DOI: 10.24850/j-tyca-14-02-05

Articles

Hydrological modeling based on the KNN algorithm: An application for the forecast of daily flows of the Ramis river, Peru

Modelado hidrológico basado en el algoritmo KNN: una aplicación para el pronóstico de caudales diarios del río Ramis, Perú

Efrain Lujano[1], ORCID: https://orcid.org/0000-0002-6543-8324
Rene Lujano[2], ORCID: https://orcid.org/0000-0002-4103-9724
Juan Carlos Huamani[3], ORCID: https://orcid.org/0000-0002-7734-945X
Apolinario Lujano[4], ORCID: https://orcid.org/0000-0002-9386-1613

[1]Escuela Profesional de Ingeniería Agrícola, Universidad Nacional del Altiplano, Puno, Peru / Escuela Profesional de Ingeniería Ambiental, Universidad Peruana Unión, Juliaca, Peru, elujano28@gmail.com

[2]Ingeniería de Sistemas, Universidad Nacional del Altiplano, Puno, Peru, rlujano13l@gmail.com

[3]Servicio Nacional de Meteorología e Hidrología (SENAMHI), Lima, Peru, jchc.imf2014@gmail.com

[4]Autoridad Nacional del Agua (ANA), Lima, Peru, apolex23@gmail.com

Corresponding author: Efrain Lujano, elujano28@gmail.com

Abstract

The forecast of river stream flows is of significant importance for the development of early warning systems. Artificial intelligence algorithms have proven to be an effective tool in hydrological modeling data-driven, since they allow establishing relationships between input and output data of a watershed and thus make decisions data-driven. This article investigates the applicability of the k-nearest neighbor (KNN) algorithm for forecasting the mean daily flows of the Ramis river, at the Ramis hydrometric station. As input to the KNN machine learning algorithm, we used a data set of mean basin precipitation and mean daily flow from hydrometeorological stations with various lags. The performance of the KNN algorithm was quantitatively evaluated with hydrological ability metrics such as mean absolute percentage error (MAPE), anomaly correlation coefficient (ACC), Nash-Sutcliffe efficiency (NSE), Kling-Gupta efficiency (KGE') and the spectral angle (SA). The results for forecasting the flows of the Ramis river with the k-nearest neighbor machine learning algorithm reached high levels of reliability with flow lags of one and two days and precipitation with three days. The algorithm used is simple but robust to make short-term flow forecasts and can be integrated as an alternative to strengthen the daily hydrological forecast of the Ramis river.

Keywords: Artificial intelligence, hydrological data-driven modelling, machine learning, Ramis river, K-Nearest Neighbor.

Resumen

El pronóstico de caudales de un río es de gran importancia para el desarrollo de sistemas de alerta temprana. Los algoritmos de inteligencia artificial han demostrado ser una herramienta eficaz en la modelación hidrológica basado en datos, pues permiten establecer relaciones entre los datos de entrada y salida de una cuenca hidrográfica, y de esta manera tomar decisiones basado en datos. Este artículo investiga la aplicabilidad del algoritmo k vecino más cercano (KNN) para el pronóstico de caudales medios diarios del río Ramis en la estación hidrométrica Ramis. Como insumo de entrada al algoritmo de aprendizaje automático KNN utilizamos un conjunto de datos de precipitación media de la cuenca y caudal medio diario de estaciones hidrometeorológicas con varios rezagos. El rendimiento del algoritmo KNN se evaluó cuantitativamente con métricas de habilidad hidrológica, como el error porcentual absoluto medio (MAPE), anomalía del coeficiente de correlación (ACC), eficiencia de Nash-Sutcliffe (NSE), eficiencia de Kling-Gupta (KGE') y ángulo espectral (SA). Los resultados para realizar pronóstico de caudales del río Ramis con el algoritmo de aprendizaje automático KNN alcanzaron altos niveles de confiabilidad, sobre todo con rezagos de caudales de uno y dos días, y precipitación con tres días. El algoritmo utilizado es simple, pero robusto para efectuar pronósticos de caudales a corto plazo, y puede ser integrado como una alternativa para el fortalecimiento del pronóstico hidrológico diario del río Ramis.

Palabras clave: inteligencia artificial, modelado hidrológico basado en datos, aprendizaje automático, río Ramis, k-vecino más cercano.

Received: 30/11/2020

Accepted: 29/09/2021

Introduction

Floods induced by excessive rainfall and overflowing rivers, are common natural hazards in the regions of Peru. The frequencies with which they occur in flood periods, causes significant losses and damage to property. This phenomenon is likely to become more prevalent with climate change, and accurate and reliable forecasts of river flows would help minimize the damage associated with flooding. Accurate short-term forecasts (hourly and daily) are important for predicting floods and developing early warning systems (Mundher, Ahmed, & Abdulmohsin, 2015). An accurate stream flow forecasting is critical to optimal flood control (Solomatine & Xue, 2004).

The use of process-based models has become essential tools to study the response of hydrological regimes (Madsen, 2000; Mendez & Calvo-Valverde, 2016), but a sufficiently representative and precise implementation can lead to invest a lot of time and cost, in addition to calibrating a large number of parameters. Since the 1930s, numerous rainfall-runoff models have been developed and the entire physical process of the hydrological cycle were formulated mathematically in conceptual models that compose a large number of parameters (Tokar & Johnson, 1999). In a data-driven hydrological modeling context, "All models are wrong and some are useful", this quote is significant due to the presence of different unresolved queries and deliberate assumptions (Remesan & Mathew, 2015).

Data-driven models, especially machine learning (ML) techniques, do not require complex physical equations and assumed parameters that process-based models require. Due to the simplicity in their implementation of ML algorithms and more accurate prediction, it has been widely applied in hydrological modeling/forecasting (Mundher *et al*., 2016; Remesan & Mathew, 2015; Solomatine & Xue, 2004), achieving good performances even with small data sets (Veintimilla-Reyes, Cisneros, & Vanegas, 2016). Data-driven models or ML models are capable of forecasting rainfall-runoff, even for a fairly complex system (Solomatine & Xue, 2004).

ML is considered a subfield of artificial intelligence (AI) and is divided into three main classes, supervised learning, unsupervised learning, and reinforcement learning (Igual & Seguí, 2017). A review of AI techniques, specifically ML supervised algorithms, have successfully demonstrated their applicability in urban flow prediction (Xie *et al*., 2020), flood prediction (Mosavi, Ozturk, & Chau, 2018; Solomatine & Xue, 2004), forecast of daily flows (Mundher *et al*., 2015), modeling and forecast of mean monthly flows (Laqui, 2010; Lujano, Lujano, Quispe, & Lujano, 2014; Mundher *et al*., 2016), in the same way, they have also been applied in the wind energy forecast based on daily wind speed data (Demolli, Dokuz, Ecemis, & Gokcek, 2019), estimation of evapotranspiration (Granata, 2019; Xu *et al*., 2018), hydro-climatological predictions (Thakur, Kalra, Ahmad, & Lamb, 2020), landslide modeling (Liu *et al*., 2021), reference evapotranspiration estimation (Alipour, Yarahmadi, & Mahdavi, 2014; Antonopoulos & Antonopoulos, 2017; Mehdizadeh, 2018), in flood risk assessment modeling (Wang *et al*., 2015), as well as to model susceptibility to rain-induced landslides (Dou *et al*., 2019), rainfall-runoff modeling (Tokar & Johnson, 1999), configuration of rating curve relationships (Jain & Chalisgaonkar, 2000).

In reference to KNN, applied to water resources variables, we found investigations applied to the precipitation forecast (Huang, Lin, Huang, & Xing, 2017), multi-model ensemble predictions of precipitation and temperature (Ahmed *et al*., 2020), for real-time flood forecasting (Liu *et al*., 2020), with weather generating model (Sharif & Burn, 2007), as well as for wind energy prediction (Yesilbudak, Sagiroglu, & Colak, 2017) and data completion (Kowarik & Templ, 2016).

Since ML algorithms are a promising approach, this paper aimed to evaluate the k-nearest neighbor machine learning algorithms for forecasting the stream flows of the Ramis river, based on known data from the hydrological system (stream flows and precipitation), in order to contribute to the development of early warning systems and strengthening of the hydrological forecast.

Materials and methods

Study area

The area in which this study was carried out is the Ramis river basin (14 769.62 km^2), which extends from the Ramis hydrometric station to the eastern mountain range in the department of Puno, Peru (Figure 1) and is the hydrographic unit with the highest contribution of flows to the highest navigable lake in the world (Titicaca). The altitude of the basin is between 3 812 and 5 749 meters above sea level, with an average slope of 22 % and a length of the main river of approximately 321 km.

According to the climatic classification of Peru (SENAMHI, 2020), the basin under study has a predominantly rainy type of climate, with dry autumn and winter. The multi-year average precipitation of the basin is 700.1 mm (Fernández, 2017), with higher accumulated precipitation in summer (December-February), with a dry autumn and winter that make the difference to the dry season. The type of land cover, according to the annual classification of the international geosphere-biosphere program (IGBP) available in google earth engine (GEE), image collection ID MODIS/006/MCD12Q1 (Friedl & Sulla-Menashe, 2015), has 0.01 % tree cover, 96.86 % grasslands dominated by herbaceous plants (< 2 m), 0.03 % permanent wetlands, 1.68 % farmland, 0.13 % urban and urbanized lands, 0.01 % permanent snow and ice, 1.25 % areas arid and 0.02 % of water bodies.

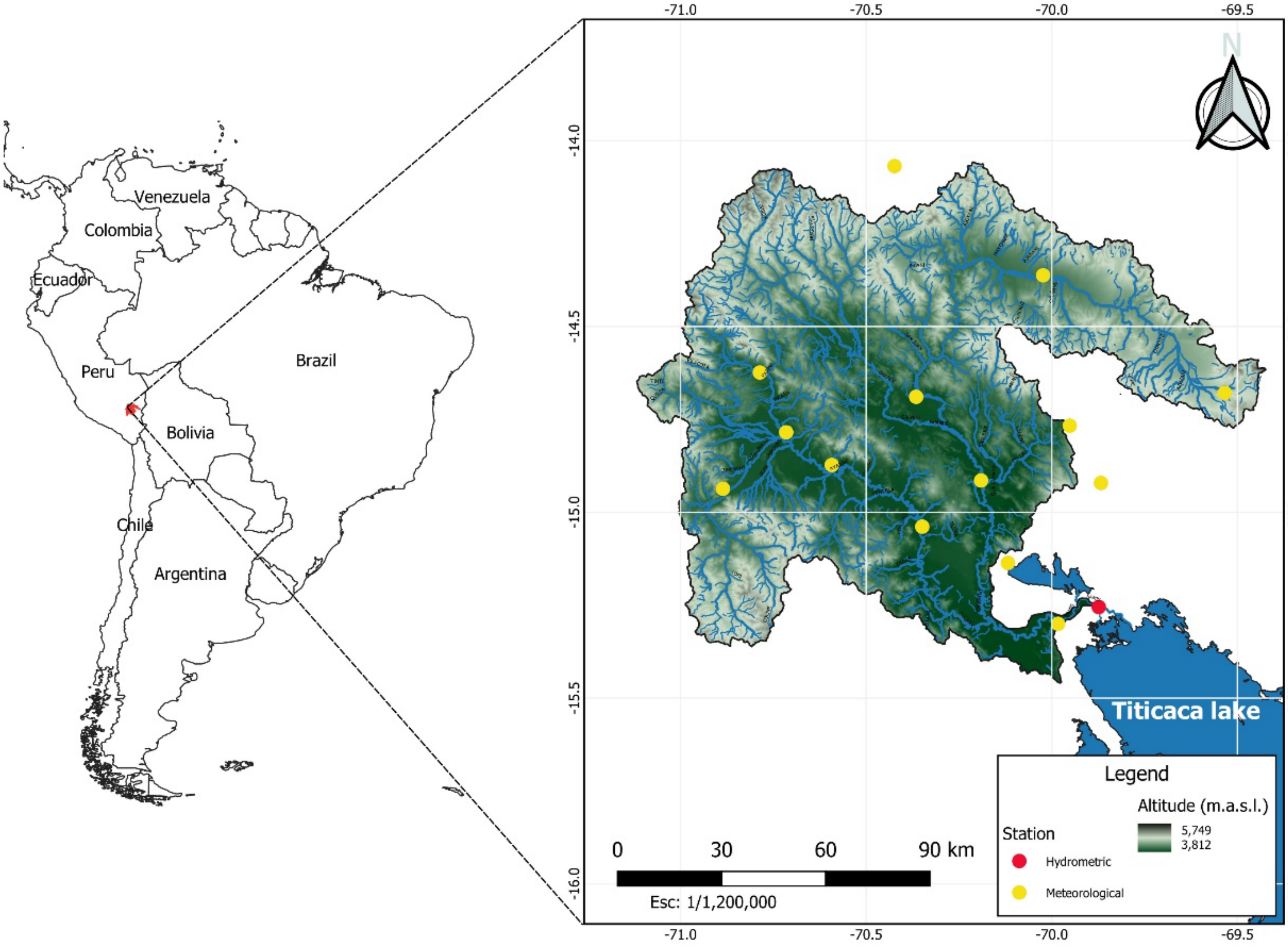


Figure 1. Location of the study area.

The daily time series of total precipitation in millimeters (mm) and mean flows in cubic meters per second (m³/s) were obtained from the

National Meteorology and Hydrology Service of Peru (SENAMHI) and the period of time used extends from September 1, 2005, to August 31, 2016. Figure 1 shows the location of the study area and the spatial distribution of 14 meteorological stations and one hydrometric station.

K-nearest neighbor (KNN)

The KNN algorithm is one of the simplest algorithms in the ML field, the idea is to memorize the training data set, and then make predictions of any new data taking as reference the data of its closest neighbors in the training set (Shalev-Shwartz, Science, Ben-David, & Science, 2013). Furthermore, KNN is a non-parametric method that can be used as a classifier (Gupta & Mittal, 2018) and a regressor (Hossny, Magdi, Soliman, & Hossny, 2020). The algorithm does not assume any type of equation or functional relationship between the input and the output (Joshi, 2020):

$$\hat{y} = \frac{\left(\sum_{i=1}^{k} y_i\right)}{k} \tag{1}$$

where $\hat{y}$ is the output value, y_i the ith nearest neighbor and k is the number of nearest neighbors.

Development the hydrological data-driven modeling

A previous and significant step in ML algorithms is the selection of the most important characteristics (variables) in order to obtain a more effective predictive model and avoid characteristics that do not contribute to the training of the model, and in this way reduce the time of training, reduce the complexity of the model, and reduce overfitting. Excessive use a large number of features in the model input leads to a perfect fit and causes the model to memorize the training data set and therefore lose generalization and obtains poor results in the validation stage (Remesan & Mathew, 2015).

There are several ways to measure the importance of characteristics (Pedregosa, Weiss, & Brucher, 2011; Remesan & Mathew, 2015), but we focus on the Pearson correlation coefficient (Tokar & Johnson, 1999) and the algorithm of importance of the permutation characteristic (Pedregosa *et al*., 2011).

The best way to incorporate input characteristics in a data-driven model is to take into account the lags of the data series (Mundher *et al*., 2016; Solomatine & Xue, 2004; Tokar & Johnson, 1999), it is like this, that the procedure is based on developing hydrological forecasting models that will use memory, that is, using retrospective flow and precipitation values to forecast the Q_t flow of the Ramis river.

So, in the first instance, the inputs have been selected based on a cross-correlation analysis between the input data set (precipitation and flow) with various lags and the output flows Q_t (Remesan & Mathew, 2015; Solomatine & Xue, 2004; Tokar & Johnson, 1999). Although the Pearson correlation technique is a suitable technique in linear systems, and the flow precipitation process is non-linear (Remesan & Mathew, 2015), its use is common and popular to select the appropriate inputs (Huang & Foo, 2002), since its foundation is to determine the strength of the relationship between the input time series and the output time series with various lags (Haugh & Box, 1977).

Consequently, to infer which characteristics have the greatest impact on flow forecasting, we use the permutation importance algorithm, implemented with the KNN predictive model with possible predictors and the predicted characteristic. The permutation importance algorithm is especially useful for nonlinear estimators and can be calculated in the training set or in the extended test or validation set when the data is tabular and a model score drop is indicative of how much the model depends on the characteristic (Pedregosa *et al*., 2011).

The importance i_j is calculated with:

$$i_j = s - \frac{1}{K}\sum_{K=1}^{K} s_{k,j} \qquad (2)$$

The permutation importance algorithm requires as input the adjusted predictive model, data set (training or validation). The reference score s of the fitted model is calculated with the data set (to verify the performance of the model, use precision in a classifier or R^2 for a regressor). For each characteristic j (column of the data set), for each repetition k in $1, \ldots, K$ randomly shuffle column j of the data set to generate a new version of the data set and calculate the $s_{k,j}$ score of the model fitted with the new version of the data set and keep K cases.

Hydrological data driven modelling

The training data must be large enough to contain the characteristics of the basin, on the contrary, an insufficient data set would not allow the model to generalize the patterns in physical phenomena (Tokar & Johnson, 1999).

The flow precipitation process was modeled using the KNN algorithm representing the current flow of the river Q_t, depending on the most important characteristics. The selection of the training data set (calibration) was considered 70 % (2808) of the total data (4012), while the remaining 30 % (1204) was considered for the test stage (validation). Rusli, Yudianto and Liu (2015) indicated that the calibration stage is carried out to understand the correlation that exists between the model parameters and the hydrological response of the basin and also to achieve the best agreement between the observed and simulated flows. To obtain the best hydrological forecast model (Equation (1)), the configurations with the most important characteristics (use of different lags / variables to model Q_t) determined by means of the correlation matrix and the permutation importance algorithm were trained and tested.

Goodness of fit metrics

The effectiveness of the models was evaluated using five different goodness of fit metrics (Table 1), mean absolute percentage error (MAPE), anomaly correlation coefficient (ACC), Nash-Sutcliffe efficiency (NSE), the efficiency of KGE' (Kling, Fuchs, & Paulin, 2012) and spectral angle (SA).

Table 1. Goodness of fit metrics

Metrics	**Ecuation¹**	**Optimal value**
Mean absolute percentage error (MAPE)	$MAPE = \frac{100}{n}\sum_{i=0}^{n}\left\lvert\frac{S_i - O_i}{O_i}\right\rvert$	0
Anomaly correlation coefficient (ACC)	$ACC = \frac{1}{n}\frac{(\sum_{i=1}^{n}(S_i - \bar{S})\,(O_i - \bar{O})}{\sigma_O\sigma_S}$	±1
Nash-Sutcliffe Efficiency (NSE)	$NSE = 1 - \frac{\sum_{i=1}^{n}(S_i - O_i)^2}{\sum_{i=1}^{n}(S_i - \bar{O})^2}$	1
Kling-Gupta efficiency (KGE')	$KGE' = \sqrt{(r-1)^2 + (\beta - 1)^2 + (\gamma - 1)^2}$ $\beta = \frac{\mu_S}{\mu_O};\ \gamma = \frac{CV_S}{CV_O} = \frac{\sigma_S/\mu_S}{\sigma_O/\mu_O}$	1
Spectral angle (SA)	$SA = arcos\left(\frac{\langle S, O\rangle}{\lVert S\rVert_2\lVert O\rVert_2}\right)$	0

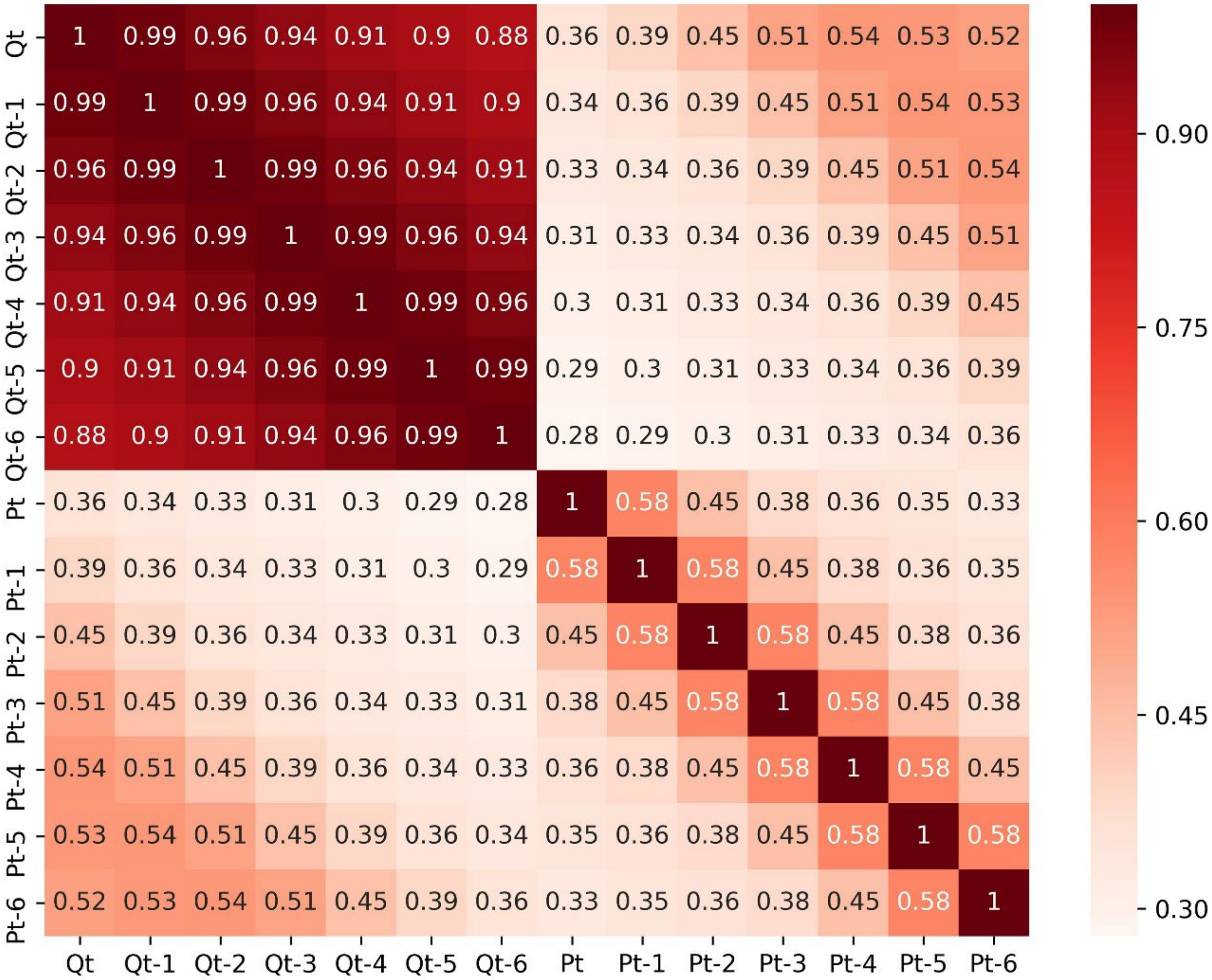
[1]Variables: S is the simulated value; $\bar{S}$ is the mean of the simulated values; O is the observed value; $\bar{O}$ is the mean of the observed value; σ is the standard deviation in m^3/s; r is the correlation coefficient between the simulated and observed value (dimensionless); β is the bias ratio (dimensionless); γ is the variability ratio (dimensionless); μ is the mean value in m^3/s; CV is the coefficient of variation (dimensionless), and the subscripts s and o represent observed and simulated values respectively.

MAPE calculates the mean absolute percentage error, and its range is 0 % ≤ MAPE ≤ inf, where 0 % indicates a lower percentage error and on the contrary indicates a higher percentage error in the data. Also, ACC is a common measure in the verification of spatial fields and measures the correlation between the variation pattern of the simulated values compared to those observed, the range varies -1 ≤ ACC ≤ 1, where -1 indicates a negative correlation, 0 indicates complete randomness, while 1 indicates a perfect correlation of the pattern of variation of the anomalies. NSE is a metric that uses the mean value as a benchmark and the range can vary from -inf < NSE <1, as the value approaches unity the better. On the other hand, KGE' (Kling *et al*., 2012) is the modified version of KGE (Gupta, Kling, Yilmaz, & Martinez, 2009) proposed to avoid cross-correlation between biases and variability relationships, the range can vary -inf < KGE' <1, values close to unity do not indicate bias. SA is an attractive measure to be used in the coincidence of spectra, it measures the angle between the two vectors in hyperspace and indicates how well the shape of the simulated and the observed series matches (not the magnitude), its range -π/2 ≤ SA <π/2 (π = pi), where values close to zero are better.

For this process, we use the Python Jupyter Notebook IDE and the hydrostats package that contains the metrics to characterize the errors between the simulated and observed time series (Roberts, Williams, Jackson, Nelson, & Ames, 2018).

Results and discussion

We find that the most important characteristic for the forecast of the stream flow Q_t, is the stream flow with a lag Q_{t-1} with a correlation coefficient equal to 0.99. As we widen the lag in Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5} and Q_{t-6}, the correlation coefficient decreases to 0.96, 0.94, 0.91, 0.90 and 0.88 respectively (Figure 2). If we analyze the relationship between precipitation and flow, the flow series Q_t and the precipitation series with lag P_{t-4}, have a higher correlation (*r* = 0.54), with respect to P_{t-1}, P_{t-2}, P_{t-3}, P_{t-5} and P_{t-6} with correlation coefficients of 0.39, 0.45, 0.51, 0.53 and 0.52, respectively. If a Q_t forecast model were developed based on Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5} and Q_{t-6} considering that they have higher values correlation, we would obtain a complex model in which characteristics that do not contribute to the training of the model would be included.



	Qt	Qt-1	Qt-2	Qt-3	Qt-4	Qt-5	Qt-6	Pt	Pt-1	Pt-2	Pt-3	Pt-4	Pt-5	Pt-6
Qt	1	0.99	0.96	0.94	0.91	0.9	0.88	0.36	0.39	0.45	0.51	0.54	0.53	0.52
Qt-1	0.99	1	0.99	0.96	0.94	0.91	0.9	0.34	0.36	0.39	0.45	0.51	0.54	0.53
Qt-2	0.96	0.99	1	0.99	0.96	0.94	0.91	0.33	0.34	0.36	0.39	0.45	0.51	0.54
Qt-3	0.94	0.96	0.99	1	0.99	0.96	0.94	0.31	0.33	0.34	0.36	0.39	0.45	0.51
Qt-4	0.91	0.94	0.96	0.99	1	0.99	0.96	0.3	0.31	0.33	0.34	0.36	0.39	0.45
Qt-5	0.9	0.91	0.94	0.96	0.99	1	0.99	0.29	0.3	0.31	0.33	0.34	0.36	0.39
Qt-6	0.88	0.9	0.91	0.94	0.96	0.99	1	0.28	0.29	0.3	0.31	0.33	0.34	0.36
Pt	0.36	0.34	0.33	0.31	0.3	0.29	0.28	1	0.58	0.45	0.38	0.36	0.35	0.33
Pt-1	0.39	0.36	0.34	0.33	0.31	0.3	0.29	0.58	1	0.58	0.45	0.38	0.36	0.35
Pt-2	0.45	0.39	0.36	0.34	0.33	0.31	0.3	0.45	0.58	1	0.58	0.45	0.38	0.36
Pt-3	0.51	0.45	0.39	0.36	0.34	0.33	0.31	0.38	0.45	0.58	1	0.58	0.45	0.38
Pt-4	0.54	0.51	0.45	0.39	0.36	0.34	0.33	0.36	0.38	0.45	0.58	1	0.58	0.45
Pt-5	0.53	0.54	0.51	0.45	0.39	0.36	0.34	0.35	0.36	0.38	0.45	0.58	1	0.58
Pt-6	0.52	0.53	0.54	0.51	0.45	0.39	0.36	0.33	0.35	0.36	0.38	0.45	0.58	1

Figure 2. Cross-correlations of Q_t with precipitation and flow lags

Then, for a definitive selection of the important characteristics for the model, the correlation analysis has been complemented by applying the algorithm of importance of the permutation characteristic (Figure 3) and it is confirmed that the flow with lag Q_{t-1} has a higher score (0.779) and is the most relevant characteristic to forecast Q_t, followed by Q_{t-2} (0.163) and the characteristics with less importance Q_{t-3} (0.046), Q_{t-4} (0.039), Q_{t-5} (0.033) and Q_{t-6} (0.037). Likewise, for precipitation with lag P_{t-3}, had a higher score (0.001) with respect to P_t, P_{t-1}, P_{t-2}, P_{t-4}, P_{t-5} and P_{t-6} with the lowest importance score (< 0.001).

Therefore, the model for the flow forecast was defined by a combination with the most important characteristics of precipitation and flow according to the results of the algorithm of importance of the permutation characteristic: 1) $Q_t = f(Q_{t-1})$, 2) $Q_t{:}\, f(Q_{t-1}, Q_{t-2})$, 3) $Q_t{:}\, f(Q_{t-1}, P_{t-3})$ and 4) $Q_t{:}\, f(Q_{t-1}, Q_{t-2}, P_{t-3})$, where Q_t is the flow to forecast, Q_{t-1} and Q_{t-2} are the flows lagged by 1 and 2 days, while P_{t-3} is the mean precipitation of the basin with a lag of 3 days. Although P_{t-3} is a less important characteristic with respect to Q_{t-1} and Q_{t-2}, we consider the model as an input variable, since Pedregosa *et al*. (2011) indicate that the characteristics that are considered of low importance could be very important for a good model and could increase the performance of the model.

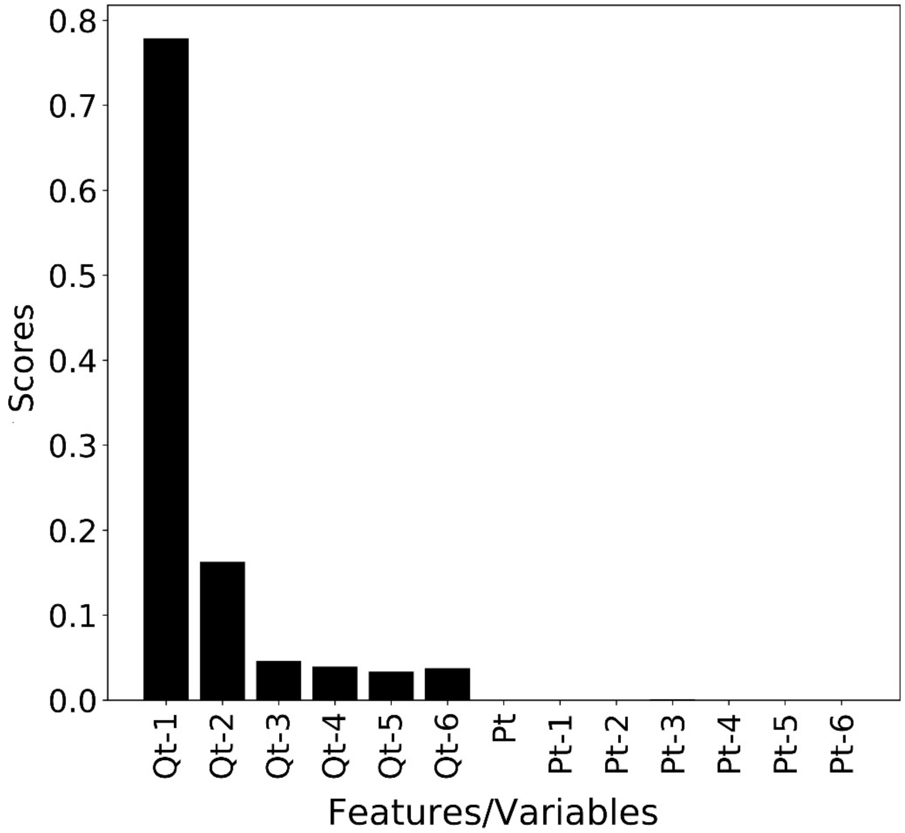


Figure 3. Importance scores of the permutation characteristic.

The results of the goodness of fit metrics for the KNN model, show the effectiveness for forecasting the flows of the Ramis river at the Ramis hydrometric station (Table 2). The predictive capacities lead to very high values of NSE (NSE = 0.96) in the validation stage, in particular $Q_t\text{:}\,f(Q_{t-1}, P_{t-3})$ is characterized by high values of ACC = 0.989, NSE = 0.979, KGE' = 0.988, lower error values (MAPE = 6.070 %) and a better match in the shape of the simulated and observed series (SA = 0.113).

Q_t: $f(Q_{t-1})$ is the least effective model showing values of ACC = 0.98, NSE = 0.965, KGE' = 0.982, error values (MAPE = 7.403 %) and a lower coincidence in the form of the simulated and observed series (SA = 0.145). Also Q_t: $f(Q_{t-1}, Q_{t-2}, P_{t-3})$ shows better performance than Q_t: $f(Q_{t-1}, Q_{t-2})$ and Q_t: $f(Q_{t-1})$, characterized by an MAPE = 6.230 %, ACC = 0.987, NSE = 0.975, KGE = 0.988 and SA = 0.122. For its part, Q_t: $f(Q_{t-1}, Q_{t-2})$, is identified by presenting better performance with respect to Q_t: $f(Q_{t-1})$, with values of ACC = 0.985, NSE = 0.972, KGE' = 0.985, error values (MAPE = 6.546 %) and a similar coincidence in the shape of the simulated and observed series (SA = 0.129). It should be noted that with a flow lagged by one day and precipitation lagged three days, the flow forecast model show a better performance compared to the models. This addition of P_{t-3} to the model is corroborated by Pedregosa *et al*. (2011) that although it is a characteristic that is considered of low importance, it produces better results and increases the performance of the model.

Table 2. Performance results of the KNN algorithm-validation stage.

Model	MAPE (%)	ACC	NSE	KGE'	SA
Q_t: $f(Q_{t-1})$	7.403	0.982	0.965	0.982	0.145
Q_t: $f(Q_{t-1}, Q_{t-2})$	6.546	0.985	0.972	0.985	0.129
Q_t: $f(Q_{t-1}, P_{t-3})$	6.070	0.989	0.979	0.988	0.113
Q_t: $f(Q_{t-1}, Q_{t-2}, P_{t-3})$	6.230	0.987	0.975	0.988	0.122

To continue with a further evaluation of the flow forecasting models, we present a series of scatter diagrams (Figure 4), in which an almost perfect fit between the observed and predicted values with the 45° line, especially for $Q_t : f(Q_{t-1}, P_{t-3})$ (Figure 4c), followed by the other models (Figure 4a, 4b and 4d). We can deduce that the data set chosen to train the KNN model, have the same statistical properties and therefore the estimated parameters do not significantly affect the forecast of Q_t in the validation period, thus, the MAPE values, ACC, NSE, KGE' and SA are similar in the training and validation period. A significant difference in the evaluation criteria of the goodness of fit in the training and validation set could correspond if the model is trained using a set of data that deviates greatly from the mean situation and significantly affects the forecast in the period of test (Antonopoulos & Antonopoulos, 2017). A set of training, validation, and test data with the same statistical properties, helps to develop the best possible model (Maier, Jain, Dandy, & Sudheer, 2010).

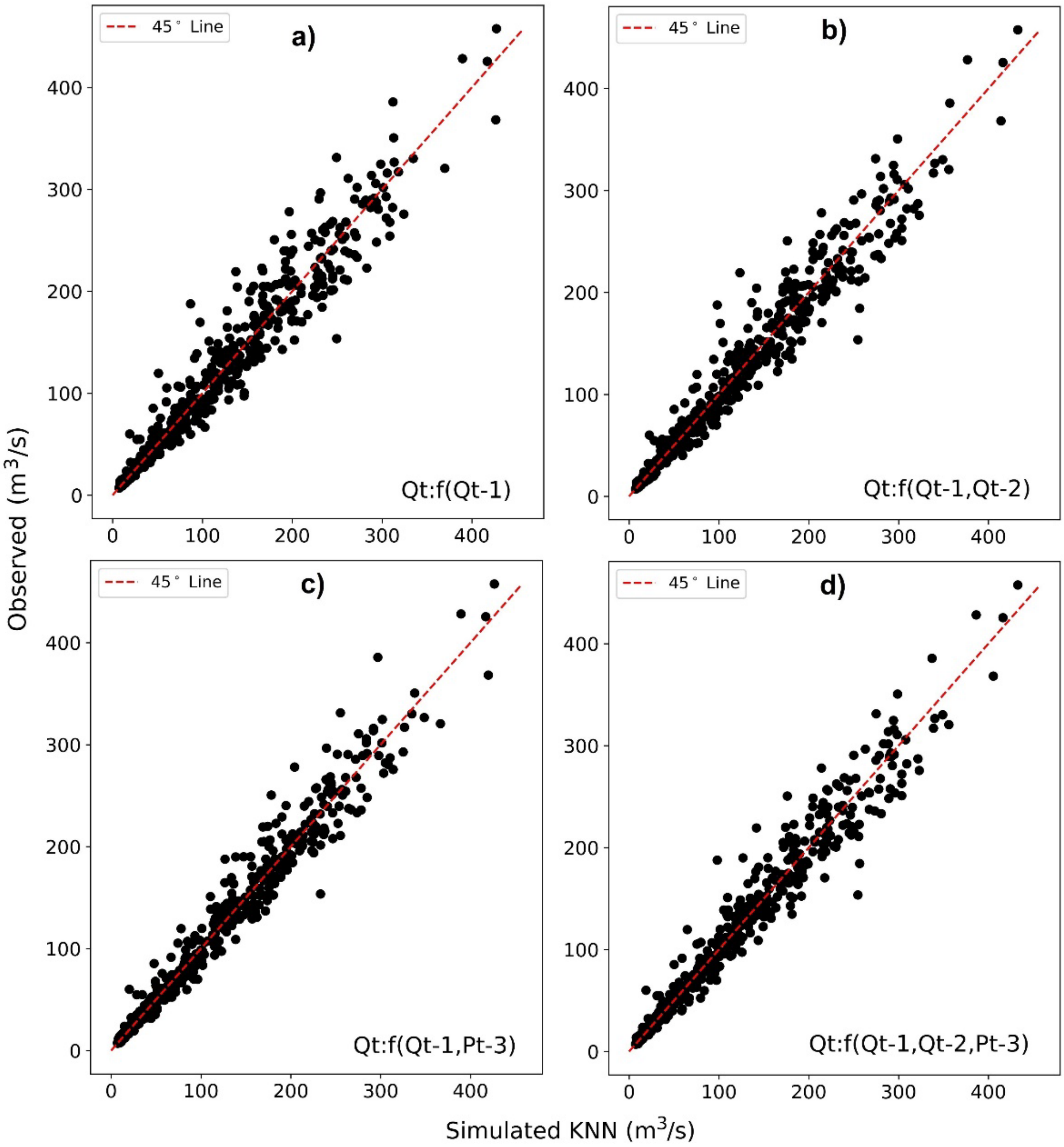


Figure 4. Scatter plot of simulated and observed flows.

Figure 5 illustrates the observed and simulated flow pattern with the KNN algorithm within the validation period. As observed, the data-driven forecast with KNN were able to closely match the actual values. The KNN algorithm is an effective tool for forecasting the daily flows of the Ramis river and has the advantage of directly providing Q_t based on past data, thus reducing investment in time and cost, which are required to implement hydrological models based on physically/conceptual hydrological processes.

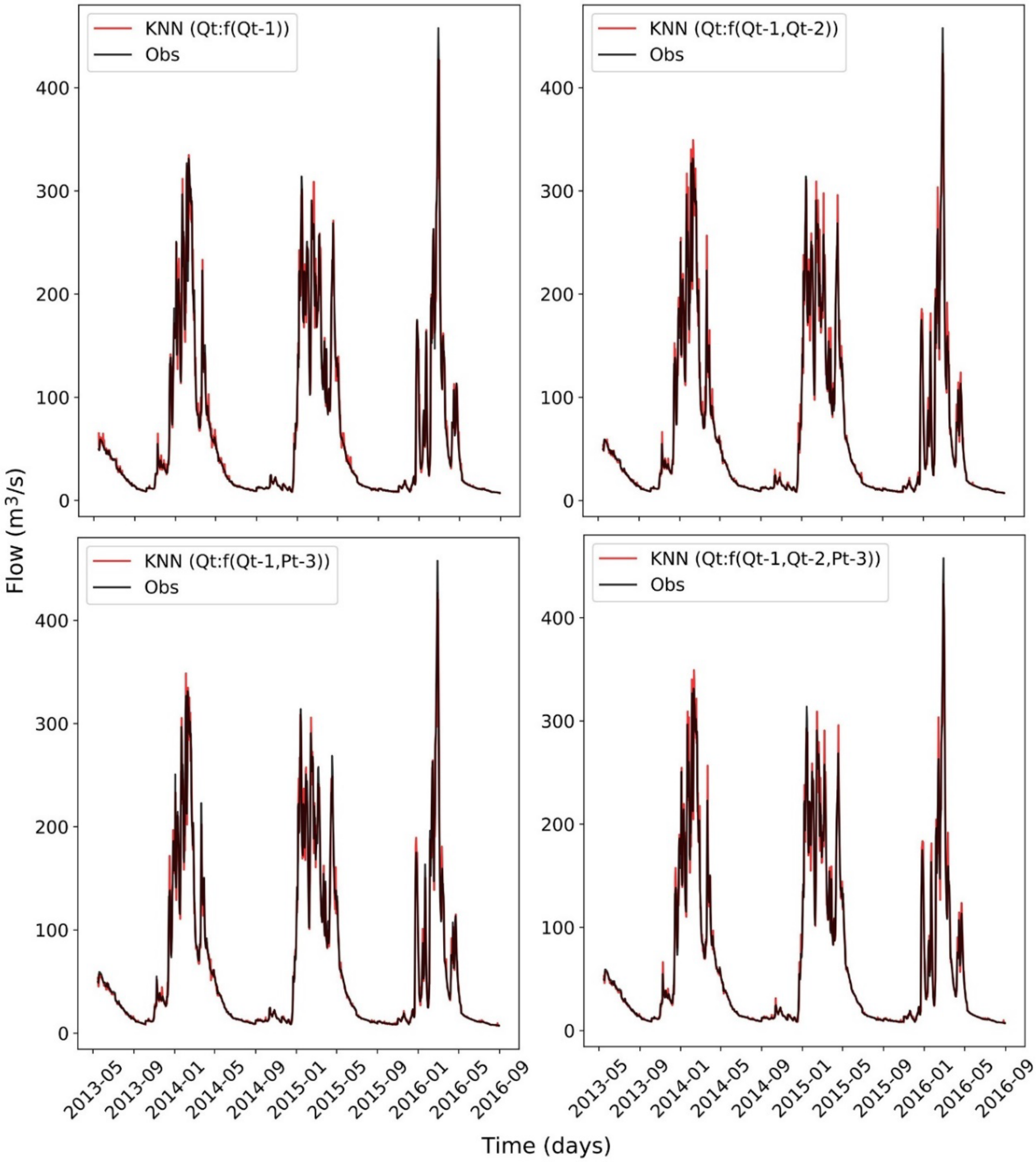


Figure 5. Hydrograph of time series observed and simulated with KNN algorithms-validation stage.

Conclusions

This study focused on hydrological modeling through the use of the KNN algorithm, exploring its applicability for forecasting mean daily flows of the Ramis river. The most important characteristics were selected in the first instance using Pearson correlation coefficient and supplemented by the importance of permutation characteristic algorithm. We found that Q_{t-1} is the most relevant characteristic for the Q_t flow forecast of the Ramis river at the Ramis hydrometric station, however, when we consider Q_{t-1} and P_{t-3} as input the model, the precision of KNN increases.

The research shows that the KNN algorithm would be a suitable approach for flow forecasting and can be integrated as an alternative for the strengthening of the daily hydrological forecast and implementation of an early warning system.

Acknowledgments

The authors wish to thank to Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI) for providing the climate information used in this research.

References

Ahmed, K., Sachindra, D. A., Shahid, S., Iqbal, Z., Nawaz, N., & Khan, N. (2020). Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. *Atmospheric Research*, 236, 104806. Recovered from https://doi.org/10.1016/j.atmosres.2019.104806

Alipour, A., Yarahmadi, J., & Mahdavi, M. (2014). Comparative study of M5 model tree and artificial neural network in estimating reference evapotranspiration using MODIS products. *Journal of Climatology*, 2014. Recovered from https://doi.org/10.1155/2014/839205

Antonopoulos, V. Z., & Antonopoulos, A. V. (2017). Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. *Computers and Electronics in Agriculture*, 132, 86-96. Recovered from https://doi.org/10.1016/j.compag.2016.11.011

Demolli, H., Dokuz, A. S., Ecemis, A., & Gokcek, M. (2019). Wind power forecasting based on daily wind speed data using machine learning algorithms. *Energy Conversion and Management*, 198(July), 111823. Recovered from https://doi.org/10.1016/j.enconman.2019.111823

Dou, J., Yunus, A. P., Tien, D., Merghadi, A., Sahana, M., Zhu, Z., Chen, C. W., Khosravi, K., Yang, Y., & Thai, B. (2019). Science of the total environment assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima volcanic island, Japan. *Science of the Total Environment*, 662, 332-346. Recovered from https://doi.org/10.1016/j.scitotenv.2019.01.221

Fernández, C. (2017). *Modelamiento hidrológico de la región hidrográfica del Titicaca*. Lima, Perú: Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI), Dirección de Hidrología.

Friedl, M., & Sulla-Menashe, D. (2015). *MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006* [Data set]. NASA EOSDIS Land Processes DAAC. Recovered from https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MCD12Q1/

Granata, F. (2019). Evapotranspiration evaluation models based on machine learning algorithms—A comparative study. *Agricultural Water Management*, 217, 303-315. Recovered from https://doi.org/10.1016/j.agwat.2019.03.015

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377, 80-91.

Gupta, V., & Mittal, M. (2018). KNN and PCA classifier with autoregressive modelling during different ECG signal interpretation. *Procedia Computer Science*, 125, 18-24. Recovered from https://doi.org/10.1016/j.procs.2017.12.005

Haugh, L. D., & Box, G. E. P. (1977) Identification of dynamic regression (distributed lag) models connecting two time series. *Journal of the American Statistical Association*, 72(397), 121-130.

Hossny, K., Magdi, S., Soliman, A. Y., & Hossny, A. H. (2020). Detecting explosives by PGNAA using KNN Regressors and decision tree classifier: A proof of concept. *Progress in Nuclear Energy*, 124, 103332. Recovered from https://doi.org/10.1016/j.pnucene.2020.103332

Huang, M., Lin, R., Huang, S., & Xing, T. (2017). A novel approach for precipitation forecast via improved K-nearest neighbor algorithm. *Advanced Engineering Informatics*, 33, 89-95. https://doi.org/10.1016/j.aei.2017.05.003

Huang, W., & Foo, S. (2002) Neural network modelling of salinity variation in Apalachicola River. *Water Research*, 36, 356-362.

Igual, L., & Seguí, S. (2017). *Introduction to data science: A python approach to concepts, techniques and applications*. New York, USA: Springer International Publishing.

Jain, S. K., & Chalisgaonkar, D. (2000). Setting up stage-discharge relations using ANN. *Journal of Hydrologic Engineering*, 5, 428-433. Recovered from https://doi.org/https://doi.org/10.1061/(ASCE)1084-0699(2000)5:4(428)

Joshi, A. (2020). *Machine Learning and Intelligence Artificial* (Springer). Recovered from https://doi.org/https://doi.org/10.1007/978-3-030-26622-6

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424-425, 264-277. Recovered from https://doi.org/10.1016/j.jhydrol.2012.01.011

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, 74(7). Recovered from https://doi.org/10.18637/jss.v074.i07

Laqui, W. F. (2010). Aplicación de redes neuronales artificiales a la modelización y previsión de caudales medios mensuales del río Huancané. *Revista Peruana Geo-Atmosférica RPGA*, 44(2), 30-44.

Liu, M., Huang, Y., Li, Z., Tong, B., Liu, Z., Sun, M., Jiang, F., & Zhang, H. (2020). The applicability of lstm-knn model for real-time flood forecasting in different climate zones in China. *Water (Switzerland)*, 12(2), 1-21. Recovered from https://doi.org/10.3390/w12020440

Liu, Z., Gilbert, G., Cepeda, J. M., Kydland, A. O., Piciullo, L., Hefre, H., & Lacasse, S. (2021). Modelling of shallow landslides with Machine Learning algorithms. *Geoscience Frontiers*, 12(1), 385-393. Recovered from https://doi.org/10.1016/j.gsf.2020.04.014

Lujano, E., Lujano, A., Quispe, J. P., & Lujano, R. (2014). Pronóstico de caudales medios mensuales del río Ilave utilizando modelos de redes neuronales artificiales. *Revista de Investigaciones Altoandinas*, 16(1), 89-100.

Madsen, H. (2000). Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. *Journal of hydrology*, 235(3-4), 276-288.

Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling and Software*, 25(8), 891-909. Recovered from https://doi.org/10.1016/j.envsoft.2010.02.003

Mehdizadeh, S. (2018). Estimation of daily reference evapotranspiration (ETo) using artificial intelligence methods: Offering a new approach for lagged ETo data-based modeling. *Journal of Hydrology*, 559, 794-812. Recovered from https://doi.org/10.1016/j.jhydrol.2018.02.060

Mendez, M., & Calvo-Valverde, L. (2016). Development of the HBV-TEC Hydrological Model. *Procedia Engineering*, 154, 1116-1123. Recovered from https://doi.org/10.1016/j.proeng.2016.07.521

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water (Switzerland)*, 10(11), 1-40. Recovered from https://doi.org/10.3390/w10111536

Mundher, Z., Ahmed, Y., & Abdulmohsin, E. H. (2015). RBFNN *versus* FFNN for daily river flow forecasting at Johor, Malasya. *Neural Computing and Applications*, 27(6). DOI:10.1007/s00521-015-1952-6

Mundher, Z., Jaafar, O., Deo, R. C., Kisi, O., Adamowski, J., Quilty, J., & El-Shafie, A. (2016). Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. *Journal of Hydrology*, 542, 603-614. Recovered from Recovered from https://doi.org/10.1016/j.jhydrol.2016.09.035

Pedregosa, F., Weiss, R., & Brucher, M. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(2011), 2825-2830.

Remesan, R., & Mathew, J. (2015). Hydrological data driven modelling. In: *Hydrological Data Driven Modelling*. Recovered from https://doi.org/10.1007/978-3-319-09235-5

Roberts, W., Williams, G. P., Jackson, E., Nelson, E. J., & Ames, D. P. (2018). Hydrostats: A Python package for characterizing errors between observed and predicted time series. *Hydrology*, 5(4). Recovered from https://doi.org/10.3390/hydrology5040066

Rusli, S. R., Yudianto, D., & Liu, J. T. (2015). Effects of temporal variability on HBV model calibration. *Water Science and Engineering*, 8(4), 291-300. Recovered from https://doi.org/10.1016/j.wse.2015.12.002

SENAMHI, Servicio Nacional de Meteorología e Hidrología. (2020). *Climas del Perú - Mapa de Clasificación Climática Nacional*. Lima, Perú: Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI), Dirección de Meteorología y Evaluación Ambiental Atmosférica.

Shalev-Shwartz, S., Science, C., Ben-David, S., & Science, C. (2013). *Understanding machine learning from theory to algorithms*. New York, USA: Cambridge University Press.

Sharif, M., & Burn, D. H. (2007). Improved K-nearest neighbor weather generating model. *Journal of Hydrologic Engineering*, 12(1), 42-51. Recovered from https://doi.org/10.1061/(ASCE)1084-0699(2007)12:1(42)

Solomatine, D. P., & Xue, Y. (2004). M5 model trees and neural networks: Application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering*, 9(6), 491-501. Recovered from https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491)

Thakur, B., Kalra, A., Ahmad, S., & Lamb, K. W. (2020). Bringing statistical learning machines together for hydro-climatological predictions - Case study for Sacramento San Joaquin River Basin, California. *Journal of Hydrology: Regional Studies*, 27, 100651. Recovered from https://doi.org/10.1016/j.ejrh.2019.100651

Tokar, B. A. S., & Johnson, P. A. (1999). Rainfall-Runoff modeling using artificial neural networks. *Journal of Hydrologic Engineering*, 4(3), 232-239. Recovered from https://doi.org/https://doi.org/10.1061/(ASCE)1084-0699(1999)4:3(232)

Veintimilla-Reyes, J., Cisneros, F., & Vanegas, P. (2016). Artificial neural networks applied to flow prediction: A use case for the Tomebamba River. *Procedia Engineering*, 162, 153-161. Recovered from https://doi.org/10.1016/j.proeng.2016.11.031

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527, 1130-1141. Recovered from https://doi.org/10.1016/j.jhydrol.2015.06.008

Xie, P., Li, T., Liu, J., Du, S., Yang, X., & Zhang, J. (2020). Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion*, 59, 1-12. Recovered from https://doi.org/10.1016/j.inffus.2020.01.002

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., Xia, Y., Xiao, J., Zhang, Y., Ma, Y., & Song, L. (2018). Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale. *Journal of Geophysical Research: Atmospheres*, 123(16), 8674-8690. Recovered from https://doi.org/10.1029/2018JD028447

Yesilbudak, M., Sagiroglu, S., & Colak, I. (2017). A novel implementation of kNN classifier based on multi-tupled meteorological input data for wind power prediction. *Energy Conversion and Management*, 135, 434-444. Recovered from https://doi.org/10.1016/j.enconman.2016.12.094