




Mathematical model for streamflow prediction in an Andean basin by Pearson correlation with ocean temperature

Modelo matemático para la predicción de caudales en una cuenca andina mediante correlación de Pearson con temperaturas oceánicas

Gustavo Adolfo Carrillo-Soto^a, Nelson Javier Cely-Calixto^b, Carlos Alexis Bonilla-Granados^c

 a. Doctorado en Hidrología, gustavocarrillo@ufps.edu.co, Grupo de Investigación en Hidrología y Recursos Hídricos - HYDROS, Universidad Francisco de Paula Santander, Cúcuta, Colombia.

 b. Maestría en Obras Hidráulicas, nelsonjaviercc@ufps.edu.co, Grupo de Investigación en Hidrología y Recursos Hídricos - HYDROS, Universidad Francisco de Paula Santander, Cúcuta, Colombia

 c. Maestría en Tecnologías para el manejo de Aguas y Residuos, carlos.bonilla@unipamplona.edu.co, Grupo de investigación Etenoha, Universidad de Pamplona, Pamplona, Colombia.

Recibido: Junio 1 de 2021 Aceptado: Octubre 8 de 2021

Forma de citar: G.A. Carrillo-Soto, N.J. Cely-Calixto, C.A. Bonilla-Granados, “Mathematical model for streamflow prediction in an Andean basin by Pearson correlation with ocean temperature”, *Mundo Fesc* vol. 11, S4, pp. 223-229, 2021

Abstract:

Streamflow prediction constitutes a fundamental tool in water resources decision-making and risk management actions. The possibility of implementing a mathematical model for the prediction of streamflow in an Andean basin was studied, identifying, through Pearson's correlation coefficient, the best correlations between the time series of mean monthly streamflow (Q_m) and the sea surface temperature (SST), considering up to 11 lags. The monthly SST data were obtained from the MODIS-NASA sensor processed on the Ocean Color platform, selecting cells of 2° longitude and 4° latitude to cover the strip +180° lon / -180° lon, from -20° lat / +20° lat, analyzing a total of 3600 cells. The water resource was characterized by the Q_m time series of the La Donjuana station on the Pamplonita river (Norte de Santander, Colombia). The time window studied was from July 2002 to December 2015 (162 months). Linear models were built for each month by selecting the lag that produced the maximum correlation and verifying values of the p statistic, which were much lower than 0.001. The models were evaluated using the mean square error and the Nash-Sutcliffe efficiency, differentiating Normal years, El Niño years, and La Niña years. Satisfactory results were found for the prediction in La Niña years (wet) with lags greater than three months. The investigation is expected to be extended to consider a greater range of latitudes and to consider other Andean basins.

Keywords: Average Monthly Flows, Ocean temperatures, Predictive Mathematical Models, Water Resources.

Autor para correspondencia:

*Correo electrónico: gustavocarrillo@ufps.edu.co



Resumen:

La predicción de caudales constituye una herramienta fundamental en la toma de decisiones para el manejo del recurso hídrico y en el manejo de la gestión del riesgo. Se estudió la posibilidad de implementar un modelo matemático para la predicción de caudales en una cuenca andina, identificando mediante el coeficiente de correlación de Pearson, las mejores correlaciones entre las series de tiempo de caudales medios mensuales y de la temperatura superficial del océano (SST), considerando hasta 11 rezagos. Los datos de la SST mensual se obtuvieron del sensor MODIS-NASA procesados en la plataforma Ocean Color, seleccionando celdas de 2° longitud y 4° latitud para cubrir la franja +180°lon / -180°lon, desde -20°lat / +20°lat, analizando un total de 3600 celdas. El recurso hídrico se caracterizó mediante los Caudales Medios Mensuales (Qmm) de la estación La Donjuana sobre el río Pamplonita (Norte de Santander, Colombia). La ventana temporal estudiada fue de julio 2002 a diciembre 2015 (162 meses). Se construyeron modelos lineales para cada mes seleccionado el rezago que producía la máxima correlación y verificando valores del estadístico p, los cuales fueron muy inferiores a 0.001. Los modelos se evaluaron mediante el error medio cuadrático y la eficiencia de Nash-Sutcliffe, diferenciando años Normales, años El Niño y años La Niña. Resultados satisfactorios fueron encontrados para la predicción en años Niña (húmedos) con rezagos superiores a tres meses. Se espera extender la investigación para considerar una mayor franja de latitudes y considerar otras cuencas andinas.

Palabras clave: Caudales Medios Mensuales, Modelos Matemático Predictivo, Recurso Hídrico, Temperatura Oceánica.

Introduction

Within applied mathematics, the use of predictive models is very useful. The prediction of river flows is a topic widely addressed by different mathematical models such as fuzzy neural networks [1], artificial neural networks [2, 3], wavelet regression [4], and more traditional methods such as geographic regionalization [5]. Additionally, models based purely on statistical correlation are being used given the increasing availability of global coverage satellite data [6, 7, 8, 9]. However, very few studies cover geographic regions of Central and South America, even more so in specific ecosystems such as the Andean basins. This study explores the possibility of constructing mathematical models that allow predicting the average monthly flow in an Andean basin by correlating them with the sea surface temperature, at a specific site, recorded in previous months. This establishes a lag condition that allows to have a predictive model. Clearly, providing a first approximation on the magnitude of average monthly flows establishes an important element for decision makers about water resources and the possibility

of implementing actions related to risk mitigation in the event of fluvial events [10]. Events of fluvial origin that threaten human lives and physical infrastructure are very frequent in Andean basins [11].

Methods

Data

The independent variable in the model corresponds to the SST obtained from the Ocean Color web platform (<http://oceancolor.gsfc.nasa.gov>), the information of which is freely available to the public [12, 13]. The strip of ocean considered corresponds to that located between latitudes $\pm 20^\circ$ latitude, covering the entire circumference of the earth. This strip is divided into cells of 2° longitude and 4° latitude, generating 3600 cells in total. The dependent variable corresponds to the average monthly flow recorded at the La Donjuana water-logging station operated by the Institute of Hydrology, Meteorology and Environmental Studies of Colombia (IDEAM). This station is located on the Pamplonita River in the Norte de Santander department (Colombia), at the geographic coordinates of $7^\circ 41' 17''$ north

latitude, and 72° 26' 20" west longitude, at an elevation of 730 meters above sea level. The time window considered has a width of 14 years corresponding to the period 2002 - 2015.

Model Selection

The 12 linear models were identified by performing an exhaustive sequential evaluation of Pearson's linear correlation coefficient, r , between the SST time series of cell C (where C takes values between 1 and 3600), for month M (where M takes values between 1 and 12) of the 14 years considered and the flow time series for the 11 months following month M, of the same 14 years considered. The following algorithm better illustrates the evaluation process, which was carried out by implementing the corresponding code in Matlab®:

Step 1: For C taking values between 1 and 3600.

Step 2: For M taking values between 1 and 12.

Step 3: Select the SST time series for cell C, for month M, which has a length of 14 records, since there is one value per year considered.

Step 4: For L taking values between 0 and 10 (Considering 11 lags, L).

Step 5: Select the time series for the Monthly Average Flow of the month M + L.

Step 6: Evaluate and store the coefficient r between both time series.

Step 7: Iterate the K lags (in Step 4).

Step 8: Iterate the M months (in Step 2).

Step 9: Iterate oceanic cell C (in Step 1).

Step 10: Identify the maximum values of r (in absolute value) obtained for each of the 12 flow time series.

Pearson's linear correlation coefficient, r , [14] is evaluated by (1)

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad (1)$$

Where:

x corresponds to the SST time series (selected in Step 3)

y corresponds to the Qm time series (selected in Step 5),

s_x is the standard deviation of variable x

s_y is the standard deviations of variable y

n is the length of the series.

Model validation

The validation of the set of linear models found, which allow predicting the mean flow (Qm) for each month, for a given lag of L months, based on the SST of a given oceanic cell, was evaluated by determining the p statistic of each model and two objective functions for the series of flows predicted for three specific years. The two objective functions considered are the Relative Mean Square Error (rMSE) and the Nash-Sutcliffe Efficiency (NSE) [15]. Equations (2) and (3) present the calculation of the objective functions considered in the present study.

$$rMSE = \frac{1}{12} \sum_{M=1}^{12} \left(\frac{Q_{simM} - Q_{obsM}}{Q_{obsM}} \right)^2 \quad (2)$$

$$NSE = 1 - \frac{\sum_{M=1}^{12} (Q_{simM} - Q_{obsM})^2}{\sum_{M=1}^{12} (Q_{obsM} - \bar{Q}_{obs})^2} \quad (3)$$

Where:

Q_{sim} is the flow simulated by the linear model of each month M,

Q_{obs} is the flow observed in the limnigraphic station during month M

(\bar{Q}_{obs}) is the average of the observed flows.

Regarding the three years considered for the validation of the models, 2009, 2015 and 2011 were considered as representative of Normal, Dry and Wet years, respectively. The last two years correspond to the macro-climatic condition of El Niño year and La Niña year. El Niño years occur every three to seven years, approximately, a noticeable warming occurs on the surface waters of the tropical Pacific Ocean, linked to major changes in the atmosphere, resulting in anomalous global climate patterns, related to a phenomenon called El Niño-Southern Oscillation (ENSO) [16]. For the case of Colombia, El Niño years result in precipitations much lower than normal values with a good correlation that allows predictions for seasonal precipitation in Colombia [17]. La Niña years, on the other hand, corresponds to El Niño Southern Oscillation (ENSO) cold phase [18], and in the case of Colombia is associated With precipitations much higher than normal values, for example La Niña period of 2010-2011 caused major catastrophes in Colombia affecting four millions people and causing economic losses of about US \$7.8 billion [19], therefore, predicting this macro-climatic conditions may constitute in an important input to reduce disaster risk [20, 21].

Results and Discussion

Figure 1 presents the model result to predict the flow of May (month 5) based on the SST of August (month 8) of the previous year, allowing a 9 months lag. Likewise, the p-statistic is presented with a value much lower than 0.01. The value of the coefficient of determination indicates that 83% of the variance in the flow values for that month can be explained by the SST(8) in ocean cell 1298.

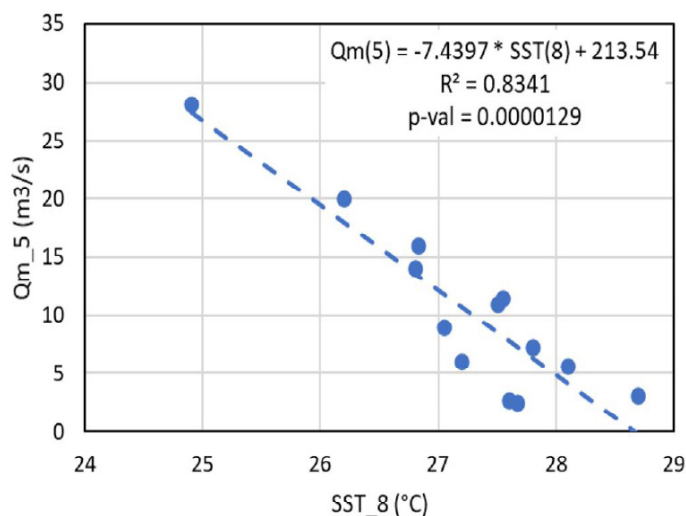


Figure 1. Model for the flow of May (month 5) as a function of the SST of August (month 8) of the previous year. 9-month lag.

Table I presents the linear models for the flow of each month, indicating the slope and the intercept of the model, as well as the oceanic cell to be used as an independent variable. Positive correlation coefficients are between 0.1 and 0.95, while negative correlation coefficients are between -0.88 and -0.96. The lags take values between 3 and 10 months, with an average value of 6.7 months. It is important to note that when the number of the month of the independent variable is greater than the number of the month of the dependent

variable, it really means that the month to be considered for the SST corresponds to the previous year.

Table I. Summary of the 12 linear models to predict the monthly flow.

Y : Month Qm		Slope	Intercept	X : Month SST		Cell, C	r coef	Lag
1	JAN	7.51	-166.74	4	APR	256	0.92	9
2	FEB	16.32	-434.08	8	AUG	3474	0.95	6
3	MAR	-6.00	176.15	7	JUL	1405	-0.95	8
4	APR	-8.25	243.32	7	JUL	679	-0.93	9
5	MAY	-7.44	213.54	8	AUG	1298	-0.91	9
6	JUN	-8.38	246.26	8	AUG	3109	-0.96	10
7	JUL	-3.14	98.56	2	FEB	2704	-0.89	5
8	AUG	-3.69	112.22	2	FEB	926	-0.95	6
9	SEP	4.86	-114.62	5	MAY	71	0.91	4
10	OCT	-13.96	402.73	7	JUL	898	-0.91	3
11	NOV	-8.28	254.07	6	JUN	146	-0.88	5
12	DEC	-11.59	328.21	5	MAY	926	-0.92	7

Regarding the results of the validation process of the model, the values of the p statistic were all well below 0.01, the maximum of the 12 values being $2.65 \cdot 10^{-4}$ indicating a good degree of statistical significance of the models. On the other hand, the results of the relative mean square error were improving when going from a normal year to a dry year and a wet year with results of 2.17, 0.88 and 0.24, respectively. A similar trend is observed in the SES values where values of -17.5, -6.1 and 0.26 were obtained for the normal, dry, and humid years; The only positive value for SES was found for the Girl Year (wet) considered, suggesting that the models found are recommended in the case of this type of year, defined by forecasts of global macro-climatic conditions.

Conclusions

A linear mathematical model was identified, consisting of a set of 12 equations, which allow forecasting the average monthly flow in an Andean basin located in Colombia, based on the sea surface temperature at specific sites for each month. The prognosis lag is between 3 and 10 months, with values of the determination coefficient (R^2) between 0.77 and 0.92, being statistically significant according to the p-statistic test. The results of the validation tests suggest the best performance of the model for wet years, decreasing in performance for dry years and normal years. It is expected to continue this type of analysis by extending the search area over the oceanic strip considered and considering a set of Andean basins.

Acknowledgements

The authors wish to express their gratitude to the Francisco de Paula Santander University (Cúcuta, Colombia), for their logistical support, as well as to the engineers Jimena Garzón

and Jessica Ramírez for their support in processing the data.

References

- [1] F.J. Chang and Y.C. Chen, “A counterpropagation fuzzy-neural network modeling approach to real time streamflow prediction”, *Journal of hydrology*, vol. 245, no.1-4, pp. 153-164, May 2001
- [2] D.I. Jeong and Y.O. Kim, “Rainfall runoff models using artificial neural networks for ensemble streamflow prediction”, *Hydrological Process*, vol. 19, no. 19, pp. 3819-3835, December 2005
- [3] L.E. Besaw, D.M. Rizzo, P.R. Bierman and W.R. Hackett, “Advances in ungauged streamflow prediction using artificial neural networks”, *Journal of Hydrology*, vol. 386, no. 1-4, pp. 27-37, May 2010
- [4] M. Küçük, and N. Ağirali oğlu, “Wavelet regression technique for streamflow prediction”, *Journal of applied statistics*, vol. 33, no. 9, pp. 943-960, November 2006
- [5] T. Razavi and P. Coulibaly, “Streamflow prediction in ungauged basins: review of regionalization methods”, *Journal of hydrologic engineering*, vol. 18, no. 8, pp. 958-975, August 2013
- [6] P. Pieper, A. Düsterhus and J. Baehr, “Improving seasonal predictions of meteorological drought by conditioning on ENSO states”, *Environmental Research Letters*, vol. 16, no. 9, pp.1-11, August 2021
- [7] W.U., Hassan and M.A. Nayak, “Global teleconnections in droughts caused by oceanic and atmospheric circulation patterns”, *Environmental Research Letters*, vol. 16, no. 1, pp. 1-12, December 2020
- [8] B. T. Jong, M. Ting and R., Seager, “El Niño's impact on California precipitation: Seasonality, regionality, and El Niño intensity”, *Environmental Research Letters*, vol. 11, no. 5, pp. 1-11, May 2016
- [9] M.B. Switanek, P.A. Troch and C.L. Castro, “Improving seasonal predictions of climate variability and water availability at the catchment scale”, *Journal of Hydrometeorology*, vol 10, no. 6, pp. 1521-1533, December 2009
- [10] V.M. Cvetkovic and J. Martinović, “Innovative solutions for flood risk management”, *International Journal of Disaster Risk Management*, vol. 2, no. 2, pp. 71-100, December 2020
- [11] P. Muñoz, J. Orellana-Alvear, P. Willems and R. Céleri, “Flash-flood forecasting in an Andean mountain catchment—Development of a step-wise methodology based on the random forest algorithm”, *Water*, vol 10, no. 11, pp. 1519, November 2018
- [12] P.J. Werdell and C.R. McClain, “Satellite remote sensing: ocean color” (No. GSFC-E-DAA-TN65587). Elsevier, 2019. Available at: https://pace.oceansciences.org/docs/werdell_and_mcclain_2019_eos3.pdf
- [13] H.M. Dierssen and K. Randolph, *Remote sensing of ocean color*. New York: Earth System Monitoring. Springer, 2013
- [14] K. Pearson, “Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs”, *Proceedings of the Royal Society of London*, vol. 60, no.

- 359-367, pp. 489-498, December 1897
- [15] J.E. Nash and J.V. Sutcliffe, "River flow forecasting through conceptual models - Part I—A discussion of principles", *Journal of hydrology*, vol. 10, no. 3, pp. 282-290, April 1970
- [16] K.E. Trenberth, *El niño southern oscillation (ENSO). Encyclopedia of Ocean Sciences (Third Edition)*, Academic Press, 2019
- [17] S. Córdoba-Machado, R. Palomino-Lemus, S.R. Gámiz-Fortis, Y. Castro-Díez and M.J. Esteban-Parra, "Influence of tropical Pacific SST on seasonal precipitation in Colombia: prediction using El Niño and El Niño Modoki", *Climate Dynamics*, vol. 44, no. 5-6, pp. 1293-1310, March 2015
- [18] A.M. Grimm, V.R. Barros and M.E. Doyle, "Climate variability in southern South America associated with El Niño and La Niña events", *Journal of climate*, vol. 13, no. 1, pp. 35-58, January 2000
- [19] N. Hoyos, J. Escobar, J.C. Restrepo, A.M. Arango and J.C. Ortiz, "Impact of the 2010–2011 La Niña phenomenon in Colombia, South America: the human toll of an extreme weather event", *Applied Geography*, vol. 39, pp. 16-25, May 2013
- [20] P.N. DiNezio, C. Deser, Y. Okumura and A. Karspeck, "Predictability of 2-year La Niña events in a coupled general circulation model", *Climate dynamics*, vol. 49, no. 11, pp. 4237-4261, December 2017
- [21] J.J. Luo, G. Liu, H. Hendon, O. Alves and T. Yamagata, "Inter-basin sources for two-year predictability of the multi-year La Niña event in 2010–2012", *Scientific reports*, vol. 7, no. 1, pp. 1-7, May 2017