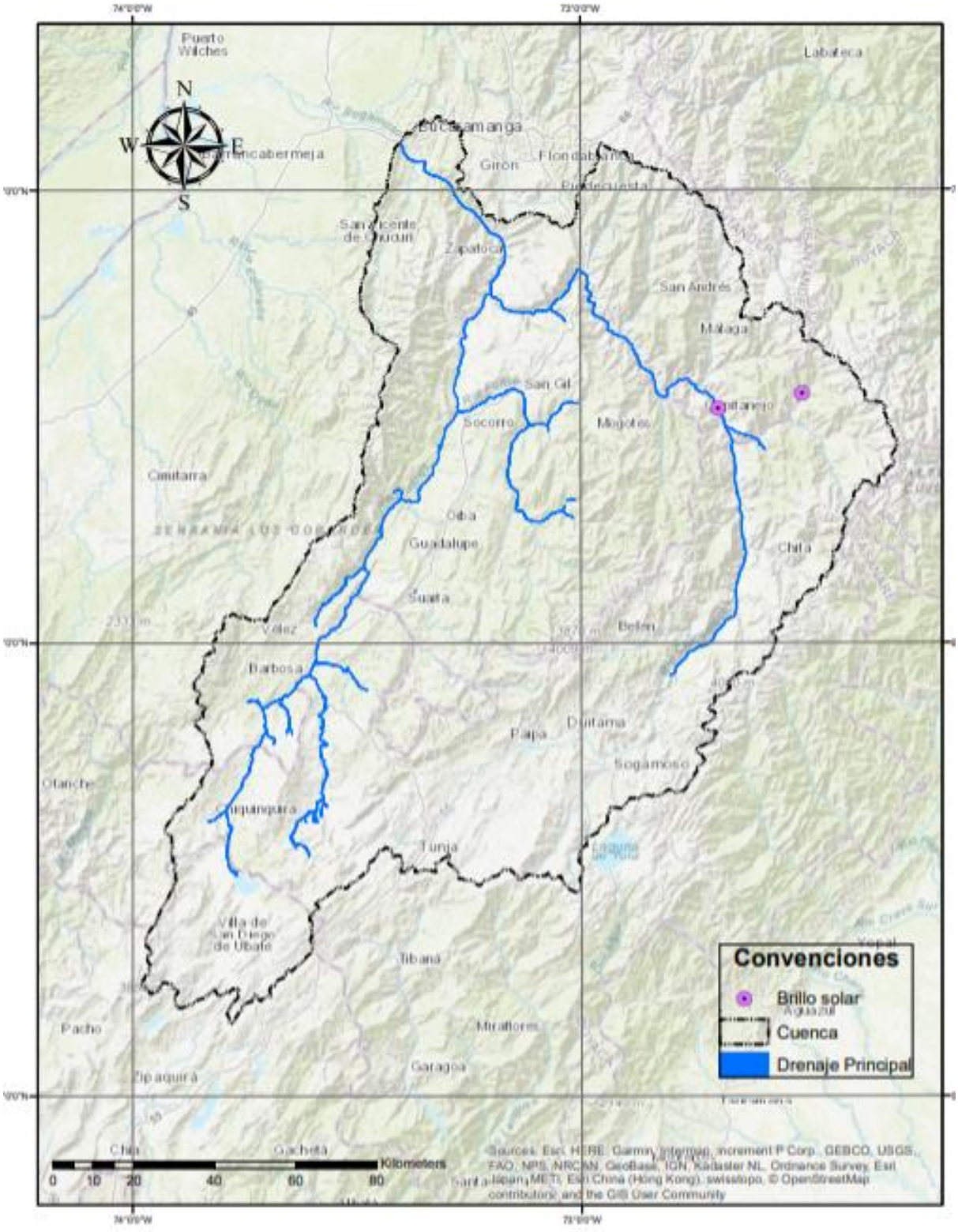
Las estaciones seleccionadas a partir de los criterios de análisis de consistencia de la información para las variables planteadas son las descritas en los siguientes mapas:

Ilustración 6. Estaciones de temperatura y precipitación



Fuente: autores

Ilustración 7. Mapa estaciones brillo solar



Fuente: autores.

5.2. Fase II: desarrollo de los modelos de pronostico

- **Evaluación de fundamentos matemáticos**

Se realizó un análisis de ventajas y desventajas de cada uno de los modelos con el fin de identificar si los criterios matemáticos exigidos por cada uno de estos se cumplen con el tipo de dato seleccionado como se observa en la siguiente tabla:

Tabla 6.Criterios matemáticos de las metodologías

Metodología	Variables	Ventajas	Desventajas
Regresión lineal simple ajustada con mínimos cuadrados	Y = Var. A predecir (Precipitación) b = Ordenada. a = Pendiente x = Var. Ind (Precipitación)	Identifica las posibles relaciones entre los cambios observados de dos variables.	Al utilizar pronósticos de regresión lineal, es como su nombre lo implica, que se supone que los datos pasados y las proyecciones a futuro caen sobre la recta
		Proporciona un medio visual para probar la fuerza de una posible relación y efectuar toma de decisiones.	La regresión analiza la relación entre la media de la variable dependiente y la independiente
		La regresión lineal es útil para el pronóstico de largo plazo de sucesos importantes.	La regresión lineal no es una descripción completa entre las variables y es sensible a valores atípicos.
Regresión lineal múltiple	Y = Var. A predecir (Precipitación) b = Ordenadas. x = Var. Ind (Precipitación, Brillo Solar, Temperatura)	La regresión lineal múltiple permite la identificación de relación entre la variable dependiente y más variables.	Asume que las variables utilizadas son independientes entre ellas
		La regresión múltiple disminuye la varianza residual. La estimación es más precisa cuanto menor es la varianza residual	Exige linealidad y normalidad en las series utilizadas.
			La complejidad que asume entre variables es muy sencilla
Regresión lineal con PCA		Los componentes utilizados se ordenan por la varianza que describen de esta forma disminuye la dimensional dad de los datos.	Este modelo asume linealidad en los datos observados.

		Este método convierte las series de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal o componentes principales.	El análisis de componentes principales utiliza los vectores propios de la matriz de covarianzas y encuentra únicamente las direcciones de ejes en el espacio de variables partiendo de los datos se distribuyen de forma gaussiana.
Modelos lineales generalizados	**Comp. Aleatoria =** Var. de respuesta (Brillo Solar, Temperatura, Precipitación). **Comp. Sistemática =** Var. Explicativa (Precipitación). **Función link =** Función del valor esperado	Esta metodología es una extensión de los modelos lineales que se aplican cuando los errores no son lineales.	Este tipo de modelos recibe como entrada de datos con distribución normal.
		MLG muestra una alternativa para las variables que no cumplen con el supuesto de modelos gaussianos.	Este modelo permite un porcentaje muy bajo de colinealidad entre las variables utilizadas.
		A partir de las diferentes funciones de vinculo y distribuciones se identifica como se adecuan los datos al modelo y el que mejor se adapte será el que minimice la desvianza residual.	
Regresión lineal con análisis de componentes principales	m = Numero de variables de entrada n = Numero de observaciones	Permite el análisis de un extenso conjunto de datos y reducir su dimensional dad con una perdida mínima de información al obtener secuencialmente la máxima variabilidad.	Dificultad de análisis de los datos resultantes
		Las componentes principales obtenidas son ortogonales entre sí, facilitando su posterior procesado y pueden tratarse independientemente	Se obtiene una combinación lineal distinta para cada conjunto de datos
		No es necesario conocer con anterioridad la correlación entre las variables para identificar si son dependientes o independientes	Solo se puede trabajar con el primer componente principal

Filtro de Kalman		La principal ventaja de esta metodología es disminuir los efectos de las señales aleatorias en las respuestas de un sistema.	El uso de derivadas parciales es una posible fuente de errores en el proceso del cálculo.
		El filtro de Kalman es capaz de escoger la forma óptima cuando se conocen las varianzas de los ruidos que afectan al sistema, además es el mejor estimador lineal si sus vectores tienen estadísticas arbitrarias y si éstas son variables aleatorias normales	Para el correcto desarrollo del filtro de Kalman es obligatorio el cumplimiento de linealidad en los datos de entrada.
		El filtro es aplicable para ambientes estacionarios y no estacionarios y permite actualizar el estado a partir del dato anterior.	El filtro de Kalman requiere el cumplimiento de normalidad de los datos de entrada y los datos de estado.

Fuente: autores

5.3. Planteamiento de combinaciones para las metodologías

A continuación, se observa las combinaciones planteadas para cada una de las metodologías.

- **Regresión Lineal Simple (RLS)**

Para la selección de las estaciones, las combinaciones se realizaron teniendo en cuenta la cercanía de las estaciones la una con la otra, todas las combinaciones tienen una estación cercana a la hidrografía y una no tan cercana, esto se realizó con el fin de verificar si este suceso es relevante con respecto a los resultados que se pueden obtener, también se tomó una situación particular (combinación 2), en donde se tienen ambas estaciones cercanas, para identificar si se encontraba una diferencia con respecto a las demás.

Tabla 7.Combinación regresión simple.

Combinación	Estación 1	Estación 2
1	24015280	24015270
2	24015260	24015250
3	24015250	24025050
4	24035250	24035320
5	24035260	24035310
6	24035240	24035250
7	24015260	24055030

Fuente: autores

- **Regresión Lineal Múltiple (RLM)**

En el modelo se seleccionaron de las variables a partir del criterio de cercanía donde se realizaron 5 combinaciones con las diferentes variables como se muestra a continuación:

Tabla 8.Combinaciones regresión múltiple

Combinación	Variables			
1	Precipitación	Precipitación	Temperatura	
	24015270	24015280	24015250	
2	Brillo solar	Brillo solar	Temperatura	Precipitación
	24035260	24035310	24035240	24035240
3	Precipitación	Precipitación	Precipitación	
	24055030	24015260	24035260	
4	Precipitación	Temperatura	Brillo Solar	
	24055030	24055030	24035260	
5	Precipitación	Temperatura	Temperatura	Brillo solar
	24015250	24015260	24025050	24035260
6	Precipitación	Brillo solar	Brillo Solar	
	24035260	24035260	24035310	

Fuente: autores

- **Modelos Lineales Generalizados (GLM)**

A diferencia que con los anteriores modelos que se utilizaron, en MLG no se utilizan diferentes estaciones para realizar las combinaciones y su respectivo analisis, en esta, se utiliza la misma estacion, pero se analizan dos variables mas que las encontradas en RLM, estas variables son evapotranspiracion y humedad relativa.

Para este modelo se realizan cinco combinaciones diferentes con cinco diferentes estaciones.

Tabla 9.Combinaciones modelos lineales generalizados

COMBINACION	VARIABLES				
1	**PT_4**	**HR_1**	**EV_4**	**TS_1**	**BS_4**
	24035240	24035240	24035240	24035240	24035260
2	**PT_4**	**HR_1**	**EV_4**	**TS_1**	**BS_4**
	24035260	24035260	24035260	24035260	24035260
3	**PT_4**	**HR_1**	**TS_1**	**BS_4**	
	24035310	24035310	24035310	24035310	
4	**PT_4**	**HR_1**	**EV_4**	**TS_1**	**BS_4**
	24035250	24035250	24035250	24035250	24035260
5	**PT_4**	**HR_1**	**EV_4**	**TS_1**	**BS_4**
	24055030	24055030	24055030	24055030	24035310

Fuente: autores

- **Filtro de Kalman (FK)**

A partir de criterio de cercanía se identificaron las estaciones de precipitación que cumplieran el porcentaje mínimo de datos para posteriormente desarrollar este modelo.

Tabla 10.Combinaciones filtro de Kalman

	Estaciones
1	24035240
2	24035260
3	24035310
4	24035250
5	24035320
6	24015260
7	24015280
8	24055030
9	24015250
10	24015270
11	24025050

Fuente: autores

- **Regresión Lineal con Análisis de Componentes Principales**

El criterio de selección de las estaciones fue por medio de las variables a analizar, se utilizaron combinaciones de estaciones de precipitación con el resto de variables para identificar comportamientos de cada una de las variables con respecto a la precipitación.

Tabla 11. Combinaciones PCA

COMBINACION	ESTACION DE PRECIPITACION	ESTACION	VARIABLE
1	24015270	24035250	BRILLO SOLAR
2	24025050	24035310	HUMEDAD RELATIVA
3	24015280	24055030	HUMEDAD RELATIVA
4	24055030	24035250	EVAPOTRANSPIRACION
5	24015260	24035320	TEMPERATURA

Fuente: autores

5.4. Verificación de criterios de desarrollo para las metodologías

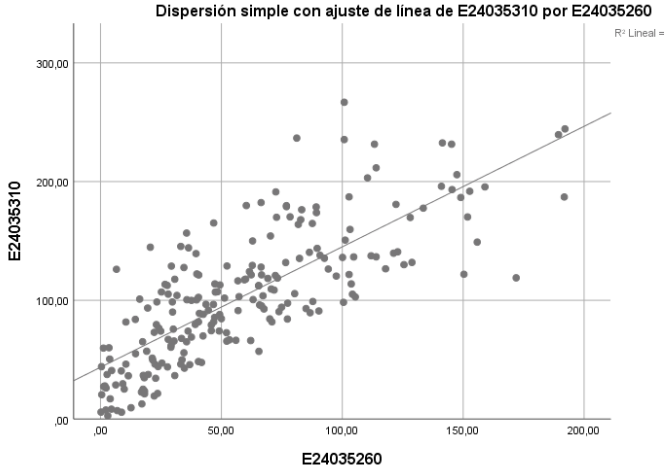
Para el desarrollo de los modelos se realizó un análisis del cumplimiento de los criterios de cada uno de los métodos propuestos a sus respectivas combinaciones descritas anteriormente. A continuación, se describe como se identificó el cumplimiento de cada uno de los criterios obligatorios en todas las metodologías.

- **Regresión lineal simple (RLS)**

En el análisis de cumplimiento se identificó que la única combinación que cumplía con estos criterios era la 5 de esta forma se procedió a desarrollar el análisis de los demás criterios obligatorios de este modelo como se muestra a continuación:

Combinación 5

Gráfico 4. Dispersión simple con ajuste entre estaciones combinación 5.



Fuente: autores.

Tabla 12. Resumen modelo combinación 5

Resumen del modelo					
Modelo	R	R cuadrado	R2 ajustado	Error estándar	D-Watson
5	0, 778	0,606	0,604	35,05626	2,104

Fuente: autores

Como se muestra en la tabla anterior r presenta un valor de 0,778 lo que indica que existe un alto grado de correlación entre las variables, el r cuadrado nos indica que el modelo explica más del 60% de la variabilidad de la variable dependiente.

Tabla 13. ANOVA combinación 5

ANOVA	
	sig.
Regresión	0,000

Fuente: autores.

Presenta un p-valor < 0,05 rechazando la hipótesis nula afirmando la significancia del modelo donde se confirma que las variables utilizadas en la combinación son aptas para realizar pronósticos.

A partir de la tabla 13 en la prueba Durbin-Watson se analizo la independencia entre los residuos, lo que quiere decir que 2,104 su resultado, acepta la hipotesis

H_0 rechazando la existencia de la autocorrelacion cumpliendo con el criterio de independencia de los residuos.

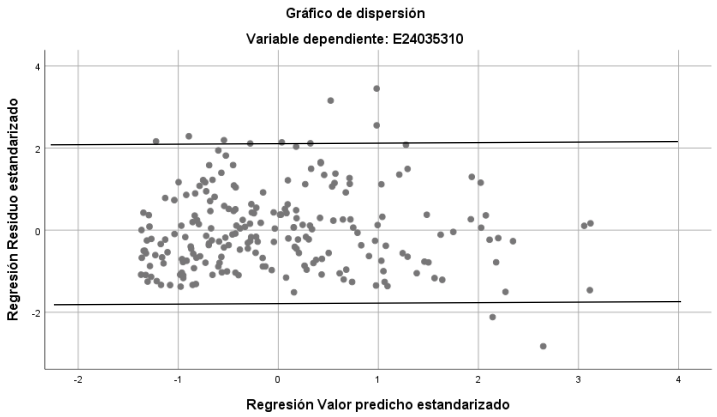
Tabla 14. Resultados de coeficientes combinación 5

Coeficientes								
Modelo		Coef. no estandarizados		Coef. Estandarizados	T	Sig.	95,0% intervalo de confianza B	
		B	Desv. Error	Beta			Lím. Inf.	Lím. Sup.
1	(Constante)	43,601	4,055		10,752	0,00	35,607	51,594
	E24035260	1,015	0,056	0,778	18,173	0,00	0,905	1,125

Fuente: autores

En los coeficientes se observa que el p valor es menos a 0,05 firmando como significativos a los coeficientes.

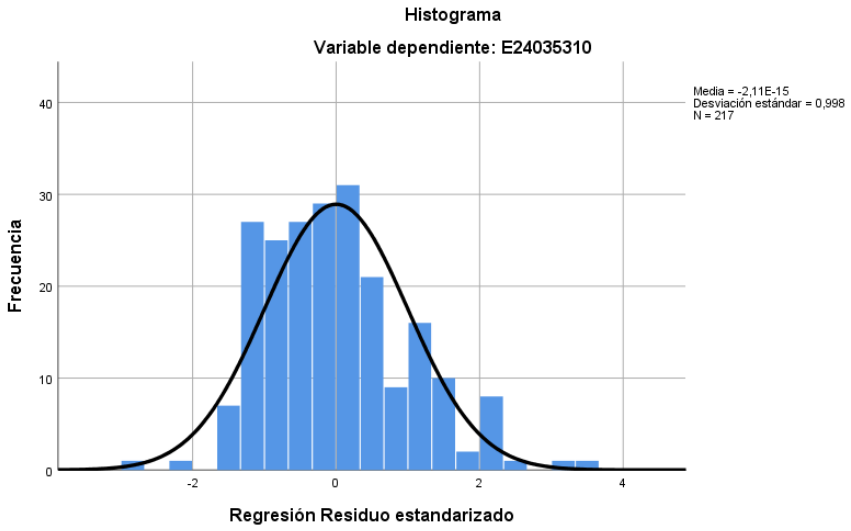
Gráfico 5. Residuos Vs. Predicción para el análisis de homocedasticidad combinación 5

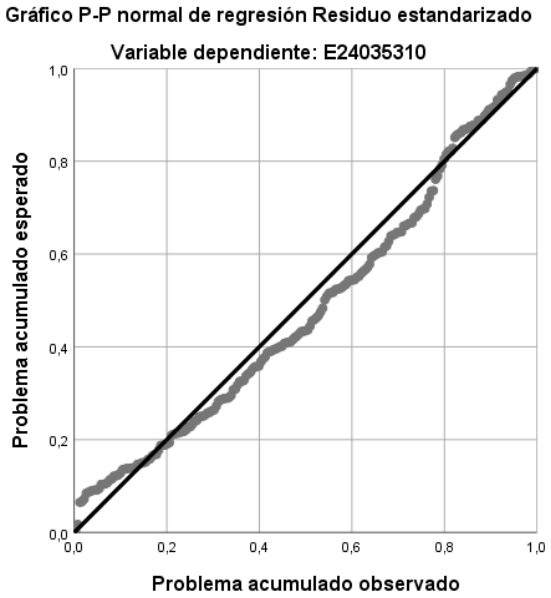


Fuente: autores

El grafico con los limites de homocedasticidad muestran incumplimiento del criterio, en la parte superior del grafico principalmente ya que sale de los limites entre 2 a -2.

Grafico 6. Identificación de normalidad combinacion 5.





Fuente: autores

En el histograma se presenta una media -2,11E-15 y desviación estándar 0,996, a partir del grafico P-P Normal no se evidencia de forma clara el no cumplimiento del criterio de normalidad en esta combinación y se procede a realizar el test de Kolmogorov al obtener un valor mayor a 0,05 se rechaza la hipótesis nula ya que los residuos no se distribuyen de forma normal.

Tabla 15. Test Kolmogorov combinación 5

	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
Standardized Residual	0,071	217	0,010

Fuente: autores

El p-valor es < 0,05 se puede comprobar que los datos no presentan una distribución normal ya que se afirma la hipótesis nula.

- **Regresión Lineal Múltiple (RLM)**

En la regresión múltiple se realizaron los análisis de cumplimiento de los criterios teniendo en cuenta las variables seleccionadas, estas fueron brillo solar y temperatura y la variable endógena precipitación. Para la selección de las combinaciones se le dio relevancia a la cercanía que tienen las estaciones entre sí, de esta forma se podían tomar datos de la misma estación o de distinta y observar si existía algún tipo de diferencia en los resultados. El planteamiento anterior se realizó de esta forma ya que se pretendía identificar si estas características influían en los resultados del pronóstico o si por el contrario esto no era relevante.

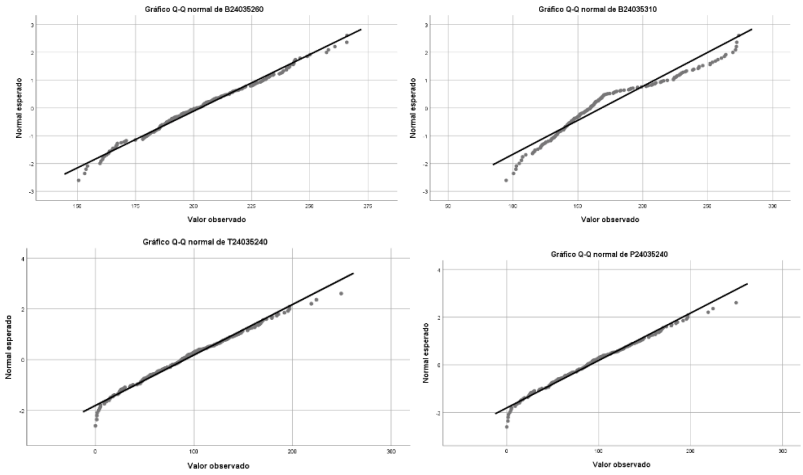
A partir del análisis anterior se puede afirmar que en las combinaciones planteadas existe mayor correlación entre la variable de precipitación con temperatura y se descarta la variable de brillo solar ya que en ninguno de los casos genera autocorrelación con la variable dependiente. A partir de esto se asume que la combinación que pasa este filtro es la 2 y se analiza si es necesario eliminan las variables de brillo solar.

Combinación 2

A partir de lo anterior se continua con el análisis de los criterios de la Regresión Lineal Múltiple para identificar el correcto cumplimiento de los supuestos básicos del modelo.

Gráfico 7. Quantil-Quantil

a) b)



c) d)

Fuente: autores

En la imagen anterior se observa los graficos de cuantil-cuantil los cuales permiten la comparacion de la distrubucion de las series utilizadas en esta combinacion con una distribucion ideal, a partir de esto se puede afirmar que los datos de (b) que representan la estacion B24035310 no cumplen con este criterio.

Tabla 16. Kolmogorov - Smirnov combinación 2

	Kolmogorov-Smirnov[a]		
	Estadístico	Gl	Sig.
B24035260	0,045	217	0,200
B24035310	0,165	217	0,000
T24035240	0,050	217	0,200
P24035240	0,050	217	0,200

Fuente: autores

Para la identificación de normalidad en las variables de una forma más acertada se analizó a partir de Kolmogorov-Smirnov, donde el p-valor es mayor de 0,05 en tres de las variables seleccionadas, lo que indica la única que no cumplió con este criterio es la variable de brillo solar B24035310.

Tabla 17. Correlación entre las variables de la combinación 2

Correlaciones					
		B24035260	**B24035310**	**T24035240**	**P24035240**
B24035260	Correlación de Pearson	1	0,762**	-0,393**	-0,393**
	Sig. (bilateral)	-	0,000	0,000	0,000
B24035310	Correlación de Pearson	0,762**	1	-0,542**	-0,542**
	Sig. (bilateral)	0,000	-	0,000	0,000
T24035240	Correlación de Pearson	-0,393**	-0,542**	1	1,000**
	Sig. (bilateral)	0,000	0,000	-	0,000
P24035240	Correlación de Pearson	-0,393**	-0,542**	1,000**	1
	Sig. (bilateral)	0,000	0,000	0,000	-

Fuente: autores

Las variables seleccionadas de brillo solar muestran resultados en la correlación de Pearson negativos por lo cual son descartados ya que no cumplen con el criterio, por el contrario, las variables de temperatura y precipitación estas variables presentan un p-valor < 0,05 por lo que se rechaza la hipótesis nula, entonces existe una asociación lineal entre estas dos variables.

Tabla 18. Resumen del modelo combinación 2

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
2	1,000[a]	1,000	1,000	0,00000

Fuente: autores.

Al realizar el análisis del modelo con las variables seleccionadas se obtuvo como resultado que el modelo explica el 100% y presenta un error estándar de estimación de 0 lo cual sucede ya que el modelo rechazo las variables de brillo solar, quedando únicamente las de temperatura y precipitación con una correlación de 1, por esta razón es necesario adicionar variables a esta combinación para poder realizar el entrenamiento.

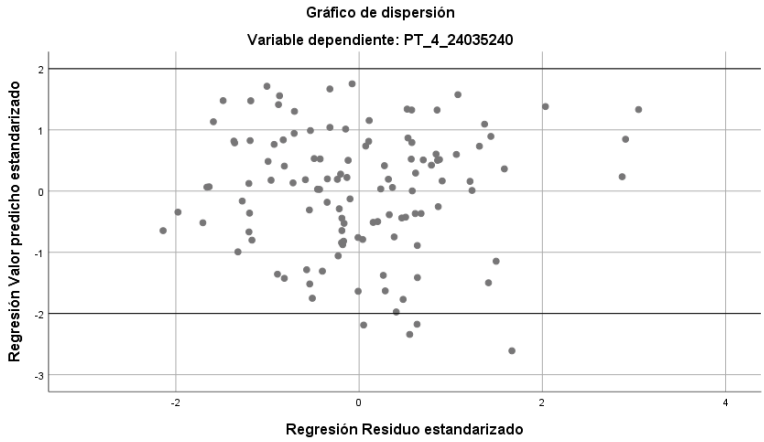
Tabla 19. Estadísticos de colinealidad de los coeficientes

Modelo		Coeficientes no estandarizados		Estadísticas de colinealidad	
		B	**Desv. Error**	**Tolerancia**	**VIF**
2	(Constante)	0,000	0,000		
	B24035260	0,000	0,000	0,419	2,386
	B24035310	0,000	0,000	0,350	2,855
	T24035240	1,000	0,000	0,706	1,417

Fuente: autores.

Las variables utilizadas no presentan colinealidad ya que sus resultados de tolerancia son < 0,10 y el VIF muestra el factor de inflación de la varianza el cual es reciproco a la tolerancia ya que cuanto mayor sea este valor mayor multicolinealidad como se observa en la variable de temperatura de presenta alta tolerancia y por lo tanto una pequeña VIF lo cual indica que no hay colinealidad.

Gráfico 8. Residuos Vs. Predicción análisis de homocedasticidad serie de precipitación



Fuente: autores

Como se observa en el grafico la dispersión se sale de los límites de 2 a -2 lo cual evidencia un incumplimiento mínimo del criterio de homocedasticidad. A partir del análisis anterior se concluye que no es viable realizar el modelo de regresión simple con las combinaciones planteadas ya que las variables no cumplen con los criterios mínimos para este proceso.

A partir del desarrollo del primer filtro del análisis de cumplimiento de criterios de este modelo es obtuvo que solo la combinación 2 cumplió con los supuestos, con las variables de precipitación y temperatura y las series de brillo solar se descartaron ya que no cumple con estos, en el segundo filtro se analizó la colinealidad, homocedasticidad, bondad de ajuste, en este análisis se identificó incumplimiento del criterio de homocedasticidad en la serie de precipitación, por esta razón se debe analizar si presenta relación lineal con otras variables diferentes a las consideradas anteriormente para realizar el entrenamiento.

A partir de lo anterior se analiza la correlación y colinealidad entre las variables de precipitación y temperatura anteriormente seleccionadas con las de humedad relativa y evaporación para identificar si se puede realizar el modelo con estas variables, para esto se utiliza las series de humedad relativa y evaporación de la

misma estación. Se realizo la correlación parcial de las variables precipitación y temperatura frente a la variable de control humedad relativa como se muestra a continuación:

Tabla 20. Correlación parcial humedad relativa

Correlaciones				
Variables de control			PT_4_24035240	TS_1_24035240
HR_1_24035240	PT_4_24035240	Correlación	1,00	-0,123
		Significación	.	0,189
	TS_1_24035240	Correlación	-0,123	1,000
		Significación	0,189	.

Fuente: autores

Y de las variables de temperatura y precipitación frente a la evaporación:

Tabla 21. Correlación parcial evaporación

Correlaciones				
Variables de control			PT_4_24035240	TS_1_24035240
EV_4_24035240	PT_4_24035240	Correlación	1,000	-0,107
		Significación	.	0,251
	TS_1_24035240	Correlación	-0,107	1,000
		Significación	0,251	.

Fuente: autores

El análisis de correlación parcial permite identificar el efecto que tiene la variable de evaporación y humedad relativa frente a las variables de precipitación y temperatura. A partir de lo anterior se puede evidenciar las variables de adicionadas no agregan valor al modelo por esta razón son descartadas y se procede a realizar el modelo con la combinación inicial.

- **Regresión lineal por análisis de componentes principales**

Para este análisis se utilizó el 70% de las series obteniendo como resultado las combinaciones que si cumplen con los criterios para el posterior entrenamiento y desarrollo del modelo. PCA es un modelo que realiza el análisis de colinealidad, correlación y a partir de este proceso determina que se debe descartar de cada uno de los datos de las estaciones que no cumplen con estos parámetros.

Las combinaciones se realizaron teniendo en cuenta todas las variables anteriormente mencionadas, la variable de precipitación se combinó con una

estación de cada variable y se realizó el respectivo análisis y de acuerdo a los resultados obtenidos por las gráficas se terminó cual combinación era la que mejores datos obtienen.

Tabla 22. Combinaciones PCA

COMBINACION	ESTACION DE PRECIPITACION	ESTACION	VARIABLE
1	24015270	24035250	BRILLO SOLAR
2	24025050	24035310	HUMEDAD RELATIVA
3	24015280	24055030	HUMEDAD RELATIVA
4	24055030	24035250	EVAPOTRANSPIRACION
5	24015260	24035320	TEMPERATURA

Fuente: autores.

- **Modelos lineales generalizados**

En el desarrollo del análisis del criterio de colinealidad se analizó principalmente el diagnostico de colinealidad, para el criterio de normalidad se analiza los gráficos de P-P Normal y el histograma de la variable dependiente y por último el criterio de homocedasticidad el grafico de valores predichos estandarizados y los residuos estandarizados donde los valores deben están dentro del rango de 2 a -2. A partir del desarrollo de este análisis a cada una de las combinaciones se obtuvo como resultado que solo la primera combinación presenta cumplimiento de todos los criterios en la mayoría de sus variables como se muestra a continuación:

Combinación 1

Tabla 23. Diagnóstico de colinealidad de la combinación 1

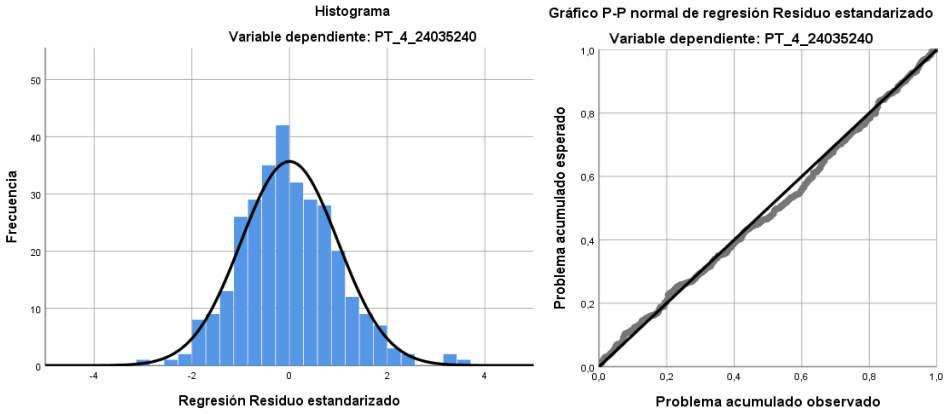
Diagnósticos de colinealidad								
Modelo	Dim	Autovalor	Índice de condición	Proporciones de varianza				
				Cte.	HR	EV	TS	BS
1	1	4,889	1,000	0,00	0,00	0,00	0,00	0,00
	2	0,096	7,143	0,01	0,02	0,02	0,01	0,01
	3	0,010	22,454	0,01	0,01	0,00	0,13	0,75
	4	0,004	36,881	0,00	0,97	0,97	0,01	0,01
	5	0,002	48,282	0,98	0,00	0,01	0,85	0,23

Fuente: autores.

En la columna de autovalores se puede observar que se presenta presencia de 4 autovalores cercanos a cero lo cual indica que las variables independientes están muy relacionadas entre sí, confirmando colinealidad entre las variables, los índices de condición se identifica incumplimiento ya que en la dim 3 presenta un valor entre 15 y 30 que afirma un posible problema de colinealidad y en la dim 4 y 5 se presenta

un valor mayor a 30 el cual es un límite de condición de no-colinealidad seria. La porción de la varianza indica el porcentaje de explicación del modelo por parte de una variable, en el caso de las dim 3, 4 y 5 se puede evidenciar que algunos coeficientes presentan valores muy altos lo que indica que explican la varianza en el modelo, lo que significa que no se cumple el criterio de no-colinealidad.

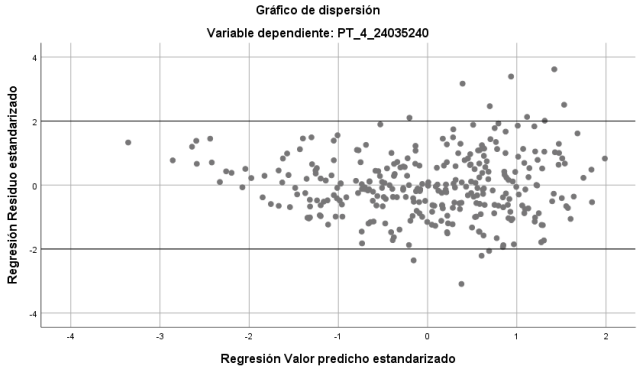
Grafico 9. Identificacion normalidad combinacion 1.



Fuente: autores.

En el histograma se presenta una media de 1,25E -16 y desviación estándar 0,994 visualmente se puede observar una distribución muy cercana a la normal, a partir del grafico P-P Normal se observa que la distribución de los datos es muy cercana a la línea de tendencia ideal lo cual indica un posible cumplimiento del criterio de normalidad por esta razón no se realiza el test de Kolmogorov-Smirnov.

Gráfico 10. Dispersión para análisis de homocedasticidad



Fuente: autores.

Este grafico representa el valor predicho estandarizado Vs el residuo estandarizado, se evidencia los límites de cumplimiento del criterio de homocedasticidad los cuales se incumplen en un porcentaje muy bajo por esta razón se afirma el cumplimiento

mínimo del criterio de homocedasticidad. A partir del análisis de colinealidad anterior se concluye que las variables de humedad relativa y evapotranspiración no se tendrán en cuenta en el desarrollo del entrenamiento y el pronóstico de este modelo.

- **Regresión lineal con análisis de componentes principales**

El análisis de componentes principales requiere el cumplimiento de supuesto obligatorios los cuales son verificados a partir del análisis del modelo por tal razón no resulta necesario verificar el cumplimiento previo.

- **Filtro de Kalman**

Para el desarrollo del modelo de filtro de Kalman solo se requiere el cumplimiento de dos criterios obligatorios los cuales son linealidad y normalidad, para la identificación del cumplimiento de este criterio dentro del modelo existe un análisis del cumplimiento de linealidad por tal razón se realiza el desarrollo del modelo con las series de las estaciones de precipitación seleccionadas.

5.5. Selección de combinaciones

De acuerdo a cada uno de los modelos, las combinaciones que cumplen con todos los criterios son:

- **Regresión lineal simple**

Las estaciones que conforman la combinación 5 son estaciones que se encuentran cercanas entre sí. A lo cual se le podría atribuir el resultado que se encontró.

- **Regresión lineal múltiple**

La combinación 2 estuvo compuesta de cuatro estaciones las cuales tenían como variable dependiente la de precipitación, las otras tres estaciones con variables independientes (temperatura y brillo solar), se observó que las estaciones se encuentran cercanas entre sí, una de las estaciones es la misma que se utilizó para la combinación 5 de RLS.

- **Modelos lineales generalizados**

La combinación 1 fue la que cumplió con los criterios. Como variable endógena se tomó una estación cuya variable era precipitación y como exógena se tomaron dos estaciones (temperatura y precipitación).

- **Análisis componentes principales**

La combinación seleccionada, es correspondiente a los datos de humedad relativa, esta variable se utilizó junto con la variable dependiente (precipitación) para desarrollar el modelo y así identificar si la relación de esta es útil para realizar pronósticos, a continuación.

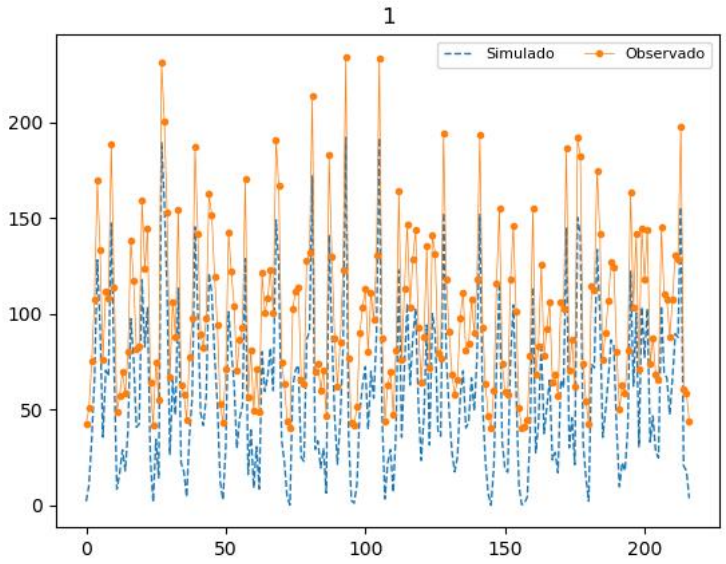
5.6. Entrenamiento de modelos

Para el desarrollo del entrenamiento se llevó a cabo la metodología planteada para cada uno de los modelos teniendo en cuenta las características.

- **Regresión Lineal Simple (RLS)**

En las siguientes graficas se observa la comparación entre los datos simulados y observados donde se evidencia el comportamiento de la precipitación

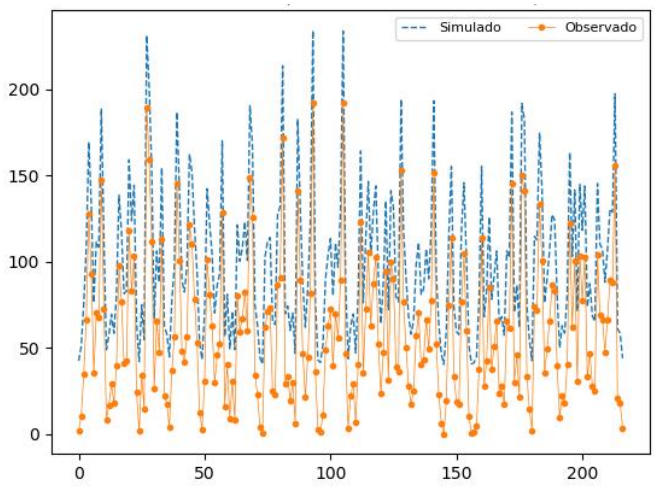
Grafico 11. Simulado Vs. Observado escenario 1



Fuente: autores.

En el primer escenario de este modelo se tuvo en cuenta los datos observados de una estacion para pronosticar los datos de la otra estacion, como se puede evidenciar en los gaficos este tipo de metodologia muestra una relacion mas cercana entre los datos simulados y los datos observados en la mayor parte de las series, aunque ignora los picos altos y bajos presentes en la serie observada de la estacion analizada.

Grafico 12. Simulado Vs. Observado escenario 2.



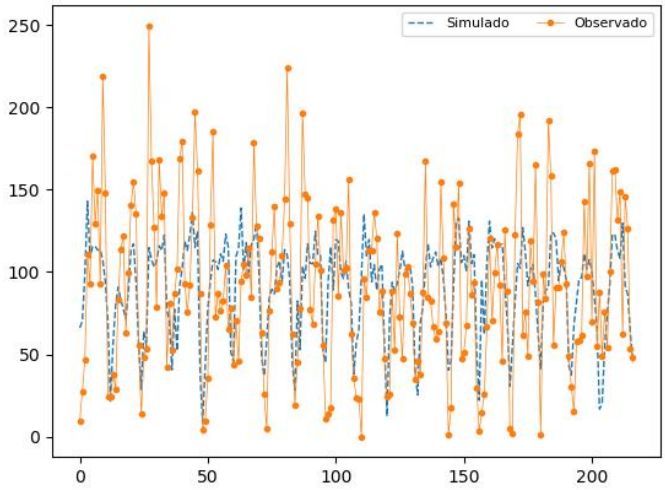
Fuente: autores.

Para el escenario dos se puede evidenciar que los valores simulados de los gaficos se escuentran muy cercanos a los picos alto en el primer grafico ignorando los picos bajos y en el segundo grafico precide valores muy cercanos a los picos bajos ignorando los picos altos.

- **Regresión Lineal Múltiple (RLM)**

Para el entrenamiento de este modelo se halló la ecuación del modelo a partir de las variables seleccionadas de la combinación que aprobó todos los criterios, utilizando el 70% de dichas series a partir de la metodología planteada en la fase II al hallar con la ecuación los datos simulados se procedió a realizar los siguientes gráficos de comparación de las series simuladas y observadas de la estación utilizada en el modelo como se muestra a continuación:

Grafico 13. Simulado Vs. Observado cmbinacion 1.



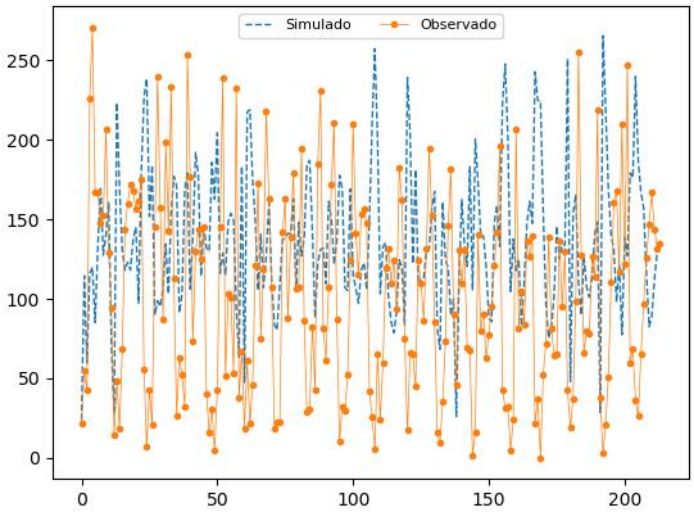
Fuente: autores.

- **Regresión Lineal con Análisis de Componentes Principales (PCA)**

De acuerdo con las cinco combinaciones que se realizaron, la combinación que más se adaptó al comportamiento de los datos observados fue la combinación 3, esta combinación fue entre una estación de precipitación y una estación de humedad relativa. Esta combinación es relevante ya que la humedad relativa es una variable que tiene relación directa con la precipitación, también se encuentra que la combinación de evapotranspiración (4) tiene similitud entre los datos simulados y los observados.
A continuación, se muestran los gráficos de la combinación cuyos resultados fueron más cercanos.

Combinación 4

Gráfico 14. Simulado Vs. Observado combinación 4



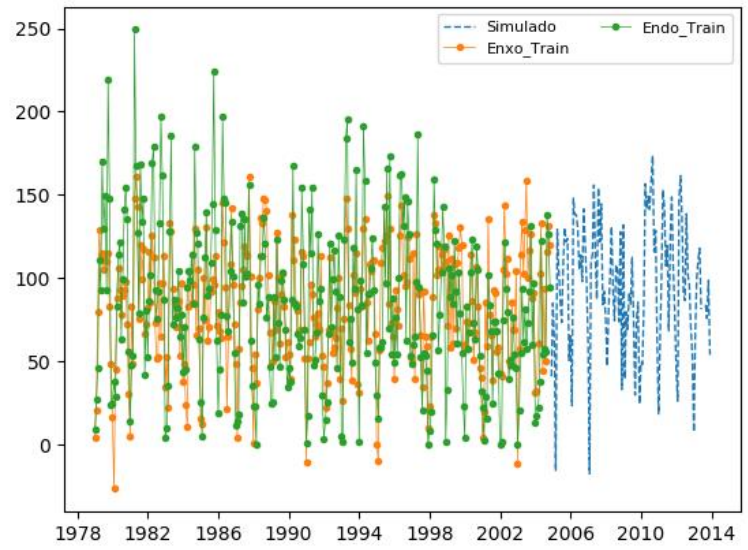
Fuente: autores.

- **Modelos Lineales Generalizados (MLG)**

En este método se tuvo en cuenta cuatro tipos de familia de distribución los cuales fueron evaluados con los datos de entrada que en este caso son los de la combinación 1, se utilizó el 70% de estos para el desarrollo del entrenamiento que se realizó a partir de la metodología planteada lo cual género como resultado la comparación del desarrollo de estas 4 aplicaciones y se procedió a seleccionar cual se adaptaba mejor a los datos del modelo a partir de este análisis se obtuvo.

A partir de esta distribución seleccionada, como la que mejor se adaptó a los datos se realiza el pronóstico de precipitación, con los datos obtenidos de este paso se procede a realizar los gráficos de comparación de las series simuladas y observadas para evidenciar la cercanía de los datos pronosticados a los datos reales de la estación que se está analizando, a partir de lo anterior se obtuvo el siguiente gráfico:

Gráfico 15. Simulado Vs. Observado 70% y pronostico combinación 2

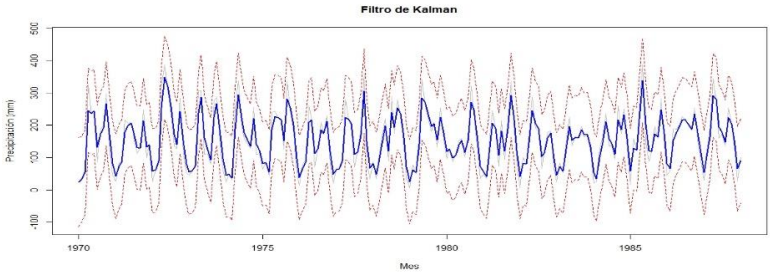


Fuente: autores.

- **Filtro de Kalman**

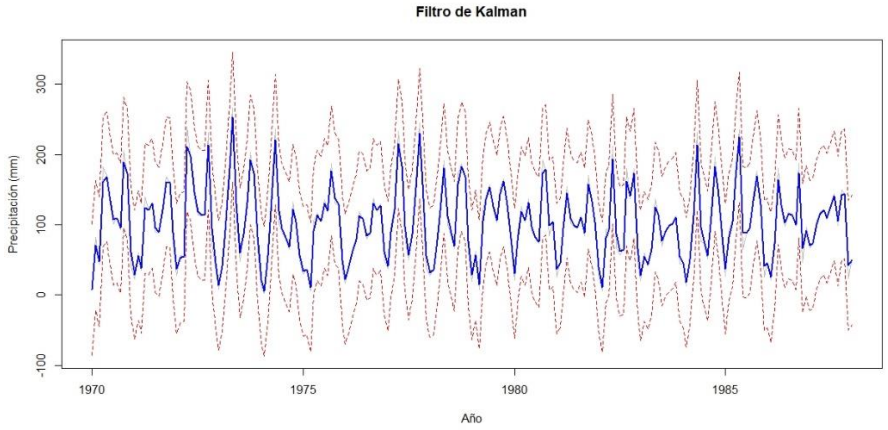
Las series de precipitación seleccionadas de la zona de estudio para este modelo son utilizadas para el entrenamiento utilizando el 70% de estas, esta fase del desarrollo del entrenamiento se realizó a partir de la metodología estipulada en la fase dos donde la línea azul representa el valor pronosticado, la gris el dato observado y las líneas rojas son el rango identificado en el modelo, para el desarrollo de este proceso se utilizaron 11 estaciones de las cuales solo cuatro de estas cumplen con el criterio de linealidad y aplican para este tipo de modelo como se observa a continuación:

Gráfico 16. Control de ruido filtro de Kalman (24015270).



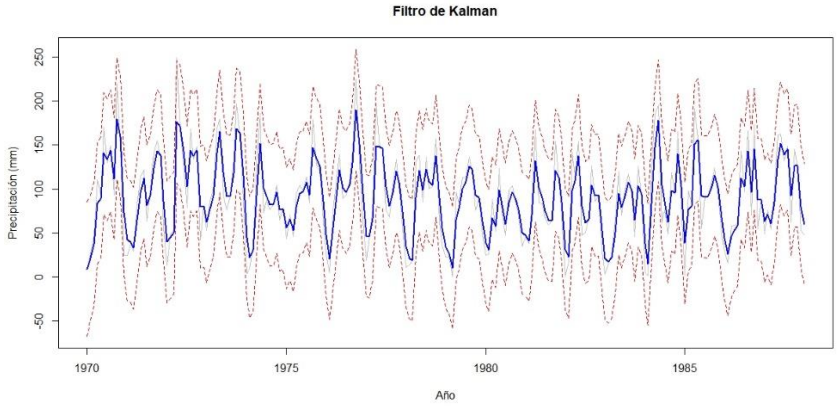
Fuente: autores

Gráfico 17. Control de ruido filtro de Kalman (24035250).



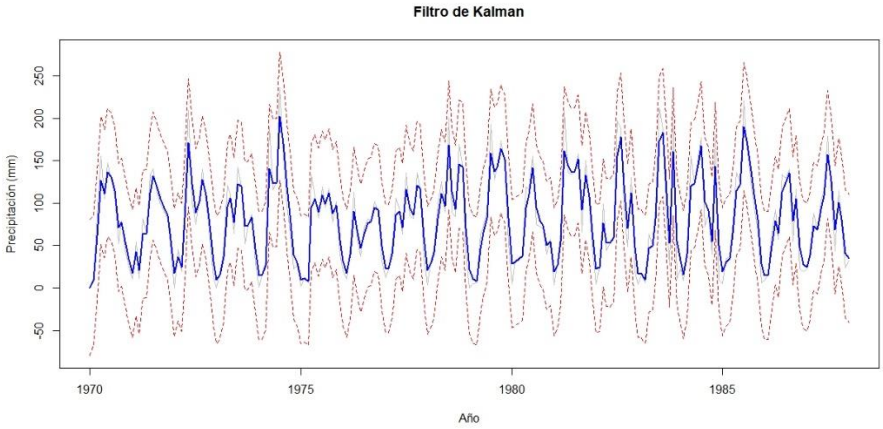
Fuente: autores.

Gráfico 18. Control de ruido filtro de Kalman (24035320).



Fuente: autores

Gráfico 19. Control de ruido de filtro de Kalman (24035250).



Fuentes: autores

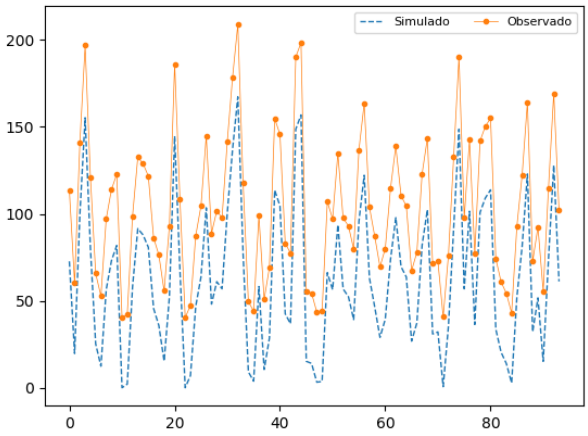
En los gráficos anteriores se puede observar como el filtro de Kalman realizo una filtración del ruido de los datos que son causados por posibles errores sistemáticos a partir de la óptima identificación de las varianzas de los ruidos que afectan el sistema. Seguido de esto se realizó una validación del modelo a partir del análisis de residuales ya que el pronóstico del siguiente periodo de los residuales y la varianza, se calculan errores estandarizados y se verifica el cumplimiento de distribución normal en la variable de estado y en la observada, en este análisis se obtuvo como resultado el incumplimiento de normalidad y linealidad por tal razón los pronósticos de precipitación obtenidos por este método no son fiables.

5.7. Desarrollo de los modelos

- **Regresión lineal simple (RLS)**

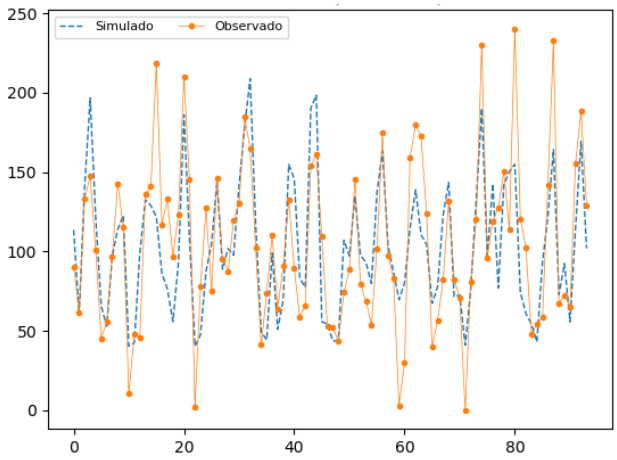
A continuación, se observa los gráficos generados a partir del desarrollo de la verificación del modelo donde se realiza la comparación de los datos simulados y los observados en los dos escenarios. Al igual que en el entrenamiento este escenario no muestra datos cercanos a los valores pronosticados ya que los datos simulados generados se mueven en la recta de forma constante indicando que esta metodología no es apta para este tipo pronóstico.

Grafico 20. Simulado Vs. Obervado escenario 1



Fuente: autores.

Grafico 21. Simulado Vs. Obervado escenario 1

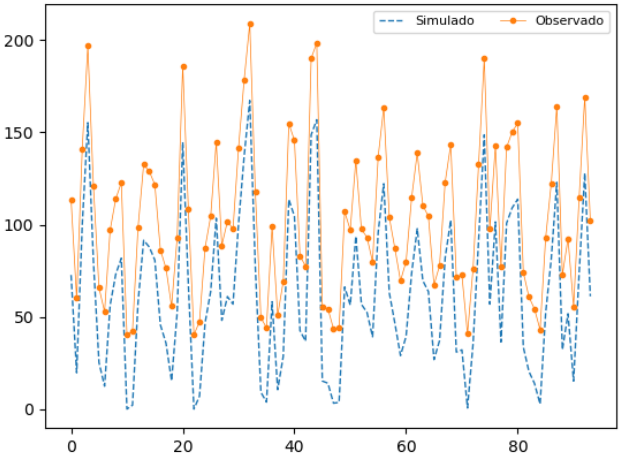


Fuente: autores.

En el segundo escenario de este modelo se tuvo en cuenta los datos observados de una estacion para pronosticar los datos de la otra estacion, como se puede evidenciar en los gaficos este tipo de metodologia muestra una relacion mas cercana entre los datos simulados y los datos observados en la mayor parte de las series, aunque ignora los pricos altos y bajos presentes en la serie observada de la estacion analizada.

- **Regresión Lineal Múltiple**

Gráfico 22. Simulado Vs. Observado combinación 2

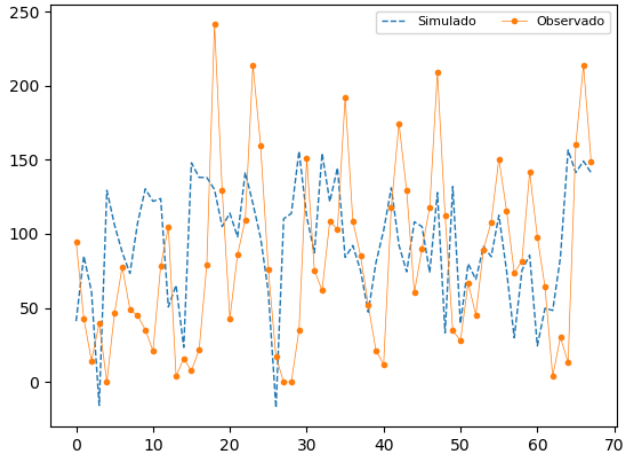


Fuente: autores

Al realizar el pronóstico de precipitación se observa que el valor pronosticado es muy cercano a los datos observados, este modelo al igual que el de regresión lineal simple no identifica correctamente los picos únicamente los valores medios.

- **Modelos lineales generalizados (GLM)**

Gráfico 23. Simulado Vs. Observado combinación 2

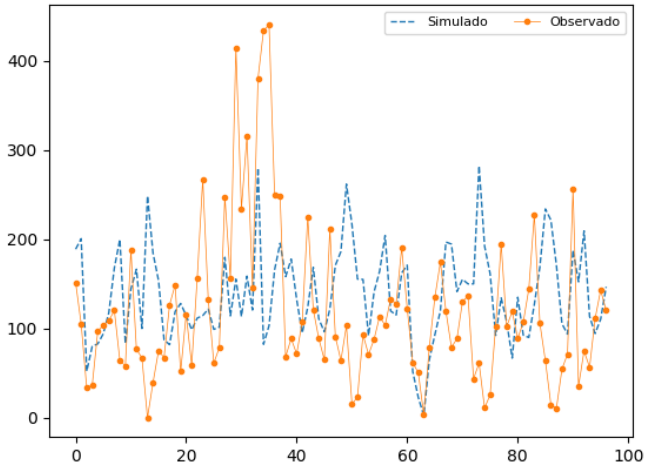


Fuente: autores

A partir de la componente sistemática utilizada se seleccionaron cuatro opciones de componentes aleatorios identificando cual se adapta a este tipo de series donde se obtuvo el modelo que se adapta la familia Gaussiana el cual utiliza una función de vinculo identidad, esta distribución presento menores errores y esto se presenta ya que este modelo utiliza variables continuas, por el contrario, la distribución Poisson requiere como entrada una variable discreta. El modelo de Gamma tiene como ventaja que descarta los valores de cero en las series y se adapta de una forma más adecuada a la variable de precipitación pues presenta variables positivas que tienes a valores mayores y adicional a esto la función de enlace identidad no realiza transformación de la variable dependiente.

- **Regresión Lineal con Análisis de Componentes Principales (PCA)**

Grafico 24. Simulado Vs. Observado combinacion 4



Fuente: autores

El entrenamiento de los datos de esta combinación se realizó con dos variables, precipitación y evaporación, se tuvo en cuenta para el entrenamiento el 70% de los datos y con el 30% se realizó el pronóstico. Según el grafico 32 se observó que los datos del pronóstico no tienen el comportamiento de los datos observados, sin embargo, en algunos picos se encuentran similitud entre los datos observados y simulados.

5.8. Fase III: análisis del desempeño de los modelos

Se desarrollaron las métricas de desempeño para evaluar los modelos desarrollados para la identificación del modelo que mejor se adapta, donde se utilizó una métrica absoluta, relativa y dimensional, a continuación, se puede observar los resultados para cada uno de los modelos.

- **Regresión lineal simple**

Tabla 24. Métricas de desempeño RLS

ESCENARIO 1	ESCENARIO 2	
Coef. De Determinación	Coef. De Determinación 1	Coef. De Determinación 2
0,61	1	1
MARE	MARE 1	MARE 2

0,179	6,895	5,699
ME	**ME 1**	**ME 2**
1,770	8,033	0,373

Fuente: autores

En la regresión simple se identifica que el escenario que presenta mejores resultados acercándose más a los datos observados, es el escenario 2, de igual forma las métricas de desempeño de este escenario muestra un modelo pobre ya que el valor del coeficiente de determinación son valores muy alejado a 1, teniendo en cuenta que 1 es el valor de un modelo perfecto y en la métrica de error medio absoluto relativo presenta valores muy alejados de cero por lo cual se puede concluir que el modelo no resulta adecuado para el tipo de datos.

- **Regresión lineal múltiple**

Tabla 25. Métricas de desempeño RLM

REGRESION LINEAL MULTIPLE (C2)		
Combinación	**Métricas de desempeño**	
2	Coef. De determinación	1
	MARE	0
	RMSE	0

Fuente: autores.

A partir del incumplimiento de los criterios obligatorios de la variable de brillo solar no es viable el resultado obtenido a partir de la regresión ya que al desarrollarse se desprecian las variables de temperatura y brillo solar pues no resultan representativas para el modelo por tal razón el resultado del pronóstico es el mismo valor de entrada y las métricas de desempeño no representan un análisis real del modelo pues las características de los datos no son aptas para su correcto desarrollo.

- **Modelos lineales generalizados**

Tabla 26. Métricas de desempeño MLG

MODELOS LINEALES GENERALIZADOS		
Modelo	**Métricas de desempeño**	
Gamma	Coef. De determinación	0,703
	MARE	Inf
	RMSE	388,6

Fuente: autores

Este modelo presenta las métricas de desempeño más cercanas a los valores de un modelo ideal pues su coeficiente de determinación presenta un valor medianamente cercano a 1 lo que significa que podría ser significativo pero el valor de error medio absoluto presenta valores muy alejados del ideal.

- **Regresión lineal con análisis de componentes principales**

Tabla 27. Métricas de desempeño PCA

ANALISIS DE COMPONENTES PRINCIPALES		
Modelo	**Métricas de desempeño**	
Combinación 3	Coef. De determinación	0,0072
	MARE	1.13
	RMSE	0.01715
Modelo	**Métricas de desempeño**	
Combinación 4	Coef. De determinación	0.01163
	MARE	1.18
	RMSE	6.9864

Fuente: autores.

Según los datos de las combinaciones seleccionadas por similitud de datos, para la combinación 3 se encontraron datos lejanos a 1 en el coeficiente de determinación, lo que significa que no hay valor significativo, sin embargo, en el MARE se encuentra cercanía en los datos. Para la combinación 4 se encentra valores similares a los de la combinación 3, lo que indico que los resultados obtenidos por este modelo no son significativos.

6. ANALISIS DE RESULTADOS

6.1. Análisis

En el análisis de los criterios de cumplimiento de cada uno de los modelos desarrollados se identifica que la mayoría de las series utilizadas en las combinaciones no cumple lo necesario para realizar los pronósticos a bases de estos datos puesto que en la mayoría de los modelos se exige el cumplimiento de linealidad, normalidad entre otros criterios y muy pocas de las series cumplen con estos requerimientos.

En la regresión simple se generar un pronóstico de la variable que fue seleccionada como dependiente (Y), donde se identifica que el R^2 presenta un valor de 0,606 lo que indica que el modelo explica mas del 60% de la variabilidad de la variable dependiente. El ANOVA presenta un p- valor de 0 el cual es < 0,05 lo cual confirma que las variables utilizadas se relacionan y son aptas para realizar pronósticos, al analizar la normalidad a partir del test de Kolmogorov se identifica que los datos no presentan una distribución normal por esta razón se evidencia que los resultados de sus pronósticos no son tan acercados a los observados.

El pronóstico con regresión lineal simple no es preciso para este tipo de datos ya que la precipitación se ve vinculada a más variables a la hora de ser pronosticada. Por tal razón fue pertinente el desarrollo del pronóstico a partir de otros modelos que contemplen más variables a la hora de ser desarrollados como la regresión múltiple ya que esta tiene en cuenta diversas variables por esta razón se analiza la correlación entre variables independientes que resultan significativas para el modelo, a partir de la combinación que cumplían con todos los criterio se desarrolló el pronóstico, a esta combinación se le descarta una de las variables iniciales (brillo solar) ya que presenta incumplimiento del criterio evidenciando resultados negativos de correlación de Pearson con las variables de temperatura y precipitación

Seguido de esto se realiza un análisis de correlación de las variables preseleccionadas de temperatura y precipitación frente a humedad relativa y evaporación, donde se identifica que estas variables no agregan valor al modelo por esta razón son descartadas. A partir de lo anterior se concluye que el resultado de este modelo no resulta significativo ya que los datos no cumplen con los criterios básicos lo cual obstaculiza el correcto desarrollo de este pues los datos de precipitación no presentan en su mayoría linealidad ni normalidad y las variables muestra correlación parcial entre ellas.

En el modelo de PCA, aunque la relación entre humedad relativa y evapotranspiración corresponde con el comportamiento de la temperatura, se

encontraron resultados no satisfactorios, puesto que los datos simulados no se encuentran cercanos a los datos tomados de las estaciones, también se soporta este resultado con los valores encontrados en las métricas de desempeño en donde los valores demuestran que los datos no son significativos.

El modelo de filtro de Kalman presenta como requerimiento inicial el cumplimiento de los criterios de linealidad y normalidad este procede a desarrollar inicialmente una corrección del ruido a los modelos que presentan un cumplimiento cercano al ideal, al desarrollar el pronóstico, corrección y suavizado no se desarrolla correctamente ya que el no cumplimiento de estos criterios genera que el pronóstico sea el promedio de los datos anteriores y se mantenga el mismo valor para cada dato hallado.

En el modelo de regresión lineal generalizado se identificó que de las cuatro familias de distribución que se utilizaron las cuales fueron Gauss, Gauss inversa, Gamma y Poisson la que presento un mejor desempeño fue la familia de gauss ya que presenta un coeficiente de determinación de 0,703 siendo el más cercano a 1 de los cinco modelos desarrollados, esto se sucede ya que esta distribución es la que mejor se adapta a las variables continuas y su correcto desempeño se ve limitado ya que las series de precipitación tomadas como datos de entrada no presentan una distribución normal de error.

7. CONCLUSIONES

- Según el análisis realizado a lo largo del proyecto, se pudo establecer que, de acuerdo con los resultados obtenidos, el modelo de Modelos Lineales Generalizados (MLG), esto se debió a que en este modelo se utiliza la familia Gamma, y esta permite que la cantidad de errores sea mínima.

- La evaluación de los fundamentos matemáticos permitió identificar las ventajas y desventajas de cada uno de los modelos teniendo en cuenta el tipo de dato que se utilizó en el proyecto (continuos) y de esta forma saber cómo se debía desarrollar cada uno de los modelos.

- La elaboración de las combinaciones para cada uno de los modelos permitió que la construcción de los algoritmos de acuerdo al mismo fuera más sencilla de analizar y concluir, para cada uno de los modelos, se realizaron escenarios que permitieron observar los diferentes aspectos de las variables, el comportamiento y los resultados de los métodos.

- La colinealidad en los datos utilizados en el modelo de regresión lineal múltiple representa un problema ya que al presentarse colinealidad parcial se aumenta el tamaño de los residuos generando coeficientes de regresión inestables pues al realizarse un cambio mínimo en los datos se producen cambios muy grandes en los coeficientes de regresión.

- Los modelos lineales parten del cumplimiento obligatorio de los criterios de distribución normal, homocedasticidad y linealidad, en la mayoría de los casos aplicados no se han cumplido en su totalidad estos criterios por lo cual se hace necesario el uso de los modelos lineales generalizados ya que este método permite el uso de distribuciones no normales en la variable de respuesta.

- El desarrollo de GLM tiene como objetivo determinar entre todos los posibles modelos cual es el más adecuado para el tipo de datos ingresados y el modelo resultante es el Gamma el cual logro explicar el 70% de la varianza, este modelo ofrece como ventaja frente a los de regresión que permite a la variable utilizar distribución diferente a la normal.

- El modelo de Gamma fue el que presento un desempeño más cercano al idear ya que este presenta una mejor adaptación a los datos continuos como los son las series de precipitación.

- La metodología de filtro de Kalman se incumple los supuestos básicos por lo cual el modelo no se desarrolla de forma óptima, pues este modelo presenta mejor uso como filtro de ruido en los datos para las series que no cumplen con normalidad y linealidad como es el caso de la variable precipitación y a partir de este filtrado utilizar otro tipo de metodologías de pronóstico para esta variable climatológica.

- En los tres modelos de regresión realizados se evidencia que los pronósticos hallados no identifican con facilidad los picos de precipitación de la zona de estudio, contrario a los valores medios de precipitación donde el valor observado y simulado se encuentran muy cercanos.

8. BIBLIOGRAFIA

[1] G. Poveda *et al.*, “Influencia de fenómenos macroclimáticos sobre el ciclo anual de la hidrología colombiana: cuantificación lineal, no lineal y percentiles probabilísticos,” *Meteorol. Colomb.*, vol. 6, no. Octubre de 2002, pp. 121–130, 2002.

[2] Isagen, “Central Hidroeléctrica Sogamoso,” *Isagen.com.co*, 2016.

[3] “Climatologia Colombiana,” pp. 1–25, 1965.

[4] ANH Agencia Nacional de Hidrocarburos, “Mapa de Cuencas,” p. 1091500, 2010.

[5] C. Arango, “Implementación del modelo wrf para la sabana de Bogotá,” p. 16, 2011.

[6] F. Bordas Pérez, “Implementación y evaluación de la Factorización de Cholesky mediante TBB y threads en arquitecturas multicore,” 2011.

[7] D. Vigo, “Linealidad.”

[8] A. N. Velasco and E. Domínguez, “Integración del concepto de variabilidad hidroclimática en pronósticos de largo plazo de resolución mensual en Colombia,” p. 93, 2016.

[9] D. Fernando, C. Madariaga, J. L. González, R. Miller, R. Lozano, and E. C. Vallejo, “Inferencia estadística Módulo de regresión lineal simple,” no. 147, p. 57, 2013.

[10] E. J. (Universidad N. de C. Hernandez, N. D. (Universidad N. de C. Duque Méndez, and J. (Universidad N. de C. Moreno-Cadavid, “Generación de pronósticos para la precipitación diaria en una serie de tiempo de datos meteorológicos Forecast generation for daily precipitation in a time series of meteorological data Geração de prognósticos para a precipitação diária em uma série de te,” *Ingenio Magno*, vol. 7, no. June, pp. 144–155, 2016.

[11] J. Nieto, “MODELO_REGRESIÓN_INEAL_MÚLTIPLE_PARA_DETERMINAR_INFLUENCIAS_DEL_ÍNDICE_NIÑO 1_2_y_LA_MJO_SOBRE_PRECIPITACIONES_EN_GYE_ENE_FEB_MAR_AB,” *Acta Oceanografica del Pacifico*. 2007.

[12] R. Granados, “Modelos de regresión lineal múltiple,” *Doc. Trab. en Econ. Apl.*, 2016.

[13] L. M. Carrascal, “Modelos generalizados lineales con spss,” 2015.

[14] S. De la Fuente Fernandez, “Análisis de Componentes Principales,” *Proy. e-Math Financ. por la Secr. Estado Educ. y Universidades*, pp. 141–151, 1999.

[15] H. Salinas P, “Análisis de componentes principales,” *Rev. Chil. Obstet. y Ginecol.*, vol. 71, no. 1, pp. 1–11, 2006.

[16] E. Leyton, “Desarrollo , implementación y prueba de un filtro de Kalman del tipo UKF para un vehículo aéreo no tripulado,” 2009.

[17] J. Dorado Ruíz, J. F., “Implementación de filtros de Kalman como método de ajuste a los modelos de pronóstico (GFS) de temperaturas máximas y mínima

para algunas ciudades de Colombia," 2018.
[18] T. Dalgleish *et al.*, "[No Title]," *J. Exp. Psychol. Gen.*, vol. 136, no. 1, pp. 23–42, 2007.
[19] Á. Jaramillo-Robledo and B. Chaves-Córdoba, "Distribución de la precipitación en Colombia analizada mediante conglomeración estadística," *Cenicafé*, vol. 51, no. 2, pp. 102–113, 2000.