Predicción de caudales medios diarios por medio de modelos de inteligencia artificial y aprendizaje automático

Juan Esteban Taborda Soto[1*]

[1]Universidad Nacional de Colombia

Resumen

Aquí voy a escribir un resumen jejeje

1 Introducción

Los caudales de los ríos fluctúan en diferentes escalas de tiempo que comprenden horas, días, semanas, meses y años. El entendimiento de estas fluctuaciones ha sido de gran importancia para los investigadores colombianos, especialmente en los ríos aferentes al Sistema Interconectado Nacional (SIN), ya que la generación de energía eléctrica por medio de dichos caudales representa cerca del 70% de la energía producida por la Asociación Colombiana de Generadores de Energía Eléctrica (ACOLGEN), la cual tiene cerca del 70% de la capacidad instalada del país (Acolgen, 2022). Los principales hallazgos respecto a los moduladores de estas fluctuaciones tienen que ver con la oscilación meridional de la Zona de Convergencia Intertropical (ZCIT) que tiene influencia sobre la variación anual y semianual del caudal (Mejia et al., 1999) junto con la actividad del Chorro del Occidente Colombiano (CHOCÓ) (Poveda & Mesa, 2000), El Niño - Oscilación del Sur (ENSO), el cual influye sobre la variación interanual (Arias et al., 2021; Poveda et al., 2020, 2011) y, la Oscilación Decadal del Pacifico (PDO) y Oscilación Multidecadal del Atlántico (AMO) que repercuten sobre las variaciones decadales (Poveda, 2004). Bajo este conocimiento se han desarrollado diferentes modelos de pronóstico estadístico de caudal medio mensual que permitan asegurar la operación de los embalses con desempeños aceptables en estas escalas de tiempo. Algunos de los métodos utilizados en estos modelos de pronóstico consisten en la aplicación de Regresión Lineal Múltiple (Poveda et al., 2002), Redes Neuronales (Poveda et al., 2002), Modelo MARS (Poveda et al., 2002; Sanchez & Poveda, 2006), Análisis Espectral Singular (Carvajal et al., 1998; Rojo-Hernández & Carvajal-Serna, 2010), Modelo ARIMA (Sanchez & Poveda, 2006), Modelos autoregresivos (Salazar Velásquez & Mesa Sánchez, 1994) y Predicción por bandas espectrales (Poveda et al., 2002).

Por otro lado, la necesidad de contar con información de los mecanismos que intervienen en la modulación de las oscilaciones intraestacionales (con periodos entre 1-90 días) ha impulsado diversos estudios que se han enfocado principalmente en las relaciones con la Oscilación de Madden-Julian (MJO) (Arias, 2005; Poveda & Mesa, 2000; Torres-Pineda & Pabón Caicedo, 2017; Grimm, 2019), la cual es el principal modo de variabilidad intraestacional en el trópico (Madden & Julian, 1971, 1972) y en modelos de pronóstico que incluyan este mecanismo como variable predictora, los cuales se han basado principalmente en métodos de Análisis Espectral Singular y Predicción por bandas espectrales (Arias, 2005; Arenas Cárdenas & Carvajal Serna, 2010; Yepes Palacio, 2012). Recientemente, se ha encontrado que además de la MJO, otras ondas atmosféricas acopladas con la convección, en particular Ondas Kelvin, Ondas Rossby y Ondas del Este, se relacionan coherentemente con la precipitación sobre la región (Giraldo-Cardenas et al., 2022; Hoyos & Taborda, n.d.;

*ORCID: https://orcid.org/0000-0002-1908-6030

Corresponding author: Juan Esteban Taborda Soto, jetabordas@unal.edu.co

Taborda-Soto, n.d.), y que en la dinámica los chorros de bajo nivel podrían jugar un papel importante en la variación intraestacional (Arias et al., 2021; Taborda-Soto, n.d.; Serra et al., 2010; Arias, 2005). En este sentido, en el presente trabajo se explora la predicción de caudales medios diarios del río Sogamoso utilizando estas variables como datos de entrada junto con variables tradicionales y diferentes modelos de inteligencia artificial y aprendizaje automático para determinar la ganancia de estos enfoques respecto a enfoques tradicionales de pronóstico climatologico y antecedente.

El documento se divide de la siguiente manera...

2 Datos

La información de la variable objetivo, la cual es el caudal diario del río Sogamoso antes de la entrada al embalse Topocoro, se calcula por medio de la información de las estaciones El Jordan, Remolino y El tablazo del IDEAM, las cuales se pueden descargar en `http://dhime.ideam.gov.co/atencionciudadano/`. El caudal por tanto, tiene un periodo desde enero del 2001 hasta diciembre del 2020, con un periodo faltante de datos desde enero del 2011 hasta diciembre del 2012.

Por otro lado, para las variables predictoras se descargó información de precipitación total, vientos horizontales y humedad especifica desde 700 a 1000 hPa del reanalisis ERA5 desde enero del 2001 hasta diciembre del 2020 que se puede obtener de la pagina web `https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5`. Con esta información se calcula la precipitación promedio sobre la cuenca y la precipitación acumulada de los últimos 7 días, y las series de advección de humedad de los chorros del CHOCÓ (región), Caribe (región) y Orinoquia (región). El periodo que comprenden estas series es desde enero del 2001 hasta diciembre de 2020.

Adicionalmente, se utilizó radiación de onda larga saliente (OLR) diaria de la NOAA que se puede obtener de la pagina web `https://psl.noaa.gov/data/gridded/data.olrcdr.interp.html`. Con esta información se calcularon las series de las ondas acopladas con la convección sobre la cuenca del río Sogamoso, primeramente calculando los campos por medio del filtro de frecuencia y numero de onda (Wheeler & Kiladis, 1999) y luego calculando las series de estos campos de ondas sobre la región aferente a la cuenca del río Sogamoso. El periodo que comprende estas series es desde febrero de 2001 hasta noviembre de 2020, debido a que en el filtro se pierden los primeros y últimos 30 días del periodo descargado. Hablar sobre los tipos de ondas

Asimismo, de la variable original se calcularon dos variables predictoras, el caudal del día antes, y la tendencia, que es la diferencia entre dos días de caudal consecutivos.

3 Metodología

La metodología se divide en 3 items: selección de variables predictoras, entrenamiento y evaluación de modelos de pronósticos, y evaluación de la ganancia de estos modelos respecto a los pronósticos de referencia: climatologico y antecedente.

3.1 Selección de variables

El desempeño de un modelo de pronostico estadístico depende en gran medida de las variables de entrada, esto significa, que las variables predictoras deben contener información relevante sobre el comportamiento de la variable que se quiere explicar. Para realizar este analisis primero se dividen los conjuntos de datos en un conjunto de entrenamiento (desde 2001 hasta 2016) y un conjunto de prueba (desde 2017 hasta 2020), y los analisis se realizan en el conjunto de entrenamiento. Dado que estas variables climatologías tienen un marcado ciclo anual, se calculan las anomalías estandarizadas por medio de ecuación 1, donde i cor-

responde a cada día de los 366 días del año. Luego de tener estas anomalías estandarizadas que serán con las que se calibraran los modelos, se realiza un analisis de correlaciones rezagadas y componentes principales para definir las mejores variables predictoras que ingresan a los modelos de pronóstico.

$$AnomEstand = \frac{dato_i - media_i}{desv_i} \tag{1}$$

3.2 Entrenamiento, validación y prueba de modelos

Como se mencionó anteriormente, el periodo de entrenamiento de los modelos corresponde a las fechas entre 2001 y 2016. En este caso se entrenaron siete (7) tipos de modelos los cuales son: regresión lineal múltiple (LR), vecino más cercano (KNN), vectores de soporte (SVR), ramdom forest (RF), SARIMAX, redes neurales artificiales tipo percertrón multicapa (MLP) y recurrentes de largo y corto plazo (LSTM). En cada uno de estos modelos se variaron los hiperparametros de los que dependen en mallas fijas y aleatorias para obtener los hiperparámetros con mejores resultados, con los cuales se definieron los modelos finales a los cuales se les evaluó por medio de una validación cruzada kfold de 5 ventanas consecutivas y el numero de muestras. Además, se calculo en el periodo de prueba (2017-2020), las metricas del r2 y el error cuadratico medio (mse) para determinar cual modelo presenta el mejor desempeño.

3.3 Ganancia respecto a pronósticos de referencia

Por ultimo, es importante reconocer que en el medio se desarrollan principalmente dos tipos de pronósticos de referencia; el primero de ellos es el pronóstico climatologico, es decir, que el pronostico de caudal para un día i corresponde al promedio de los días i de todos los años de registro; y el segundo es el pronostico antecedente, el cual hace referencia a pronosticar el caudal del día i como el caudal observado el día anterior, esto dada la memoria que tienen los caudales de los ríos. Por tanto, en este paso se evaluar la ganancia en la reducción del error cuadrático medio al pronosticar con los modelos de pronostico respecto a estos dos tipos de pronostico de referencia, esto se realiza por medio de métrica del Skill Score (SS), la cual se presenta en la ecuación 2.

$$SS = 1 - \frac{E_{pron}}{E_{ref}} \tag{2}$$

4 Resultados

A continuación se presentan los resultados de la selección de variables, calibración y evaluación de los modelos de pronóstico, y la evaluación de los pronósticos con respecto a los pronósticos de referencia.

4.1 Selección de variables

En la Figura 1 se presentan las series de cada una de las variables estandarizadas con los parámetros de medias y desviación estándar el periodo de entrenamiento, además, se presenta la serie de anomalías de caudal, la cual es la variable que se pretende pronosticar. De esta imagen se puede observar que las anomalías en el periodo de prueba presentan valores mayores a los encontrados en el periodo de entrenamiento en todas las variables, pero que se ve más marcado en las variables de caudal y precipitación. Además, también se observa que en variables como el caudal se presentan variaciones interanuales.

Por otro lado, es bien sabido que cuando hablamos de pronósticos de series de tiempo, las relaciones entre las variables explicativas respecto a la variable objetivo pueden tener

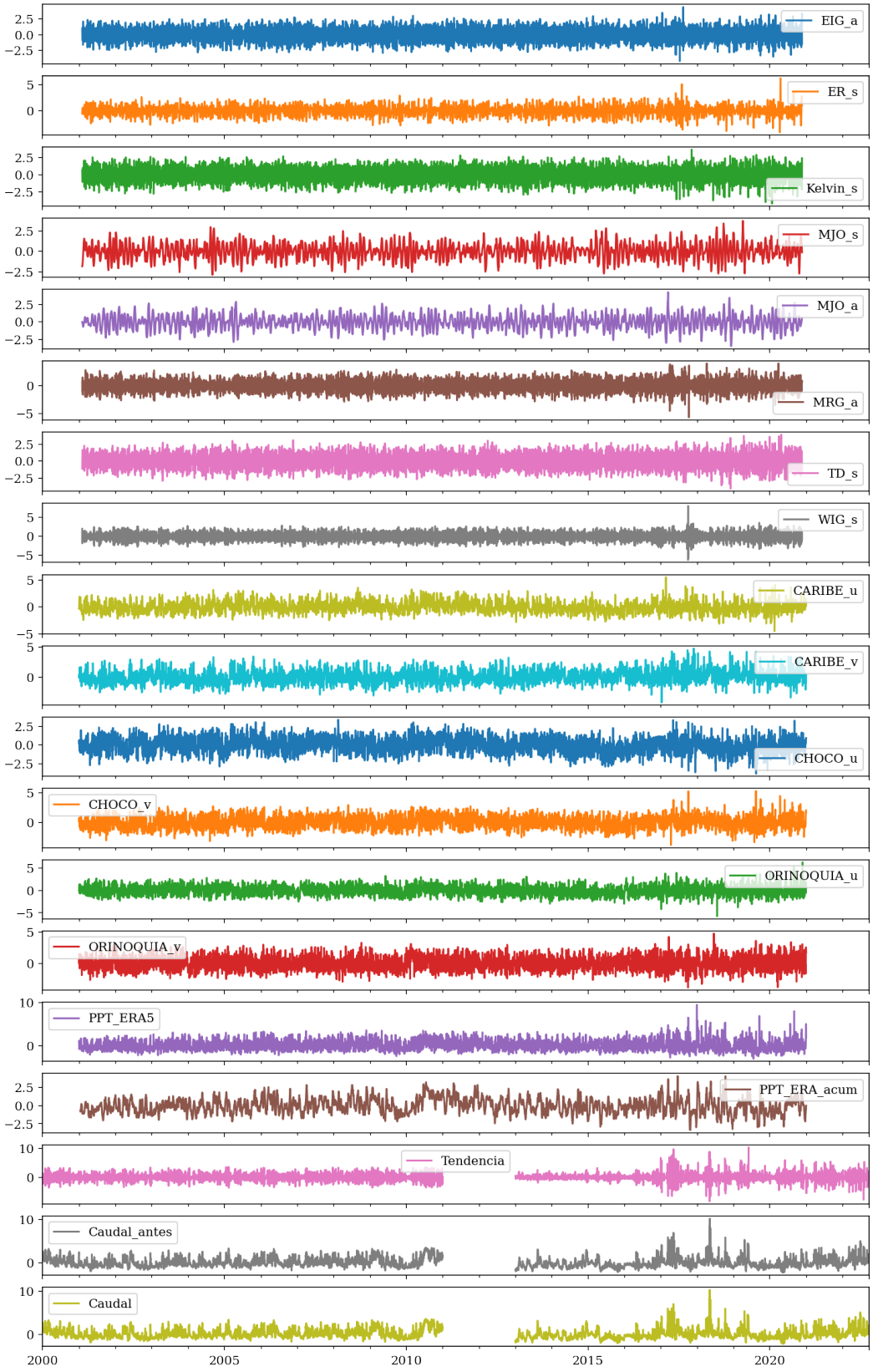

Figure 1. Series de anomalías estandarizadas de cada una de las variables predictoras y objetivo.

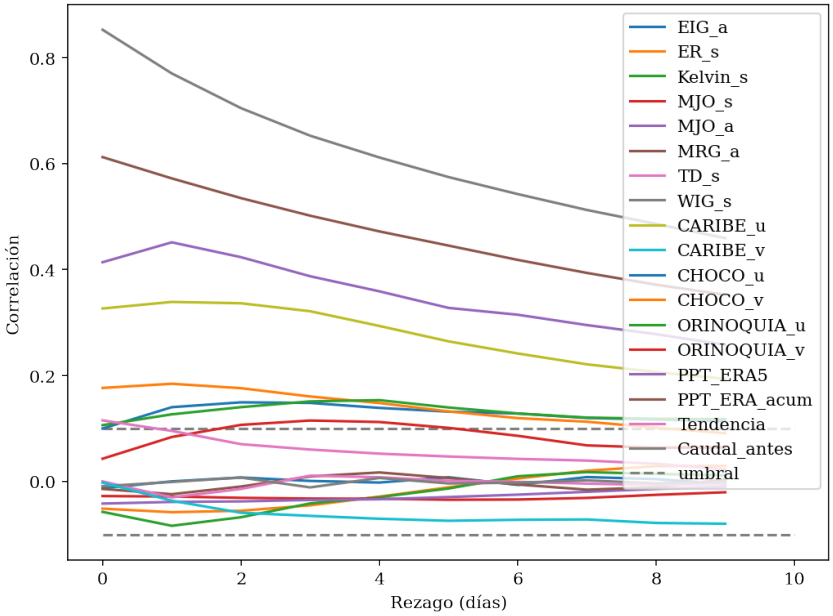

Figure 2. Correlaciones rezagadas entre las variables explicativas y objetivo.

cierto rezago, por lo que se hace necesario desarrollar una analisis de correlaciones rezagadas con cada uno de los índices para determinar cuales son los rezagos óptimos para las variables que ingresan a los modelos. En este sentido, en la Figura 2 se presentan las correlaciones rezagadas de cada una de las variables con la variable objetivo. Se puede observar que las variables de la precipitación, las ondas kelvin, MRG y TD, el chorro del caribe zonal y los chorros del chocó tienen un rezago de 1 día; Los chorros del chocó, y las ondas Rossby, tienen rezago de 2 y el chorro de la orinoquia rezago de 3; y además, variables tienen las mayores correlaciones con rezago 0, destacablemente el caudal del día antes y la precipitación acumulada. También es importante mencionar que estos rezagos por ejemplo en la precipitación son congruentes con el tiempo de concentración de la cuenca que es aproximadamente de 1.5 días.

Con base en los rezagos anteriores se calculan las correlaciones entre los indices y las anomalías de caudal, donde resaltan altas correlaciones con el las anomalías de caudal antecedente (0.85), precipitación acumulada (0.61), precipitación del día anterior (0.45), caribe zonal con rezago 1 (0.31), entre otros, pero además se presentan altas relaciones entre las variables predictoras como las anomalías de caudal antecedente y la precipitación.

En este sentido, se decide realizar un analisis de componentes principales con estas variables para obtener variables independientes entre ellas, que ingresen a los modelos de pronostico. En la Figura 3 la matriz de correlación de las 18 componentes principales con las anomalías de caudal. Se puede observar que las componentes principales PC1, PC2, PC18, PC9, PC6, PC16, PC7 y PC8 superan el umbral de 0.1 de correlación y por tanto serán las variables que se utilizarán como variables predictoras en los modelos. Estas variables presentan una buena separación de las anomalías de caudal como se puede observar en la Figura 4, dónde en el espacio de las componentes principales PC1 y PC2, para valores negativos de ambas variables se tienen anomalías negativas de caudal, y para valores positivos de ambas variables se tienen anomalías positivas de caudal.

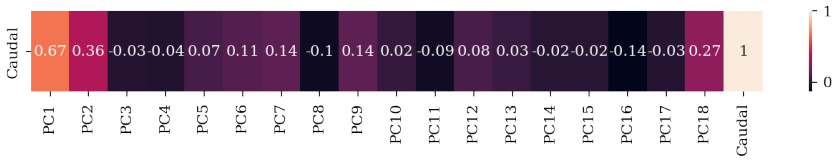

Figure 3. Correlaciones rezagadas entre las pcs y la variable objetivo.

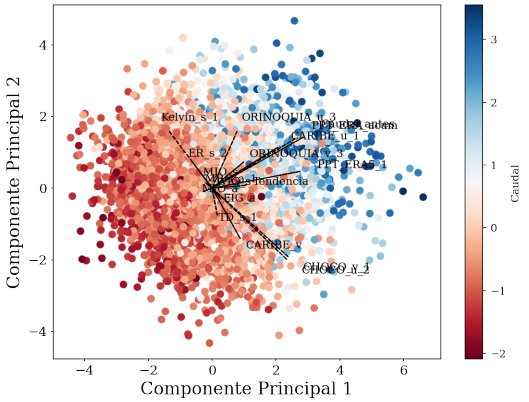

Figure 4. Anomalías de caudal en el espacio de las PC1 y PC2.

4.2 Entrenamiento, validación y prueba de modelos

De acuerdo con las variables definidas en la subsección anterior se plantean 7 modelos de pronostico a ser evaluados, estos son: regresión lineal múltiple (LR), vecino más cercano (KNN), vectores de soporte (SVR), ramdom forest (RF), SARIMAX, redes neurales artificiales tipo percertrón multicapa (MLP) y recurrentes de largo y corto plazo (LSTM) Hablar sobre los hiperparametros y su busqueda. Todos estos modelos, excpecto SARIMAX y LSTM, se evaluan en el periodo de entrenamiento por medio de una validación cruzada kfold de ventanas sucesivas y la métrica del r2, y además, se presenta la curva de entrenamiento para determinar si es posible que el modelo continué aprendiendo de más datos.

En la Figura 5 se pueden observar la validación cruzada de los modelos en la cual la forma de la función es similar en todos los modelos, donde se presenta un mejor desempeño en la ventana 4 de validación, del orden de 0.8. Además, los desempeños en las demás ventanas son del orden de 0.6, aunque con mejores resultados en la regresión lineal, la red neuronal MLP y SVR. Dado que estos tres modelos presentan resultados similares, se calculan las curvas de aprendizaje que se presentan en la Figura 6, de donde se puede observar que el modelo de regresión lineal y SVR aunque presentan buen desempeño, las curvas muestran que no pueden aprender más, mientras la red neuronal podría seguir aprendiendo con más muestras. Asimismo, se observa problemas de varianza en RF y KNN, siendo el primero más critico.

Por otro lado, dado que el modelo SARIMAX y LSTM requieren de continuidad porque son modelos que guardan memoria, estos se calibraron desde 2013 hasta 2016. Los resultados de la distribución de los residuales del modelo SARIMAX es similar al de regresión lineal, donde ambos se distribuyen de manera normal con algunos alejamientos en los eventos extremos. Asimismo, el modelo corrido para LSTM, muestra que no podría seguir aprendiendo con los hiperparametros que se le han sobreimpuesto.

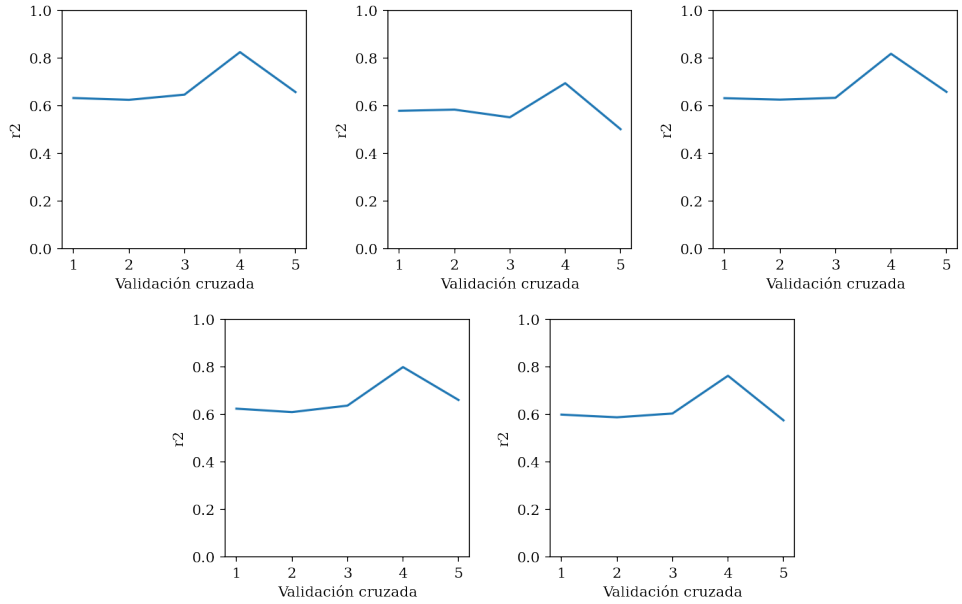

Figure 5. Validación cruzada en términos del r2 para los modelos de LR, KNN, SVR, MLP y RF (de izquierda a derecha y arriba hacia abajo).

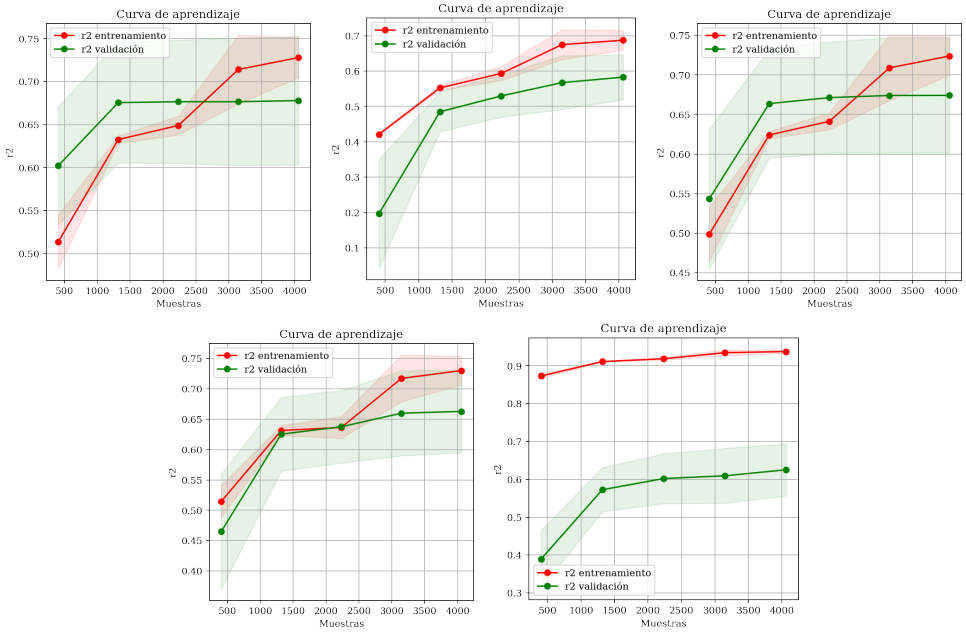

Figure 6. Curva de aprendizaje en términos del r2 para los modelos de LR, KNN, SVR, MLP y RF (de izquierda a derecha y arriba hacia abajo).

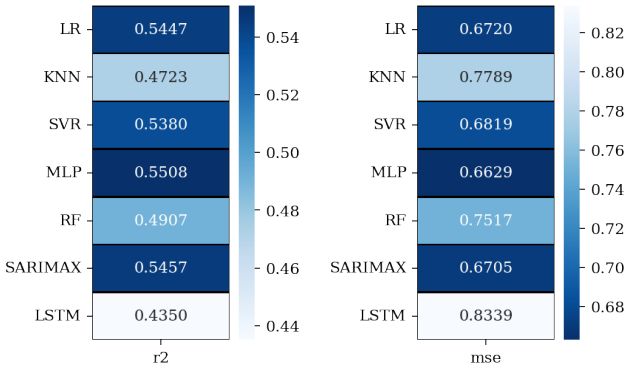

Figure 7. Curva de aprendizaje en términos del r2 para los modelos de LR, KNN, SVR, MLP y RF (de izquierda a derecha y arriba hacia abajo).

Para evaluar equitativamente el desempeño de los modelos se pronostican las anomalías de caudal en el periodo de prueba (2017-2020) y se calcula las métricas de r2 y error cuadrático medio (mse), las cuales se consignan en la Figura 7. Se puede observar que de acuerdo con estas métricas, los mejores 4 modelos en orden decendente son MLP, SARIMAX, LR y SVR, que son los mismos modelos en lo que el desempeño en el entrenamiento era superior.

4.3 Ganancia respecto a pronósticos de referencia

Pronosticar las anomalías de caudal surgió como una necesidad debido a que las variables climáticas tienen un marcado ciclo anual. En este punto, es necesario desestandarizar las anomalías de caudal, es decir, volver a las unidades originales del caudal para determinar el valor real del pronostico. En la Figura 8 se presenta los pronósticos de cada modelo, junto con el caudal observado y la climatología. Se puede observar que las deficiencias en algunos modelos como KNN que presentan bajos desempeños se debe a que los picos más altos no son alcanzados y estos pesan más en el error.

También se observa que modelos como LSTM captura los extremos pero se equivoca más en otros rangos de valores, pero en general, se ve una representación de los caudales mayor que la climatología, pero para evaluar esto más contundente mente en la Figura 9 se presenta la métrica de ganancia con respecto a la climatología y al caudal antecedente. De esta ultima figura, se observa que

5 Conclusiones

Se podrían probar nuevos modelos LSTM.

References

Acolgen. (2022). *Capacidad instalada en colombia.* Retrieved 2022-08-30, from `https://acolgen.org.co/`

Arenas Cárdenas, J. S., & Carvajal Serna, L. F. (2010). Desarrollo de un modelo de predicción de caudales semanales asociado a la variabilidad intraestacional en colombia. *Escuela de Geociencias y Medio Ambiente*.

Arias, P. A. (2005). *Diagnostico y predicción de la variabilidad intra-anual de la hidrología colombiana* (Unpublished master's thesis). Universidad Nacional de Colombia. Sede Medellín. Facultad de Minas.

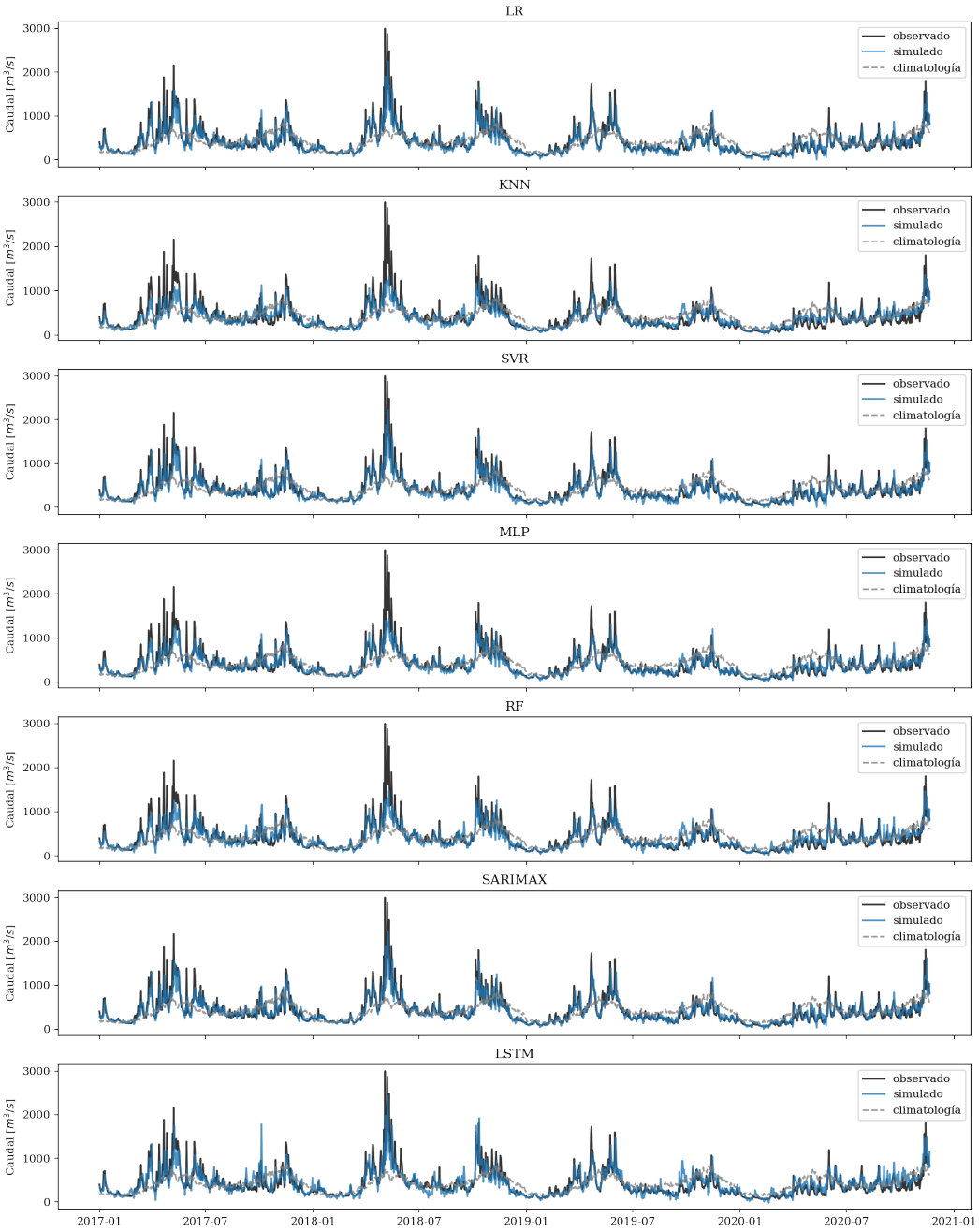

Figure 8. Series pronosticadas por cada uno de los modelos en el periodo de prueba.

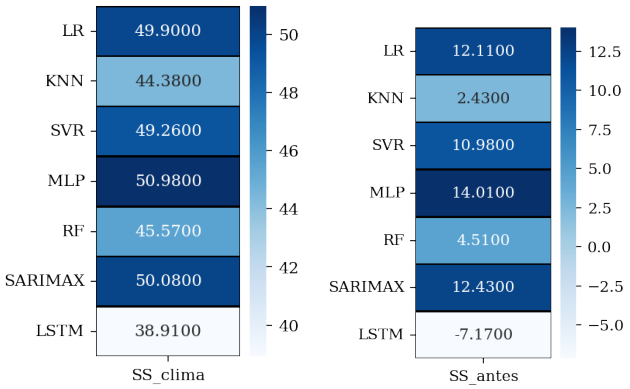

Figure 9. Curva de aprendizaje en términos del r2 para los modelos de LR, KNN, SVR, MLP y RF (de izquierda a derecha y arriba hacia abajo).

Arias, P. A., Garreaud, R., Poveda, G., Espinoza, J. C., Molina-Carpio, J., Masiokas, M., ... van Oevelen, P. J. (2021). Hydroclimate of the andes part ii: Hydroclimate variability and sub-continental patterns. *Frontiers in Earth Science*, *8*. Retrieved from `https://www.frontiersin.org/article/10.3389/feart.2020.505467` doi: 10.3389/feart.2020.505467

Carvajal, L., Salazar, J., Mesa, O., & Poveda, G. (1998, 01). Hydrological prediction in colombia using singular spectral analysis and the maximum entropy method. *Ingeniería hidráulica en México*, *13*, 7-16.

Giraldo-Cardenas, S., Arias, P. A., Vieira, S. C., & Zuluaga, M. D. (2022). Easterly waves and precipitation over northern south america and the caribbean. *International Journal of Climatology*, *42*(3), 1483-1499. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.7315` doi: https://doi.org/10.1002/joc.7315

Grimm, A. (2019, 07). Madden–julian oscillation impacts on south american summer monsoon season: precipitation anomalies, extreme events, teleconnections, and role in the mjo cycle. *Climate Dynamics*, *53*. doi: 10.1007/s00382-019-04622-6

Hoyos, C., & Taborda, J. (n.d.). *The influence of equatorially trapped waves on precipitation variability in the amazon basin and northern south america.*

Madden, R. A., & Julian, P. R. (1971). Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. *Journal of Atmospheric Sciences*, *28*(5), 702 - 708. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/28/5/1520-0469_1971_028_0702_doadoi_2_0_co_2.xml` doi: 10.1175/1520-0469(1971)028⟨0702:DOADOI⟩2.0.CO;2

Madden, R. A., & Julian, P. R. (1972). Description of global-scale circulation cells in the tropics with a 40–50 day period. *Journal of Atmospheric Sciences*, *29*(6), 1109 - 1123. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/29/6/1520-0469_1972_029_1109_dogscc_2_0_co_2.xml` doi: 10.1175/1520-0469(1972)029⟨1109:DOGSCC⟩2.0.CO;2

Mejia, J., Mesa, O., Poveda, G., Velez, J., Hoyos, C., Mantilla, R., ... Botero, B. (1999, 01). Distribución espacial y ciclos anual y semianual de la precipitación en colombia. *Dyna (Medellin, Colombia)*, *127*, 7-26.

Poveda, G. (2004, 01). La hidroclimatología de colombia: Una síntesis desde la escala inter-decadal hasta la escala diurna. *Rev. Acad. Colomb. Cienc*, *28*, 201-222.

Poveda, G., Alvarez, D. M., & Rueda, O. A. (2011). Hydro-climatic variability over the andes of colombia associated with enso: a review of climatic processes and their impact on one of the earth's most important biodiversity hotspots. *Climate Dynamics*, *36*(11), 2233–2249.

Poveda, G., Espinoza, J. C., Zuluaga, M. D., Solman, S. A., Garreaud, R., & van Oevelen, P. J. (2020). High impact weather events in the andes. *Frontiers in Earth Science*, *8*. Retrieved from https://www.frontiersin.org/article/10.3389/feart.2020.00162 doi: 10.3389/feart.2020.00162

Poveda, G., & Mesa, O. (2000, 06). On the existence of lloro (the rainiest locality on earth): Enhanced ocean-land-atmosphere interaction by a low level jet. *Geophysical Research Letters*, *27*, 1675-1678. doi: 10.1029/1999GL006091

Poveda, G., Mesa, O., Carvajal, L., Hoyos, C., Mejia, J., Cuartas, L., & Pulgarín. (2002, 01). Predicción de caudales medios mensuales en ríos colombianos usando métodos no lineales. *Meteorología Colombiana*, *6*, 101-110.

Rojo-Hernández, J. D., & Carvajal-Serna, L. F. (2010). Predicción no lineal de caudales utilizando variables macroclimáticas y análisis espectral singular. *Tecnología y ciencias del agua*, *1*(4), 59–73.

Salazar Velásquez, J. E., & Mesa Sánchez, O. J. (1994, ene.). Aplicación de dos modelos no lineales a la simulación de series hidrológicas. *Avances en Recursos Hidráulicos*(02), 27–47. Retrieved from https://revistas.unal.edu.co/index.php/arh/article/view/91916

Sanchez, J., & Poveda, G. (2006, 01). Aplicación de los métodos mars, holt-winters y arima generalizado en el pronóstico de caudales medios mensuales en ríos de antioquia. *Meteorología Colombiana*, *10*, 36-46.

Serra, Y. L., Kiladis, G. N., & Hodges, K. I. (2010). Tracking and mean structure of easterly waves over the intra-americas sea. *Journal of Climate*, *23*(18), 4823 - 4840. Retrieved from https://journals.ametsoc.org/view/journals/clim/23/18/2010jcli3223.1.xml doi: 10.1175/2010JCLI3223.1

Taborda-Soto, J. E. (n.d.). *Variabilidad intraestacional de la precipitaci´on sobre el norte de sudamérica: diagnóstico y conexiones.*

Torres-Pineda, C., & Pabón Caicedo, J. D. (2017, 03). Variabilidad intraestacional de la precipitación en colombia y su relación con la oscilación de madden-julian. *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales*, *41*, 79. doi: 10.18257/raccefyn.380

Wheeler, M., & Kiladis, G. N. (1999). Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. *Journal of the Atmospheric Sciences*, *56*(3), 374 - 399. Retrieved from https://journals.ametsoc.org/view/journals/atsc/56/3/1520-0469_1999_056_0374_ccewao_2.0.co_2.xml doi: 10.1175/1520-0469(1999)056⟨0374:CCEWAO⟩2.0.CO;2

Yepes Palacio, L. J. (2012). *Variabilidad climática intraestacional y su efecto sobre la precipitación en colombia: Diagnóstico y pronóstico* (Unpublished master's thesis). Escuela de Geociencias y Medio Ambiente.