DYNA
http://dyna.medellin.unal.edu.co/

UNIVERSIDAD NACIONAL DE COLOMBIA

Colombian monthly energy inflows predictability

Andrés Felipe Hurtado-Montoya [a] & Nicolás Alberto Moreno-Reyes [b]

[a] *ISAGEN, Medellín, Colombia. afhurtado@isagen.com.co*
[b] *Universidad EAFIT, Medellín, Colombia. namorenor@eafit.edu.co*

Received: May 6th, 2024. Received in revised form: September 13th, 2024. Accepted: September 27th, 2024.

Abstract
Streamflow forecasting is essential for water resources management in several social and economic strategic sectors, involving space-temporal variability modeling of the hydrological processes and the influence of several climatic phenomena. Furthermore, high water-dependent sectors such as the Colombian electricity market, require not only the expected streamflow values but also the occurrence probability or reliability bands of such forecast inflows necessary in robust risk analyses. We propose a mathematical approach for monthly streamflow forecasting in Colombia and quantify its predictability, incorporating climate model outcomes as a time series of macroclimatic indexes and punctual hydro-climatological stations. The methodology integrates parametric and non-parametric models, exogenous variables analysis, and uncertainty estimation through stochastic modeling. This research will contribute to the Colombian hydrology understanding and provide elements for risk analysis, planning, and decision-making in social and economic sectors involved with water resources management.

Keywords: streamflow forecasting; uncertainty; Colombian streamflow's predictability.

Predictibilidad de los aportes mensuales de energía en Colombia

Resumen
El pronóstico de caudales es esencial en la gestión de los recursos hídricos en varios sectores sociales y económicos estratégicos e involucra la modelación de variabilidad espacio temporal de los procesos hidrológicos y la influencia de varios fenómenos climáticos. Además, la alta dependencia del agua en sectores como el mercado eléctrico colombiano requiere (solo los valores de caudal esperados, sino también su probabilidad de ocurrencia o bandas de confianza de tales pronósticos necesaria en análisis robustos de riesgo. Se propone un enfoque matemático para el pronóstico mensual de caudales en Colombia y la cuantificación de su predictibilidad, incorporando resultados de modelos climáticos como series temporales de índices macroclimáticos y estaciones hidroclimatológicas puntuales. La metodología integra la aplicación de modelos paramétricos y (paramétricos, el análisis de variables exógenas y la estimación de la incertidumbre mediante modelos estocásticos. Esta investigación contribuirá en el entendimiento de la hidrología colombiana y brindará elementos para el análisis de riesgo, la planificación y la toma de decisiones en los sectores sociales y económicos involucrados con la gestión de los recursos hídricos.

Palabras clave: pronóstico de caudales; incertidumbre; predictibilidad de los caudales colombianos.

1 Introduction

Space-temporal variability understanding of the hydrological process is an imperative goal because of its environmental, social, economic, and cultural implications. Research in hydro-climatological topics is necessary to improve the water resource management of the country, where future water availability estimation through streamflow forecasting is a central task for the strategic planning of electrical, agricultural, tourist, and drinking water distribution sectors. Water resources management relies on hydrological historical data and future projections, where precipitation and streamflow time series, statistical analysis, and modeling have allowed a better understanding of the country's hydro-climatological processes. Furthermore, hydrological forecasting is an essential task for planning, and improving it is a continuous challenge to gain confidence in the decision-making process involving strategic national sectors.

Thanks to its abundant water resources, hydroelectricity makes up about 70% of the generation matrix, where different reservoirs help manage the natural inter-annual

How to cite: Hurtado-Montoya, A.F., and Moreno-Reyes, N.A., Colombian monthly energy inflows predictability. DYNA, 91(234), pp. 24-33, October - December, 2024.

variability linked to the hydro-climatological behavior. This high dependence on water availability and the uncertainty associated with the evolution of the climatic system led to significant levels of vulnerability to deal with through reservoir management, streamflow forecasting, and uncertainty quantification. Government planning institutions, the system operator, and the hydroelectric power plant owners perform the monthly streamflow forecasting for the Colombian electricity market to guarantee reliability through energetic resources management (mainly water, coal, and gas). Despite the availability of climate data and monthly streamflow time series of rivers used for the hydroelectricity generation in Colombia, streamflow forecasting uncertainty estimation is still subject to improvement. A mathematical approach incorporating elements such as the time of the year, the prediction horizon, and the climatic events will improve the forecasting activities using probabilistic methodologies.

This research starts with the importance of improving water availability forecasting and determining the reliability of the results. Both are essential tasks for planning in different highly water-dependent sectors, where not only the expected values are necessary but also the probability of occurrence of other plausible scenarios. Moreover, apart from the practical applications of this research, another goal is related to the continuous efforts to understand the Colombian hydro-climatological processes through mathematical modeling.

The next two subsections present an overview of the Colombian hydro-climatology variability and the mathematical modeling applied to hydrological time series. Then, section 2 describes the data and procedure to obtain the time series of interest. Section 3 corresponds with the methodology, and describes the models evaluated and the forecasting process. Finally, analysis and conclusions about the model performance and forecasting uncertainty in Colombia are given in sections 4 and 5.

1.1 Colombian hydro-climatology

Several factors modulate Colombian hydro-climatology, such as the tropical dynamics, the Intertropical Convergence Zone (ITCZ) displacement during the year [1,21], the humidity inflow from the Amazon basin and the Pacific and Atlantic oceans, the hydrological surface process, and the Andes physiography. Additionally, various climatic phenomena operating in different spatial and temporal scales affect the Colombian hydro-climatology, like the Pacific Decadal Oscillation (PDO), the North Atlantic Oscillation (NAO), the Quasi-Bienal Oscillation (QBO), El Niño-Southern Oscillation (ENSO), the Chocó low-level jet, the Mesoscale Convective Systems (MCS), the Madden-Julian Oscillation (MJO), and the tropical easterly waves [23,26]. The influence of this hydroclimatic complexity in water resources distribution, the access to numerous hydro-climatological databases, and the computational capabilities justify the research efforts to shape a better conceptual framework of the hydrological processes and their predictability using mathematical modeling. This research aims to contribute to the knowledge of Colombian hydrology through monthly streamflow predictability estimation using probability theory approaches.

Various hydrology and climate research form the conceptual framework for streamflow characterization and modeling in Colombia. Although ITCZ characterizes the streamflow annual cycle in the Colombian regions, other climate phenomena also influence the space-temporal variability. Associated with the trade winds, the Chocó low-level jet [25] is a wind current with considerable influence on Colombian climatology, specifically in the central and western zones. Furthermore, the Chocó jet interacts dynamically and thermodynamically with the MCS [17] in the region [23,27,29-31]. Another significant jet stream for the hydrological characterization is the San Andrés jet [27,31], which is also associated with the trade wind transporting humidity from the Atlantic Ocean to Colombia. Climate change due to natural variability and anthropogenic actions, the PDO, and the NAO [23,35] operate on the interdecadal scale. The QBO (26 months) [14,30] and the ENSO, the main phenomena, operate on the interannual scale. The MJO (40-60 days) [32] and the weekly variability (5-7 days) relate to tropical westerly waves operate on the intra-annual scale. Finally, the tropical dynamics and the country's physiographic are relevant on the diurnal scales [23,26]. The influence of these phenomena acting at different temporal scales shows the level of complexity related to the water resources distribution study and modeling in Colombia.

The ENSO [1] is the global scale climate event that affects the most the Colombian hydro-climatology, from the monthly to interannual scales [25-28,33]. The ENSO is a quasi-periodic phenomenon with an average recurrence of four years that varies between two and seven years [40]. Most of the research regarding Colombian water resources incorporate the ENSO as the most important predictive variable [9-11,28,42]. However, The ENSO forecasting has limitations due to the climate system complexity, which are higher when forecasting goes further the northern hemisphere spring, time of the year known as the spring barrier [39,43] when there is high uncertainty about the evolution of the climate tropical system. Nevertheless, new approaches, more data, and better computational resources continuously improve the ENSO forecast [8], as modeling incorporating Machine Learning [13], information theory [22], and including some climatic variables [20] to overpasses the spring barrier.

1.2 Mathematical modeling

Streamflow forecasting is one of the most relevant tasks in water resources management [9,11,36,44], letting us anticipate the hydrological processes with a certain level of reliability. Physically-based hydrological and mathematical models are two ways to approach monthly streamflow forecasting. In the first case, simulation of the physical processes inside the basin demands high computational resources and amounts of information unavailable most of the time. In the second case, even though physical laws govern the hydrological processes, their non-linear complexities involving many variables justify the use of mathematical models, achieving versatility and including many more basins in the analysis. Hydrological forecasting

studies typically use mathematical or data-driven models, but most do not incorporate adequate measures of uncertainty [18].

Between statistical models, linear regression and autoregressive models have been traditionally used in streamflow prediction. Autoregressive models, which preserve the mean, variance, and autocorrelation structures of the series for a defined number of lags, have been widely used since the beginning of the sixties to the generation of time series [37], and Box and Jenkins [5] improved the mathematical framework since 1970. Fulfilling the model assumptions, they become parsimonious models suitable for time series modeling through a simple mathematical approach like in [1], where an autoregressive model is more accurate than a neural network model.

Mathematical models are functional at incorporating efficiently secondary data like macroclimatic index time series from climate models' outcomes. Most of the streamflow forecasting research in Colombia involves exogenous variables related to the ENSO characterization, and among the mathematical models used are those based on linear regression, neural networks, Multivariate Adaptive Regression Spline (MARS), autoregression, spectral analysis, and entropy [7,36]. It is also relevant to highlight the limitations of the analogous-based models [41], the importance of multivariate series when using dynamic system theory [6], and the usefulness and versatility of Machine Learning [45]. Otherwise, non-parametric forecasting is an appropriate methodology to deal with error propagation when using exogenous variables [34].

Despite the number of methodologies and models used to forecast the water resources in Colombia, limitations in uncertainty modeling justify investigative efforts to improve water resources management. Some common approaches to estimate the uncertainty are probabilistic methods based on historical models' performance [19], the construction of confident bands using independent components [44], the use of autoregressive modeling to generate probabilistic scenarios [16], autoregressive residual representation using the t-student distribution [12], parental methods [2], and bootstrap-based models [1,24,38]. However, it is necessary to improve and develop methodologies to quantify the uncertainty in the forecasting process. It requires considering the geographical zone, the spatial and temporal scales, the month at the beginning of the forecast, the forecasting horizon, and the active climatic phenomena. Stochastic modeling and the estimation of confidence intervals will provide more and better elements for risk analysis, planning, and decision-making in the sectors involved with water resources management.

The ENSO is the most influential climate event in Colombian hydrology, and most of the monthly streamflow forecasting models incorporate it through projected data of the Pacific Ocean temperature. However, climate forecasting also involves high uncertainty, which propagates to streamflow forecasting that usually is not quantified. This research aims to quantify monthly streamflow forecasting uncertainty in Colombia by incorporating measuring station data, secondary time series, modeling, and evaluation using efficiency, parsimony, and physical representability criteria. In the short and medium term, streamflow forecasting is commonly a time series where confidence intervals or another uncertainty measure rarely integrate the forecasting values. The lack of knowledge or information about the models' predictive capacity depreciates the forecasting task and increases the risk of making erroneous decisions. Moreover, the hydrological process's complexity and modeling limitations in streamflow forecasting make it essential to include the expected values and the quantification of its uncertainty.

The advantages of probabilistic approaches are related to the need to highlight the uncertainty of the streamflow forecasting process and the importance of having tools for better management of the resources through risk analysis. The methodology proposed will be applicable in different sectors involving hydrological forecasting, specifically in the Colombian electrical market, improving the understanding of spatial and temporal variability of Colombian water resources. The Colombian electrical market is an economic sector with a high dependence on water resources from rivers that flow to several reservoirs in the country. This condition has led to the instrumentation of several basins in the country, with more than 30 monthly streamflow time series available for studying the spatial and temporal water resource distribution. Furthermore, this worthy data and additional climate data justify the mathematical modeling to have more reliable forecasting results.

2. Data

Hydroelectric projects in Colombia (Figs. 1, 2) represent 66% of the power system capacity. We used the data from 41 monthly streamflow time series related to the rivers that supply the main hydroelectric projects in Colombia, accounting for more than 90% of the inflows. Record length varies depending on the time series, ranging from 1938 to 2023 for the longest and from 1982 to 2023 for the shortest series (Figs. 1, 2, Table 1).

Energy inflows, it is hydrological inflows to hydroelectric power plants, are also known as SIN time series and are usually represented in Colombia in GWh units or as a percentage of the historical mean. XM, as administration of the Colombian energy market, updates the SIN time series daily, and it has been available since 2000.

To have a much longer SIN time series necessary for robust statistical analysis, we reconstruct it using the monthly streamflow data since 1950 by converting the streamflow time series (monthly average in cubic meters per second) to power using the conversion factor (energy power and streamflow ratio) associated to each hydropower plant and then to energy according with the seconds of the month,

$$E = s \sum_{i=1}^{n} f_i \cdot Q_i$$

where E is the total monthly energy inflow in $[GWh]$, Q_i is the monthly streamflow value of the river i associated with the hydroelectric power plant i in $[m^3/s]$, f_i is its corresponding energy power and streamflow ratio in $[GW/m^3/s]$, n is the number of rivers, and s is a constant containing the seconds of the month. Fig. 3 shows the multi-annual monthly means calculated in the 1982-2023 period, where all the streamflow series have records.

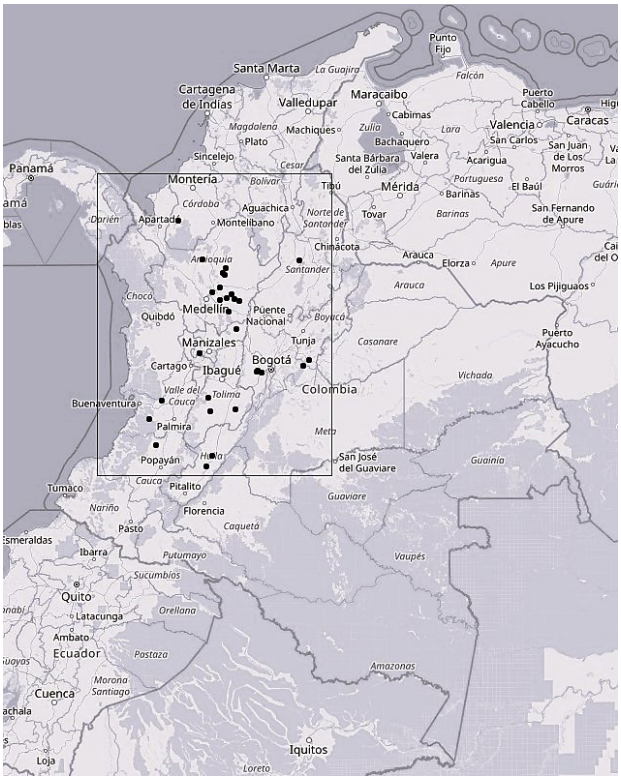

Figure 1. Location of the Colombian hydroelectric projects with installed capacity greater than 20 MW.
Source: The authors.

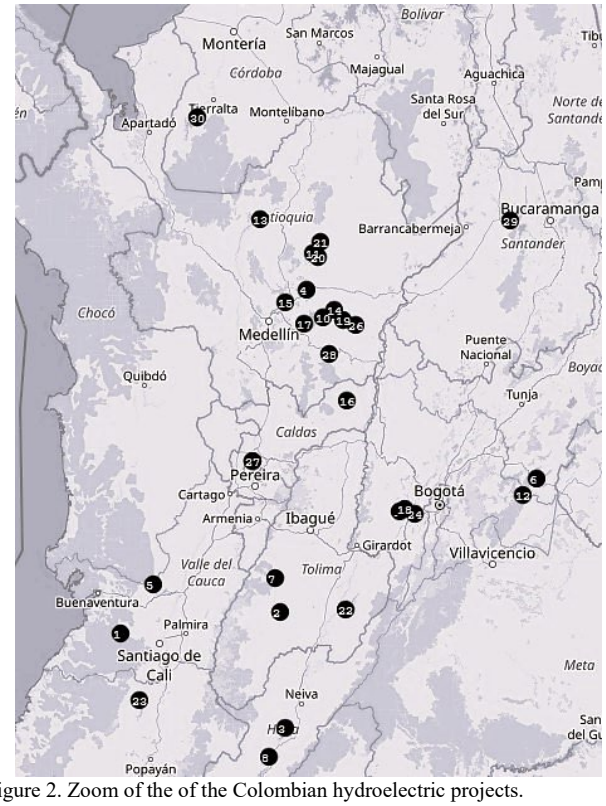

Figure 2. Zoom of the of the Colombian hydroelectric projects.
Source: The authors.

Table 1.
Monthly streamflow time series used.

ID	Hydroelectric Plant	Streamflow time series	Beginning year
1	Albán	Alto Anchicayá	1976
		Digua	1976
2	Amoyá	Amoyá	1974
3	Betania	Betania	1961
4	Carlos Lleras	Carlos Lleras	1972
5	Calima	Calima	1946
6	Chivor	Batá	1978
7	Cucuana	Cucuana	1972
		San Marcos	1972
8	El Quimbo	El Quimbo	1961
9	Esmeralda	Campoalegre	1980
		Chinchina	1961
		Estrella	1972
		San Eugenio	1980
10	Guatapé	Nare	1956
11	Guatrón	Concepción	1955
		Nechí Pajarito Dolores	1955
		Guadalupe	1938
		Tenche	1955
12	Guavio	Guavio	1963
13	Ituango	Ituango	1982
14	Jaguas	San Lorenzo	1956
15	La Tasajera	Grande	1942
16	Miel	Miel	1963
		Guarinó	1980
		Manso	1966
17	Minas	Escuela de Minas	1956
18	Pagua	Bogotá (regulado)	1965
		Blanco	1972
		Chuza	1967
19	Playas	Guatapé	1959
20	Porce 2	Porce 2	1973
		Quebradona	1942
21	Porce 3	Porce 3	1973
22	Prado	Prado	1955
23	Salvajina	Salvajina	1947
24	Salto II	Bogotá (regulado)	1965
25	Samper	Bogotá (regulado)	1965
26	San Carlos	San Carlos	1965
27	San Francisco	San Francisco	1980
28	San Miguel	San Miguel	1974
29	Sogamoso	Sogamoso	1959
30	Urrá	Urrá	1960

Source: The authors.

From January 1950 to December 2023, the SIN inflows time series as a percentage of the average conditions is constructed by dividing each energy month's value by the long-term monthly mean. In months with some missing streamflow data, SIN values in percentage are estimated using available data, and the E value by multiplying it with the corresponding historical monthly mean.

2.1 Standardization

Most of the hydroelectric streamflow time series in Colombia has a clear bimodal annual cycle, with two maximums at April-May and October-November because of the pass of the ZCIT two times in a year through the center of Colombia. Moreover, the east region of Colombian, with two of the largest hydropower plants (Chivor and Guavio), exhibits a unimodal annual cycle with a maximum in June-

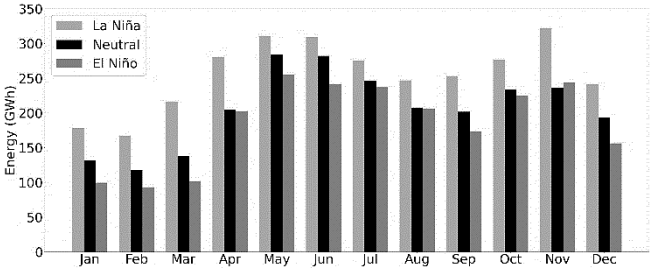

Figure 3. Monthly multi-annual means of the SIN time series for the three ENSO climatological conditions.
Source: The authors.

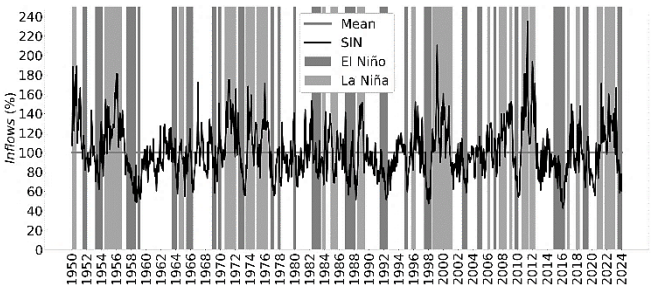

Figure 4. SIN time series in percentage of the average conditions.
Source: The authors.

July, which entails for the total energy inflows one period of high inflows (April-November) and another of low inflows (December-March) which are known as the wet and dry seasons for the Electricity Colombian Market.

To capture not the known intra-annual variability but the inter-annual variability associated with climatic events as the ENSO, the E monthly time series is standardized by removing the multi-annual mean of the month and dividing it by the standard deviation of the month. With this transformation, time series are scaled with zero mean and unit standard deviation without an annual cycle. Fig. 3 shows the reconstructed SIN time series, while Fig. 4 shows the monthly multi-annual means for the three ENSO climatological conditions.

2.2 ENSO time series

The ENSO is the principal climatological phenomenon affecting Colombian water resources availability and should be incorporated in the forecasting models to represent inter-annual climate fluctuations. One of the most common climatic variables used in ENSO monitoring is the Oceanic Niño Index (ONI), calculated by the NOAA (National Oceanic and Atmospheric Administration) agency. The ONI, available since 1950, tracks the running 3-month average sea surface temperatures in the east-central tropical Pacific between 120°-170°W (denominated as El Niño 3.4 region) to estimate the anomaly (deviation from mean conditions). ONI index data is available in https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php.

Along with the ONI, whose data is historical, El Niño 3.4 anomaly forecasts made by different global climate agencies and collected by the Research Institute for Climate and Society (IRI) were used. Every month, the IRI publishes forecasts for the next nine months of about 25 agencies, 16 of which use dynamic climate models and nine statistical models. Historical data on this forecasting activity has been available since 2002 in https://iri.columbia.edu/our-expertise/climate/forecasts/enso/current/?enso_tab=enso-sst_table.

3. Methodology

Different mathematical models commonly used in hydrological time series forecasting and widely employed in Colombia were implemented, as well as the use of the Root Mean Square Error (RMSE) as the error metric for the models' performance analysis. We evaluated simple and multiple linear regression, simple and multiple linear regression using robust techniques, analogous series, non-parametric regression, neural networks, decision trees, and autoregressive models. The SIN time series length is 888 months long (1950-2023), and it was divided into calibration (1950-2012) and test (2013-2023) periods, representing 85% and 15%. We sensitized and adjusted the model parameters in the calibration period and evaluated the model in the test period, estimating the error variance to posterior uncertainty analysis.

Analogous series is the simplest model analyzed and commonly used in Colombia in medium and long-term energetic analysis. The first step is to compare the last τ (model parameter) observed months with the historical record to determine the most similar period with the present condition (last τ months). After that, the forecast is the observed data following that period. We also incorporated the ONI time series to determine the most similar historical period, accounting for the streamflow and the climate signal, and using the Euclidean distance as a similarity criterion.

Robust techniques applied to simple and multiple linear regression involve discarding outliers or atypical data according to different criteria: data points with high residuals or far away according to the Euclidean distance (kNN) or the Mahalanobis distance. Another technique evaluated consists of the estimation of the parameters of the regression through a modification of the covariance matrix, employing statistical measures for the correlation and standard deviation less sensitive to outliers like the Kendall and Spearman for the correlation and the MAD for the standard deviation.

We also fitted autoregressive models to the data. These models are also linear, in which the number and values of the parameters are a function of the linear structure dependence of the data. An autoregressive model $AR(p)$ is defined as:

$$Y_t = \beta + \phi_1 \cdot Y_{t-1} + \cdots + \phi_p \cdot Y_{t-p} + \epsilon_t$$

where β, ϕ_1, $\cdots$, ϕ_p are the model parameters and ϵ_t is assumed to be a white noise process. An autoregressive model with exogenous variable $ARX(p)$ is defined as:

$$Y_t = \beta + \phi_1 \cdot Y_{t-1} + \cdots + \phi_p \cdot Y_{t-p} + \sum_{j=1}^{m} \gamma_j \cdot X_{t,j} + \epsilon_t$$

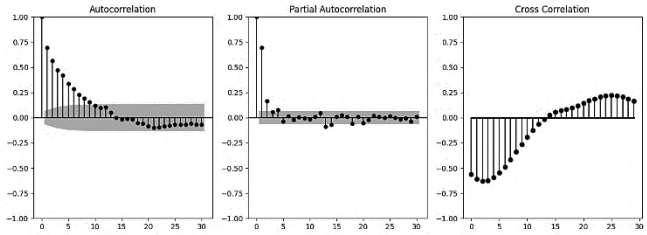

Figure 5. Time series' dependence structure.
Source: The authors.

where $X_{t,j}$ are the exogenous regressors. Model parameters were estimated using Conditional Maximum Likelihood.

Fig. 5 shows the correlation structure of the data. The autocorrelation diagram exhibits a quick decay, with correlations statistically significant observed until ten months. However, the partial autocorrelation diagram suggests that only the first lag represents almost all the linear dependence structure of the series. The pattern observed in the autocorrelation and partial autocorrelation diagrams is typical of AR models, and due to the high correlation between the inflows and the ONI time series, we also evaluated autoregressive models with exogenous variables (ARX). Although analyzed, different autoregressive models like AR with climatically conditioned parameters or SARIMA models using non-standardized data led to a good forecasting performance (results not shown), but AR1 and ARX1 models are superior according to parsimony criteria.

Forecast confidence interval were estimated with two different approaches. In the first one, we used the error variance estimated in the test period, calculated after the Box-Cox transformation used to stabilize the variance over the residuals. Otherwise, in the second one and using the empirical function of the residuals, we used Bootstrap to generate 1000 series of residuals to simulate 1000 possible outputs, re-calibrate the model, and obtain a distribution of the parameters. This methodology starts estimating the autoregressive coefficients β and ϕ in the equation $Y_t = \beta + \phi \cdot Y_{t-1} + \epsilon_t$, the computation of the residuals ϵ_t, and the definition of its empirical distribution function of the centered residuals. If the residuals are not centered, then $\tilde{e}_t = \epsilon_t - \hat{e}$ and $\hat{e} = \frac{1}{n-p}\sum_{t=p+1}^{n} \hat{\epsilon}_t$ equations are used.

Using Bootstrap to draw i.i.d. ϵ_t^* resamples from the empirical distribution of the residuals $\hat{F}_{\tilde{\epsilon}}$ leads to one forecast for each residual resampling defining the recursion $Y_t^* = \beta + \phi \cdot Y_{t-1}^* + \epsilon_t^*$. It allows computing B future observations where the mean represents the future expected value, it is the forecast, and the 5% and 95% percentiles define the confidence band.

Continuing with more sophisticated models capable of capturing the nonlinearities present in the time series, we calibrated non-parametric regressions, neural networks, and decision trees models. In all the cases and according to the train and test periods, optimal parameters estimation using the train data consisted of a bias and variance trade-off. For the case of non-parametric regression, we evaluated different kernels and bandwidth values, in neural networks we considered different configurations varying the number of neurons and layers, while for the decision trees the deep was the hyper-parameter calibrated.

Table 2.
Algorithm.

1:	Estimate β and ϕ in $Y_t = \beta + \phi \cdot Y_{t-1} + \epsilon_t$ using Conditional Maximum Likelihood Estimation
2:	Compute the residuals ϵ_t
3:	Define the empirical distribution function of the centered residuals $\hat{F}_{\tilde{\epsilon}}$
4:	if residuals ϵ_t are centered:
5:	$\tilde{e}_t = \epsilon_t$
6:	if not:
7:	$\tilde{e}_t = \epsilon_t - \hat{e}$ and $\hat{e} = \frac{1}{n-p}\sum_{t=p+1}^{n} \hat{\epsilon}_t$
8:	end if
9:	for $i = 1\ to\ B$:
10:	Draw a i.i.d. ϵ_t^* resample from $\hat{F}_{\tilde{\epsilon}}$
11:	Compute $Y_{ti}^* = \beta + \phi \cdot Y_{t-1}^* + \epsilon_t^*$
12:	end for
13:	$Y_{t\_forecast} = mean(Y_{ti}^*)$
14:	$Y_{t_5\%} = percentile(Y_{ti}^*, 0.05)$
15:	$Y_{t_{95}\%} = percentile(Y_{ti}^*, 0.95)$

Source: The authors.

In all the models described above, the target variable is the standardized energy inflows, and the predictive variables are the past standardized energy inflows and the past and forecast ONI time series. We analyzed different numbers of lags for the predictive variables, especially for the non-linear models, and calibrated monthly models, which is 12 models (one per month), trying to obtain a better forecasting performance.

4. Results

As expected, all the models evaluated (Table 2) presented better results than the simplest forecasting method, the monthly multi-annual mean (RMSE = 0.26 and RMSE = 0.25 for the calibration and test periods). Additionally, the error metric in the test period is higher using non-linear models than simpler linear models. It is due to a high linear dependence between the SIN's time series with its lags and with the exogenous variable, rather than a non-linear dependence, or possibly due to the lack of other variables that adequately represent the non-linear dependence structure. Besides, the monthly models that consider an independent model for each month do not improve the forecasting performance despite increasing the number of parameters 12 times.

4.1 Autoregressive model

Due to the linear dependence structure of the data and the fact that it is a time series, the ARX1 is the best of all the models studied. This model presents better adjustment, has few parameters, and allows simulating the stochastic nature of the SIN time series, preserving the historical dependence structure and incorporating the climate signal through the exogenous variable. Furthermore, the ARX1 model is more suitable since it retains temporal and climate coherence when forecasting ahead of one month.

Table 3.
Results.

Model/RMSE	Standard Model		Monthly Model	
	Calibration	Test	Calibration	Test
Analogous		0.90		
Analogous ONI		0.91		
Simple linear regression	0.71	0.81	0.69	0.82
Multiple linear regression	0.68	0.76	0.64	0.78
Deep data residuals	0.68	0.75	0.65	0.76
Deep data residuals Knn	0.68	0.76	0.65	0.76
Deep data residuals Mahal.	0.68	0.75	0.65	0.77
Robust - Kendall	0.69	0.76	0.66	0.76
Robust - Spearman	0.68	0.76	0.64	0.77
Robust - MAD	0.68	0.76	0.70	0.77
Robust - Kendall-MAD	0.70	0.76	0.69	0.77
Robust - Spearman-MAD	0.68	0.76	0.71	0.77
AR1	0.71	0.81		
ARX1	0.68	0.75		
Non-parametric regression	0.71	0.81	0.63	0.82
Mult. non parametric reg.	1.00	1.02	0.60	0.77
Decision Trees	0.64	0.82	0.52	0.80
Neural Networks	0.70	0.75	0.64	0.77

*CLIMATOLOGICAL RMSE: 0.99 y 1.02
Source: The authors.

Fig. 6 shows the forecasting results in the test period for a horizon of one month. After fitting the ARX1 model with the standardized SIN time series, as described in section 3, the residuals distribution obtained follows a normal distribution (Fig. 7). Similarly, with the Box-Cox transformation, normality in residuals is met (Fig. 8), but the new model is adjusted after two transformations, increasing the complexity of the model. Moreover, the residuals met the non-autocorrelation hypothesis with and without the Box-Cox transformation (Fig. 9).

Neither the Gaussian residual distribution nor data transformation is necessary with the bootstrapping approach, in which the simulations start from the empirical distribution of the residuals. Fig. 6 shows a comparison of the simulation in the test periods. Both forecast values and confidence intervals are similar, making the bootstrapping method more suitable as it is less restrictive. Although not perfectly, the forecast follows the tendency and variability of the observed data, and more importantly, the forecast uncertainty represented by the shaded area adequately contains the observed data for the level of significance assumed (5%).

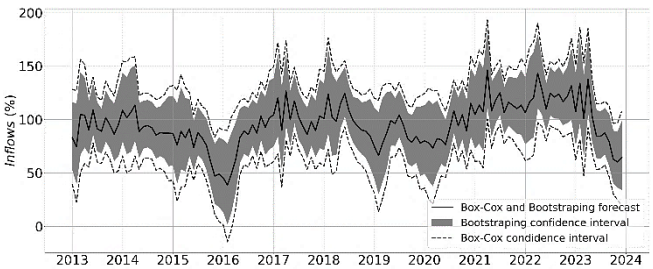

Figure 6. ARX1 forecasting results in the test period.
Source: The authors.

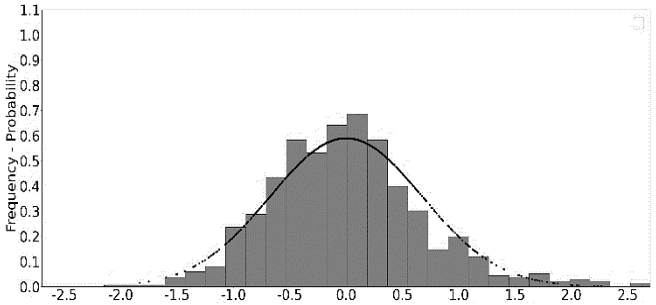

Figure 7. Residuals distribution of the ARX1 model.
Source: The authors.

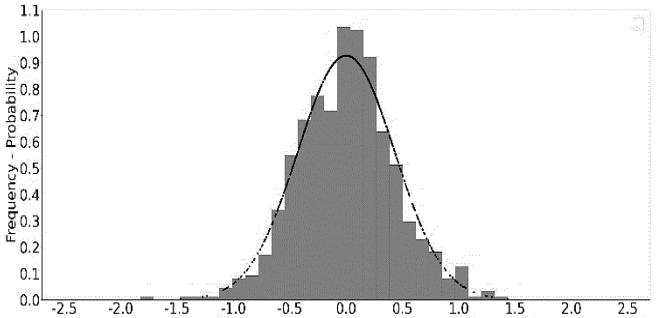

Figure 8. Residuals distribution of the ARX1 model after the Box-Cox transformation.
Source: The authors.

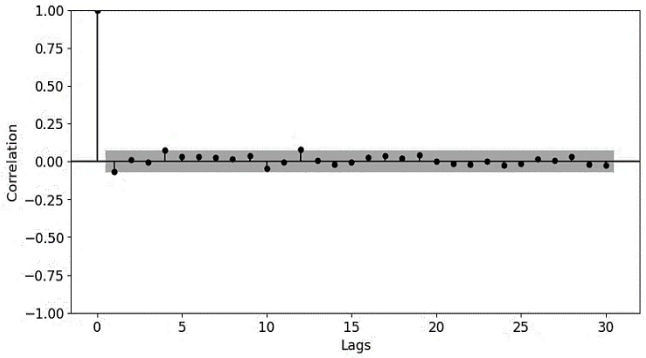

Figure 9. Residuals partial autocorrelation of the ARX1 model after the Box-Cox transformation.
Source: The authors.

Figs. 10 and 11 are related to the uncertainty level of the forecast process. Fig. 8 compares the AR1 and ARX1 models of how the error variance increases with the forecast horizon. Naturally, in both cases, the error variance increases with the time horizon, and it is evident how the exogenous variable influences this increase to be less pronounced, maintaining the error variance significantly lower concerning the AR1 model (without exogenous variables). Moreover, Fig. 9 shows the level of uncertainty according to the target month, estimated as the quotient between the uncertainty length and the multi-annual average value of the forecast month.

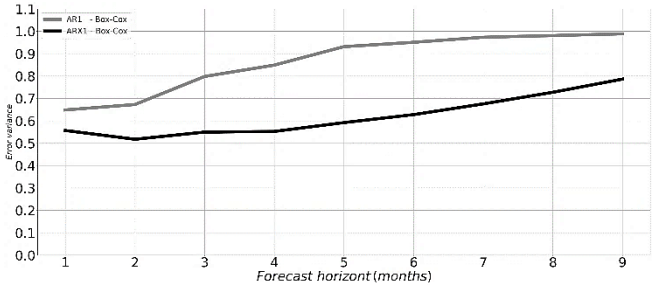

Figure 10. Error variance as a function of the forecast horizon.
Source: The authors.

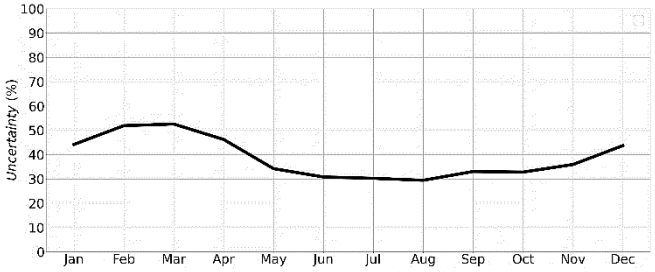

Figure 11. Forecast uncertainty for each month of the year.
Source: The authors.

4.1 El Niño 2023-2024

In May 2023, began one of the strongest El Niño events in recorded history according to the NOAA ONI index. Only five El Niño events have reached two degrees temperature anomaly umbral, and between them is the El Niño 2023-2024 with high global impacts. Particularly in Colombia, since May 2023, a prolonged water availability deficit was observed.

To visualize the model performance in the 2023-2024 event, Fig. 12 shows the one-month ahead forecast for the April-December 2023 period, while Fig. 13 shows the forecasting exercise for the April-December 2023 period using the March 2023 observed inflows and the 9 ONI projected values. The model follows the low streamflow tendency because of the serial dependence and the expected continuity of the El Niño event, as suggested by the climate agencies and incorporated in the exogenous values. As expected in the 9-month forecasting exercise, confidence bands increase with the forecasting horizon, reflecting both the natural uncertainty associated with several months ahead forecast and the duration and intensity of the event.

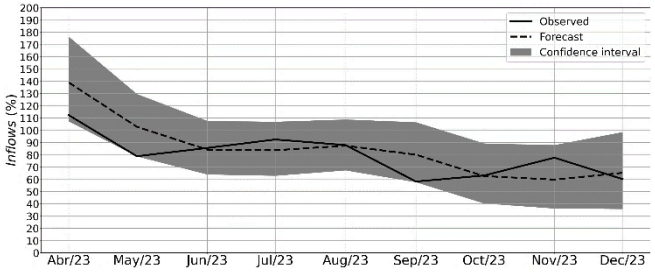

Figure 12. One-month ahead April-December 2023 forecast.
Source: The authors.

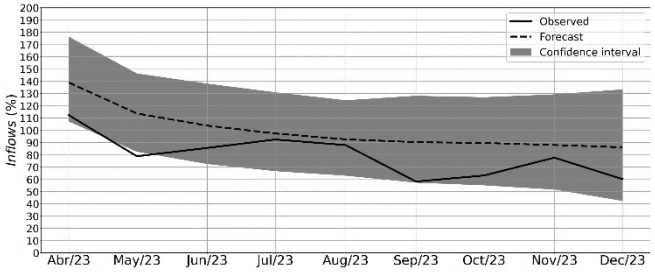

Figure 13. April-December 2023 forecast using the March 2023 observed inflows and the 9 ONI projected values.
Source: The authors.

4.2 La Niña 2024-2025

According with the official CPC ENSO probability forecast, based on a consensus of CPC and IRI forecasters available in https://iri.columbia.edu/our-expertise/climate/forecasts/enso/current/ and published on March 14, El Niño 2023-2024 was expected to finish between April and May, followed by two or three months of neutral conditions and of a La Niña 2024-2025. Fig. 14 shows the December 2023 - August 2024 forecast using the November 2023 observed inflows and the 9 ONI projected values. Consistent with the El Niño event, the forecast suggests system inflow below the historical mean until April. However, the observed inflows have been below forecasted values due to the limitation of the model by not incorporating other variables that could explain the intense system water deficit, such as those related to the Amazon basin and Atlantic Ocean activity. In addition to the above, ENSO patterns and their influence on Colombia may be changing because of the temperature records of the last 2 years, making it more difficult the forecasting exercise. However, the reliability bands capture the variability of the series according to the selected significance level.

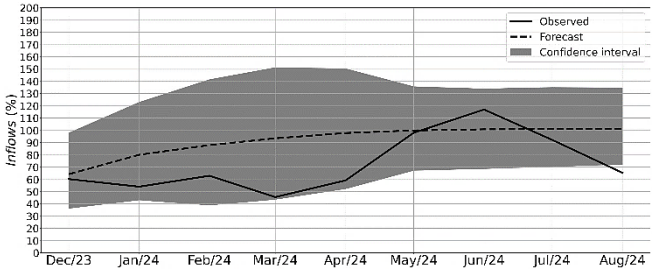

Figure 14. December 2023 - August 2024 forecast using the November 2023 observed inflows and the 9 ONI projected values.
Source: The authors.

5. Conclusions

We studied the Colombian monthly streamflow predictability through the evaluation of different forecasting mathematical models, the use of robust and non-parametric techniques, and the incorporation of exogenous data, trying to extract the maximum amount of information only from the streamflow time series and estimating reliability bands using probabilistic approaches, providing tools for a better water

resources management in the country.

After analyzing the correlation structure of the data employing techniques such as the partial autocorrelation function, monthly energy inflow forecasting through different mathematical models was evaluated. The autoregressive model with exogenous variables presented the best results due to the inflows' high dependence on the previous month, the linear dependence structure of the series, and the climate information contained in the ONI series. Besides, confidence interval estimation using Bootstrap from the empirical distribution of the residuals allows for a reliable estimate of uncertainty without assuming any residual distribution. The stochastic structure of the model finally adopted permits the estimation of multiple plausible scenarios when forecasting, which is an essential condition in the study of uncertainty and risk analysis. The methodology for estimating confidence bands allows us to represent the Colombian hydro-climatology complexity related to its geography, physiography, and hydro-climatological season.

Three aspects determined the viability of incorporating exogenous variables in the forecasting process. Firstly, a statistical analysis of the energy inflow series allowed us to conclude about a significant dependence on the ONI exogenous variable. Secondly, the exogenous data, with predicted ONI values for a horizon of 9 months, significantly impact the energy inflow forecast, especially when predictions are made for horizons longer than three months. Finally, using parsimony criteria, despite having more parameters, the exogenous variable incorporation leads to an increase in the performance or predictability of the model.

Machine learning models implemented did not show better performance than autoregressive models. It could be because the time series does not exhibit nonlinearities, it is much less than the linear dependence, or there is not enough available exogenous data to incorporate in the models. In summary, autoregressive models are more appropriate to capture the time series dynamics.

The results obtained in this research will contribute to the understanding of the hydro-climatological processes in Colombia via mathematical modeling, as the forecasting and reliability bands estimation methodology will provide tools for decision-making in sectors dealing with the uncertainty associated with the distribution of the water resources.

References

[1] Al-Saati, N.H., Omran, I.I., Salman, A.A., Al-Saati, Z., and Hashim, K.S., Statistical modeling of monthly streamflow using time series and artificial neural network models: Hindiya Barrage as a case study. Water Practice & Technology 16(2), pp. 681–691, 2021. DOI: https://doi.org/10.2166/wpt.2021.012.

[2] Bezerra, B., Veiga, A., Barroso, L.A., and Pereira, M., Stochastic long-term hydrothermal scheduling with parameter uncertainty in autoregressive streamflow models. IEEE Transactions on power systems, 32(2), 2017. DOI: https://doi.org/10.119/TPWRS.2016.2572722.

[3] Beyaztas, U., Arikan, B.B., Beyaztas, B.H., and Kahya, E., Construction of prediction intervals for palmer drought severity index using bootstrap. Journal of Hydrology 559, pp. 461-470, 2018. DOI: https://doi.org/10.1016/j.jhydrol.2018.02.021.

[4] Bjerknes, J., Atmospheric teleconnections from the equatorial pacific. Monthly Weather Review, 97(3), pp. 163-172, 1969. DOI: https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2

[5] Box, G.E.P., and Jenkins, G.M., Time series analysis: forecasting and control. Holden-Day, 1976. ISBN: 0816211043, 9780816211043.

[6] Cao, L., Mees, A., and Judd, K., Dynamics from multivariate time series. Physica D: Nonlinear Phenomena, 121(1-2), pp. 75–88, 1998. DOI: https://doi.org/10.1016/S0167-2789(98)00151-1.

[7] Carvajal, L.F., Salazar, J.E., Mesa, O.J., y Poveda, G., Predicción hidrológica en Colombia mediante análisis espectral singular y máxima entropía. Ingeniería Hidráulica en México, XIII, (1), pp. 7–16, 1998.

[8] Chen, D., and Cane, M.A., El Niño prediction and predictability. Journal of Computational Physics, 227(7), pp. 3625-3640, 2008. DOI: https://doi.org/10.1016/j.jcp.2007.05.014.

[9] Córdoba, S., Palomino, R., Raquel, Gámiz, S., Castro, Y., and Esteban, M., Influence of tropical Pacific SST on seasonal precipitation in Colombia: prediction using El Niño and El Niño Modoki. Climate Dynamics, 44(5-6), pp. 1293–1310, 2015. DOI: https://doi.org/10.1007/s00382-014-2232-3.

[10] Córdoba, S., Palomino, R., Raquel, Gámiz, S., Castro, Y., and Esteban, M., Seasonal streamflow prediction in Colombia using atmospheric and oceanic patterns. Journal of Hydrology, 538, pp. 1–12, 2016. DOI: https://doi.org/10.1016/j.jhydrol.2016.04.003.

[11] Dracup, J.A., and Gutie, F., An analysis of the feasibility of long-range streamflow forecasting for Colombia using El Niño–Southern Oscillation indicators. Journal of Hydrology, 246, pp. 181–196, 2001. DOI: https://doi.org/10.1016/S0022-1694(01)00373-0.

[12] Favereau, M., Lorca, A., Negrete-Pincetic, M., and Vicuña, S., Robust streamflow forecasting: a student's t-mixture vector autoregressive model. Stochastic Environmental Research and Risk Assessment 36, pp. 3979–3995, 2022. DOI: https://doi.org/10.1007/s00477-022-02241-y.

[13] Ham, Y.G., Kim, J.H., and Luo, J.J., Deep learning for multi-year ENSO forecasts. Nature, 573(7775), pp. 568–572, 2019. DOI: https://doi.org/10.1038/s41586-019-1559-7.

[14] Holton, J.R., and Hakim, G.J., An introduction to dynamic meteorology. 5th Ed., 2012. DOI: https://doi.org/10.1016/C2009-0-63394-8.

[15] Jaramillo, A., y Chaves, B., Distribución de la precipitación en Colombia analizada mediante conglomeración estadística. Cenicafé, 51(2), pp. 102–113, 2000.

[16] Kelman, J., Vieira, A., and Rodriguez, J.E., El Niño influence on streamflow forecasting. Stochastic Environmental Research and Risk Assessment, 14, pp. 123–138, 2000. DOI: https://doi.org/10.1007/PL00009776.

[17] Laing, A.G., and Fritsch, J.M., Mesoscale convective complexes in Africa. Monthly Weather Review, 121(8), pp. 2254–2263, 1993. DOI: https://doi.org/10.1175/1520-0493(1993)121<2254:MCCIA>2.0.CO;2.

[18] Yan, L., Feng, J., Hang, T., and Zhu, Y., Flow interval prediction based on deep residual network and lower and upper boundary estimation method. Applied Soft Computing Journal, 104, art. 107228, 2021. DOI: https://doi.org/10.1016/j.asoc.2021.107228.

[19] L'Heureux, M.L., Tippett, M.K., Takahashi, K., Barnston, A.G., Becker, E.J., Bell, G.D., Di Liberto, T.E., Gottschalck, J., Halpert, M.S., Hu, Z.Z., Johnson, N.C., Xue, Y., and Wang, W., Strength outlooks for the El Niño-Southern oscillation. Weather and Forecasting, 34(1), pp. 165–175, 2019. DOI: https://doi.org/10.1175/WAF-D-18-0126.1.

[20] McPhaden, M.J., Tropical Pacific Ocean heat content variations and ENSO persistence barriers. Geophysical Research Letters, 30(9), pp. 1995–1998, 2003. DOI: https://doi.org/10.1029/2003GL016872.

[21] Mejía, J.F., Mesa, O.J., Poveda, G., Vélez, J.I., Hoyos, C., Mantilla, R., Barco, J., Cuartas, L.A., Montoya, M., y Botero, B., Distribución espacial y ciclos anual y semianual de la precipitación en Colombia. DYNA, 127, pp. 7-26, 1999.

[22] Meng, J., Fan, J., Ludescher, J., Agarwal, A., Chen, X., Bunde, A., Kurths, J., and Schellnhuber, H.J., Complexity-based approach for El Niño magnitude forecasting before the spring predictability barrier. Proceedings of the National Academy of Sciences of the United States of America, 117(1), pp. 177–183, 2020. DOI: https://doi.org/10.1073/pnas.1917007117.

[23] Mesa, O.J., Poveda, G., y Carvajal, L.F., Introducción al clima de Colombia. Universidad Nacional de Colombia, Medellín, primera edición, 1997. ISBN: 9586281442, 9789586281447.
[24] Palm, B.G., Bayer, F.M., and Cintra, R.J., Prediction intervals in the beta autoregressive moving model. Communications in statistics – simulation and computation. 52(8), pp. 3635–3656, 2023. DOI: https://doi.org/10.1080/03610918.2021.1943440.
[25] Poveda, G., Retroalimentación dinámica entre El Niño Oscilación del Sur y la hidrología de Colombia. Tesis de Doctorado, Universidad Nacional de Colombia, sede Medellín, 1998.
[26] Poveda, G., La hidroclimatología de Colombia: una síntesis desde la escala interdecadal hasta la escala diurna. Revista Academia Colombiana de Ciencias, 28(10), pp. 201–222, 2004. DOI: https://doi.org/10.18257/raccefyn.28(107).2004.1991.
[27] Poveda, G., Gil, M.M., and Quiceno, N., El ciclo anual de la hidrología de Colombia en relación con el ENSO y la NAO. Bulletin de l'Institut Français d'Études Andines, 27(3), pp. 721–731, 1998.
[28] Poveda, G., Jaramillo, A., Gil, M.M., Quiceno, N., and Mantilla, R., Seasonality in ENSO-related precipitation, river discharges, soil moisture, and vegetation index in Colombia. Water Resources Research, 37(8), pp. 2169–2178, 2001. DOI: https://doi.org/10.1029/2000WR900395.
[29] Poveda, G., and Mesa, O.J., On the existence of Lloró (the rainiest locality on Earth), Enhanced ocean-land-atmosphere interaction by a low-level jet. Geophysical Research Letters 27(11), pp. 1675-1678, 2000. DOI: https://doi.org/10.1029/1999GL006091.
[30] Poveda, G., and Mesa, O.J., Feedbacks between hydrological processes in tropical South America and large-scale ocean-atmospheric phenomena. Journal of Climate, 10(10), pp. 2690–2702, 1997. DOI: https://doi.org/10.1175/1520-0442(1997)010<2690:FBHPIT>2.0.CO;2.
[31] Poveda, G., y Mesa, O.J., La corriente de chorro superficial del oeste ("del Chocó") y otras dos corrientes de chorro en Colombia climatología y variabilidad durante las fases del ENSO. Ciencias de la Tierra, 23(89), pp. 517–528, 1999. DOI: https://doi.org/10.18257/raccefyn.23(89).1999.2848.
[32] Poveda, G., Mesa, O.J., Salazar, L.F., Arias, P.A., Moreno, H.A., Vieira, S.C., Agudelo, P.A., Toro, V.G., and Álvarez, J.F., The diurnal cycle of precipitation in the tropical Andes of Colombia. Monthly Weather Review, 133(1), pp. 228–240, 2005. DOI: https://doi.org/10.1175/MWR-2853.1.
[33] Poveda, G., Vélez, J.I., and Mesa, O.J., Atlas Hidrológico de Colombia. Universidad Nacional de Colombia, Medellín, 2000.
[34] Rodríguez, N., y Siado, P., Un pronóstico paramétrico de la inflación colombiana. Revista Colombiana de Estadística, 26(2), pp. 89-128, 2003.
[35] Rogers, J.C., The association between the North Atlantic Oscillation and the Southern Oscillation in the Northern Hemisphere. Monthly Weather Review, 112(10), pp. 1999-2015, 1984. DOI: https://doi.org/10.1175/1520-0493(1984)112<1999:TABTNA>2.0.CO;2.
[36] Rojo, J.D., Carvajal, L.F., and Velásquez, J.D., Streamflow prediction using a forecast combining system. IEEE Latin America Transactions, 13(4), pp. 1035–1040, 2015. DOI:10.1109/TLA.2015.7106354.
[37] Thomas, H., and Fiering, M., Mathematical synthesis of streamflow sequences for the analysis of river basins by simulations. Design of Water Resource Systems, Edited by Mass et al., Harvard University Press, Cambridge, pp. 459-493, 1962. DOI: https://doi.org/10.4159/harvard.9780674421042.c15.
[38] Thombs, L.A., and Schucany, W.R., Bootstrap prediction intervals for autoregression. Journal of the American Statistical Association, 85(410), pp. 486-492, 1990. DOI: https://doi.org/10.2307/2289788.
[39] Torrence, C., and Webster, P.J., The annual cycle of persistence in the El Niño/Southern Oscillation. Quarterly Journal of the Royal Meteorological Society 124(550), pp. 1985–2004, 1998. DOI: https://doi.org/10.1002/qj.49712455010.
[40] Trenberth, K.E., General characteristics of El Niño-Southern Oscillation. Teleconnection Linking Worldwide Climate Anomalies. Cambridge University Press, New York, 1991.
[41] Van Den Dool, H.M., Searching for analogues, how long must we wait? Tellus A: Dynamic Meteorology and Oceanography, 46(3), pp. 314-324, 1994. DOI: https://doi.org/10.1034/j.1600-0870.1994.t01-2-00006.x.
[42] Waylen, P., and Poveda, G., El Nino-Southern Oscillation and aspects of western South American hydro-climatology. Hydrological Processes, 16(6), pp. 1247–1260, 2002. DOI: https://doi.org/10.1002/hyp.1060.
[43] Webster, P.J., The annual cycle and the predictability of the tropical coupled Ocean-Atmosphere system. Meteorology and Atmospheric Physics, 56, pp. 33–55, 1995. DOI: https://doi.org/10.1007/BF01022520.
[44] Westra, S., Sharma, A., Brown, C., and Lall, U., Multivariate streamflow forecasting using independent component analysis. Water Resources Research, 44(2), pp. 1–11, 2008. DOI: https://doi.org/10.1029/2007WR006104.
[45] Yang, T., Asanjan, A.A., Welles, E., Gao, X., Sorooshian, S., and Liu, X. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research. Journal of the American Water Resources Association, (53), pp. 2786–2812, 2017. DOI: https://doi.org/10.1002/2017WR020482.

A.F. Hurtado-Montoya, received the BSc. Eng. in Civil Engineering in 2006 and the MSc. in Engineering (Water Resources) in 1999, both from the Universidad Nacional de Colombia Medellín Campus, and the MSc in Applied Mathematics from the Universidad Eafit Medellin, Colombia. He has a sepecialization in Turbomachenary. As a researcher in the field of hydroclimatology he has published articles related to the spatio-temporal variability of precipitation and climate change in Colombia. He has taught mathematics, hydrology, design of hydraulic structures and general and transient hydraulics. He has participated in the identification, structuring and design of various hydroelectric plants. Currently, he is an energy transactional professional at Isagen. His research interests include time series forecasting and hydroclimatology,
ORCID: 0000-0003-4928-3227.

N.A. Moreno-Reyes, is BSc. in Mathematics, Dr. in statistics conferred by the State University of Campinas. Currently, he serves as a professor at EAFIT University in Medellín, Colombia. His research focus lies within the field of probability and statistics, with a current emphasis on Bayesian inference, stochastic processes, and statistical applications.
ORCID: 0000-0002-9371-1615.