Ambiente & Água - An Interdisciplinary Journal of Applied Science
ISSN 1980-993X – doi:10.4136/1980-993X
www.ambi-agua.net
E-mail: ambi.agua@gmail.com

Machine-learning methods for hydrological imputation data: analysis of the goodness of fit of the model in hydrographic systems of the Pacific - Ecuador

ARTICLES doi:10.4136/ambi-agua.2708

Received: 11 Feb. 2021; Accepted: 12 Apr. 2021

Diego Heras; Carlos Matovelle*

Center for Research, Innovation and technology transfer. Environmental Engineering. Catholic University of Cuenca, Avenida de las Americas, EC 010101, Cuenca, Azuay, Ecuador. E-mail: dherasb@ucacue.edu.ec
*Corresponding author. E-mail: cmmatovelleb@ucacue.edu.ec

ABSTRACT

Computational methods based on machine learning have had extensive development and application in hydrology, especially for modelling systems that do not have enough data. Within this problem, there are data series that are missing, and that should not necessarily be discarded; this is achieved by means of the imputation of the same ones, obtaining complete sets. For this reason, this research proposes a comparison of computer-learning techniques to identify those best suited for hydrographic systems of the Pacific of Ecuador. For the elaboration of this investigation, the hydro-meteorological records of the monitoring stations located in the watersheds of the Esmeraldas, Cañar and Jubones Rivers were used for 22 years, between 1990 and 2012. The variables that were imputed were precipitation and flow. Automatic learning machines of the Python Scikit_Learn module were used; these modules integrate a wide range of automated learning algorithms, such as Linear Regression and Random Forest. Finally, results were obtained that led to a minimum useful mean square error for Random Forest as an automatic machine-learning imputation method that best fits the systems and data analyzed.

Keywords: data imputation, hydrographic systems, machine learning.

Métodos de aprendizado de máquina para dados de imputação hidrológica: análise da qualidade de ajuste do modelo em sistemas hidrográficos do Pacífico - Equador

RESUMO

Métodos computacionais baseados em aprendizado de máquina tiveram amplo desenvolvimento e aplicação em hidrologia, especialmente para modelagem de sistemas que não possuem dados suficientes. Dentro deste problema faltam séries de dados que não devem ser necessariamente descartadas. Isso é feito por meio da imputação das mesmas obtendo-se conjuntos completos. Por este motivo, esta pesquisa propõe uma comparação de técnicas de aprendizagem computacional para identificar aquelas mais adequadas aos sistemas hidrográficos do Pacífico do Equador pelo interesse representado pelo estudo destes sistemas por complementaridade hidrológica. Para a elaboração desta investigação foram utilizados os registros hidrometeorológicos das estações de monitoramento localizadas nas bacias dos rios Esmeraldas, Cañar e Jubones durante 22 anos, compreendidos entre 1990 e 2012. As variáveis

imputadas foram precipitação e vazão. Foram utilizadas máquinas de aprendizagem automática do módulo Python Scikit_Learn; esses módulos integram uma ampla gama de algoritmos de aprendizagem automatizados, como Linear Regression e Random Forest. Finalmente, foram obtidos resultados que levaram a um erro quadrático médio útil mínimo para Random Forest como um método de imputação de aprendizado de máquina automático que melhor se ajusta aos sistemas e dados analisados.

Palavras-chave: aprendizado de máquina, imputação de dados, sistemas hidrográficos.

1. INTRODUCTION

In recent years, methods based on machine learning have advanced considerably and have been applied in several areas of science and technology. Within hydrology, they have been widely applied in the development of basin behavior models, especially those that do not have enough information to apply physical models. In this aspect, another problem is that much of the available information is incomplete, and the series that are available are useless, further reducing the data for the work of hydrological modelling.

Hydrologists and water managers have made use of observed relationships between rainfall and runoff to predict streamflow ever since the creation of the rational method in the 19th century (Beven, 2012), a properly designed monitoring network with optimal data allows us to know the relationship between these behaviours and to be able to apply this in studies of water interest. However, streamflow and rainfall records suffer from missing observations, mostly resulting from unexpected causes including the loss of records, sensor problems, or disruption of data collection (Ng *et al*., 2009).

In the area of analysis, one of the problems is that there are not enough nor adequate monitoring systems, and from those that exist a large amount of missing data is evident. This makes the process of modelling these watersheds complicated and inaccurate; the application of this type of study generates knowledge of the area and its subsequent exploitation for different activities linked to water. These data would result in an incorrect response of hydrological models, but it is illogical to ignore abnormal or missing values if there is limited data available; substantial uncertainty in hydrologic and water quality modelling can be driven by these missing records (Kim *et al*., 2015). There are several methods to solve the problem of missing observations from statistics based on linear regressions that have already been validated in other investigations. These depend on the amount of data existing in the series and the data and relationships that they have with neighbouring weather stations (Mwale *et al.*, 2012; Rees, 2009). For these reasons, authors such as (Adeloye, 2009) indicate that regression methods might only be applicable when all predictors exist.

Artificial neural networks (ANNs), regression trees, and support vector machines have been shown to be powerful tools for predictive modelling and exploratory data analysis, particularly in areas that do not meet the conditions for using traditional statistical methods (Shortridge *et al*., 2016). These methods have mathematical formulations that require a high cost of computational processes, but are very effective when there are non-linear relationships to use traditional statistical methods (Dawson *et al.*, 2010). These strengths make them very useful, especially in countries with poor monitoring traditions, where gaps of information in climatological and hydrologic time series are ubiquitous (Campozano *et al*., 2014).

2. MATERIALS AND METHODS

The analysis stations are located in three main river basins of the coastal zone of Ecuador. These are the Cañar River, the Jubones River and the Esmeraldas River. These basins have

around 318 weather stations and 106 hydrological stations. The three systems have been chosen because they have the largest monitoring network in the country and represent a significant area of analysis (see Figure1).

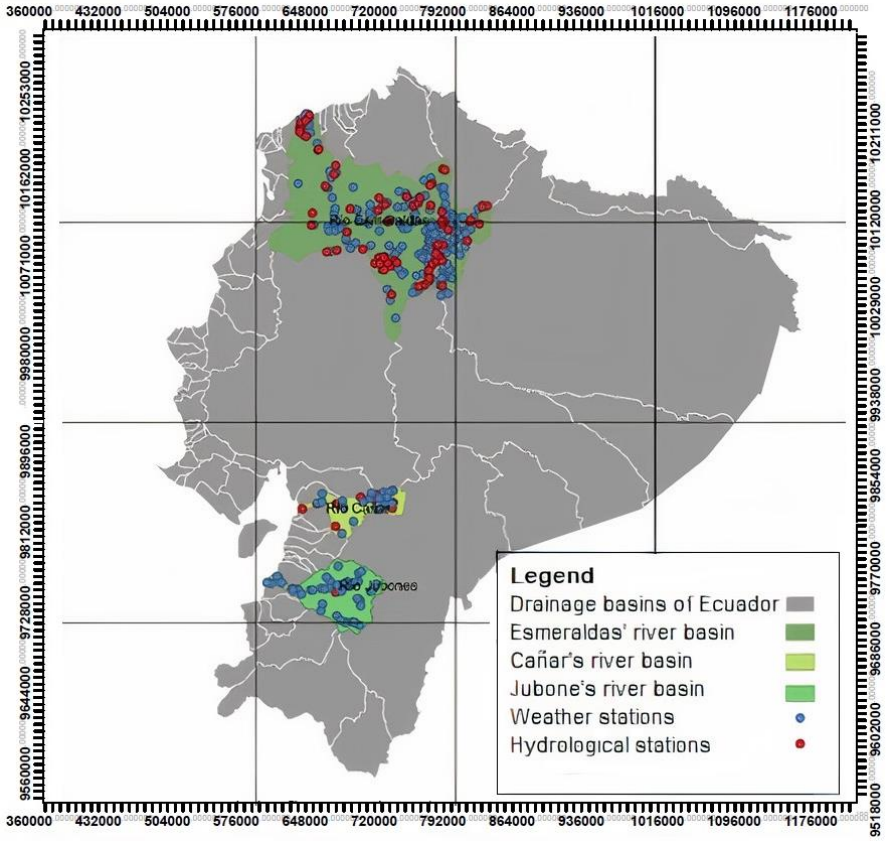

Figure 1. Location of the hydrographic systems for the analysis.

The sample of the stations to be analyzed was obtained by simple random sampling. One meteorological station and one hydrological station were selected per basin, as well as two nearby reference stations for the case of meteorological stations and a reference station for the example of hydrological stations. These nearby reference stations were selected as predictors at the time of analysis of data.

2.1. Data Imputation Methods

For the development of the imputation model with the uses of machine learning, we work with a pattern search to optimize parameters and later cross-validation for the periods of analysis of the research (Kim *et al*., 2015) (Carpenter, 1999[1]). This imputation method is based on supervised learning models, that is, the machine is presented with the response information at the same time as the input information, with which the machine will learn to arrive at the

[1] CARPENTER, J. **Personal Communication**. Virginia Tech, 1999.

answer through an iterative process. Within the operation of a machine learning, we have an input data vector that transfers it to the network where the complexity of the training is determined, thus obtaining a vector of data output as a result of the model (Guo *et al*., 2015).

The process is divided into training data, test data, and validation data; this division is established through cross-validation that allows an adequate distribution of the data between test data and validation so that the model does not over-fit in the trained data and have deficiencies in the validation data (see Figure 2). Followed by the supervised learning machine, this information is processed and a linear model based on the least square's method is established. The multiple iterations that the learning machine performs with the training and test data allow us to identify the best linear model, which will allow the imputation of hydro-meteorological data.

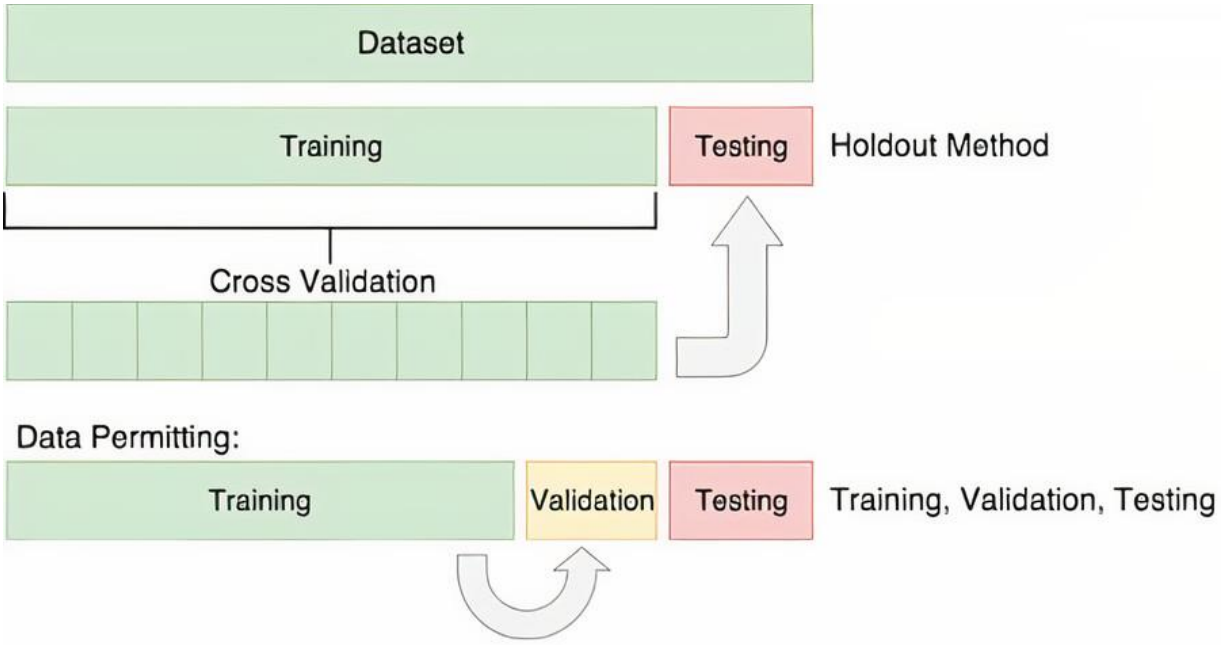

Figure 2. Cross-validation operation, between training, testing and validation data.
Source: www.towardsdatascience.com

The Tansig Function is used as a transfer function since it gives efficient results within hydrological studies (Kim *et al*., 2015; Akhter, 2017), the function is as Equation 1 follows:

$$y = f \sum_{i=1}^{N} w_i x_i + b \quad (1)$$

Where x_i is the input in the network, y is the output in the network, N is the number of neurons in the input vector, w_i is the connection weight between input and output, f is the transfer function, and b is the bias term.

To analyze the weight of each calculation, the neural networks use a back-propagation algorithm, where the error in the output data and the observed data are analyzed. It is a type of supervised learning based on the generalization of the delta rule (Veintimilla-Reyes and Cisneros, 2015; Hsu *et al.*, 1995; Bisoyi *et al*., 2019). This algorithm updates weights by moving along the gradient descent of the error function, which allows the steepest decreasing change. The advantages of this algorithm are its ability to adjust the learning rate by updating the learning rate parameter and it also guarantees less oscillation with the momentum constant (Kim *et al.*, 2015). The process is repeated until the error is minimized. This method is widely used in hydrological studies (Dawson and Wilby, 2001).

In this research, we analyze and compare two methods. The first method is autonomous learning based on linear regression, which integrates statistical models for relating responses to linear combinations of predictor variables (Ahmad *et al*., 2010; Srivastava *et al.*, 2013). The second method is the random forest algorithm, which is widely used for the study of water resources. Applications falling under this category include streamflow modeling using data-driven rainfall-runoff models, while streamflow imputation of missing values is also generating increased interest (Tyralis *et al.*, 2019).

Random Forest is a supervised Machine Learning algorithm based on a stochastic model that relates a result to explanatory variables or characteristics. Each decision within the tree can be viewed as a set of conditions, organized hierarchically and applied successively to the data set. For regression applications, they provide independent numerical predictions of the phenomenon of interest. In the end, the result corresponds to the mean forecast of all individual trees (Muñoz *et al*., 2018).

3. RESULTS AND DISCUSSIONS

The machine of autonomous learning based on linear regression and in random decision forests produced models that allowed the imputation of missing data in the hydro-meteorological records of the stations located in the study basins, i.e., the stations of the basins of the Esmeraldas, Cañar and Jubones Rivers. The following models are calibrated to meet the imputation of missing data from each station within the period of records comprising 22 years, from 1990 to 2012, for both meteorological stations and hydrological stations. It should also be taken into account that we worked only with the hydro-meteorological stations near the hydrographic basin, and that in the analysis of correlations they maintained between them a correlation value greater than or equal to 0.75.

The analysis of the best regressions obtained for each imputed data in the selected meteorological stations is presented (see Figure 3). These models were established with the Linear Regression learning machine of the Sklearn Python library, and their data sets were applied for cross-validation as a fundamental pillar for the validation of results (Hastie *et al*., 2017). The analyses have a relationship between the test values and the predicted values. As a result of these results, the linear models allowed imputation of missing data in the hydro-meteorological records. In the figure, it can be seen that there is a linear relationship for each of the data and the stations. This relationship has to be validated by statistical indicators of goodness-of-fit between observed and predicted data (Tyralis *et al.*, 2019; Zambrano-Bigiarini, 2017; 2011). These analyses are presented in Section 3.1, where they are compared between the two methods presented in this research.

Table 1 shows the equations of the linear models that have been obtained for each station with the stations with which it has been correlated with the previous spatial analysis. This line regression model is obtained by the Machine Learning Linear Regression algorithm between test values and the predicted values of the stations.

Similarly, the analysis of the allocation models based on the Random Forest learning machine (see Figure 4) is presented. The relationship between the values of the data of the reference station and the data of the analysis station is evaluated; each parameter of the model is calibrated so the statistical indicator is the most reliable.

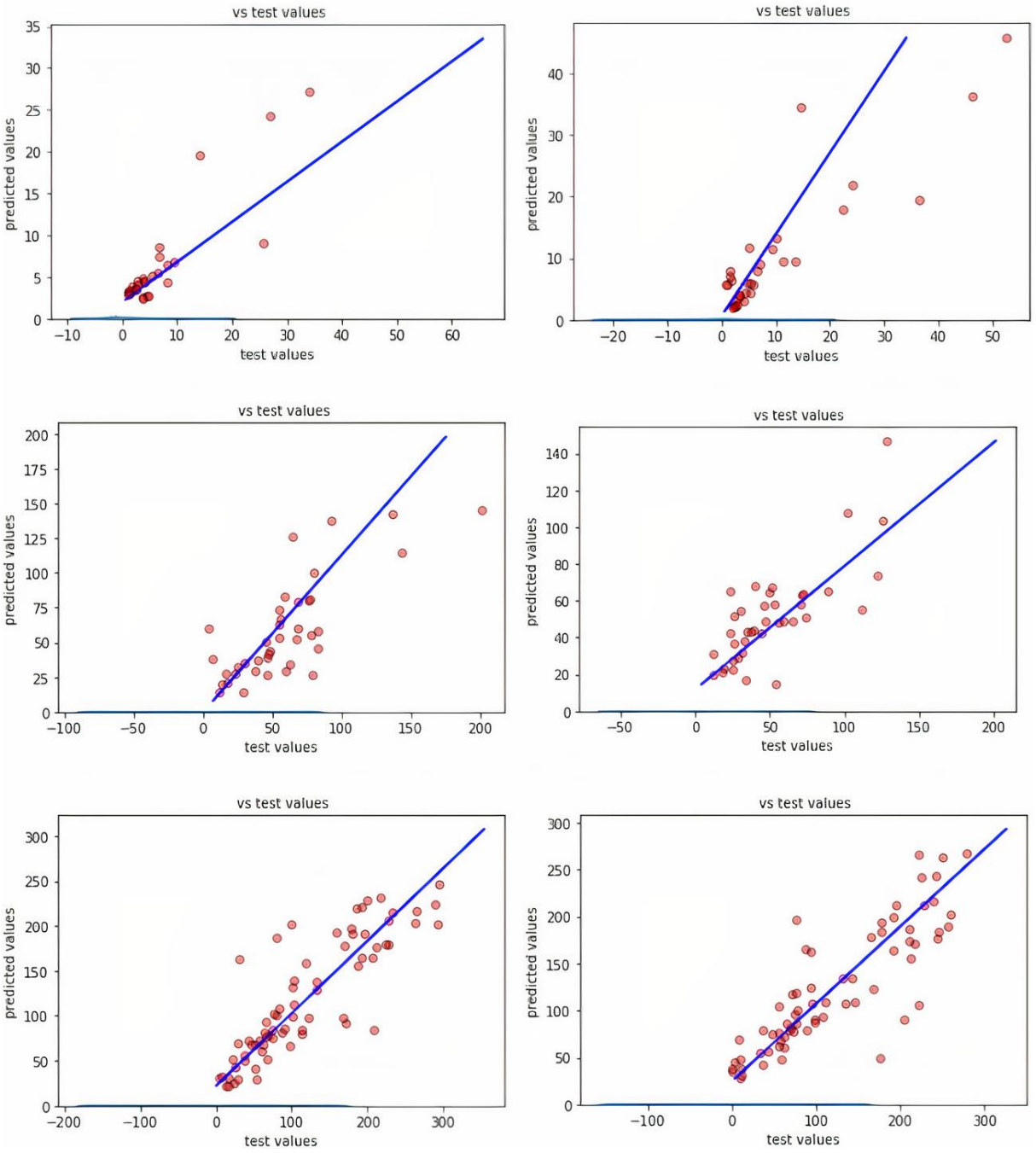

Figure 3. Linear regression line obtained by the Machine Learning Linear Regression algorithm between test values and the predicted values of the stations.

Table 1. Linear models for each weather station.

Weather station	Linear Model
M003	20.91 + 0.8150 M364
M364	22.68 + 0.8370 M003
M411	1.59 + 1.0939 M31
M031	10.52 + 0.6960 M411
M040	24.75 + 0.2145 M185 + 0.5375 M292
M185	23.25 + 0.3488 M040 + 1.2632 M292

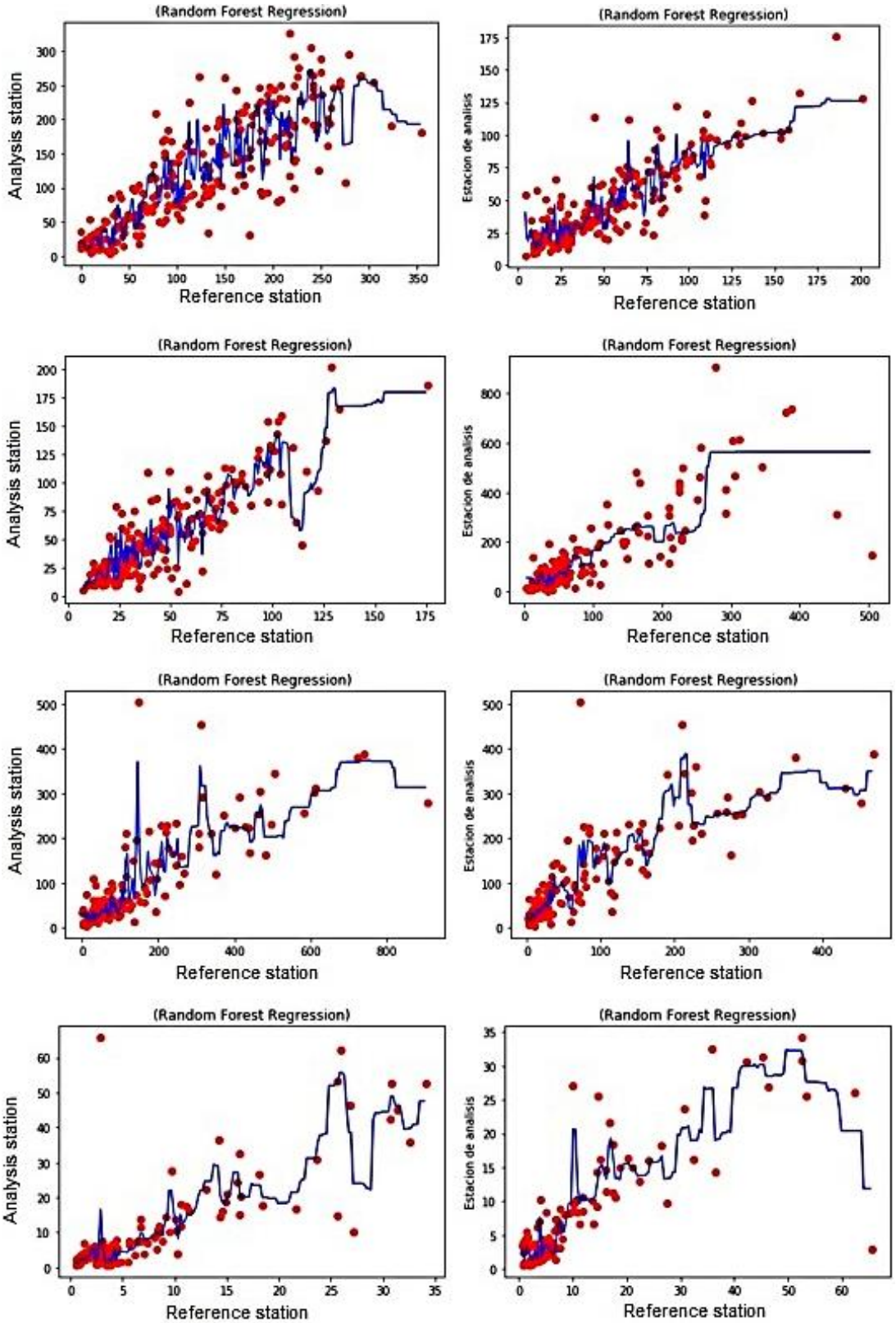

Figure 4. Predictive model obtained by the Machine Learning Random Forest algorithm between test values and the predicted values of the stations.

3.1. Analysis of statistical index for model validation

To perform an interpretation, comparison, and analysis of applicability for large hydrographic systems where the basin has many variations in both meteorological and hydrological behaviour, it is necessary to obtain indicators of goodness-of-fit that indicate

which is the optimal model. The statistical significance of the performance statistics is an aspect that is generally ignored that helps in reducing subjectivity in the proper interpretation of the model performance (Ritter and Muñoz-Carpena, 2013). To obtain these indicators, the data observed in the meteorological stations have been compared with the simulated data (imputed through the techniques analyzed). This process allows us to validate which of the two methods has a better fit, and therefore, which would be more efficient and result better in subsequent applications.

The indicators have been analyzed, and the efficiency coefficient of Nash and Sutcliffe (Nash and Sutcliffe, 1970) has received considerable attention in hydrological modelling (Gupta and Kling, 2011; Moussa, 2010). It has already been used for the imputation of missing data and it is generally used in other fields of science (Schaefli and Gupta, 2007). It is also tested by the Kling-Gupta Index of Efficiency (KGE) (Galleguillos *et al.*, 2017), which uses the values between -1 and 1, with the value 1 as an ideal and positive values greater than 0.5 as sufficiently robust correctors. The Mean Square Error is a standard indicator for this type of analysis (Gupta *et al*., 2009).

To determine the best method for data imputation, the indices are compared (Table 2). It is observed that the best method in the three reliability analyses is the Random Forest. The Mean Square Error is the one that indicates the highest relation of the imputed values and, observed with 0.01, the NSE and KGE index has values of 0.76 and 0. 67 respectively, indicating good data adjustments.

Table 2. Comparison of indexes of the evaluated methods.

LINEAR REGRESSION				RANDOM FOREST			
STATION	**MSE**	**NSE**	**KGE**	**STATION**	**MSE**	**NSE**	**KGE**
M003	0	0.75	0.55	M003	0	0.78	0.68
M364	0	0.75	0.55	M364	0	0.78	0.68
M411	0.02	0.71	0.52	M411	0	0.78	0.68
M031	0.1	0.7	0.50	M031	0	0.78	0.68
M040	0	0.75	0.45	M040	0.06	0.7	0.65
M185	0.32	0.68	0.4	M185	0.07	0.7	0.65
M292	0.08	0.65	0.45	M292	0.02	0.73	0.66
H172	0.05	0.6	0.45	H172	0	0.78	0.68
H173	0	0.65	0.43	H173	0	0.78	0.68
MEAN	**0.06**	**0.69**	**0.48**	**MEAN**	**0.01**	**0.76**	**0.67**

The evaluations of the indices verify that the models have given good results. Reliable indices are appreciated with the tools of artificial intelligence, although reviewing each accurate indicator the NSE index for Linear Regression is 0.69 compared to 0.76 for Random Forest; the two values show right adjustments between the data (Waseem *et al*., 2017). The KGE index between the models has more noticeable differences between Linear Regression (0.48) and Random Forest (0.67), the Random Forest indicator has given values that establish (Näschen *et al*., 2018; Pool *et al.*, 2018) the use of that model as the best for analysis of large river basins with variations of the Pacific climatology.

For several decades the need to have complete time series to validate subsequent studies has meant that many studies are done, and various techniques have been used. In Aissia *et al.* (2017), a review of multivariate methods and their application to reduce the loss of information is made; these range from simple relationships such as linear regressions that are based on spatial approximations, artificial intelligence techniques, and even much more innovative methods, such as those presented by Williams *et al.* (2018), who formed two methods with Bayesian structures to generate an algorithm that represented the signal of the time series of

temperatures. In Teegavarapu (2019) research, spatial interpolation methods were evaluated, although Euclidean distances were substituted to improve fit indicators' goodness.

In another study carried out by Chen *et al.* (2019), several techniques were tested to impute precipitation data with the premise that having complete series improves analyses within hydrographic basins. After using several methods, they decide to perform the hydrological model with the best fit; This leads us to consider that there is no "best" technique, but rather that the analysis must be based on several determining factors such as the type of hydrometeorological variable, the years of the series, the time scale, spatial variations, conditioning factors or external phenomena.

The apparent difference compared to traditional methods is that the response to abnormal weather patterns can be better exploited, which is of great interest for rainfall patterns as variable as that of the Pacific in Ecuador, which is influenced by various external phenomena.

4. CONCLUSIONS

In this study, we propose the analysis of two traditional methods of artificial intelligence to study the accuracy and use for imputation of missing data in Hydrographic Systems on the slope of the Pacific in Ecuador. The two methods proposed and evaluated were Linear Regression and Random Forest, which were tested in the three most representative Hydrographic Systems of the country, in the basins of the Esmeraldas, Cañar and Jubones Rivers, intending to cover the extension of the surface and the water from north to the south of Ecuador.

After carrying out a preliminary analysis of the data, we worked with 9 test stations in the three systems, observing goodness-of-fit indicators for each of the stations and each model, we worked with Medium Squared Error (MSE), Nash and Sutcliffe (NSE) and Kling-Gupta Index of efficiency (KGE). The values obtained after the goodness-of-fit analysis mark ranges for good efficiency analyses, but the Random Forest model has the three best indicators both on average and for each of the analysis stations. It is important to highlight the importance of carrying out this type of analysis in watersheds of the Pacific slope of Ecuador, since available information is scarce and the hydro-meteorological behaviour is different from the Amazon slope. They are also very large systems in extension but with hypsometries marked by very marked altitudinal differences in a small area of land and with a lower amount of water compared to the Amazon slope.

After the data analysis and the discussion process with authors who have carried out similar works worldwide for several decades, it can be seen that techniques that can reproduce atypical effects should be evaluated in the first place and then validated before their application to the management of water resources.

5. REFERENCES

ADELOYE, A. The relative utility of regression and artificial neural networks models for rapidly predicting the capacity of water supply reservoirs. **Environmental Modelling and Software**, v. 24, n. 10, p. 1233–1240, 2009. https://doi.org/10.1016/j.envsoft.2009.04.002

AHMAD, S.; KALRA, A.; STEPHEN, H. Estimating soil moisture using remote sensing data: A machine learning approach. **Advances in Water Resources**, v. 33, n. 1, p. 69–80, 2010. https://doi.org/10.1016/j.advwatres.2009.10.008

AISSIA, M.-A. B.; CHEBANA, F.; OUARDA, T. B. M. J. Multivariate missing data in hydrology–Review and applications. **Advances in Water Resources**, v. 110, p. 299-309, 2017. https://doi.org/10.1016/j.advwatres.2017.10.002

AKHTER, M. Application of ANN for the Hydrological Modeling. **International Journal for Research in Applied Science & Engineering Technology**, v. 5, n. 7, p. 203–213, 2017.

BEVEN, K. Rainfall-Runoff Modelling. *In*: ABRAHART, R.; KNEALE, P. E.; SEE, L. M. (eds.). **Neural Networks for Hydrological Modeling**. 2nd ed. New York: John Wiley & Sons, 2012. https://doi.org/10.1201/9780203024119.ch9

BISOYI, N.; GUPTA, H.; PADHY, N. P.; CHAKRAPANI, G. J. Prediction of daily sediment discharge using a back propagation neural network training algorithm: A case study of the Narmada River, India. **International Journal of Sediment Research**, v. 34, n. 2, p. 125–135, 2019. https://doi.org/10.1016/j.ijsrc.2018.10.010

CAMPOZANO, L.; SÁNCHEZ, E.; AVILÉS, Á.; SAMANIEGO, E. Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the Ecuadorian Andes. **Maskana**, v. 5, n. 1, p. 99–115, 2014. https://doi.org/10.18537/mskn.05.01.07

CHEN, L.; XU, J.; WANG, G.; SHEN, Z. Comparison of the multiple imputation approaches for imputing rainfall data series and their applications to watershed models. **Journal of Hydrology**, v. 572, p. 449–460, 2019. https://doi.org/10.1016/j.jhydrol.2019.03.025

DAWSON, C. W.; HARPHAM, C.; WILBY, R. L.; CHEN, Y. Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China. **Hydrology and Earth System Sciences**, v. 6, n. 4, p. 619–626, 2010. https://doi.org/10.5194/hess-6-619-2002

DAWSON, C. W.; WILBY, R. L. Hydrological modelling using artificial neural networks. **Progress in Physical Geography**, v. 25, n. 1, p. 80–108, 2001. https://doi.org/10.1191/030913301674775671

GALLEGUILLOS, M.; ZAMBRANO, M.; PUELMA, C.; JOPIA, A. Evaluación espacio-temporal del déficit hídrico para las cuencas de Chile a partir de información satelital. *In:* **Escenarios Hídricos 2030**. 2017. Available at: http://escenarioshidricos.cl/multimedia/ Access: 2018.

GUO, H.; JEONG, K.; LIM, J.; JO, J.; KIM, Y. M.; PARK, J.-P.; KIM, J. H.; CHO, K. H. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. **Journal of Environmental Sciences**, v. 32, p. 90–101, 2015. https://doi.org/10.1016/j.jes.2015.01.007

GUPTA, H. V.; KLING, H. On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. **Water Resources Research**, v. 47, n. 10, p. 2–4, 2011. https://doi.org/10.1029/2011WR010962

GUPTA, H. V.; KLING, H.; YILMAZ, K. K.; MARTINEZ, G. F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. **Journal of Hydrology**, v. 377, n. 1–2, p. 80–91, 2009. https://doi.org/10.1016/j.jhydrol.2009.08.003

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. Springer, 2017.

HSU, K.; GUPTA, H. V.; SOROOSHIAN, S. Artificial neural network modeling of the rainfall-runoff process that arises and based on Background and Scope. **Water Resources**, v. 31, n. 10, p. 2517-2530, 1995. https://doi.org/10.1029/95WR01955

KIM, M.; BAEK, S.; LIGARAY, M.; PYO, J.; PARK, M.; CHO, K. H. Comparative studies of different imputation methods for recovering streamflow observation. **Water**, v. 7, n. 12, p. 6847–6860, 2015. https://doi.org/10.3390/w7126663

MOUSSA, R. When monstrosity can be beautiful while normality can be ugly: assessing the performance of event-based flood models. **Hydrological Sciences Journal**, v. 55, n. 6, p. 1074–1084, 2010. https://doi.org/10.1080/02626667.2010.505893

MUÑOZ, P.; ORELLANA-ALVEAR, J.; WILLEMS, P.; CÉLLERI, R. Flash-flood forecasting in an andean mountain catchment-development of a step-wise methodology based on the random forest algorithm. **Water**, v. 10, n. 11, 2018. https://doi.org/10.3390/w10111519

MWALE, F. D.; ADELOYE, A. J.; RUSTUM, R. Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi - A self organizing map approach. **Physics and Chemistry of the Earth**, v. 50–52, p. 34–43, 2012. https://doi.org/10.1016/j.pce.2012.09.006

NÄSCHEN, K.; DIEKKRÜGER, B.; LEEMHUIS, C.; STEINBACH, S.; SEREGINA, L. S.; THONFELD, F.; VAN DER LINDEN, R. Hydrological modeling in data-scarce catchments: The Kilombero floodplain in Tanzania. **Water**, v. 10, n. 5, p. 1–27, 2018 https://doi.org/10.3390/w10050599

NASH, J. E.; SUTCLIFFE, J. V. River flow forecasting through conceptual models part I — A discussion of principles. **Journal of Hydrology**, v. 10, n. 3, p. 282–290, 1970 https://doi.org/https://doi.org/10.1016/0022-1694(70)90255-6

NG, W. W.; PANU, U. S.; LENNOX, W. C. Comparative Studies in Problems of Missing Extreme Daily Streamflow Records. **Journal of Hydrologic Engineering**, v. 14, n. 1, p. 91–100, 2009. https://doi.org/10.1061/(asce)1084-0699(2009)14:1(91)

POOL, S.; VIS, M.; SEIBERT, J. Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. **Hydrological Sciences Journal**, v. 63, n.13–14, p. 1941–1953, 2018. https://doi.org/10.1080/02626667.2018.1552002

REES, G. Hydrological data. *In:* WMO. **Manual on Low-flow Estimation and Prediction**. Geneva: WMO, 2009. (Operational Hydrology Report, n. 1029).

RITTER, A.; MUÑOZ-CARPENA, R. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. **Journal of Hydrology**, v. 480, p. 33–45, 2013. https://doi.org/10.1016/j.jhydrol.2012.12.004

SCHAEFLI, B.; GUPTA, H. V. Do Nash values have value? **Hydrological Processes**, v. 21, p. 2075–2080, 2007. https://doi.org/10.1002/hyp

SHORTRIDGE, J. E.; GUIKEMA, S. D.; ZAITCHIK, B. F. Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. **Hydrology and Earth System Sciences**, v. 20, n. 7, p. 2611–2628, 2016. https://doi.org/10.5194/hess-20-2611-2016

SRIVASTAVA, P. K.; HAN, D.; RAMIREZ, M. R.; ISLAM, T. Machine Learning Techniques for Downscaling SMOS Satellite Soil Moisture Using MODIS Land Surface Temperature for Hydrological Application. **Water Resources Management**, v. 27, n. 8, p. 3127–3144, 2013. https://doi.org/10.1007/s11269-013-0337-9

TEEGAVARAPU, R. S. V. Precipitation Imputation with Probability Space-based Weighting Methods. **Journal of Hydrology**, v. 581, 2019. https://doi.org/10.1016/j.jhydrol.2019.124447

TYRALIS, H.; PAPACHARALAMPOUS, G.; LANGOUSIS, A. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. **Water**, v. 11, n. 5, p. 910, 2019. https://doi.org/10.3390/w11050910

VEINTIMILLA-REYES, J.; CISNEROS, F. Predicción de Caudales Basados en Redes Neuronales Artificiales (RNA) para Períodos de Tiempo Sub Diarios. **Revista Politécnica**, v. 35, n. 3, p. 42–49, 2015.

WASEEM, M.; MANI, N.; ANDIEGO, G.; USMAN, M. A Review of Criteria of Fit for Hydrological Models. **International Research Journal of Engineering and Technology**, v. 4, n. 11, 2017.

WILLIAMS, D. A.; NELSEN, B.; BERRETT, C.; WILLIAMS, G. P.; MOON, T. K. A comparison of data imputation methods using Bayesian compressive sensing and Empirical Mode Decomposition for environmental temperature data. **Environmental Modelling and Software**, v. 102, p. 172–184, 2018. https://doi.org/10.1016/j.envsoft.2018.01.012

ZAMBRANO-BIGIARINI, M. Goodness-of-fit Measures to Compare Observed and Simulated Values with hydroGOF Installation Installing hydroGOF. p. 1–5, 2011. Available at: https://mran.microsoft.com/snapshot/2017-05-24/web/packages/hydroGOF/vignettes/hydroGOF_Vignette.pdf Access: 2018.

ZAMBRANO-BIGIARINI, M. **HydroGOF**. 2017. Available at: http://hzambran.github.io/hydroGOF/. Access: 2017.