Revista de Climatología

Vol. 23 (2023): 58-70 ISSN 1578-8768

Análisis de series temporales en estaciones meteorológicas para la predicción de la precipitación en la ciudad de Manizales, Colombia

Camilo Andrés Pulzara Mora[1]*, Juan David Losada Losada[2]

[1] Universidad Internacional de Valencia. España.

[2] Facultad de Ciencias e Ingeniería. Universidad de Manizales. Colombia

*Autor para correspondencia: Camilo Andrés Pulzara Mora, capulzaram@unal.edu.co

(Recibido: 19-06-2023. Publicado: 29-07-2023.)

DOI: 10.59427/rcli/2023/v23.58-70

Resumen

En este artículo se analizaron diferentes series de tiempo utilizando los modelos ARIMA, SARIMA, SARIMAX, Prophet y Neural Prophet con el fin de predecir la precipitación en la ciudad de Manizales-Colombia, haciendo uso de datos proporcionados por el SIMAC. Adicionalmente, los resultados obtenidos por los modelos para las predicciones de los últimos 7 días, muestran valores del error cuadrático medio (RMSE) y del error absoluto medio (MAE) alrededor de 19, indicando que los valores predichos presentan un buen ajuste frente a modelos más robustos como redes neuronales. Por otro lado, el modelo Prophet alcanzó un valor de RMSE igual a 19.06313 y un MAE de 16.24064, donde se evidenciaron errores más bajos que los modelos estocásticos implementados en este trabajo. Por otra parte, los valores predichos por la librería Prophet pueden ser de gran utilidad para el desarrollo de mejores prácticas en la gestión de análisis y riesgo de deslizamientos de tierra en el área. Finalmente, con base en este análisis se desarrolla un sistema de alerta temprana basado en el A25.

Palabras claves: *Series de tiempo, precipitación, ARIMA, SARIMA, SARIMAX, Prophet, Neural Prophet.*

Abstract

In this article, different time series are analyzed using the ARIMA, SARIMA, SARIMAX, Prophet, and Neural Prophet models in order to predict precipitation in the city of Manizales, Colombia, using data provided by SIMAC. Additionally, the results obtained by the models for the predictions of the last 7 days show root mean square error (RMSE) and mean absolute error (MAE) values around 19, indicating a good fit of the predicted values against more robust models such as neural networks. On the other hand, the Prophet model achieved an RMSE value of 19.06313 and a MAE of 16.24064, demonstrating lower errors compared to the stochastic models implemented in this work. Furthermore, the predicted values from the Prophet library can be highly useful for the development of best practices in landslide analysis and risk management in the area. Lastly, based on this analysis, an early warning system based on the A25 is developed.

Keywords: *Time series, precipitation, ARIMA, SARIMA, SARIMAX, Prophet, Neural Prophet.*

1. Introducción

En Colombia, existen diferentes ciudades que atraviesan la cordillera de los Andes, donde se presentan tormentas y diluvios, asociados al ascenso de masa del aire producto del choque entre ellas. Además, la ciudad de Manizales que se encuentra ubicada en la parte central de la cordillera de los andes se ve influenciada por el fenómeno del niño y el frente intertropical (CIOH, 2010). Por otro lado, debido a sus condiciones geológicas y topográficas, el crecimiento de la población en zonas de riesgo y a la falta de políticas de planeación territorial, esta ciudad del departamento de Caldas se ha visto afectada por sucesos naturales como deslizamientos e inundaciones que han generado un impacto social y económico devastador para su población. Por este motivo, Van Westen dedica un primer estudio sobre la probabilidad de ocurrencia de un deslizamiento a razón de la lluvia acumulada de los 25 días en esta ciudad (van Westen & Erlien, 1995). Adicionalmente, se han desarrollado proyectos en temas de prevención de desastres naturales, tales como inundaciones y deslizamientos con el fin de conocer los factores de causalidad (Grajales García, 2021), (Hardoy & Velásquez Barrero, 2014). Además, por ley en el Decreto 919 de 1989 y la ley 99 de 1993, se hace referencia sobre la parte de prevención y atención de desastres para garantizar la seguridad de los ciudadanos (Decreto Ley 919 de 1989, 1997). Por esta misma razón, en la última década, se han implementado en las entidades locales y académicas nuevos proyectos para realizar investigaciones en áreas de alto riesgo. De igual manera, se han establecido indicadores y áreas críticas que deben ser analizadas y monitoreadas constantemente. Sin embargo, los fenómenos naturales son muy difíciles de pronosticar debido a que los sucesos están relacionados con otros factores atmosféricos, tales como cambios en la presión, temperatura, humedad del lugar, dirección del viento e intensidad de la radiación. Por este motivo, tener un conocimiento sobre los cambios del clima se vuelve un requisito principal a la hora de evaluar los riesgos en la infraestructura, medio ambiente, y en la sociedad misma.

En este sentido, las series de tiempo toman un papel importante para la predicción del modelamiento de información meteorológica y en diversas áreas donde se presenten fenómenos que varíen en términos de intervalos de tiempo (Collischonn et al., 2005), (Hung et al., 2009), (Mahsin et al., 2012). Además, existen diferentes objetivos que incluyen: la comprensión y la descripción de un mecanismo de generación, la predicción de valores futuros y el control óptimo de un sistema. Por último, la naturaleza de una serie de tiempo se basa en que sus observaciones pueden estar correlacionadas o ser dependientes, donde, y su orden es significativo (Wei, 1991). La predicción de la lluvia o inundaciones como eventos probabilísticos es un tema esencial para la planificación del recurso hídrico. Este tipo de variables se miden de manera longitudinal en el tiempo. De esta manera, el análisis de series de tiempo de eventos con valores discretos se vuelve apropiado para monitorear el comportamiento hidrológico (Ansari, 2013). De igual manera, la precipitación hace parte de los componentes complejos y desafiantes del ciclo hidrológico para modelar y pronosticar, debido a los fenómenos ambientales y variaciones aleatorias en el espacio - tiempo (Htike & Khalifa, 2010).

Diferentes métodos de Machine Learning e inteligencia artificial son frecuentemente utilizados para la predicción de lluvias, y de los cuales muchos analistas en el tema u operarios se han beneficiado debido a sus resultados (Abhishek et al., 2012), (Gorlapalli et al., 2022). Sumado a esto, se han llevado a cabo diferentes estudios sobre la evolución espacial y temporal de la precipitación (Lu et al., 2019). Como ejemplo, algunos de los modelos más importantes son el modelo ARIMA que combina redes neuronales para predecir la lluvia por mes, el modelo SARIMA que permite el análisis de lluvia utilizando la estacionalidad en la serie de tiempo y el test de Dickey Fuller para determinar la estacionariedad (Jibril et al., 2017) y finalmente el modelo SARIMAX para predecir subseries utilizando el método de wavelet para obtener información sobre el tiempo y la frecuencia de la señal (Farajzadeh & Alizadeh, 2018). Este trabajo contribuye al pronóstico de la precipitación en la ciudad de Manizales, donde los resultados se utilizan para construir un sistema de detección temprana de deslizamientos de tierra. Por esta razón, se realiza un aporte en el estudio de series de tiempo utilizando los modelos ARIMA, SARIMA, SARIMAX, Prophet y Neural Prophet, con los datos climáticos proporcionados por el SIMAC. Finalmente, se destaca que no se encuentran muchas investigaciones en esta región en particular, lo cual abre las puertas a nuevos resultados que pueden ser la base para mejorar las predicciones de la precipitación.

2. Metodología

Recolección de Datos:

En este estudio se utilizó la base de datos proporcionada por el Sistema Integrado de Monitoreo Ambiental de Caldas (SIMAC) de Manizales, de las 13 estaciones meteorológicas, la cual registra información cada 5 minutos de las siguientes variables: fecha, hora, temperatura (°C), precipitación (mm), velocidad del viento (m/s), presión (mmHg), humedad (%) y radiación solar (W/m2) (CORPOCALDAS & Universidad Nacional, 2022).

Tratamiento de los datos:

Se realizó un procedimiento para la imputación de los valores faltantes utilizando los métodos MissForest y MICE, con el fin de determinar el modelo que se ajusta mejor al comportamiento de los datos climáticos. Después de modificar algunos hiperparámetros se observó que los valores predichos para cada modelo presentan un comportamiento y tendencia diferente, tal como se muestra en la figura 1. En la figura 1-b se puede observar que los valores predichos dependen de la escala de la variable que se imputa, por esta razón las magnitudes se acercan al punto máximo de la precipitación. Este resultado comprueba que el modelo se basa en los k-vecinos más cercanos (KNN) del dato faltante para estimar su valor. Por otro lado, en la figura 1-d el modelo MissForest imputa la variable utilizando un bosque aleatorio para estimar los datos nulos y en una etapa de mejora iterativa corrige los errores utilizando la media o la moda, repitiendo los pasos hasta converger en las predicciones.

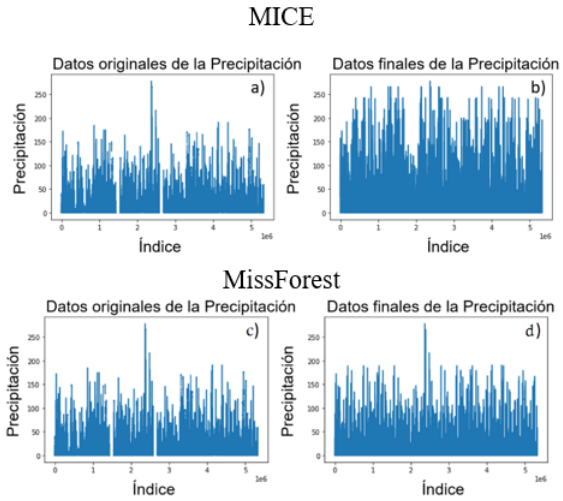

Figura 1: Precipitación vs Índice – Datos originales y Datos finales aplicando los modelos MICE y MissForest.

Los resultados de los valores predichos para la variable de la temperatura, utilizando ambos modelos, se presentan en la figura 2. Donde, se puede observar una diferencia menos significativa entre los dos modelos, debido a que los valores predichos oscilan en un intervalo menor (10°C y 30 °C), y el modelo MICE está sesgado por los valores extremos de cada variable meteorológica, cambiando el comportamiento o la tendencia natural en los datos, bajo el criterio experto. Por esta razón, el método más viable a utilizar es el modelo MissForest.

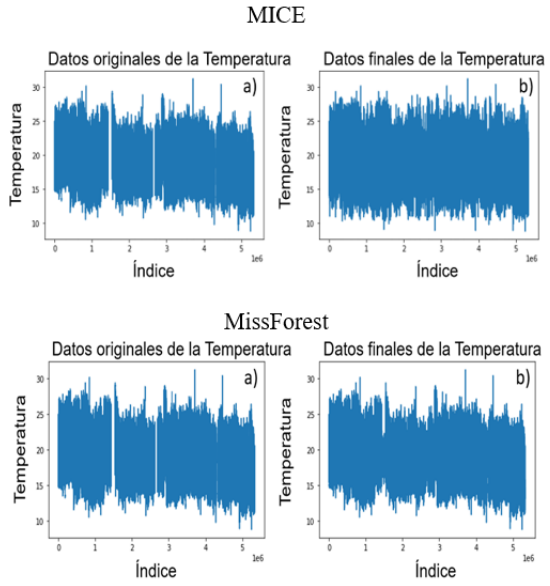

Figura 2: Temperatura vs Índice – Datos originales y Datos finales aplicando el modelo MICE y MissForest.

3. Resultados

El análisis de los modelos de series de tiempo se divide en dos partes. En la primera parte, se presentan los modelos ARIMA, SARIMA y SARIMAX. En donde para determinar si la serie es estacionaria, se utilizó la prueba de Dickey-Fuller (ADF) y se obtuvo el valor p. Además, se analizó la función de autocorrelación y la autocorrelación parcial para entender el comportamiento de la serie y determinar los hiperparámetros del modelo (p, d, q).

En la segunda parte, se implementó una primera aproximación y optimización en los modelos estocásticos y probabilísticos (Prophet y Neural Prophet) para determinar los valores predichos que mejor se ajusten a los datos meteorológicos. Este proceso se llevó a cabo utilizando métricas de evaluación tales como RMSE, MSE y MAE, seleccionando el resultado con el menor valor.

Hipótesis Nula

La serie de tiempo es no estacionaria.

Con el fin de determinar el comportamiento de la serie de tiempo, fue necesario implementar el método Augmented Dickey Fuller (ADF). Este método expande la Ecuación de Dickey Fuller, que se muestra en la ecuación 1, para incluir un proceso regresivo de alto orden en el modelo.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} + ... + \phi_p \Delta Y_{t-p} + e_t \quad (1)$$

Donde:
y_{t-1}=lag 1 de la serie de tiempo.
ΔY_{t-1}=Primera derivada de la serie de tiempo (t-1)

Dado que la hipótesis nula asume la presencia de raíz unitaria, es decir α=1, el valor p obtenido debe ser menor que el nivel de significancia ($< .05$) y el estadístico de prueba menor que los valores críticos, para rechazar la hipótesis nula. De esta manera, se infiere que la serie es estacionaria (Mushtaq, 2011). Los valores estadísticos obtenidos de la prueba ADF, el valor – p y valores críticos para cada estación, se muestran en la tabla 1.

Tabla 1: Valores críticos.

Valores críticos		
1 %	5 %	10 %
-3.430	-2.862	-2.567

En la tabla 2 se observa que el valor-p de cada una de las 13 estaciones (p-valor = 0) es menor que el nivel de significancia, y además el ADF estadístico es mucho menor a los valores críticos del 1 %, 5 %, 10 % que se muestran en la tabla 3. Esto implica que las series temporales son estacionarias, rechazando así la hipótesis nula.

Tabla 2: Valores obtenidos para el ADF estadístico y el valor - p para las 13 estaciones

Estación	**ADF Estadístico**	**p-valor**
Alcázares	-16.05967	0.000
Aranjuez	-17.53631	0.000
Bosques del Norte	-15.18559	0.000
Chec Uribe	-17.54788	0.000
El Carmen	-16.61556	0.000
EMAS	-15.82624	0.000
Hospital de Caldas	-16.70738	0.000
La Nubia	-15.95407	0.000
La Palma	-16.26608	0.000
Milán	-16.60355	0.000
Obs. Vulcanológico	-15.49576	0.000
Posgrados	-16.75009	0.000
Yarumos	-15.33693	0.000

Autocorrelación (ACF) Y Autocorrelación Parcial (PACF)

Estos métodos se utilizan para determinar la estacionalidad y los parámetros del proceso autorregresivo y media móvil (p, q) de la serie de tiempo, analizando los valores que están dentro del umbral de significancia en la gráfica de la función de autocorrelación (ACF) y autocorrelación parcial (PACF). Los resultados para la estación Alcázares se muestran en la figura 3.

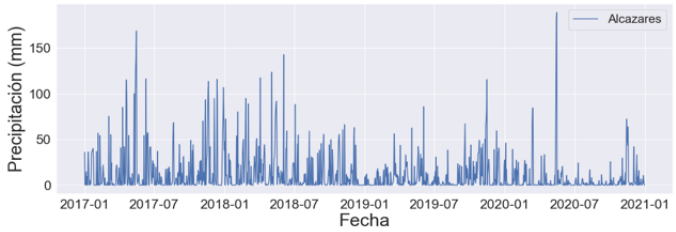

Figura 3: Precipitación vs fecha – Estación Alcázares.

En la figura 3, se observa como la serie de tiempo no tiene una tendencia, y presenta fluctuaciones que incrementan y disminuyen aleatoriamente en los valores de la precipitación. Además, el tamaño de las variaciones es constante en el tiempo, lo cual es otro indicador de la estacionariedad en la gráfica.

En la figura 4, la función de autocorrelación (ACF) cae rápidamente y tiende a cero, lo que sugiere que los valores de la precipitación no están correlacionados con sus valores anteriores, lo cual es una característica de una serie de tiempo estacionaria. Por otro lado, los valores de la ACF no son estadísticamente significativos ya que se encuentran por debajo del umbral que se representa por la región de color azul, lo cual indica que hay presencia de ruido blanco. Es importante resaltar que en los lags de la gráfica se evidencia un comportamiento parecido a una función senoidal, donde los valores oscilan de manera positiva y negativa, dentro del umbral de significancia. Esto se podría explicar debido a la presencia de estacionalidad en la serie de tiempo. Adicionalmente, se encuentran picos significativos aproximadamente en el lag 1, 30 y 60, que representan posiblemente una estacionalidad (Hyndman & Athanasopoulos, 2018).

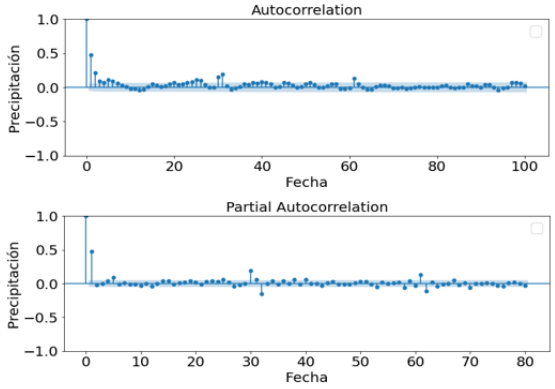

Figura 4: ACF y PACF - Estación Alcázares.

El comportamiento del ACF y PACF, representa un modelo ARMA (1,1), o en su defecto un modelo ARIMA (101). Adicionalmente, es posible notar que, el proceso ARMA (1,1) contiene el MA (1) como caso especial. Además, el PACF del proceso ARMA (1,1) también disminuye exponencialmente como el ACF, dependiendo de los signos y magnitudes de $\phi 1$ y $\theta 1$, de la ecuación de autocorrelación. Por último, para ARMA (1,1) el modelo sigue la función de autocorrelación ($\rho\kappa$) que se presenta en la ecuación 2:

$$\rho_k \begin{cases} 1 & k = 0 \\ \frac{(\phi_1 - \theta_1)(\phi_1 \theta_1)}{1 + \theta_1^2 - 2\phi_1 \theta_1} & k = 1 \\ \phi_1 \rho_{k-1} & k \geq 2 \end{cases} \tag{2}$$

En la ecuación 2 del modelo ARMA, $\phi 1$ representa el término de la media móvil y $\theta 1$ corresponde a la parte autorregresiva. Adicionalmente, la figura 4 confirma que cuando tanto la función de autocorrelación (ACF) como la función de autocorrelación parcial (PACF) disminuyen, se obtiene un modelo ARMA mixto. Es importante destacar que debido al efecto combinado de $\phi 1$ y $\theta 1$, el PACF del proceso ARMA (1,1) incluye una variedad más amplia de formas que el PACF del proceso MA (1), que solo contiene dos posibilidades (Wei, 1991). De acuerdo a lo mencionado anteriormente se ha desarrollado un modelo ARMA para predecir la precipitación teniendo en cuenta la parte estacional y teniendo en cuenta las variables exógenas de los datos. Por lo tanto, se ha implementado una primera aproximación para el modelo de serie de tiempo con la configuración del proceso autorregresivo (p) igual a 1 y el proceso de media móvil (q) igual a 1, los cuales se describen a continuación.

Modelo Baseline

Varios hiperparámetros fueron probados inicialmente, realizando algunos experimentos durante la programación para su ajuste final. Adicionalmente, dependiendo de los resultados de los valores estadísticos de ADF, valor p, la función ACF y PACF se seleccionó el modelo Baseline.

model = sm.tsa.arima.ARIMA(train['Precipitacion'],order=(1,0,1))
smodel = sm.tsa.statespace.SARIMAX(train['Precipitacion'],order=(1,0,1), seasonal_order = (1,0,1,6))
smodel = sm.tsa.statespace.SARIMAX(train['Precipitacion'],order=(1,0,1), seasonal_order = (1,0,1,6), exog = train[['Temperatura']])

Las métricas utilizadas para determinar el menor error de los modelos implementados son el RMSE, MAE y MSE. Además, se ha obtenido el error total de cada modelo entrenado y validado, calculando la media de las métricas individuales de las estaciones y dividiéndolas entre 13. Finalmente, se ha seleccionado el modelo que presenta el menor valor de las métricas entre los cinco modelos evaluados. En la figura 5, se puede observar la gráfica de los datos de entrenamiento (línea azul) y los datos predichos (línea naranja) del modelo ARIMA (1,0,1) para la estación Alcázares, este modelo presenta el menor resultado de la métrica RMSEtrain = 19.08935 tal como se muestra en la tabla 3.

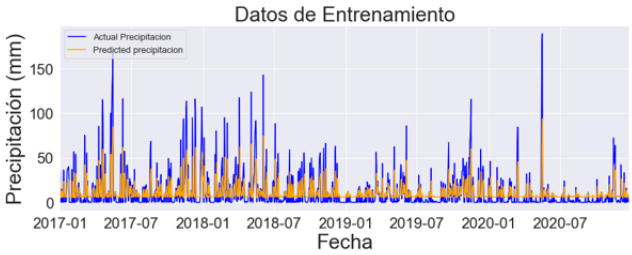

Figura 5: Datos de Entrenamiento para la estación Alcázares, utilizando el Modelo ARIMA_{101}.

Tabla 3: Métricas de evaluación para los datos de entrenamiento.

Entrenamiento	RMSE	MAE	MSE
ARIMA_{101}	19.08935	12.12777	369.58516
SARIMA_{101}	19.24187	12.06760	375.48963
SARIMAX_{101}	19.29660	11.94998	377.68039
$\text{Prophet}_{DÍAS}$	20.83209	13.76298	442.65334
$\text{Neural Prophet}_{DÍAS}$	21.08390	13.43836	452.82367

A continuación, en la tabla 4 se presentan los resultados de los datos de testeo correspondientes a la predicción de los últimos siete días del mes de diciembre del año 2021 para cada modelo implementado:

Tabla 4: Métricas de evaluación para los datos de testeo.

Test	RMSE	MAE	MSE	AIC
ARIMA_{101}	27.07459	23.70040	870.11410	12690.11469
SARIMA_{101}	27.70922	23.83590	924.10553	12705.52531
SARIMAX_{101}	28.34202	24.37122	958.88549	12829.05908
$\text{Prophet}_{DÍAS}$	24.91720	21.37104	756.58101	-
$\text{Neural Prophet}_{DÍAS}$	30.38424	24.75415	1115.48134	-

Contrario a los datos de entrenamiento, el menor valor de error de las métricas RMSE, MAE y MSE, se encuentran en el modelo Prophet, tal como se muestra en la tabla 4. En particular, el RMSE se ha utilizado para penalizar los errores grandes, y se ha encontrado que el modelo Neural Prophet presenta diferentes variaciones en los datos predichos que no se ajustan bien a los datos de entrenamiento ni a los de validación. No obstante, se sabe de antemano que el RMSE y el MAE se dan en las mismas unidades de la variable respuesta, lo que implica que los errores son altos en magnitud ya que la precipitación oscila entre 0 y 150 mm. Con el fin de reducir el error en las métricas de evaluación fue necesario optimizar los hiperparámetros de los modelos.

Optimización de hiperparámetros

ARIMA, SARIMA y SARIMAX

Para la optimización de los hiperparámetros de los modelos ARIMA, SARIMA y SARIMAX, se implementó la función auto ARIMA, que busca la combinación de los parámetros p,d,q que minimiza el valor del AIC para evitar el sobreajuste, los resultados se muestran en la tabla 5.

Tabla 5: Resultados de las métricas de evaluación - Modelo ARIMA$_{d=0,s=True}$.

ARIMA d = 0 Seasonal = True	**Parámetro (pdq)**	**RMSE**	**MAE**	**MSE**
Alcázares	(100)	7.47173	6.44317	55.82679
Aranjuez	(200)	18.07889	15.49646	326.84650
Bosques del norte	(101)	16.59625	15.30626	275.43558
CHEC Uribe	(101)	43.47141	36.48689	1889.76429
El Carmen	(101)	16.77893	14.43663	281.53257
EMAS	(100)	11.73780	8.89895	137.77597
Hospital de Caldas	(100)	21.16828	18.71273	448.09638
La Nubia	(400)	18.45964	15.90361	340.75846
La Palma	(201)	41.67653	35.19205	1736.93333
Milán	(303)	29.75427	26.09933	885.31662
Obs. Vulcanológico	(200)	39.50590	31.97134	1560.71687
Posgrados	(100)	24.59690	21.06550	605.00755
Yarumos	(201)	28.92923	25.36730	836.90051
Total		**24.47890**	**20.87540**	**721.60857**

Después de realizar el análisis estadístico de la estacionariedad, utilizando las pruebas ADF y el valor-p donde se rechaza la hipótesis nula, se decidió cambiar algunos hiperparámetros dentro del modelo auto ARIMA. Por consiguiente, el término de derivación d, el cual representa la n-ésima derivada, se modificó dependiendo del comportamiento de cada estación meteorológica. Finalmente, con el fin de disminuir el valor de las métricas y el AIC, los valores de autoregresión y de media móvil tomaron valores entre cero y cinco.

En la tabla 6, se observan los resultados de las métricas finales para los modelos ARIMA, SARIMA y SARIMAX. Así mismo, se identificó que el mejor modelo es ARIMA bajo la configuración de diferenciación cero (d = 0) y presencia de estacionalidad. Por último, los errores de las métricas de evaluación para el modelo ARIMA toma valores de RMSE = 24.47890, MAE = 20.87540, MSE = 721.60857 y un AIC = 12696.14238.

Tabla 6: Resultado del promedio de las métricas de evaluación final – Modelos ARIMA, SARIMA y SARIMAX.

Modelos	**RMSE**	**MAE**	**MSE**	**AIC**
ARIMA $_{d=1,Seasonal=False}$	25.31393	21.15138	784.65185	12709.60054
ARIMA $_{d=0,Seasonal=True}$	24.47890	20.87540	721.60857	12696.14238
SARIMA $_{D=1,m=12}$	27.22153	21.93400	865.01840	13005.91031
SARIMA $_{D=0,m=12}$	24.57647	20.92591	729.32498	12727.42554
SARIMAX $_{D=1,m=6}$	28.32908	23.33140	967.88948	13075.36138
SARIMAX $_{D=0,m=6}$	24.50772	20.88507	707.97779	12693.24569

Se analizaron los residuos o errores de los modelos para las 13 estaciones, los cuales corresponden a la diferencia entre valores reales y valores predichos. Adicionalmente, se muestra un estudio de la estación Alcázares para explicar los resultados obtenidos. Por otra parte, el criterio para seleccionar un buen modelo se basa en obtener ruido blanco aleatorio de los residuos, es decir, aquellos que presentan una distribución con media cero. A continuación, en la figura 6 se muestran los resultados residuales usando gráficas de Q-Q, histograma, análisis cuantil y del residuo estándar, del modelo ARIMAd = 0, Seasonal = True, de la estación.

En la figura 6, se puede determinar que existe una autocorrelación positiva de los residuos en el correlograma (figura 6-d) y conforme incrementa el eje x toma valores cercanos a cero. Por otra parte, según el gráfico de la estandarización residual (figura 6-a), los valores oscilan cerca a la media. Además, en la figura 6-c se muestra que las cantidades de los residuos en el gráfico normal Q-Q no se ajustan bien a la línea recta diagonal que representa el comportamiento de una distribución normal, debido a los valores extremos donde los puntos se desvían en la

parte superior de la gráfica. En otras palabras, no se cumple por completo el supuesto de una distribución normal, y esto se corrobora en el histograma donde la curva es asimétrica con sesgo positivo. Por consiguiente, el modelo puede mejorarse, hasta satisfacer las propiedades de media cero.

El resumen de salida que se obtiene al implementar la función auto ARIMA devuelve diferentes cantidades significativas de información estadística. En la figura 7 sólo se observa la columna de coeficientes la cual muestra la importancia de cada característica y cómo influye en la serie temporal. El valor p indica el impacto de cada función de peso, en este caso, cada valor es igual a cero, por lo tanto, se dejan los coeficientes dentro del modelo ($p < .05$).

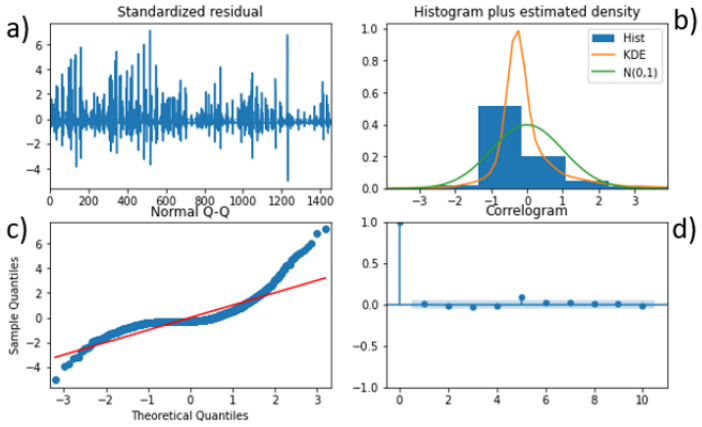

Figura 6: Diagnóstico de residuos del modelo ARIMA$_{d=0,s=True}$ para la estación Alcázares.

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	5.9100	0.855	6.914	0.000	4.235	7.585
ar.L1	0.4727	0.010	45.361	0.000	0.452	0.493
sigma2	362.5956	7.257	49.962	0.000	348.371	376.820

Figura 7: Resumen de parámetros del modelo ARIMA$_{d=0,s=True,(100)}$ - Estación Alcázares.

En la figura 8, se presenta el resultado de la predicción para los 7 días siguientes después de la fecha del 2020-12-24, para la estación Alcázares. Adicionalmente, la línea de tendencia de los valores predichos es aproximadamente constante en el tiempo, donde la precipitación oscila entre 5 y 12 mm. De igual manera, se observa el umbral en color gris que representa el rango de valores dentro del cual podría encontrarse el valor de la predicción para la precipitación, fuera del rango se clasifica como un valor atípico. Por otro lado, el valor de la métrica RMSE se muestra en la Tabla 5, donde se puede observar el puntaje más bajo de las 13 estaciones para el modelo ARIMA$_{d=0,s=True}$ con una configuración de (1,0,0) y un RMSE = 7.47173.

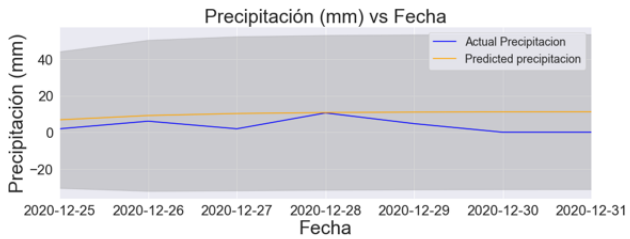

Figura 8: Predicción 7 días - Modelo ARIMA$d = 0, s = True, (100)$ para la estación Alcázares.

Prophet y Neural Prophet

Los modelos probabilísticos Prophet y neural Prophet fueron optimizados utilizando diferentes configuraciones en los parámetros de entrada. Además, las variables exógenas (temperatura, presión, radiación, velocidad, humedad) fueron añadidas, con el fin de determinar si podrían influenciar en la variable objetivo y a partir de las métricas de evaluación obtener un valor menor de error. Por otro lado, el efecto de los días feriados fue reemplazado por los días en donde una catástrofe por deslizamiento de tierra sucedió en la ciudad de Manizales. Finalmente, estos sucesos se modelan de forma análoga a los regresores futuros para el modelo Neural Prophet. El código implementado es el siguiente:

```
holidays = pd.DataFrame({
            'holiday': 'eventosmanizales',
            'ds': pd.to_datetime(['2017-04-19']),
            'lower_window': 0,
            'upper_window': 1,
})
    m.add_regressor('Temperatura')
    m.add_regressor('Presion')
    m.add_regressor('Velocidad')
    m.add_regressor('Humedad')
    m.add_regressor('Radiacion')
```

En la figura 9, se puede observar el comportamiento de los valores predichos de entrenamiento, que se ajustan muy bien a los valores de precipitación reales por debajo de 50 mm. Adicionalmente, se identifica un pico cercano a 130 mm en los valores predichos (Línea de color azul) que el modelo entrena y predice con un valor de precipitación alta. Se infiere que, debido al efecto del evento de desastre agregado al modelo Prophet, se presenta como una anomalía que se ve reflejada en el entrenamiento de los datos. Así mismo, se presenta para las demás estaciones meteorológicas.

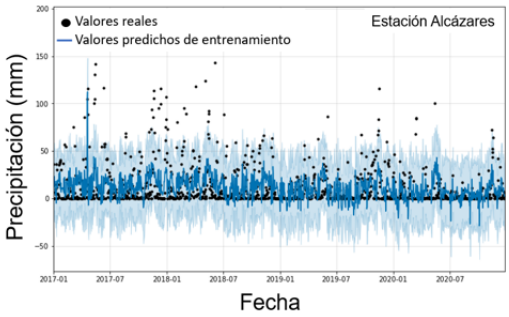

Figura 9: Valores predichos de entrenamiento vs valores reales usando el modelo Prophet multivariado con 5 variables para la estación Alcázares.

Las métricas de evaluación final se muestran en la tabla 7. Adicionalmente, los modelos fueron llevados desde 3 variables hasta 5 variables exógenas, para encontrar el mejor ajuste. Por otro lado, se implementaron diferentes hiperparámetros en el modelo Prophet como el tipo de la estacionalidad, que toma los valores de aditivo y multiplicativo. Además, las fechas de los desastres por deslizamiento de tierra fueron incorporadas. Por último, para el modelo Neural Prophet, las redes neuronales se entrenaron con 8 capas ocultas, un epochs de 200 y 16 dimensiones de capas ocultas de AR-Net.

Tabla 7: Resultados del promedio de las métricas de evaluación final para el modelo Prophet.

Modelos	**RMSE**	**MAE**	**MSE**
Prophet Multivariado 3 Variables	19.55671	17.07854	460.28991
Prophet Multivariado 5 Variables	19.06313	16.24064	430.72827
Prophet Multivariado 5 Variables Multiplicative	20.58596	17.63104	504.82400
Neural Prophet Multivariado 3 Variables	29.60179	23.35640	1037.28838
Neural Prophet Multivariado 5 Variables	28.40272	22.55186	966.35207
Neural Prophet Multivariado 5 Variables Multiplicative	27.56973	22.93964	920.99573

Se determina una disminución en el valor de las métricas de evaluación al utilizar 5 variables exógenas, donde la variable objetivo se ve afectada por la influencia de estas. Finalmente, se puede analizar que el pronóstico de la precipitación con menor valor del RMSE es el modelo Prophet multivariado, con estacionalidad aditiva y con 5 variables exógenas, el valor de RMSE = 19.06313, MAE = 16.24064 y MSE = 430.72827.

En la figura 10, se observa el resultado del modelo con menor error entre los hiperparámetros implementados para la estación Alcázares. De igual manera, se evidencia un buen ajuste y que los datos predichos están muy cerca a los datos reales. Así mismo, la línea azul tiene un comportamiento creciente y decreciente en el tiempo, que varía entre 0 y 10 mm de precipitación. Finalmente, el valor del RMSE para la estación Alcázares es de 5.41035, con un MAE = 4.67015 y un MSE = 29.27193.

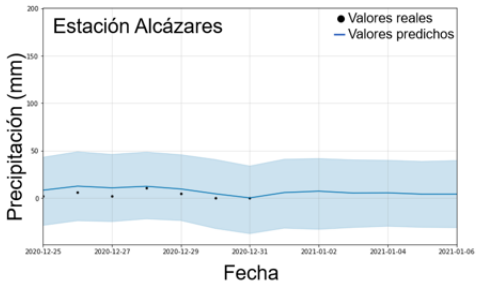

Figura 10: Valores de test predichos vs reales usando el modelo Prophet multivariado con 5 variables para la estación Alcázares.

4. Discusión

El sistema de alarma que se implementó en el SIMAC en la sede de la ciudad de Manizales se basa en el artículo escrito por Westen y Terlien (Van Westen & Erlien, 1995). Por otra parte, el deslizamiento de tierra puede predecirse a mediano o largo plazo, por medio de la estadística, la probabilidad y con la obtención de los datos históricos recolectados de la red meteorológica. A raíz de esto, es posible crear un boletín con los indicadores de la precipitación acumulada (mm) de los 25 días anteriores, la cual se le conoce como A25.

Westen y Terlien realizaron una investigación sobre la cantidad de lluvia diaria en la estación de agronomía ubicada en Cenicafé (Norte - Manizales), para determinar alguna correlación con las características del suelo. Posteriormente, como resultado se obtuvo presencia de deslizamiento cuando el valor del A25 alcanzó un límite de 200 mm de precipitación acumulada durante los últimos 25 días (Van Westen & Erlien, 1995), (Vélez et al., 2010). Finalmente, como consecuencia de este estudio, las entidades de análisis de riesgo establecieron los niveles de alerta siguiendo los umbrales de la tabla 8:

Tabla 8: Estrategia del semáforo de acuerdo con el color.

Color Del Semáforo	**Precipitación Acumulada - A25**
Alerta Amarilla	$\geq 200\ mm\ y < 300\ mm$
Alerta Naranja	$\geq 300\ mm\ y < 400\ mm$
Alerta Roja	$\geq 400\ mm$

De acuerdo con los resultados el modelo que se seleccionó para predecir la precipitación fue Prophet, donde se obtuvieron los valores más bajos en las métricas de evaluación entre los 5 modelos implementados, para este trabajo.

El objetivo es prevenir y alertar algún tipo de deslizamiento en las diferentes estaciones, con la ayuda del indicador A25, se sumaron los últimos 18 valores de la precipitación de la base de datos y los 7 días de la predicción. Además, el programa imprime el tipo de alerta de acuerdo con su color para cada estación meteorológica y el valor acumulado de los últimos 25 días de la precipitación (mm), tal como se observa en la Figura 11.

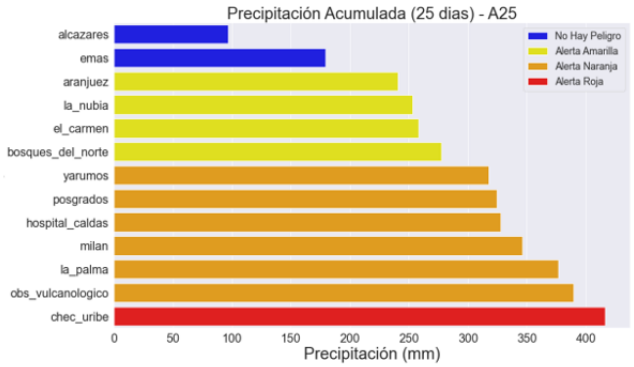

Figura 11: Sistema de alerta para la estación Posgrados.

Implementación del Sistema de anomalías

Se implementó un análisis de anomalías para detectar patrones de la precipitación (mm) que se desvían de su comportamiento, con el fin de alertar sobre tendencias anormales por día. Adicionalmente, este sistema está basado en los valores de la media, la ventana de tiempo y la desviación estándar de los datos de lluvia por cada estación meteorológica. Finalmente, la media se calcula para los datos de cada estación, la ventana = 25 y σ toma valores entre 2 y 3. El código implementado para encontrar los valores superiores es el siguiente:

$$data\left['superior'\right] = \frac{data[0] \cdot rollling(window = wind)}{mean()} - (sigma * data[0] \cdot rolling(window = wind).std())$$

A continuación, en la figura 12 se presenta el resultado obtenido por el análisis de las anomalías.

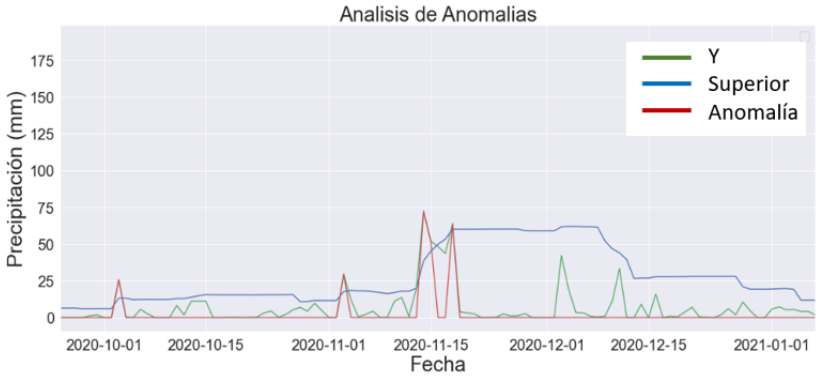

Figura 12: Análisis de anomalías para la estación Alcázares.

5. Conclusiones

Haciendo uso de las pruebas estadísticas que determinan el tipo de serie para las 13 estaciones meteorológicas, se determinó el valor p < 0.05 y un valor ADF menor a los valores críticos del 1 %, 5 % y 10 %. Lo anterior permite sugerir que las series son estacionarias. A partir de los modelos de imputación implementados para los valores nulos, se puede concluir que MICE presenta valores de imputación que se ven afectados por los vecinos más cercanos donde se evidencia la mayor cantidad de datos predichos con altas magnitudes; Por otro lado, MissForest utiliza un bosque para estimar los valores nulos y en una etapa iterativa corrige los errores utilizando la media o la moda, por esta razón presenta una mejor tendencia respecto a cada variable climática del conjunto de datos estudiados. Para los 3 modelos ARIMA, SARIMA y SARIMAX, de acuerdo con el criterio de información de Akaike, se implementó la función auto ARIMA minimizando el valor AIC para las 13 estaciones meteorológicas. Finalmente, ARIMA con derivada cero (d = 0) y presencia de estacionalidad (s = True), se ajustó más a los datos de precipitación; así mismo, los valores obtenidos en las métricas son RMSE = 24.47890, MAE = 20.87540 y MSE = 721.60857. Se evidencia que Prophet se ajusta mejor que el modelo Neural Prophet y los modelos estocásticos, usando las 5 variables exógenas y las fechas de los desastres por deslizamientos de tierra. Se reporta un valor de

RMSE obtenido para Prophet de 19.06313 y para Neural Prophet de 27.56973. En comparación con los modelos estocásticos, la librería Prophet prevé parámetros más intuitivos que son fáciles de usar y de mejorar para la predicción de la precipitación. La ventaja de estos modelos radica en que la duración de la implementación es más corta que los modelos ARIMA, SARIMA y SARIMAX.

6. Referencias bibliográficas

Abhishek, K., Kumar, A., Ranjan, R., & Kumar, S. (2012). A rainfall prediction model using artificial neural network. Proceedings - 2012 IEEE Control and System Graduate Research Colloquium, ICSGRC 2012, 1, 82–87.

Ansari, H. (2013). Forecasting Seasonal and Annual Rainfall Based on Nonlinear Modeling with Gamma Test in North of Iran. International Journal of Engineering Practical Research, 2(1), 16–29.

CIOH. (2010). Circulación general de la atmósfera en colombia. In Bicentenario de la Independencia de Colombia.

Collischonn, W., Haas, R., Andreolli, I., & Tucci, C. E. M. (2005). Forecasting River Uruguay flow using rainfall forecasts from a regional weather-prediction model. Journal of Hydrology, 305(1–4), 87–98.

CORPOCALDAS, & Universidad Nacional, C. (2022). CDIAC - Centro de Datos e Indicadores Ambientales de Caldas. cdiac.manizales.unal.edu.co/ Decreto Ley 919 de 1989. (1997).

Farajzadeh, J., & Alizadeh, F. (2018). A hybrid linear–nonlinear approach to predict the monthly rainfall over the Urmia Lake watershed using wavelet-SARIMAX-LSSVM conjugated model. Journal of Hydroinformatics, 20(1), 221–231.

Gorlapalli, A., Kallakuri, S., Sreekanth, P. D., Patil, R., Bandumula, N., Ondrasek, G., Admala, M., Gireesh, C., Anantha, M. S., Parmar, B., Yadav, B. K., Sundaram, R. M., & Rathod, S. (2022). Characterization and Prediction of Water Stress Using Time Series and Artificial Intelligence Models. Sustainability, 14(11), 6690.

Grajales García, J. A. (2021). Landslide hazard assessment by climatic events in the basin of Quebrada El Rosario – Manizales using application ALICE.

Hardoy, J., & Velásquez Barrero, L. S. (2014). Re-thinking "Biomanizales": Addressing climate change adaptation in Manizales, Colombia. Environment and Urbanization, 26(1), 53–68.

Htike, K. K., Khalifa, O. O. (2010). Rainfall forecasting models using Focused Time-Delay Neural Networks. International Conference on Computer and Communication Engineering, ICCCE'10, May, 11–13.

Hung, N. Q., Babel, M. S., Weesakul, S., & Tripathi, N. K. (2009). An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrology and Earth System Sciences, 13(8), 1413–1425.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. Principles of Optimal Design, 504.

IDEAM. (2020). SIMAC. IDEAM. https://idea.manizales.unal.edu.co/reporte-meteorologico.html

Jibril, Y. K., Abdulkarim, K., & Nathan, S. A. (2017). Time Series Analysis And Forecasting Of Monthly Rainfall Data In Zaria, Nigeria. 3rd YUMSCIC, November, 310–313.

Lu, Y., Jiang, S., Ren, L., Zhang, L., Wang, M., Liu, R., & Wei, L. (2019). Spatial and Temporal variability in precipitation concentration over mainland China, 1961-2017. Water (Switzerland), 11(5).

Mahsin, M., Akhter, Y., & Begum, M. (2012). Modeling Rainfall in Dhaka Division of Bangladesh Using Time Series Analysis. Journal of Mathematical Modelling and Application, 1(5), 67–73.

Matplotlib: Python plotting. (2022). https://matplotlib.org/

Mushtaq, R. (2011). Augmented Dickey Fuller Test. SSRN Electronic Journal, 1–19.

NumPy. (2019). https://www.numpy.org/

Pandas. (2022). https://pandas.pydata.org/docs/

Python Data Analysis Library - pandas. (2022).

Sci-kit learn: machine learning in Python. (2022).

SciPy.org. (n.d.). https://www.scipy.org/

van Westen, C. J., & Erlien, M. T. J. (1995). An approach towards deterministic landslide hazard analysis in GIS: a case study from Manizales, Colombia. Earth Surface Processes and Landforms, 21, 853–868.

Vélez, J. J., Mejía, F., Pachón, A., & Vargas, D. (2010). An Operative Warning System of Rainfall-Triggered Landslides at Manizales, Colombia. Proceedings of World Water Congress and Exhibition IWA, 1, 19–24.

Wei, W. W. S. (1991). Time Series Analysis: Univariate and Multivariate Methods. In International Journal of Forecasting (pearson, Vol. 33, Issue 1).

Welcome to Python.org. (2019). https://www.python.org/