Environmental Challenges 10 (2023) 100680

Contents lists available at ScienceDirect

Environmental Challenges

journal homepage: www.elsevier.com/locate/envc

# Assessing a data-driven approach for monthly runoff prediction in a mountain basin of the Central Andes of Argentina

Sofia Andrea Teverovsky Korsic[a,c,d,*], Claudia Notarnicola[b], Marcelo Uriburu Quirno[a], Leandro Cara[e]

[a] *Comisión Nacional de Actividades Espaciales (CONAE), Buenos Aires, Argentina*
[b] *Institute for Earth Observation, EURAC Research, European Academy of Bozen, Bolzano, Italy*
[c] *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina*
[d] *Departamento de Posgrado, Doctorado en Ciencias Aplicadas, Universidad Nacional de Lujan, Buenos Aires, Argentina*
[e] *Instituto Argentino de Nivología, Glaciología y Ciencias Ambientales (IANIGLA) – CONICET, Mendoza, Argentina*

## ARTICLE INFO

*Keywords:*
Support vector regression
Runoff prediction
Remote sensing
Machine learning techniques
Snow cover area

## ABSTRACT

In the semi-arid Central Andes of Argentina, the water from snowmelt runoff plays a fundamental role as a provider of ecosystem services. Nowadays, the global climate change has an observable negative impact on this area, due, principally, to the decrease in both liquid and solid rainfall, with the consequent decrease in water availability. In this context, runoff prediction acquires vital importance for the integrated water resources management. The aim of this study is to investigate the performance of the Support Vector Regression (SVR) technique in predicting monthly discharges with 1-month lead-time in the Tupungato River basin in the Central Andes of Argentina. This methodology has never been applied before in this mountainous region. Different inputs, like meteorological data and satellite-based snow cover area estimates from MODIS, were analyzed in order to identify the suitable inputs predictors to forecast monthly streamflow. The results were compared against the results derived from a Classification and Regression Tree (CART) model and, also, against an Auto-regressive Integrated Moving-average (ARIMA) model. Different metrics were used to evaluate the performance of the SVR tests in reproducing streamflow observations at the basin outlet. The coefficient of determination for each of the analyzed tests lays between 0.75 and 0.89 in the validation set. The comparison with the other models showed a significant improvement in performance of SVR in respect of CART and ARIMA model. SVR models proved a promising approach to support water management and decision making for productive activities, potentially also in other basins in the region.

## 1. Introduction

The water supplied by the rivers draining the drier hillside of the Central Andes (i.e. the Argentine hillside) strongly determined the settlement of the populations. Irrigated areas (locally known as oases) cover 4.8% of the area while concentrate 98% of the population. In this regard, the Mendoza province has an extensive irrigation infrastructure, necessary to carry out the typical agricultural activities in this area (Abraham et al., 2005). These productive and industrial activities, also the local domestic consumption, are highly dependent on the availability of snowmelt water draining from the Andes range.

Besides, Mendoza's productive oases are areas vulnerable to climate change. Climate change has significant consequences, like the glaciers retreat, less surface of snow accumulation, loss in ice masses and, consequently, alterations of the hydrographs of the Andean rivers.

Accordingly, the problem of water scarcity in arid and semi-arid regions, not only in this region, but also in the world, encourages the exploration of new methodologies that improve the integrated water resources management.

Effective water resources management in the Central Andes requires access to accurate streamflow forecasting. In accordance with this, the knowledge of the streamflow regime is an important key to understand a wide range of problems and activities related to the river system, like hydropower generation or flood control, and, also, providing a better understanding of the runoff fluctuations due to climate change (Ferguson, 1999; Dong, 2018).

[*] Corresponding author at: Comisión Nacional de Actividades Espaciales (CONAE), Buenos Aires, Argentina.
*E-mail address:* skorsic@sec.conae.gov.ar (S.A.T. Korsic).

https://doi.org/10.1016/j.envc.2023.100680
Received 25 October 2022; Received in revised form 10 January 2023; Accepted 11 January 2023

Hydrological models attempt to capture the complex behavior of the variables involved in the streamflow production by using historical data. They can be expressed using probabilistic, deterministic or stochastic approaches, depending of the study purpose (Raghavendra and Deka, 2014). However, there are some constraints related to the spatial and temporal availability of hydrometeorological data in mountain areas like the Andes range, due to the difficulties in accessing, installing, and maintaining in situ networks. Therefore, the general scarcity of the in-situ data on these areas increases the relevance of remote sensing data, especially on the monitoring and variability fluctuations of the snow cover area (SCA).

The SCA has a large impact on these basins as an input for improving runoff forecasting. For instance, Maza et al. (1995) performed flow simulations in the Tupungato River basin using LANDSAT remote sensing data as an input to the deterministic SRM model with reliable performance.

The use of machine learning techniques for prediction of water availability has been expanded to successfully solve forecasting problems (Wang et al., 2009). The capability of these techniques for analyzing long series and large-scale data has increased the interest among researchers in water resources studies. Both parametric (e.g. Bayesian approach) and non-parametric methods, like Artificial Neural Network (ANN) or regression tree (e.g. CART), can handle the non-linearity between input and output variables, but the principal limitation is the need of a large training database for a robust performance (Bhattacharya and Solomatine, 2005; Apaydin et al., 2020).

In Argentina, Dolling and Varas (2002) developed a model based on ANN for monthly streamflow prediction (spring and summer flows). The model was tested on a mountain watershed of San Juan Province, using 30 inputs variables as precipitation, temperature, relative humidity, effective sunshine hours, maximum snow depth, etc. The results obtained had a better performance than alternatives procedures. Moreover, Pierini et al. (2012) contributed with a daily flow forecast using an ANN algorithm and historical runoff data in the Colorado River basin, improving the results obtained through the use of an autoregressive model (AR).

Within these machine learning techniques, the support vector machine (SVM) has proved to be much more robust in several applications related to forecasting of hydrologic time series (Maity et al., 2010; Callegari et al., 2015). SVM is a data-driven model based on learning systems that can handle different kinds of input, reaching a good performance even when few data are available. The application of this technique for regression, called support vector regression (SVR), was introduced by Vasnik in 1995 (Asefa et al., 2006) and improved by Drucker et al. (1997). Since then, the application of this technique in diverse research fields has attracted much attention.

Working with SVR has several benefits. Firstly, this method can manage different kinds of inputs and does not require to have a large amount of data to achieve a good performance, as in the case of other non-parametric methods like ANN or Genetic Algorithms, which need a large training dataset. The SVR technique was tested also in water resources engineering, rainfall-runoff modeling (Dibike et al., 2001) and bio-physical parameters retrieval (Pasolli et al., 2011). Concerning the runoff forecasting, several approaches have been presented for testing the capability of the SVR on hydrological applications (Asefa et al., 2006; Behzad et al., 2009). In the European Alps, Callegari et al. (2015) and De Gregorio et al. (2017) analyzed the performance of monthly river discharge forecasting using the SVR technique on several basins, exploiting diverse input features and improving the results obtained with simple linear alternatives. In a snow dominated watershed of Iran, Sedighi et al. (2016) investigated the capability of the SVR and ANN models to simulate the rainfall-runoff processes using remote sensing data as input.

This paper assesses the performance of the support vector regression (SVR) for the runoff prediction with 1-month lead time in a snow regime catchment of the Central Andes of Argentina, by taking advantage of the availability of remote sensing data. The lead time was selected based on the fact that 1-month lead forecasts are useful from the point of view of water managers for most of the local activities.

The objectives of the current study are:

(1) To develop a novel 1-month lead forecast model to predict the water resource availability in a mountain catchment.
(2) To assess a proper feature selection of the input parameters to be used in the complex environment considering the peculiarity of mountain areas.
(3) To test the efficiency of the machine learning approach in respect of other models

Although the methodology was tested in other catchments over the world, the originality of this research leads on the developing of 1-month lead forecasting using easily accessible variables. Instead, other works applied machine learning techniques using several variables that are usually unavailable, especially in less developed countries. This idea is based on the fact that in the Andes range the number of meteorological stations is limited and the need of assessing water availability is high. In this way, also the possibility to use remote sensing data on hydrological applications is outstanding.

The result of this research contributes to support the water management of the region, particularly for the development of productive activities like irrigation agriculture, hydroelectricity generation, and for local domestic consumption. In fact, the proposed approach is developed by using readily available observations of variables like air temperature, precipitation, discharge, and SCA from MODIS Terra and Aqua satellites, which allows the application to basins with limited ground measurements. Moreover, the method can be easily recalibrated to be adapted to different snow dominated catchments.

This method has not yet been tested in Argentinian basins and will potentially be a significant contribution to the water management of the region.

## 2. Study area and dataset

### 2.1. Study area

The selected area for this study is the Tupungato River basin in the upper part of the Mendoza River basin in the homonymous province, in Argentina (Fig. 1). The Mendoza River basin concentrates the largest part of the province population. With a drainage area of approximately 8035 km$^2$ and an elevation ranging from 2300 to 6500 masl, the upper basin of the Mendoza River (including the Tupungato basin) is the

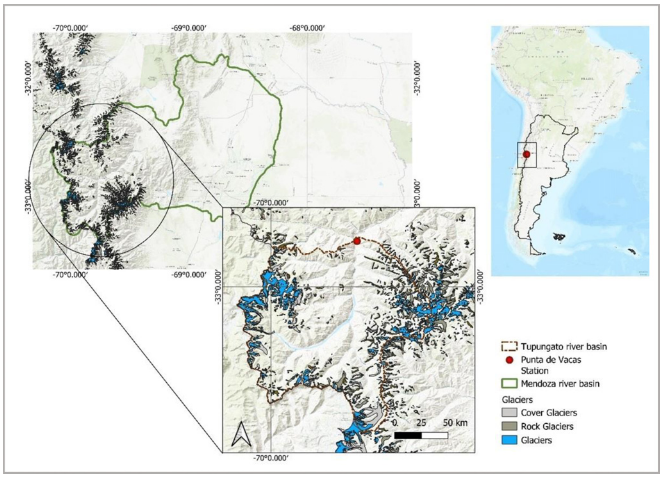

**Fig. 1.** Study Area. Top left: the map shows the entire Mendoza River basin. Top right: The location of the study area in Argentina. Center: the map shows the Tupungato River basin, used for this study.

main water supplier for the development of most of the agricultural, industrial, and human activities.

The hydrology is characterized by a snow-glacier regime. The largest amount of solid precipitation occurs during the southern winter months (from June to September). At the beginning of spring and during the warm season, the accumulated snow melts, increasing the flow rate of the river and shaping a unimodal hydrograph with a maximum value in December and January (Bruniard and Moro, 1994). Towards the end of the southern summer, the water contribution of the glaciers situated in the upper basin becomes more important, especially during dry years when the snow contribution decreases considerably (Masiokas et al., 2006).

### 2.2. Dataset

The dataset used for this study covers the period from December 2001 to June 2017. It includes time series of discharge, meteorological variables and snow cover area from MODIS as detailed below:

- Discharge time series (Q)

In this study, the discharge time series from Punta de Vacas station was used. This station is sited right upstream from the confluence of the Tupungato River with the Mendoza River, at an elevation of about 2500 masl. The data correspond to the monthly mean flow rates. In order to improve the results, monthly mean flow rates were processed to compute the average discharge of each one of the twelve months of the year (Qav) and used as a seasonal time series.

- Snow cover area (SCA)

The Snow Cover Area can be estimated from optical sensors due to the difference in reflectance between the visible and infrared bands of the spectrum. The most widely used index for snow detection is the Normalized Snow Difference Index (NDSI) (Riggs et al., 2015). This index is usually calculated at the top of the atmosphere , based on the digital numbers (DN) measured on the frequency of channel 2 and 6 (MODIS data).

The SCA map is a Boolean product derived from NDSI, using a threshold to identify the snow presence in each pixel of information per image (Bergeron et al., 2014, Roy et al., 2010).

The main issue about snow detection through optical sensors is the loss of information due to cloud coverage. In this work, we used a fully filled product derived from the combination of two identical optical sensors (MODIS Terra & Aqua) with a daily temporal resolution (one image per satellite per day). The mentioned product uses the Aqua satellite product (MYD10A1) to fill the Terra satellite product (MOD10A1) gaps, day by day. The information lost by both sensors is filled with data of previous days (Cara et al., 2016). Cara, 2018 checked the product accuracy using two ground stations of snow measurement (Toscas and Horcones Stations, in the Mendoza River basin), through a confusion matrix analysis. Obtaining a correct overall rating above 91% (94%) in 5039 (3063) images analyzed for Toscas and Horcones respectively (Cara, 2018).

- Meteorological data

The meteorological variables used were air temperature (T) and precipitation (PP), also from de Punta de Vacas station located in the basin outlet. Since the meteorological data were available at daily steps, an aggregation had to be performed in order to get the monthly mean air temperatures and the monthly cumulative precipitation values.

## 3. Short introduction to SVR theory

This section briefly presents some theoretical concepts of the SVR. A more detailed explanation can be found in Smola and Schölkopf, 2004. The idea of this machine learning technique is to select the regressive hyperplane that best fits the selected training data. For this, a margin distance is considered ($\varepsilon$), so that the samples are in a band or tube around the hyperplane, i.e. they are at a distance of less than $\varepsilon$ from the hyperplane. These samples will constitute the support vectors. The samples that fall outside the range are those that control the error made by the regression ($\zeta$) (Fig. 2).

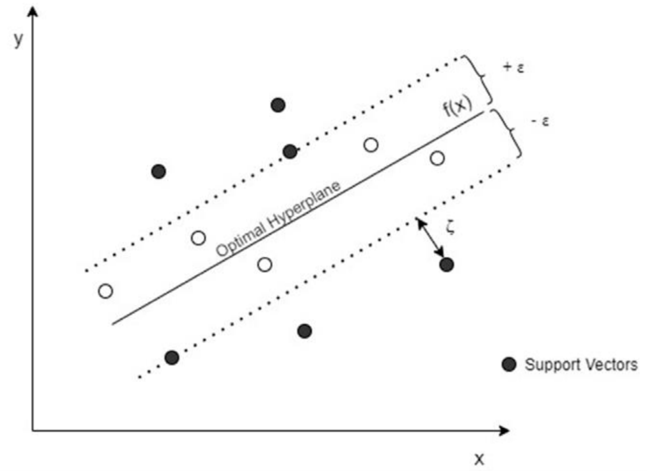

**Fig. 2.** Example of a linear vector regression. The black dots correspond to the support vectors, the optimal hyperplane corresponds to the regression line. $\varepsilon$ and $-\varepsilon$ are the size of the $\varepsilon$-insensitive tube.

In case of a linear function, the formula is the following:

$$\mathrm{f}(\mathrm{x}) = \langle w, x \rangle + b \tag{2}$$

Where $w$ are the support vector weights with the same dimensions as those of the explanatory variables x, and b is a scalar that is estimated based on the conditions of Karush-Kuhn-Tucker (KKT). The simplest function is found by minimizing the sum of absolute value of the errors above a certain threshold $\varepsilon$.

In this way, for the i$^{\text{th}}$ sample the $\varepsilon$-insensitive loss function ($\mathrm{L}_{\varepsilon \mathrm{i}}$) is introduced as:

$$L_{\varepsilon i} \ = \begin{cases} 0 & , \ if \ \left| f\left(x_i\right) - y_i \right| \le \ \varepsilon \\ \left| f\left(x_i\right) - y_i \right| - \varepsilon, & otherwise \end{cases} \tag{3}$$

In order to find f(x), the following function (called the regulated risk function) has to be minimized:

$$R_{reg}(f) = \frac{1}{2}\ w^2 + C\ \frac{1}{N} \left. \sum_{i=1}^{N} L_{\varepsilon i}(f \right) \tag{4}$$

In the previous equation, C is an indicator of the complexity of the function f(x) and determines the trade-off between this complexity and the tolerance of deviations larger than $\varepsilon$. This minimization illustrates the main idea of the structural risk minimization theory (Behzad et al., 2009).

In general, f(x) is not linear and has the following general form:

$$f(x) = \sum_{i=1}^{N} \propto_i \langle \Phi\left(x_i\right),\ \Phi(x) \rangle + b \tag{5}$$

where $\propto_\mathrm{i}$ (with $i = 1 \cdots \mathrm{N}$) are constant coefficients. Under certain conditions (Smola and Schölkopf, 2004) the scalar product $\Phi(x_i),\ \Phi(x)$ can be rewritten using nonlinear functions named "kernel". They can be represented as $K(x_{i\,,}\ x) = \Phi(x_i),\ \Phi(x)$. Some of the common kernels are: linear, polynomial, sigmoid and radial basis function (RBF).

## 4. Methodology

According to the presented objectives, the methodology can be summarized in three steps: (1) Feature selection and data splitting, (2) Setup of model parameters, (3) Implementation of SVR.

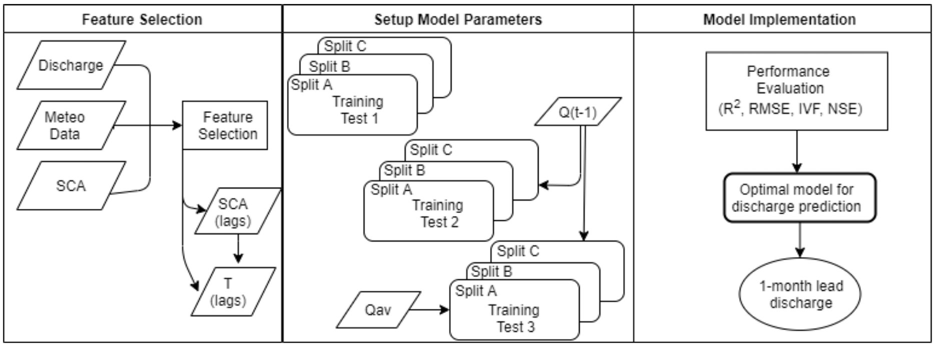

**Fig. 3.** Methodology flowchart. In the Feature Selection we selected the best variables (Discharge, Meteo Data and SCA) and their correspondent lags to build models. In the next step we setup de model parameters for all test (represented with the rectangle) and splits (A,B and C). Finally we implemented the three models and evaluated the performance.

To find the final model that best simulates runoff, different tests were carried out with different inputs (tests 1 to 9). The training and validation set were generated by considering three different splits of the total data set (the splits are named A, B and C). The aim of this procedure is to assess the influence of data selection on the result. With each split, several tests were carried out by exploiting different inputs, resulting in three different tests in each split.

Fig. 3 shows the scheme that was applied in the procedure to perform the test with the different splits and inputs.

### 4.1. Feature selection and data splitting

The selection of the model inputs in the SVR techniques values plays a significant role (Smola and Schölkopf, 2004).

Three methods were used to select the input variables of the different tests for model definition: linear model, correlation function and recursive feature selection (RFE). First, an exploratory analysis of the correlation between the observed streamflow and the SCA at different lags was carried out by analyzing the correlation function of the potential inputs. The linear model and the correlation function were applied to establish a first selection of input variables and to avoid redundancy in the input parameters, identifying the best T and SCA lags related to the runoff. Then, to improve the selection a RFE was done to the dataset, including the meteorological data (T, PP). The RFE is a popular wrapper method used to eliminate irrelevant features from the training set and so identify the best input combination of the model (Chandrashekar and Sahin, 2014).

Furthermore, to test the influence of the historical runoff data and to improve the overall prediction performance, the observed seasonal behavior described by the average monthly mean flow ($Q_{av}$) was included as an additional input, as well as the $Q_{t-1}$ series.

A usual strategy for model training is to split the data in two subsets, called training and validation set. The objective of the training set is to optimize the parameter vector by solving the algorithm problem. The k-fold cross validation technique is often used for split data (Chandrashekar and Sahin, 2014; Callegari et al., 2015). In this technique, the data is randomly divided in training and validation set. In this experiment, instead of this technique three combinations of training and validation sets were selected to enhance the training process. The idea behind these combinations is to include the whole range of streamflow fluctuations along the annual cycle, over the hydrological year. Hence, the splits were selected as follows:

- Split A: It was selected in a way that ensures that both high and low flows are included in the training and validation hydrographs, resulting in sets of ten and six years, respectively.
- Split B: The training set was composed by the first five years and the last six of the time series. The validation set included the five intermediate years.
- Split C: A simple split-sample method was applied, consisting of the first ten years for training and the last six for validation.

Finally, with the selected variables, three different tests (1,2,3) were implemented for each split (A, B, C) with different combinations of inputs.

### 4.2. Model parameter selection

For the implementation of the SVR it was necessary to tune the model parameters. For this purpose, a suitable kernel function and the corresponding hyper parameters of the model were selected. The kernel function is intended to perform a linear separation by transforming the data into a higher dimensional feature space (Langhammer and Česák, 2016). The performance of the SVR depends considerably on this selection (Behzad et al., 2009). Hence, in the training phase of the SVR, four types of kernel functions were compared: linear, polynomial, radial and sigmoidal. The coefficient of determination was calculated for each type of Kernel function, in order to select the most suitable.

Also, there are two basic versions of SVR to select: epsilon-SVR and nu-SVR. The difference between the two is that while the nu has control over how many data vectors from the data set becomes support vectors, the other version does not have this control.

Finding optimal hyper-parameters ($\gamma$, $\varepsilon$ and $\mathbb{C}$) is still a challenge and there is no compromise regarding the best parameters setting (Raghavendra and Deka, 2014).

For the purpose of tuning the parameters for the SVR model, a k-fold cross validation technique was applied on the training set of the three splits, and the best combinations of hyper-parameters were selected for each one during the model calibration (Kuhn, 2014).

### 4.3. Implementation of the SVR

The last step of this approach was the testing of different inputs for each split, and the evaluation of the model efficiency and the impact of including different variables. The e1071 R package was used to run all models (Meyer et al., 2019) with the selected parameters, as described in Section 4.1.

For each data split (A, B and C), three tests were run with the selected inputs.

For the model performance evaluation, the following metrics were used: the coefficient of determination ($R^2$), the mean average error (MAE), the Root Mean Square Error (RMSE), and the Index of Volumetric Fit (IVF). The IVF computes the degree of volumetric agreement between the observed and the modeled flows (Tan and O'Connor, 1996).

Another performance criterion used to express the model accuracy is the Nash Sutcliffe Efficiency coefficient (NSE) (Nash and Sutcliffe, 1970). This coefficient is dimensionless and ranges from minus infinity (poor model) to one (perfect model). A unit value of NSE means a perfect performance. A value of zero indicates that the measured mean ($\overline{Q}$) is as good a predictor as the model, while negative values indicate that the measured mean is a better predictor than the model itself. It is known that the NSE coefficient is sensitive to extreme values. Given that water resource management is particularly interested in periods of low flows, the Nash-Sutcliffe efficiency was also calculated on the logarithmic transformation of modeled and observed discharge series (NSE(ln)). As a result, the peaks are flattened while the low flows are kept at approximately the same level, so that the influence of low flows is increased in comparison to that of the flow peaks.

### 4.4. Comparison with the CART algorithm

In order to compare the SVR results with another machine learning technique, a Classification and Regression Tree (CART) algorithm was implemented. The CART algorithm is a non-parametric approach that builds a decision tree through repeated dataset division. The algorithm is constructed by splitting the data set using all predictor variables. The splitting rule is based on the squared residuals minimization algorithm, decreasing the Gini coefficient and variance (Lee and Kim, 2020). The primary output of CART is a hierarchy binary structure, which classifies the data set into groups. Further information can be found in Breiman, et al., 2014.

For this comparison purpose, the *rpart* library of the R software was used to applicate the CART algorithm to the dataset (Therneau and Atkinson, 2019). The algorithm was run for all splits and inputs. The analysis of variance (ANOVA) method was used to grow the tree. For all tests, the principal node of the tree selected by the algorithm was $SCA_{(t-5)}$.

### 4.5. Comparison with an ARIMA model

The Autoregressive Integrated Moving Average (ARIMA) model takes into account the dependence of the previous measurements in a moving-average form. In order to choose the appropriate ARIMA model, the first step was to calculate the autocorrelation function (ACF) and the partial autocorrelation function (PACF) to identify the order of the autoregressive model. The final model was chosen using the auto.arima function, included in the forecast library in R (Hyndman and Khandakar, 2008).

In this case, the training set was composed by data from December 2001 to June 2011 and the testing set from August 2011 to July 2017 (in correspondence with split C).

Afterwards, the residuals were analyzed, and the Ljung-Box test was performed to validate the model with suitable results, confirming the accuracy of the ARIMA model order.

## 5. Results

### 5.1. Feature selection

The feature selection was an important step to understand the contribution of each input feature in the model and to select the most significant variable to build different tests and evaluate the performance of the SVR. The correlation function between the streamflow and the lagged SCA presents a maximum at a 5-month ($R^2$=0.67). Also, a linear model regression was analyzed with all input variables and lags ($p < 0.05$). In this case, the $SCA_{t-5}$ resulted also the best predictor ($p = 0.0005$), in conjunction with the 1-month lag temperature ($p$-value=0.0003), and the 2-month lag SCA ($p = 0.02$).

The results obtained with the recursive feature elimination algorithm (RFE) showed that the SCA with a 5-month ($SCA_{t-5}$) lag was the most explicative variable. The precipitation did not show a correlation with the discharge; therefore it was discarded.

Thus, considering these results, different experiments were done with the inputs features and the addition of historical runoff data ($Q_{(t-1)}$ and $Q_{(av)}$), resulting in the following tests:

Test 1: $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$
Test 2: $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5),}$ $Q_{(t-1)}$
Test 3: $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5),}$ $Q_{(t-1),}$ $Q_{(av)}$

**Table 1**
Coefficient of determination ($R^2$) of the different kernel functions and two types of regression analyses, Eps-regression and nu-regression.

| | Eps - Regression | | | Nu- Regression | | |
|---|---|---|---|---|---|---|
| Type of Kernel | Split A | Split B | Split C | Split A | Split B | Split C |
| Linear | 0.51 | 0.57 | 0.55 | 0.60 | 0.60 | 0.58 |
| Polynomial | 0.57 | 0.49 | 0.58 | 0.59 | 0.52 | 0.60 |
| Radial | **0.76** | **0.75** | **0.71** | **0.76** | **0.75** | **0.71** |
| Sigmoidal | 0.55 | 0.51 | 0.11 | 0.14 | 0.11 | 0.15 |

### 5.2. SVR model parameters

As explained in Section 4.2, a type of kernel function and model selection parameter was performed.

Table 1 illustrates the coefficient of determination ($R^2$) of the kernel function tested in the training set (split C), assessing the epsilon-regression and nu-regression performances.

Although the $R^2$ have similar values in the nu and epsilon-regression, in the Radial function the epsilon-regression was selected. This regression method is one of the most commonly used modeling methods (Sedighi et al., 2016).

Concerning the analysis of the different types of kernel functions, we concluded that the best performance was obtained with the radial function (or RBF) (Table 1). Thus, all simulations were done using this kernel function, whose mathematical form is:

$$K(x, z) = \exp(\gamma x - z^2) \tag{6}$$

In the previous equation, $\gamma$ is a user-defined constant. The sigma parameter is part of the selected radial function. The C parameter represents the trade-off between the flatness of the regression function and the quantity up to which deviations greater than a predetermined error are tolerated.

### 5.3. Model implementation

As it can be seen in Table 2, the coefficient of determination fluctuates between 0.75 and 0.89, where the test C.3 has the best performance, with also lower values of RMSE and MAE, and the best NSE coefficient. The best IVF was reached by the tests A2 and A.1. The NSE(ln) values resulted between 0.82 and 0.97, improving the NSE values on the validation set, except on split C.

### 5.4. Comparison with the classification and regression tree (CART) model

Fig. 4 summarizes the results of the SVR in comparison with the CART model for all tests.

It can be seen that the SVR model has a better performance for all tests and splits than the RT model. The test B.1 results on similar values for the three metrics and we do not observe a change or an improvement with the addition of the $Q_{(t-1)}$ and $Q_{av}$, as it can be seen on the SVR model.

### 5.5. Comparison with ARIMA

The resulted ARIMA model had a (1,0,1) x (2,1,0) structure. This result was consistent with the fact that the time series has a seasonal component, which explains why a seasonal ARIMA model was used. Furthermore, the results were in accordance with what can be observed in the ACF and PACF, which shows significant values at lags of an integer number of 12 months (Fig. 5). Table 3 shows the different performance metrics for this model. The selected training and validation sets were the same as those used by the C split with the SVR and CART model.

Finally, with the purpose of comparing the temporal behavior of all the models with the observed data, Fig. 6 presents the hydrographs re-

**Table 2**
Performance metrics for all tests (1, 2, 3) and splits (A, B, C) in the training set (T) and in the validation set (V).

| Test | Inputs | T/V | $R^2$ | RMSE | MAE | IVF | NSE | NSE(ln) |
|---|---|---|---|---|---|---|---|---|
| A.1 | $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$ | T | 0.81 | 9 | 5.35 | 0.95 | 0.79 | 0.85 |
| | | **V** | **0.75** | **10** | **6.47** | **1.02** | **0.75** | **0.82** |
| A.2 | $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$, | T | 0.89 | 7 | 4.19 | 0.96 | 0.88 | 0.90 |
| | $\mathbf{Q_{(t-1)}}$ | **V** | **0.83** | **9** | **5.06** | **1.01** | **0.82** | **0.89** |
| A.3 | $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$, | T | 0.92 | 6 | 3.45 | 0.95 | 0.91 | 0.94 |
| | $Q_{(t-1)}$, $\mathbf{Q_{(av)}}$ | **V** | **0.89** | **7** | **4.34** | **0.96** | **0.87** | **0.93** |
| B.1 | $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$ | T | 0.76 | 10 | 6.05 | 0.96 | 0.75 | 0.85 |
| | | **V** | **0.77** | **10** | **6.00** | **0.89** | **0.75** | **0.81** |
| B.2 | $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$, | T | 0.85 | 8 | 4.90 | 0.96 | 0.84 | 0.90 |
| | $\mathbf{Q_{(t-1)}}$ | **V** | **0.89** | **7** | **4.34** | **0.91** | **0.86** | **0.87** |
| B.3 | $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$, | T | 0.90 | 7 | 3.92 | 0.97 | 0.90 | 0.93 |
| | $Q_{(t-1)}$, $\mathbf{Q_{(av)}}$ | **V** | **0.86** | **8** | **4.39** | **0.95** | **0.86** | **0.87** |
| C.1 | $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$ | T | 0.83 | 10 | 5.90 | 0.94 | 0.81 | 0.87 |
| | | **V** | **0.77** | **9** | **6.53** | **1.22** | **0.72** | **0.70** |
| C.2 | $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$, | T | 0.89 | 8 | 4.60 | 0.98 | 0.89 | 0.91 |
| | $\mathbf{Q_{(t-1)}}$ | **V** | **0.83** | **7** | **4.64** | **1.09** | **0.84** | **0.81** |
| C.3 | $T_{(t-1)}$, $SCA_{(t-2)}$, $SCA_{(t-5)}$, | T | 0.91 | 7 | 4.11 | 0.97 | 0.91 | 0.93 |
| | $Q_{(t-1)}$, $\mathbf{Q_{(av)}}$ | **V** | **0.89** | **5** | **4.08** | **1.08** | **0.90** | **0.87** |

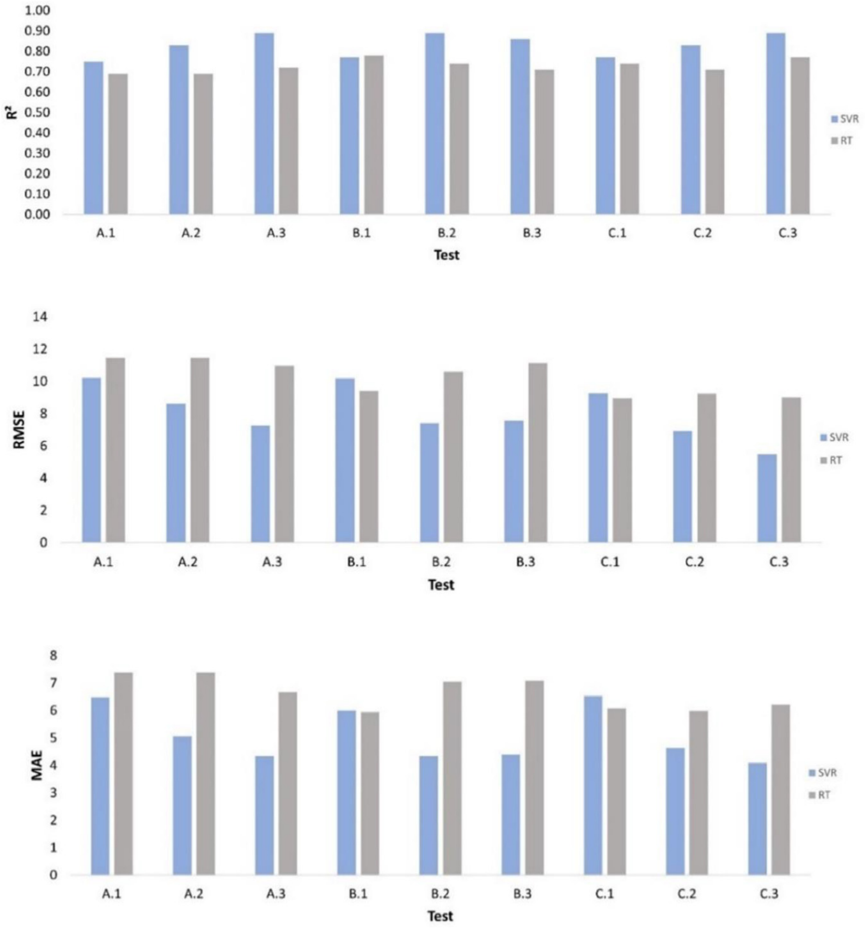

**Fig. 4.** Coefficient of determination ($R^2$), RMSE and MAE of all splits and models. SVR model in blue and RT (CART) model in gray.

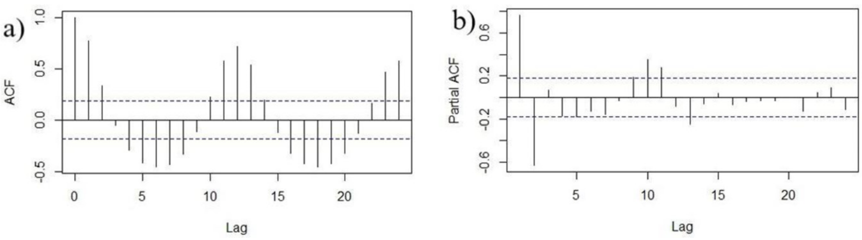

**Fig. 5.** (a) The autocorrelation function of the runoff series (b) The partial autocorrelation function of runoff series in Punta de Vacas Station.

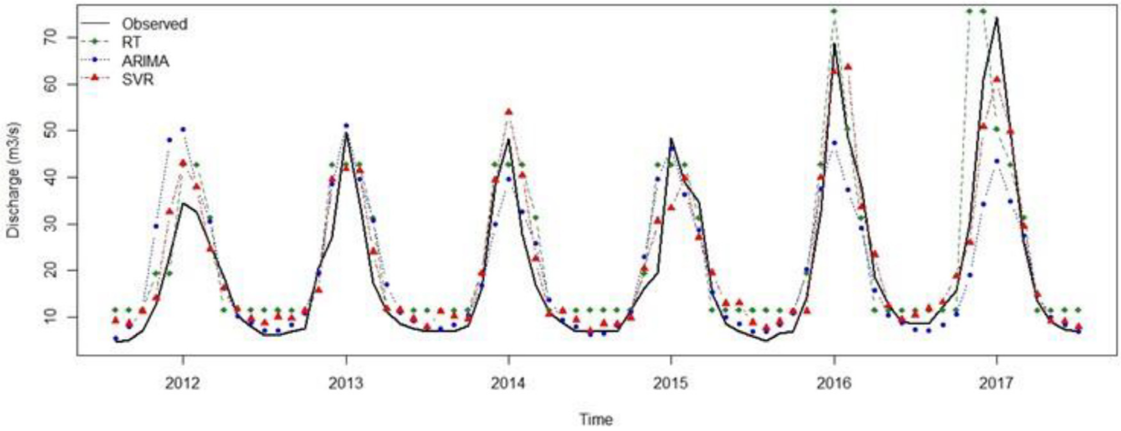

**Fig. 6.** Comparison of flow hydrographs: observed (black line), SVR (red line) ARIMA (blue line), RT (green line).

**Table 3**
Performance metrics of the ARIMA model with fitted parameters.

| ARIMA Model | $R^2$ | RMSE | MAE | IVF | NSE |
|---|---|---|---|---|---|
| (1,0,1) x (2,1,0) | 0.72 | 12 | 7.83 | 1.02 | 0.72 |

sulting from the application of the SVR model, RT model (Test C.3) and the ARIMA model over the validation period.

## 6. Discussion

### 6.1. SVR model for 1-month lead forecasting

Several aspects shall be considered to evaluate the performance of the previous tests. The selection of the performance metrics depends on the applicability of the forecast. In this work we used some regression metrics ($R^2$, RMSE and MAE), but we also added hydrological metrics. To make a valuable contribution to the water management, it is necessary to estimate the hydrograph peaks and the time correctly, and to have an idea of the volume of the available water.

As it can be seen in Table 2, the results obtained in the training set, for all the tests and metrics, were slightly better than the validation set. In particular the IVF, that represents the volumetric agreement with the observed data, shows values between 0.89 and 1.22 for all models and splits.

Although model A.1 showed the lowest coefficient of determination value (0.75), the IVF was 1.02, which means a volume overestimation of only 2%. This means that a satisfactory performance was obtained also in the test that performed worst. In the other extreme, test C.1 reached the worst IVF (1.22, i.e. an overestimation of 22%) and an $R^2$ of 0.77.

Furthermore, the NSE in the validation set for all models exceeded 0.75, while two thirds of them exceeded 0.80. The highest NSE was 0.90 (test C.3). The other performance metrics ($R^2$, RMSE and MAE) showed consistency with the described results and are in agreement with experiments done in other catchments with this technique, using different model inputs (Callegari et al., 2015; Guo et al., 2011; Maity et al., 2010; Zhang et al., 2018). In terms of NSE(ln), performances were high, even slightly higher than NSE in most of the cases, which is desirable in situations in which low flows are more critical than peak flows, as is the case of many water resource management applications.

### 6.2. Feature selection input parameters

The feature selection has an important role on the final forecasting performance. This was also verified by other studies with diverse catchment characteristics, where atmospheric patterns (Asefa et al., 2006; Callegari et al., 2015), rainfall (Chanklan et al., 2018) or physical catchment attributes (Zhang et al., 2018) have an strong influence on the runoff. Nevertheless, in this study only the temperature, SCA and historical runoff information were considered. The feature selection results showed a low correlation between the precipitation and the discharge, therefore this variable was discarded as input for the model. The SCA with a lag time of 5 months resulted to have the strongest influence on the runoff. This time lag is in concordance with climatological studies over the arid and semi-arid regions of the Andes range, that evidenced the correlation between snow cover and runoff in climate studies over the area (Masiokas et al., 2010).

### 6.3. Comparison with other methods

Most of the literature related to hydrological forecasting with machine learning techniques compared the SVR with ANN (Chanklan et al., 2018; Guo et al., 2011; Sedighi et al., 2016; Wang et al., 2009). In this work, we chose a CART model in order to compare two types of supervised machine learning techniques for regression.

Compared with the CART, the SVR showed a better performance in all cases, showing an improvement of 75% or higher.

In Fig. 4 we can observe that the addition of $Q_{(t-1)}$ and $Q_{av}$ did not have an impact on the performance metrics of the RT. The $R^2$ resulted between 0.69 and 0.78 without significant variations. In contrast, this addition significantly improves the results on the SVR in all splits, observing an increase in $R^2$ and a notable decrease in RMSE. Both variables have a good correlation with the streamflow, which explains the expected results.

Furthermore, the selection of the splits (A, B and C) did not have a significant impact on the model performance. Indeed, the quality of the results was virtually the same irrespective of the splits, which is a good indicator of the robustness of the SVR model.

Fig. 7 illustrates the results of three selected tests (A.1, B.2 and C.3) compared with the observed data, and the correlation between the discharge estimated with the models and the observations, in order to contrast the shape of the hydrographs, their peaks, recessions and time shifts.

The SVR model forecast does not match the largest peak discharges (underestimation of about 30–40%), except for the C.3 model, the one that performed best. It should be noted that the split does not contain the extreme discharge value of 100 $m^3/s$, like split A and B.

However, all SVR models respected the temporal phase very closely, without significant time shifts over the entire validation period. Also, it is important to highlight that the IVF resulted between 0.95 and 1.22, a performance of volumetric prediction that is acceptable for water resources management

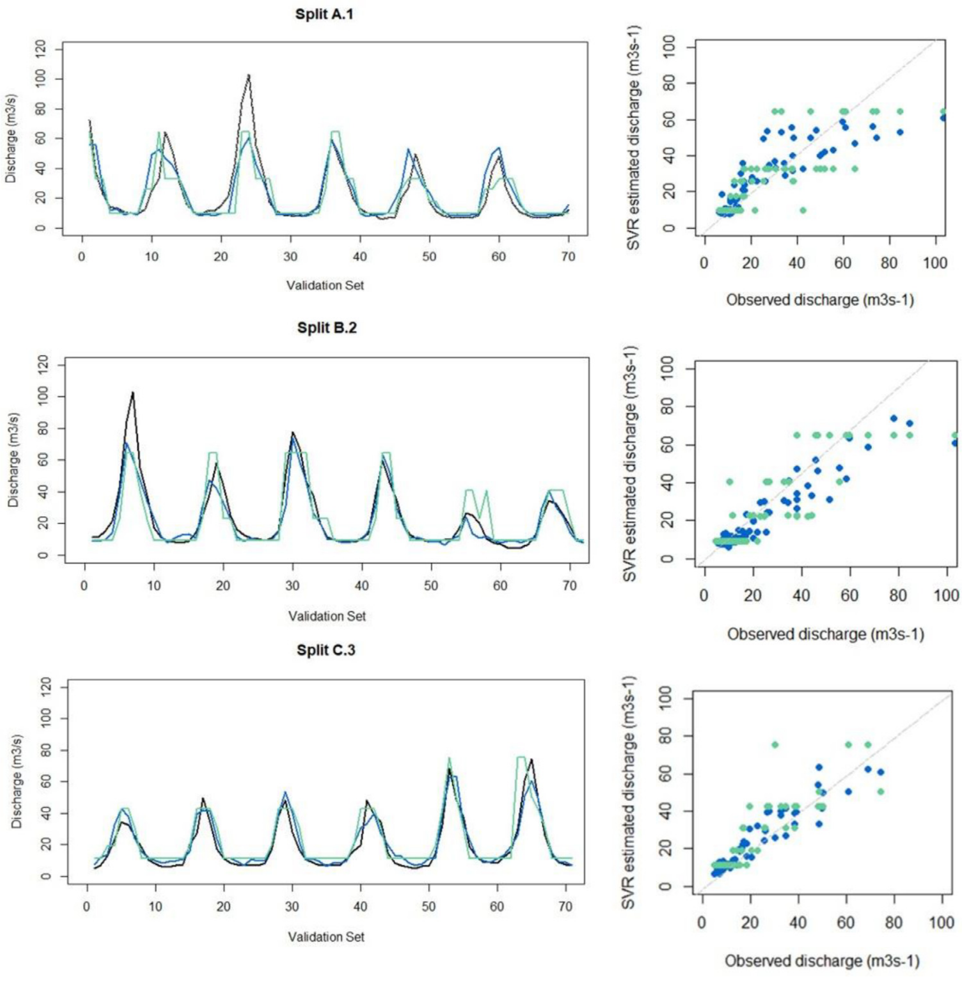

**Fig. 7.** Left: Hydrographs of the different validation set (months) compared with the observed discharge. Blue line: SVR model; green line: RT model and black line: observed discharge. Right: Respective scatter plots. Blue dots correspond to the discharge estimated with the SVR model and green dots correspond to the discharge estimated with the RT model.

Regarding the ARIMA model used as a benchmark, the results of all SVR models were exceeded, as is shown in Table 3.

It is worth noting that these results were obtained for a 1-month lead time forecasting. However, it is also possible to do different tests over different lead times, in order to find model structures for longer lead-time forecasts. A test was done with the same three splits used for the 1-month lead forecasting and using the variables with no less than two months of lag as inputs. In this case, $T_{(t-2)}$, $SCA_{(t-2)}$ and $SCA_{(t-5)}$ were used. The results (Table 4) show also good agreement for a 2-months lead time, in the three different splits. The results (Table 4) show also good agreement for a 2-months lead time, in the three different splits. As also observed in the 1-month lead case, the split selection did not show a significant effect on the model performance.

The present study leaves room for further application to longer lead times. For several productive uses of the water resource, longer lead times allow for an enhanced decision-making process, with the potential for an optimized water management.

**Table 4**
Results of a 2-months lead forecast model.

| Model inputs: $T_{(t-2),}$ $SCA_{(t-2)}$, $SCA_{(t-5)}$ | | | | | |
|---|---|---|---|---|---|
| Split | $R^2$ | RMSE | MAE | IVF | NSE |
| A | 0.75 | 10.22 | 6.47 | 1.02 | 0.75 |
| B | 0.77 | 10.19 | 6.00 | 0.89 | 0.75 |
| C | 0.77 | 9.27 | 6.53 | 1.22 | 0.72 |

## 7. Conclusion

In this paper, the potential of the Support Vector Machine for flow forecasting was explored in a mountain basin with snowmelt regime in the Central Andes range of Argentina. The performance results indicate that the SVR outperforms the Classification and Regression Tree (CART) model and the auto-regressive model used as a benchmark. The results are encouraging and demonstrate that this technique is promising in flow forecasting.

The selected inputs proved to be efficient, not only for 1-month lead forecasting, but also for 2-months lead time.

The SVR has several advantages over other forecasting techniques. Once the parameters are calibrated, the model is easy and fast to run. Moreover, there is no need to have a large amount of data to obtain suitable results, as it is the case with artificial neural networks or other machine learning data-driven models. This characteristic is particularly important in mountainous basins, where harsh weather conditions and high altitude make data collection challenging and expensive and, consequently, data are often scarce.

In recent years, the area has endured a large water deficit, which entails great economic, social, and environmental impacts. The current context of climate change threatens the productive model, based on irrigated agriculture that depends on snowmelt water in the upper basins.

This forecasting methodology, with inputs that are easily accessible and available for modeling, constitutes a great advance for the management of the water resources of the study area and, potentially, of other basins in the Andes with a snow-glacier regime.

Even though the developed model achieved good results for 1 and 2-months lead forecasting, there are applications that require a longer lead-time forecast.

Future steps within this research line include implementation of models for longer lead times and in neighbor basins. Furthermore, the glacier contribution to the runoff can be added as a model input. This can improve the water resource management. Decisions made with longer anticipation can lead to an optimized provision of water for uses that are often competing and/or conflicting.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## References

Abraham, E., Abad, J., Lora Borrero, B., Salomón, M., Sánchez, C., Soria, D., 2005. Caracterizacion y valoracion hidrologica de la cuenca del rio Mendoza mediante elaboración de modelo conceptual de evaluacion, In. Actos del Congreso Argentino del Agua (1),, 1–14.

Apaydin, H., Feizi, H., Sattari, M.T., Colak, M.S., Shamshirband, S., Chau, K.W., 2020. Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. Water 12 (5), 1–18. doi:10.3390/w12051500, (Switzerland).

Asefa, T., Kemblowski, M., McKee, M., Khalil, A., 2006. Multi-time scale stream flow predictions: the support vector machines approach. J. Hydrol. 318 (1–4), 7–16. doi:10.1016/j.jhydrol.2005.06.001.

Behzad, M., Asghari, K., Eazi, M., Palhang, M., 2009. Generalization performance of support vector machines and neural networks in runoff modeling. Expert Syst. Appl. 36 (4), 7624–7629. doi:10.1016/j.eswa.2008.09.053.

Bergeron, J., Royer, A., Turcotte, R., Roy, A., 2014. Snow cover estimation using blended MODIS and AMSR-E data for improved watershed-scale spring streamflow simulation in Quebec, Canada. Hydrol. Process. 28 (16), 4626–4639. doi:10.1002/hyp.10123.

Bhattacharya, B., Solomatine, D.P., 2005. Neural networks and M5 model trees in modelling water level-discharge relationship. Neurocomputing 63, 381–396. doi:10.1016/j.neucom.2004.04.016, SPEC. ISS..

Bruniard, E.D., Moro, C.O., 1994. Los Regímenes Fluviales de Alimentación Sólida en la República Argentina : Ensayo de Elaboración de un Modelo Hidroclimático de la Vertiente Oriental de los Andes. Academia Nacional de Geografía, p. 81 (Publicación especial ; 7).

Callegari, M., Mazzoli, P., Gregorio, L.D., Notarnicola, C., Pasolli, L., Petitta, M., Pistocchi, A, 2015. Seasonal river discharge forecasting using support vector regression: a case study in the Italian alps. Water 2494–2515. doi:10.3390/w7052494.

Cara, L., 2018. Desarrollo de una plataforma web para el procesamiento digital de imágenes satelitales enfocada al estudio del hidroclima. Master thesis. Universidad Nacional de Cordoba. doi:10.13140/RG.2.2.33116.64642.

Cara, L., Masiokas, M., Viale, M., Villalba, R., 2016. Análisis de la cobertura nivel de la cuenca superior del río Mendoza a partir de imágenes MODIS. Meteorológica 41 (1), 21–36.

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Comput. Electr. Eng. 40 (1), 16–28. doi:10.1016/j.compeleceng.2013.11.024.

Chanklan, R., Kaoungku, N., Suksut, K., Kerdprasop, K., Kerdprasop, N., 2018. Runoff prediction with a combined artificial neural network and support vector regression. Int. J. Mach. Learn. Comput. 8 (1), 39–43. doi:10.18178/ijmlc.2018.8.1.660.

De Gregorio, L., Callegari, M., Mazzoli, P., Bagli, S., Broccoli, D., Pistocchi, A., Notarnicola, C., 2017. Operational river discharge forecasting with support vector regression technique applied to alpine catchments: results, advantages, limits and lesson learned. Water Resour. Manag. 32. doi:10.1007/s11269-017-1806-3.

Dibike, Y.B., Velickov, S., Solomatine, D., Abbott, M.B., 2001. Model induction with support vector machines : introduction and applications. J. Comput. Civ. Eng. 15 (3), 208–216. doi:10.1061/(ASCE)0887-3801(2001)15:3(208).

Dolling, O.R., Varas, E.A., 2002. Utilisation des réseaux des neurones artificielles pour la prédiction des écoulements. J. Hydraul. Res. 40 (5), 547–554. doi:10.1080/00221680209499899.

Dong, C., 2018. Remote sensing, hydrological modeling and in situ observations in snow cover research: a review. J. Hydrol. 561, 573–583. doi:10.1016/j.jhydrol.2018.04.027, (Amst)March.

Drucker, H., Surges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1997. Support vector regression machines. Adv. Neural Inf. Process. Syst. 1, 155–161.

Ferguson, R.I., 1999. Snowmelt runoff models. Prog. Phys. Geogr. 23 (1999), 205–227. doi:10.1191/030913399672720559.

Guo, J., Zhou, J., Qin, H., Zou, Q., Li, Q., 2011. Monthly streamflow forecasting based on improved support vector machine model. Expert Syst. Appl. 38 (10), 13073–13081. doi:10.1016/j.eswa.2011.04.114.

Hyndman, R.J., Khandakar, Y., 2008. Automatic Time Serie Forecasting: the forecast Package for R. J. Stat. Softw. 27, 0–23. doi:10.18637/jss.v000.i00.

Kuhn M. (2014). Futility analysis in the cross-validation of machine learning models. arXiv:1405.6974v1. https://doi.org/10.48550/arxiv.1405.6974

Langhammer, J., Česák, J., 2016. Applicability of a nu-support vector regression model for the completion of missing data in hydrological time series. Water 8 (12). doi:10.3390/w8120560, (Switzerland).

Lee, E., Kim, S., 2020. Characterization of runoff generation in a mountainous hillslope according to multiple threshold behavior and hysteretic loop features. J. Hydrol. 590, 125534. doi:10.1016/j.jhydrol.2020.125534, September.

Maity, R., Bhagwat, P.P., Bhatnagar, A., 2010. Potential of support vector regression for prediction of monthly streamflow using endogenous property. Hydrol. Process. 24 (7), 917–923. doi:10.1002/hyp.7535.

Masiokas, M.H., Villalba, R., Luckman, B.H., Le Quesne, C., Aravena, J.C., 2006. Snowpack variations in the central andes of Argentina and Chile, 1951 –2005 : large-scale atmospheric influences and implications for water resources in the region. J. Clim. 19, 6334–6352. doi:10.1175/JCLI3969.1.

Masiokas, M.H., Villalba, R., Luckman, B.H., Mauget, S., 2010. Intra-to multidecadal variations of snowpack and streamflow records in the andes of Chile and Argentina between 30° and 37°S. J. Hydrometeorol. 11 (3), 822–831. doi:10.1175/2010JHM1191.1.

Maza, J., Fornero, L., Yañez, H.Anonymous, 1995. Simulación matemática de la fusión nival y pronóstico de escurrimiento. Bull. Inst. Fr. Etudes Andin. 24 (3), 651–659.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-3.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10 (3), 282–290. doi:10.1016/0022-1694(70)90255-6.

Pasolli, L., Notarnicola, C., Bruzzone, L., 2011. Estimating soil moisture with the support vector regression technique. IEEE Geosci. Remote Sens. Lett. 8 (6), 1080–1084. doi:10.1109/LGRS.2011.2156759.

Pierini, J.O., Gómez, E.A., Telesca, L., 2012. Prediction of water flows in Colorado River, Argentina. Lat. Am. J. Aquat. Res. 40 (4), 872–880. doi:10.3856/vol40-issue4-fulltext-5.

Raghavendra, S., Deka, P.C., 2014. Support vector machine applications in the field of hydrology: a review. Appl. Soft Comput. J. 19, 372–386. doi:10.1016/j.asoc.2014.02.002.

Riggs, G. A., Hall, D. K., Román, M. O., 2015. MODIS snow products collection 6 user guide. *National Snow and Ice Data.* Center, Boulder, CO, USA, p. 66.

Roy, A., Royer, A., Turcotte, R., 2010. Improvement of springtime streamflow simulations in a boreal environment by incorporating snow-covered area derived from remote sensing data. J. Hydrol. 390 (1–2), 35–44. doi:10.1016/j.jhydrol.2010.06.027.

Sedighi, F., Vafakhah, M., Reza, M., 2016. Rainfall – runoff modeling using support vector machine in snow-affected watershed. Arab. J. Sci. Eng. doi:10.1007/s13369-016-2095-5.

Smola, A. J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and computing 14 (3), 199–222.

Tan, B.Q., O'Connor, K.M, 1996. Application of an empirical infiltration equation in the SMAR conceptual model. J. Hydrol. 185 (1–4), 275–295. doi:10.1016/0022-1694(95)02993-1.

Therneau, T., & Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1–15. Available online: https://cran.r-project.org/package=rpart

Wang, W.C., Chau, K.W., Cheng, C.T., Qiu, L., 2009. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. J. Hydrol. 374 (3–4), 294–306. doi:10.1016/j.jhydrol.2009.06.019.

Zhang, Y., Chiew, F.H.S., Li, M., Post, D., 2018. Predicting runoff signatures using regression and hydrological modeling approaches. Water Resour. Res. 54 (10), 7859–7878. doi:10.1029/2018WR023325.