Building a smart hydro-informatics system for flood forecasting and warning, a real case study in atlas region -Kingdom of Morocco-

Mohammed El Idrissi [1] ⓘ · Omar El Beqqali [1] · Jamal Riffi [1]

Abstract

Many floods had pounded several Moroccan regions and had claimed a heavy toll of human life and immense damages; the most terrible event is the one that destroyed the area of Ourika in the surrounding of Marrakech City in 1995. We aim through this study to propose an effective flood forecasting and warning system (FFWS) for better management of flooding situations based on hydrology sensors for data gathering and least square support vector regression algorithm (LSSVR) optimized with PSO (particle swarm optimization) for monthly flood prediction. Since we expect that the future river flow values are heavily controlled by the past ones, as well as no autocorrelation in the residuals is required by the least square method that we use in this study, the model accepts time series lagged river flow values in the input and predicts river flow time series in the output. The smart proposed system allows to collect and predict the hydraulic situation in Aghbalou catchment and then forecasts whether a flood should be expected. The obtained results disclose the great quality of prediction of the LSSVR-PSO model with 0.8707 of R-squared score using a specific combination of cross-validation dataset.

Introduction

River flow forecasting is considered as a strong pillar in the water resource management plan, whether to precise the irrigation water demand or to avoid flood catastrophes. Predicting the river flow may provide early flood indicators, and thus avoid huge losses in lives, infrastructures, and assets. Because of the lack of efficient early flood forecasting and warning systems (FFWS), the Ourika region in the surrounding of Moroccan Marrakech City had known natural disasters in 1995 and 2015 causing hundreds of lives lost, and huge

damages of infrastructures have arisen. The crucial need for early flood warning systems pushes the scientific community to adopt many statistic and scientific methods in the hydrology domain to evaluate and predict river flows by searching relationships and patterns in the available data. So far, a limited number of studies have been carried out in the Moroccan context to model and understand flood situations; one of these studies was conducted by El Khalki et al. (2018) to model flood processes and compare two meteorological models namely AROME and ALADIN in terms of quantitative precipitation forecasting; their proposed study reported the usefulness of AROME model for flood forecasting compared to ALADIN. Despite this result, the proposed study reveals some limitations related to the density of the network of sensors used in data collection which generate more gaps in the tested data. In a general context, the well-known statistical methods that have been widely used in the hydrology domain are ARIMA (autoregressive integrated moving average) and SARIMA (seasonal autoregressive integrated moving average) (Ahlert and Mehta 1981; Yang et al. 2017). These techniques are sometimes used by researchers to model river flows (Nigam et al. 2009; Ghimire and Bhola 2017) or to analyze ground water quality (Katimon et al. 2018; Hanh et al. 2010), whereas others have successfully used them to model many

---

Responsible Editor: Broder J. Merkel

✉ Mohammed El Idrissi
mohammed.elidrissi29@usmba.ac.ma

Omar El Beqqali
omar.elbeqqali@usmba.ac.ma

Jamal Riffi
riffi.jamal@gmail.com

[1] Informatics, Signals, Automatic and Cognitivism Laboratory, Sidi Mohamed Ben Abdellah University, P.B. 1796, 30003 Fez-Atlas, Morocco

several evaporation processes (Mohan and Arumugam 1995; Shimi et al. 2020). Despite the strength of artificial neural networks (ANN), they still have some points of weaknesses such as the overfitting and the underfitting phenomenon. Furthermore, the optimization process of the ANN model is susceptible to remain stuck in the local optimum instead of the global optimum during the training process; another limitation is the huge time and amount of data needed for the training task. Consequently, statistical methods have recently replaced ANN thanks to their flexible nature and their ability to capture nonlinearity in the analyzed data. Support vector regression (SVR) is the regression method algorithm of the support vector family that shown its great ability in regression tasks due to the offered parallel processing, auto training without overfitting or underfitting issues, and the optimal results that can provide. SVR uses the quadratic programming (QP) method to solve the error function; this method presents certain complexity in nonlinear data. These weaknesses among others are resolved by the least square support vector regression algorithm (LSSVR) which is an improved version of SVR. Besides its high speed of processing, LSSVR adopts the least squared method to deal with optimization problems, so the complexity is widely minimized.

The objective of this work is to apply the LSSVR machine learning (ML) method optimized with PSO algorithm to predict monthly river flow; the model is tested in Aghbalou catchment using recorded data between 1969 and 2014, taking into account three parameters: recorded rainfall, recorded water level on the river coast, and river flow historical data.

Regression method

In this section, the family of support vector algorithms is briefly introduced notably support vector machine (SVM), support vector regression (SVR), least square support vector machine (LSSVM), and least square support vector regression (LSSVR) as a regression method adopted in data analysis and hydraulic situation prediction in the studied catchment.

Support vector machine (SVM)

In 1963, Vapnik and Lerner (1963) proposed the SVM algorithm as a supervised machine learning algorithm intended to resolve classification and discrimination problems. According to Dehghani et al. (2013), SVM is considered as one of the most powerful and popular algorithms of flood prediction; this algorithm is a generalization form of linear classifiers that are used both for classification and regression tasks. Therefore, it has been widely applied in several domains such as bioinformatics, computer vision, finance, and operational research. Support vector machine algorithm is based on two essential ideas:

– The first key idea is the maximum margin, which is the maximal distance between the separation hyperplane and the nearest data samples as shown in Figure 1; the nearest samples are called support vectors. The objective of an efficient SVM classifier is to find the optimal separation hyperplane that maximizes the margin.

– In the case of nonlinear classification as illustrated in Figure 2, SVM looks for a possible linear separation in an n-dimensional space, by transforming the representation space of the input data into a multidimensional space using the kernel function (KF); the KF does not require explicit knowledge of the transformation function needed for space transformation. The KF transforms the scalar product into a simple punctual evaluation of a function in the multidimensional space; this operation is known as the kernel trick.

Support vector regression (SVR)

SVR is another algorithm of the support vector family, but this time in the regression context. This algorithm gives the flexibility to define the acceptable error ϵ to stay inside the margin [−a, +a]; it next finds the most appropriate line or hyperplane to separate the data into classes. The error metric in the simple regression is always supposed to be minimized. SVR, however, looks for maintaining the error factor between two thresholds superior $+a$ and inferior $-a$; the best regression model is the one that gives the majority of data points close to the best hyperplane (illustrated by a green line in Figure 3).

Least square support vector machine (LSSVM)

LSSVM is a supervised machine learning algorithm and the least square version of SVM. This model analyzes data to recognize patterns and relationships. For the first time, Suykens and Vandewalle (1999) conceived LSSVM to perform classification and regression tasks by resolving linear systems instead of convex quadratic programming (QP). This algorithm belongs to a family of learning methods known as kernel-based; in other words, it uses linear classifiers to resolve nonlinear classification tasks by transforming the input data space into multidimensional data space, thus allowing to use linear systems and improve performance. This transformation of space dimensions is ensured thanks to the kernel trick technique. Two kernel types can be used by support vector models: stationary and non-stationary. The mathematical definitions of experimentally tested kernels are summarized in the following list; each kernel has its
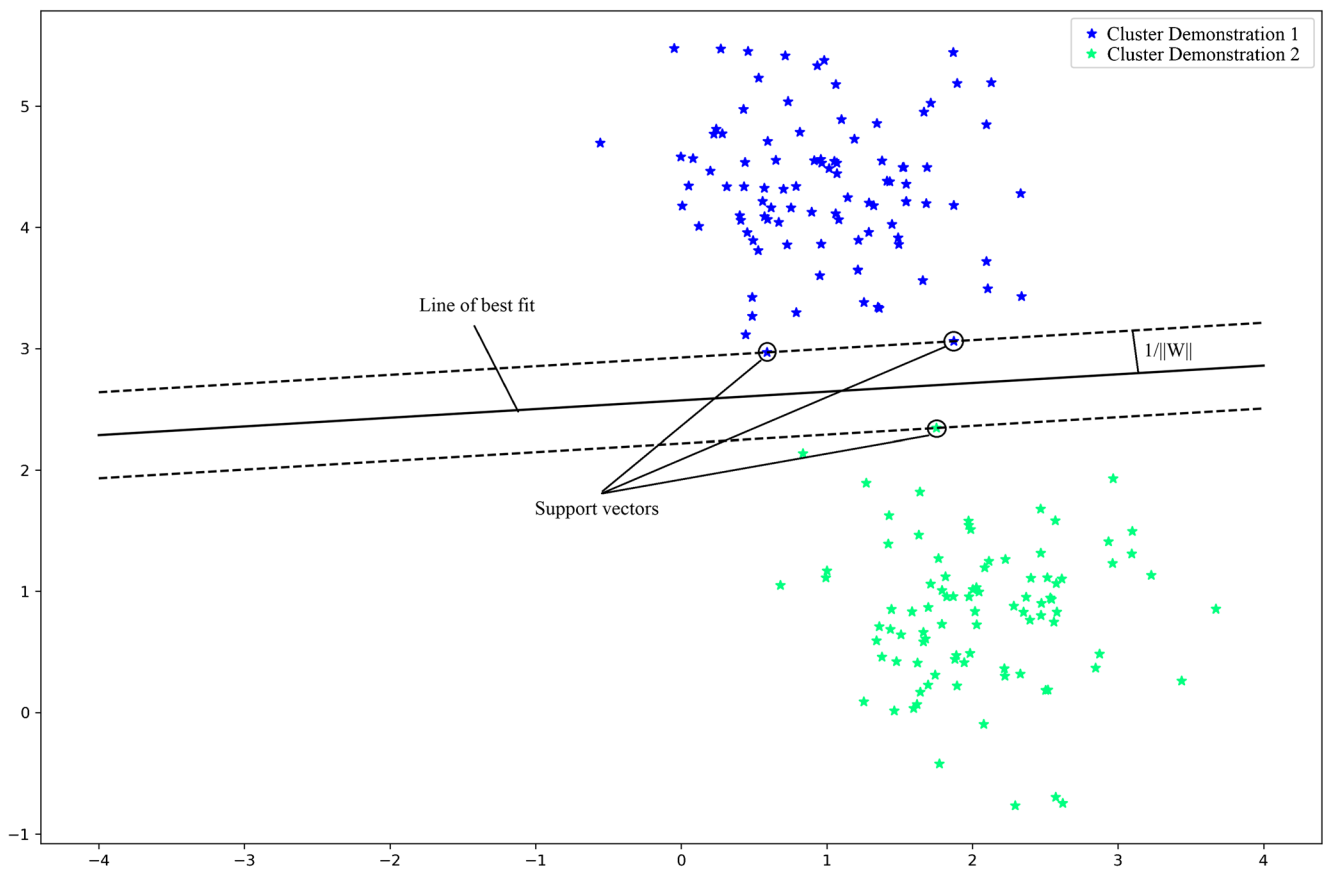
Fig. 1 Illustration of SVM linear classifier (the solid line is the best line of classification); ||W|| refers to the normal vector to the hyperplane; 1/||W|| is the distance between the line of the best fit and the hyperplanes of the support vectors; the dataset of the demonstration is divided into two clusters (number of samples = 250, two centers are generated, the standard deviation of the clusters = 0.60)

advantages and disadvantages in terms of complexity and convergence speed; it is up to the user to choose and define the hyper-parameters of the appropriate kernel depending on its context.

Stationary kernels

Gaussian radial basis function (RBF) kernel : $K\left(x, x^{'}\right)$

$$= \exp\left(-\gamma\left\|x{-}x^{'}\right\|^2\right) \tag{1}$$

where $\gamma > 0$, or sometimes $\gamma = 1/2\sigma^2$, where σ^2 is the variance.

Laplacian RBF kernel : $K\left(x, x^{'}\right) = \exp\left(-\dfrac{\left\|x{-}x^{'}\right\|}{\sigma}\right) \tag{2}$

where ||x-x'|| represents the Manhattan distance.

Sigmoid kernel : $K\left(x, x^{'}\right) = \mathrm{Tanh}\left(\alpha x^T.x^{'} + c\right) \tag{3}$

where α is the slope and c is the interception parameter.

Non-stationary kernels

Linear kernel : $K\left(x, x^{'}\right) = x^T.x' \tag{4}$

Polynomial kernel : $K\left(x, x^{'}\right) = \left(\alpha X^T.x^{'} + \lambda\right)^d \tag{5}$

Least square support vector regression (LSSVR)

LSSVR was firstly introduced by Suykens et al. (2011) to improve certain characteristics of SVR like complexity minimization as well as processing optimization. SVR deals with regression problems using the quadratic programming (QP) technique. QP is known by its ability to generate more complexity in the model and requires a lot of time and resources in processing. These issues are resolved by LSSVR using linear equations instead of quadratic programming (Gestel et al. 2001). LSSVR calculates the least square of the error function instead of Vapnik's linear ε-insensitive loss function given by Eq. (6); the error according to this method equals to 0 if the difference between the observed value
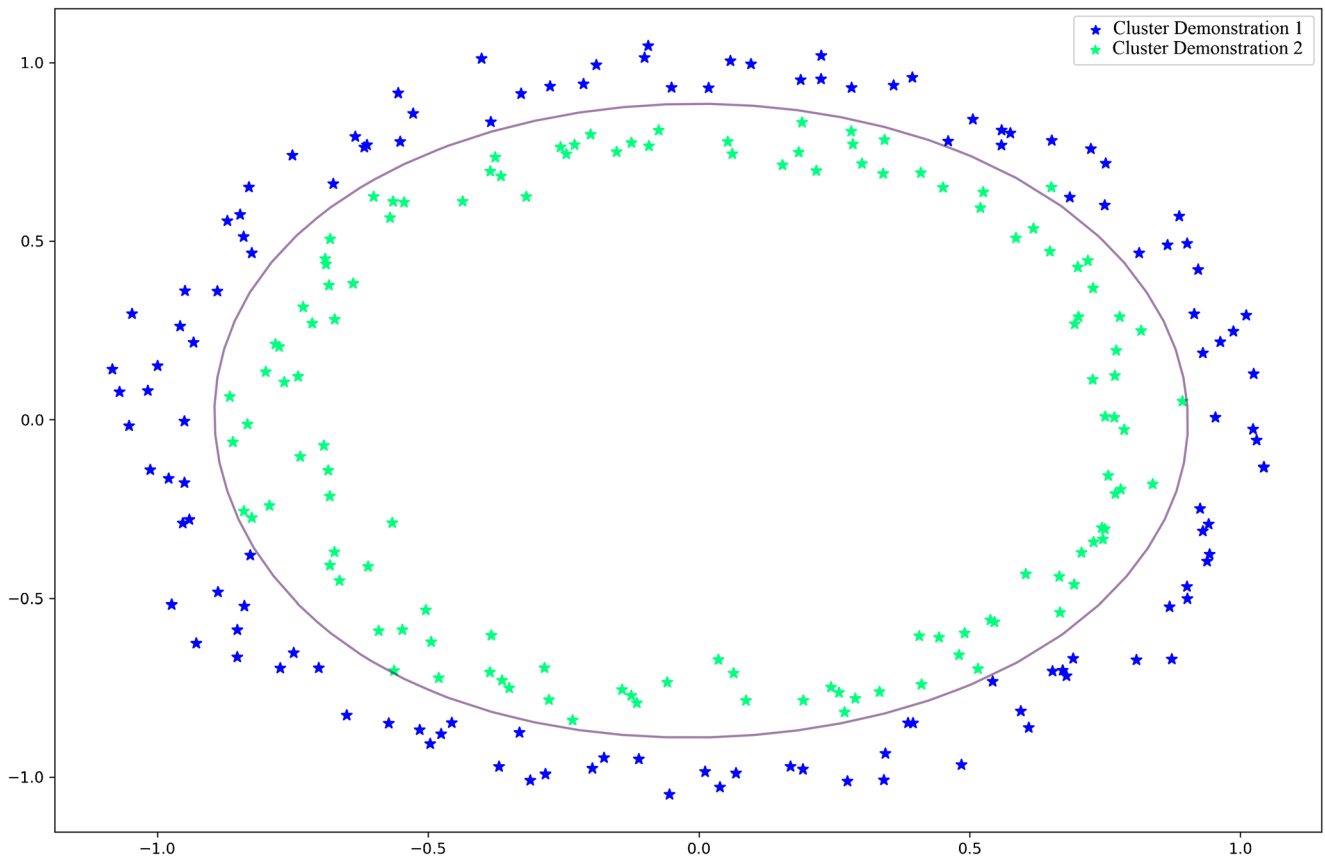
Fig. 2 Illustration of the kernel trick technique to separate nonlinear data points; the dataset of the demonstration is divided into two clusters (number of samples = 250, standard deviation of Gaussian noise added to the data = 0.05)

y and the predicted value $f(x)$ is less than ε; otherwise, the error function is given by the absolute value of the difference between those values.

$$L(x, y, f) = \max(0, |y - f(x)| - \varepsilon) \tag{6}$$

LSSVR hyper-parameters play an important role in the time series prediction; for this reason, we chose particle swarm optimization (PSO) algorithm to build optimal values of LSSVR, notably the weight vector ω and the error e. The LSSVR nonlinear function as proposed by Rana et al. (2019) is given by Eq. (7):

Fig. 3 Illustration of SVR model
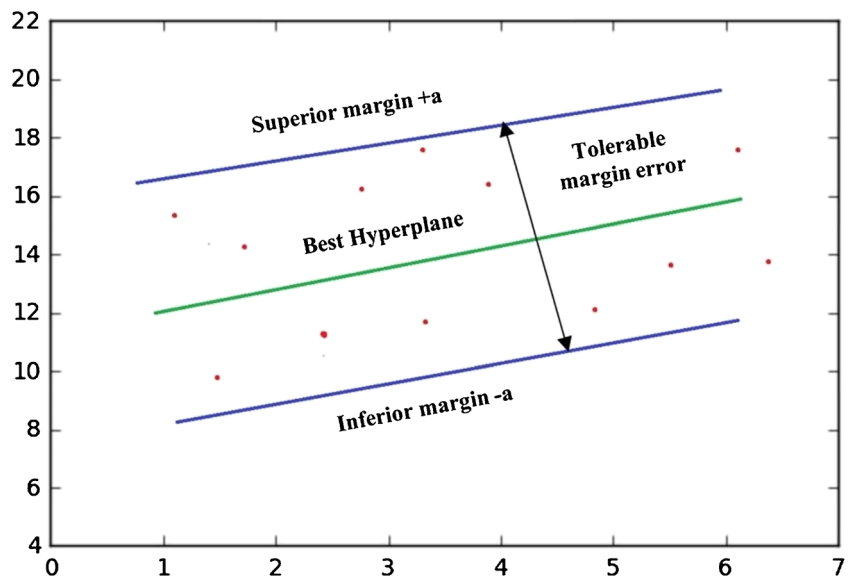
$$y(x_j) = \omega^T . \varphi(x_j) + b \tag{7}$$

The input x_j of Eq. (7) is a time series of the water level on the coast of the Aghbalou river and the amount of rainfall. The output $y(x_j)$ is the river flow, ω^T and b consecutively are the transpose of the weight vector and bias, and $\varphi(x)$ is a nonlinear mapping function. Equation (8) has to be optimized in order to minimize the cost function:

$$\min Cost = \frac{1}{2}\omega^T\omega + \frac{\gamma}{2}\sum_{j=1}^{n}e_j^2 \tag{8}$$

where e_j and γ, respectively, are the error function and the regularization parameter. Equation (8) is subject to Eq. (9):

$$y(x_j) = \omega^T . \varphi(x_j) + b + e_j \text{Avec } \{j = 1, 2, \dots, n\} \tag{9}$$

To find a couple of solutions (ω, e), the Lagrange multipliers (Kisi and Parmar 2016) are used; the Lagrange function is given by Eq. (10):

$$L = Cost - \sum_{j=1}^{n}\rho_j\left\{\omega^T.\varphi(x_j) + b + e_j - y(x_j)\right\} \tag{10}$$

where ρ_j is the coefficient of Lagrange multipliers. To solve the regression problem, it is mandatory to use one of the kernel functions previously described.

Particle swarm optimization (PSO)

PSO is a metaheuristic-based algorithm that was first inspired by birds' behavior while searching for foods. It was proposed by Eberhart and Kennedy (1995) to iteratively optimize numerical functions without using the gradient descendent technique. The optimization process is based on moving the particles (represented by (x_i, x'_i) positions) with a certain velocity; the best optimization solution is obtained when the particle swarm converges to the best values. The algorithm begins with a randomly initialized n particles (aka solution population), as well as the positions $x_{i,d}$ and the velocity v of the particles within a permissible range given that $i \in [1,2,3\dots,n]$ and d are the dimension of the vectors in the search space. Each particle has its optimal position known as *pbest*; its experience is shared with the other particles to help them to find the best position. The optimal solution for the global population is known as *gbest*. The acceleration of convergence toward the best values of *pbest* and *gbest* is calculated in each iteration, and the experience is evaluated using the fitness function. The velocity is calculated using Eq. (11):

$$v_{id}(j+1) = \omega\, v_{id}(j) + c_1 rand_1(p_{id} - x_{id}) + c_2 rand_2\left(p_{gd} - x_{id}\right) \tag{11}$$

where ω is the inertial parameter, $C1$ and $C2$ are the acceleration parameters, and $rand_i\{i \in [1,2]\}$ is a random variable between 0 and 1.

The new position of the particle is updated using Eq. (12) and the previously calculated velocity:

$$x_{id}(j+1) = x_{id}(j) + v_{id}(j+1) \tag{12}$$

The flowchart of the early flood warning system based on LSSVR-PSO is given in Figure 4.

Figure 4 describes the step-by-step process of building the optimal LSSVR-PSO model and making the prediction of river flow using the optimal hyper-parameters. A summary of the parameters used by PSO in the optimization process is given in Table 1:

- The historical data are first divided into training and validation datasets based on the *k-fold* CV technique as described in the "Data preprocessing" section.
- The PSO parameters and LSSVR hyper-parameters are initialized, the kernel function used is RBF, the sigma square (σ^2) parameter varies between 0.001 and 30, the penalty factor γ varies between 0.01 and 300, the number of iterations is set up to 30, the inertial parameter w is set up to 0.5, the acceleration coefficients $C1$ and $C2$ values consecutively are 0.8 and 0.9, the number of particles is set up to 50, and the target error is 0.1 (10%).
- In each iteration i, the model LSSVR-PSO receives 12 lagged data points in the input, and the current 12 values are considered as an output of the model; a forecasting result is obtained, an offset of 12 data points is made in each PSO iteration, and new forecasting results are provided by the model; this process is repeated until all the predictions are made. The next step is the validation of the obtained forecasts by calculating the validation error. Here we consider the mean absolute error (MAE) given by Eq. (14) as an objective function. The best individual and global positions of the particles are then updated based on the obtained fitness value of the particle in the ith iteration, the velocity is next calculated using Eq. (11), and the position of each particle is updated according to Eq. (12).
- When the stopping criteria are fulfilled, the optimal values are used to build the optimal LSSVR-PSO model, and the model is tested using the test sub-dataset.
- The new rainfall values and the water level on the coast are then presented to the optimal pre-trained model to make river flow forecasting. The predicted value is passed through the pre-defined thresholding rules, and the final hydraulic situation is predicted.
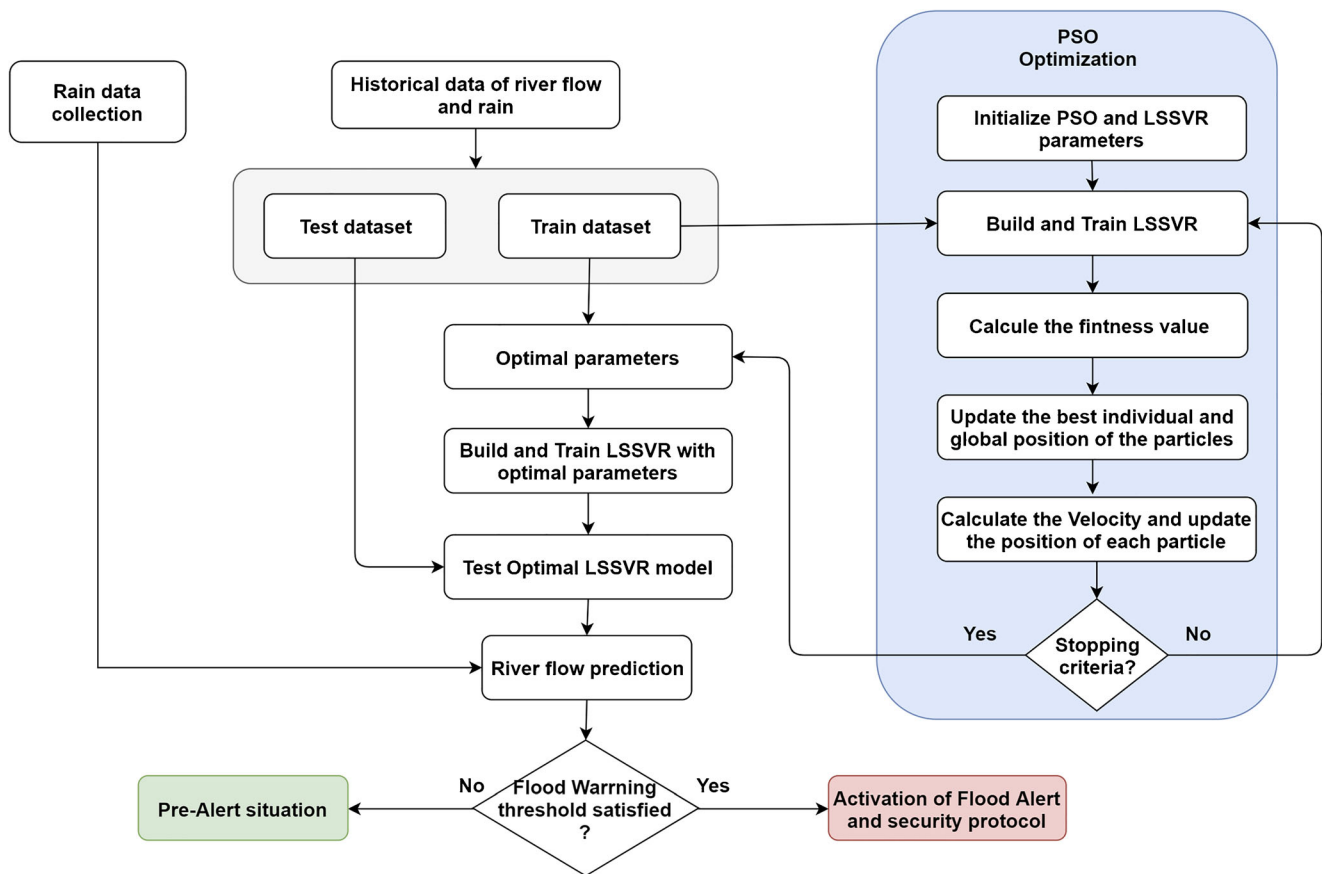
Fig. 4 FFWS flowchart based on LSSVR-PSO

Geo-localization of the study catchment

Aghbalou catchment, the subject of this study, is situated in the Ourika region, Al Haouz Province of Morocco. The geographical position of the Aghbalou catchment is given in Figure 5, and the geographical coordinates are 7°44′55.739″ West, 31°18′58.518″ North. The area covered by the catchment is about 503km^2. The catchment is crossed by two rivers, Ourika and Assif El Mal, whose approximate lengths, respectively, are 61.78km and 300km.

Table 1   Summary of parameters used in the optimization process

Parameter	Value
RBF sigma square parameter $\sigma2$	Varies between 0.001 and 30
Penalty factor γ	varies between 0.01 and 300
Inertial parameter w	0.5
Acceleration coefficients C1 and C2	0.8 and 0.9, respectively
Number of particles	50
Target error	10%
Offset data points	12

Experiment

Material used in this study

–   A rain gauge (Figure 6): This device consists of a funnel with a diameter of 400 cm. The collected precipitations are directed to the tipping troughs equipped with a double reed magnetic contact to measure the amount of precipitation. The receiving ring collects the precipitation on an area of 400 cm^2, and the collected water is driven toward the auger of the volumetric transducer. When the increment volume is reached, which corresponds to 0.2 mm of rain, the bucket tilts and empties automatically; the second bucket moves to the fill position. It will flip, in turn, when it is full. A contact notifies each flip.

–   Water level gauge (Figure 7): This device broadcasts very short microwave pulses to the water, and the water surface reflects the pulses to be received by the antenna system of the ultrasonic sensor. The time between signal transmission and reception is proportional to the water level.
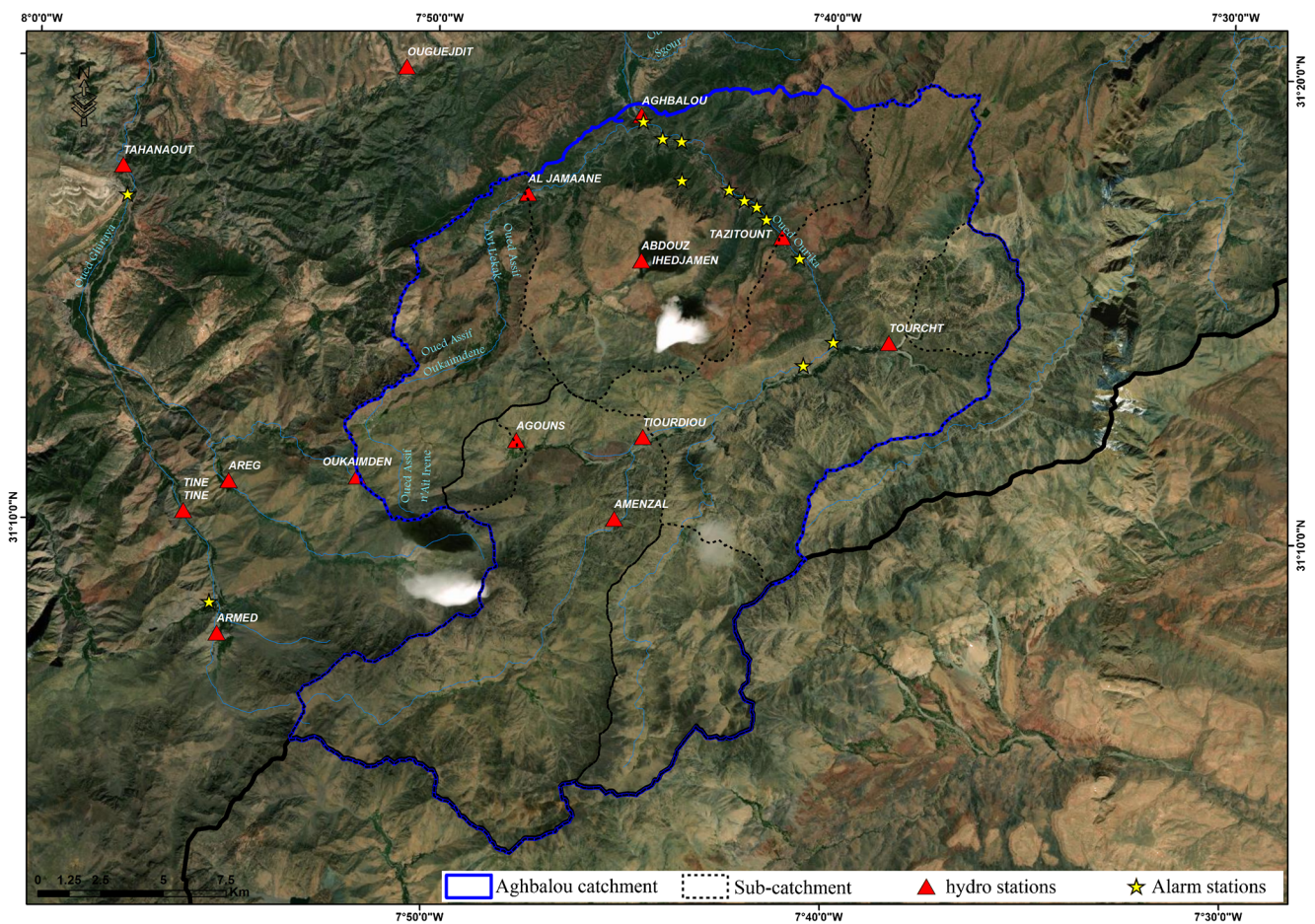
Fig. 5 Geographical map of AGHBALOU catchment

– Datasets: The model is tested in the Aghbalou catchment using recorded data between 1969 and 2014, taking into account three parameters: recorded rainfall, recorded water level on the river coast, and river flow historical data.


Fig. 6 Tipping bucket rain gauge

Data preprocessing

We will use the *k-fold* cross-validation (*k-fold* CV) technique to improve the accuracy of prediction of the conceived model. Therefore, the dataset is fairly divided into *n* sub-datasets, then *n-1* sub-datasets are used to train the model, and the remaining sub-dataset is used for validation. This process is repeated iteratively *n* times until all sub-datasets have been used for training and validation; this technique is very recommended as it provides the ability to use all the sub-datasets to train, test, and validate the model.

Analogously, the available dataset is divided into four (04) sub-datasets DS_i, $i = \{1,2,3,4\}$, three (03) of them are used to train the model, and one is used for the validation; this cycle is repeated four times until all the DS_i have been used for the validation task. Table 2 below gives more details about the statistics of the Aghbalou catchment.

The DS_1, DS_2, DS_3, and DS_4 are the obtained sub-datasets; each DS_i contains the recorded amount of rainfall and river flow in a specific period. *RFmean*, *RFsd*, *RFsc*, *RFmin*, and *RFmax*, respectively, are the mean of the river flow, the standard deviation, the skew, the min value, and the max value of the recorded river flow.

Fig. 7 Ultrasonic water level sensor and bracket in Aghbalou river

Ultrasonic water level sensor

Bracket
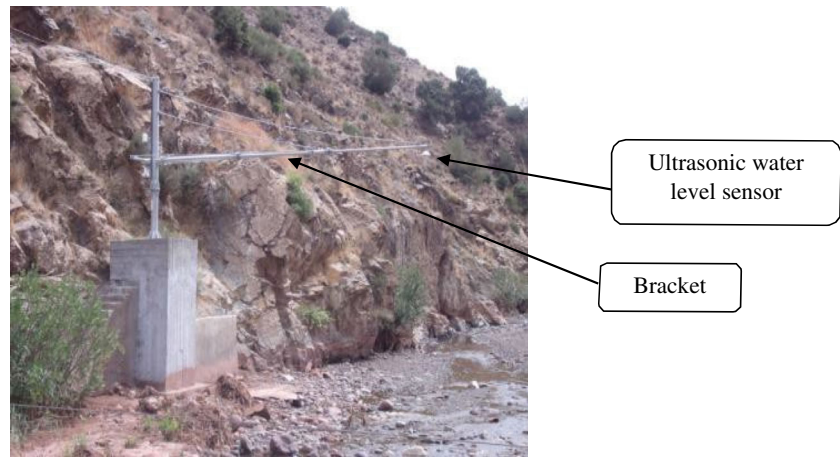
Collected data on the river flow is seen as time series; this means that the actual and future values have a close relationship with the past ones. Before applying the LSSVR-PSO model, we should understand how previous values of river flow impact the future ones, as well as test the stationary of the data time series. To check the stationary of the river flow time series, we use the following quick statistical method. We split the time series dataset into three groups and compare their means and variances; if the difference is significant, the time series can be non-stationary. The result of the check gives for the means, mean1 = 5.152950, mean2 = 6.328561, and mean3 = 3.045606, and for the variances, variance1 = 72.248783, variance2 = 137.164262, and variance3 = 12.857035. It is clear that the difference between the mean and the variance of each partition is significant, so the time series may be non-stationary and need transformation. Before doing that, let us take the plot of the dataset given by Figure 8 below to see how time series values vary over time.

The histogram of river flow time series is shown in Figure 9; the distribution of the values does not follow any familiar form, so this is another sign of non-stationarity of the river flow time series.

Now, let us see the effect of the log transform on the time series and check if the log transform can delete non-stationarity effects on the time series and then compare the new statistical means and variances of groups again. Figure 10 shows the histogram of the time series after log transformation; it is obvious that the distribution of the values

follows a Gaussian form. The new check of stationarity gives for each partition the following means and variances: mean1 = 0.651062, mean2 = 0.927963, mean3 = 0.435230, variance1 = 2.246677, variance2 = 1.720167, variance3 = 1.561821. These values become approximatively similar which leads us to say that the new form of the time series is stationarity; this will represent the input of the LSSVR model.

The calculation of the optimal value of the lag is carried out using many classical information criteria (IC). In our study, we choose Akaike (AIC) (Akaike 1973), Schwarz (SC) (Schwarz 1978), known also as Bayesian information criterion (BIC), and Hannan-Quinn (HQ) (Hannan and Quinn 1979). The first twelve (12) correlation values obtained using AIC, SC, and HQ are summarized in Table 3.

The optimal correlation value of river flow time series is obtained at Lag 12; furthermore, this optimal value is given by the AIC model. The selection of the input combination is an important step in ML model development. Thus, to evaluate the correlation effect of the lagged value previously obtained, the autocorrelation function (ACF) is used. Figure 11 plots the correlation coefficients for the first 24 lags, we can observe clearly the periodicity of the ACF, and lag = 12 corresponds to the period.

In this study, we use three common metrics to evaluate the performance of the proposed LSSVR-PSO model; these metrics are widely used to evaluate water resource management models (Santhi et al. 2001; Kalteh 2013; Landeras et al. 2009). The indexes include root

Table 2   Statistics of AGHBALOU catchment

| Datasets	Periods	RFmean	RFsd	RFsc	RFmin	RFmax
DS1	08/1969–10/1980	6.342857477	9.458465109	2.783758941	0.107626667	51.38973729
DS2	11/1980–01/1992	4.868536265	10.3497873	4.593609858	0.030719355	75.38709677
DS3	02/1992–04/2003	4.725397783	9.606718681	4.600895977	0.054316484	70.88
DS4	05/2003–07/2014	3.348097932	3.524720758	1.860900269	0.1171257	19.81218697

Fig. 8 River flow historical dataset
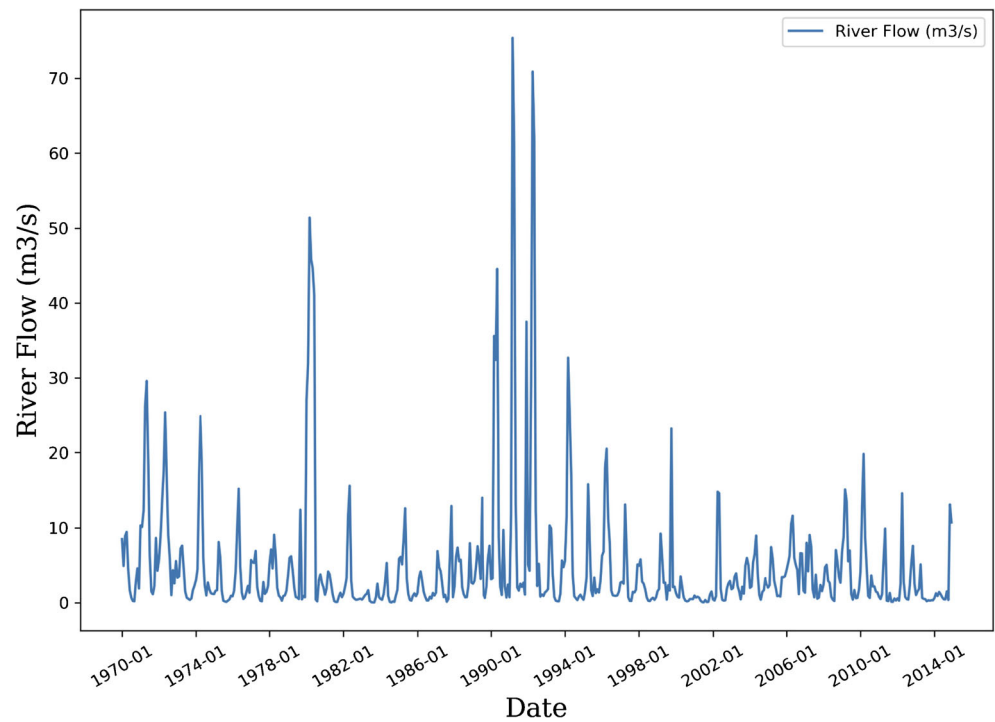
mean square error (RMSE) given by Eq. (13), mean absolute error (MAE) expressed by Eq. (14), and determination coefficient (DC) formulated by Eq. (15) (Adnan et al. 2017).

where N_{RF} is the number of river flow observed data points in the time series. RF_o is the instantly observed river flow, and RF_f refers to the river flow instantly predicted by LSSVR-PSO.
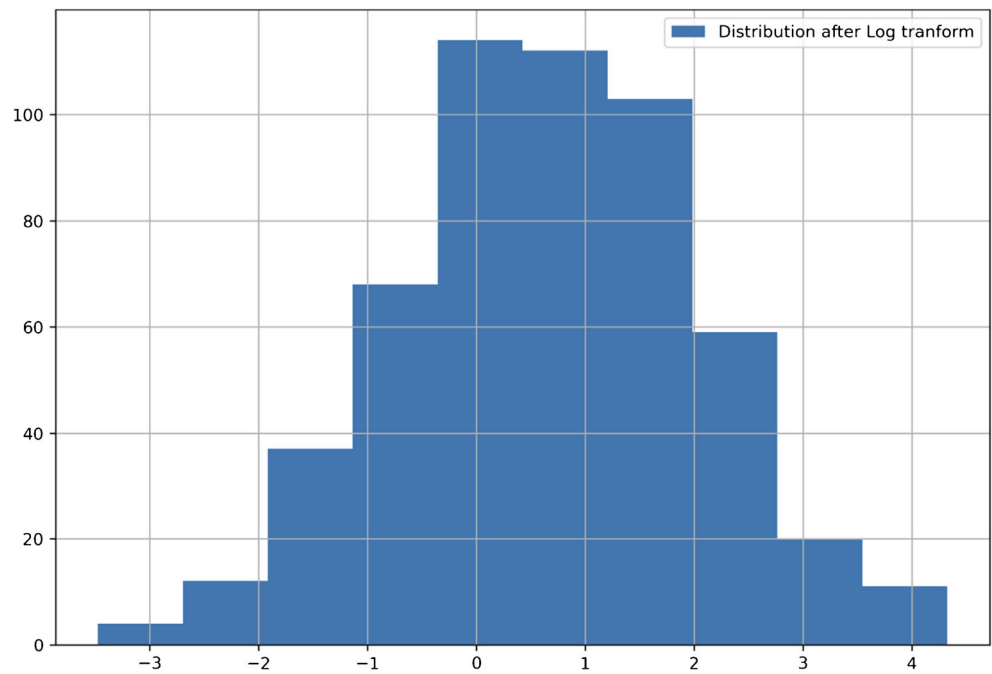
$$RF_{RMSE} = \sqrt{\frac{1}{N_{RF}} \Sigma_{t=1}^{N_{RF}} \left(RF_o - RF_f\right)^2} \qquad (13)$$

$$RF_{MAE} = \frac{1}{N_{RF}} \Sigma_{t=1}^{N_{RF}} \left| RF_o - \overline{RF_o} \right| \qquad (14)$$

Fig. 9 Histogram of the river flow time series
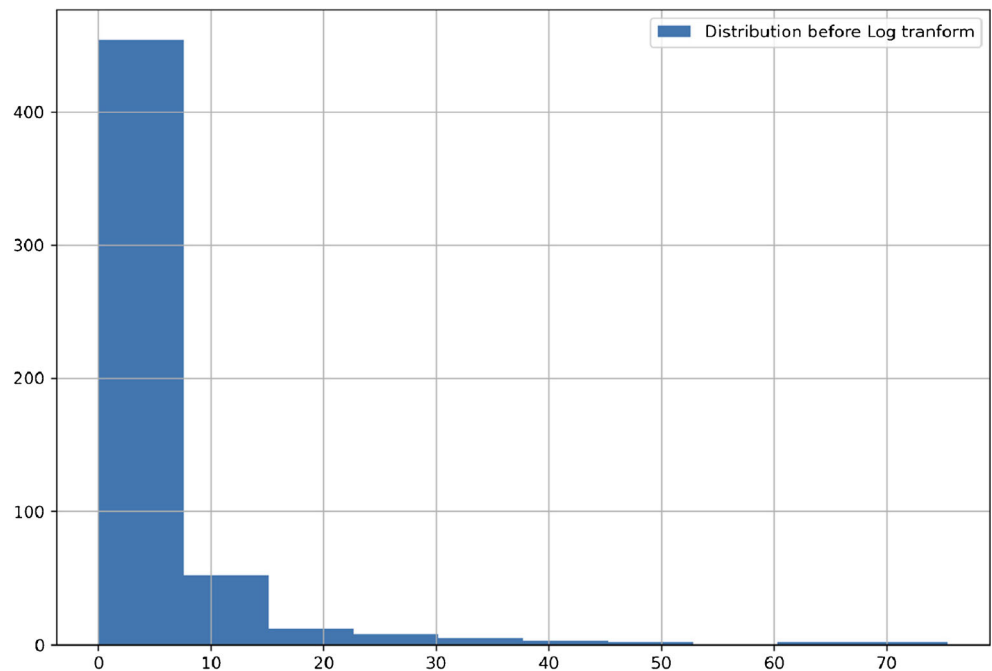
Fig. 10 Histogram of the river flow time series after log transformation

where $\overline{RF_o}$ is the mean value of the observed river flow.

$$RF_{DC} = \left[\frac{\sum_{t=1}^{N_{RF}} \left(RF_o - \overline{RF_o} \right) \left(RF_f - \overline{RF_f} \right)}{\sqrt{\sum_{t=1}^{N_{RF}} \left(RF_o - \overline{RF_o} \right)^2 \left(RF_f - \overline{RF_f} \right)^2}} \right]^2 \qquad (15)$$

where RF_f is the mean value of the river flow predicted by LSSVR-PSO.

The kernel function (KF) is an important factor that has to be carefully chosen so as to guarantee the good performance of the model. In this study, RBF-KF is used, the optimal value of RF_{DC} is obtained after a fine-tuning of RBF-KF hyper-parameters notably penalty factor gamma (γ), and the optimal RF_{DC} is found at $\gamma = 120.01$. To check the impact of the hyper-parameter γ on the DC metric, let us take Figure 12 which illustrates how DC is evolving when γ is varying between 0.01 and 300 using the combination {DS1, DS3, DS4} for training and DS2 for validation. It is obvious from Figure 12 that the DC coefficient varies in squared root depending on the variation of the γ factor; the variation function is:

$$DC = \sqrt{\gamma} \qquad (16)$$

The best value of DC is obtained at $\gamma = 120.01$.

As we mentioned earlier in the "Least square support vector machine (LSSVM)" section, there are multiple kernel functions, and the user should select the appropriate KF depending on his context. Starting from this idea, a performance comparison is performed between widely used kernels notably linear, polynomial, RBF (Gaussian), Laplacian, and sigmoid in terms of RMSE, MAE, and DC indexes. The comparison is carried out using the combination {DS1, DS2, DS3} for training and DS4 for validation. Table 4 shows clearly the optimality of the RBF kernel compared to the remaining kernels; it can efficiently minimize RMSE (3.6473) and MAE (2.0394), as well as maximize DC (0.8613).

In light of the performed experiment, the obtained results can be summarized in Table 5.

It is obvious from Table 5 that when the combination {DS1, DS2, DS3} is used for training and DS4 for validation, we obtain the worst validation performance (RMSE = 3.6473)

Table 3 The first twelve correlation values of river flow time series using AIC, SC, and HQ

Lag	AIC	SC	HQ
0	7.188356	7.196441	7.191521
1	6.612094	6.628264	6.618424
2	6.552724	6.576980*	6.562219
3	6.554794	6.587135	6.567455
4	6.556359	6.596786	6.572185
5	6.550534	6.599047	6.569526
6	6.553840	6.610438	6.575997
7	6.557618	6.622301	6.582940
8	6.559320	6.632089	6.587808
9	6.562037	6.642891	6.593689
10	6.557279	6.646218	6.592097
11	6.515062	6.612087	6.553045
12	6.494212*	6.599322	6.535360*

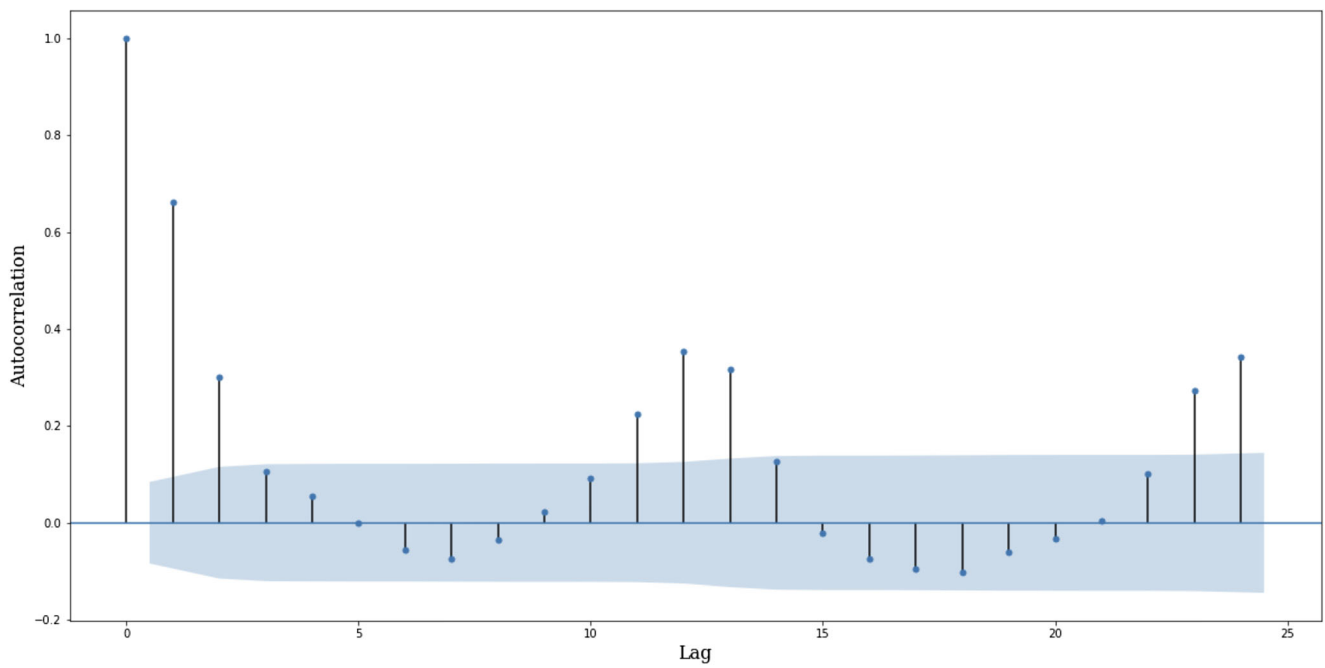*Optimal correlation value of river flow time series

Fig. 11 Autocorrelation function of river flow time series of Aghbalou catchment for the first 12 lags (months of the year)

compared to the other combinations. However, when the dataset *{DS1, DS3, DS4}* is used for training and *DS2* for validation, we got the best validation performance (RMSE = 2.9102). Regarding MAE, the combination *{DS2, DS3 DS4}* for training and *DS1* for validation provides the best MAE (1.5866); this is due to the strong correlation between river
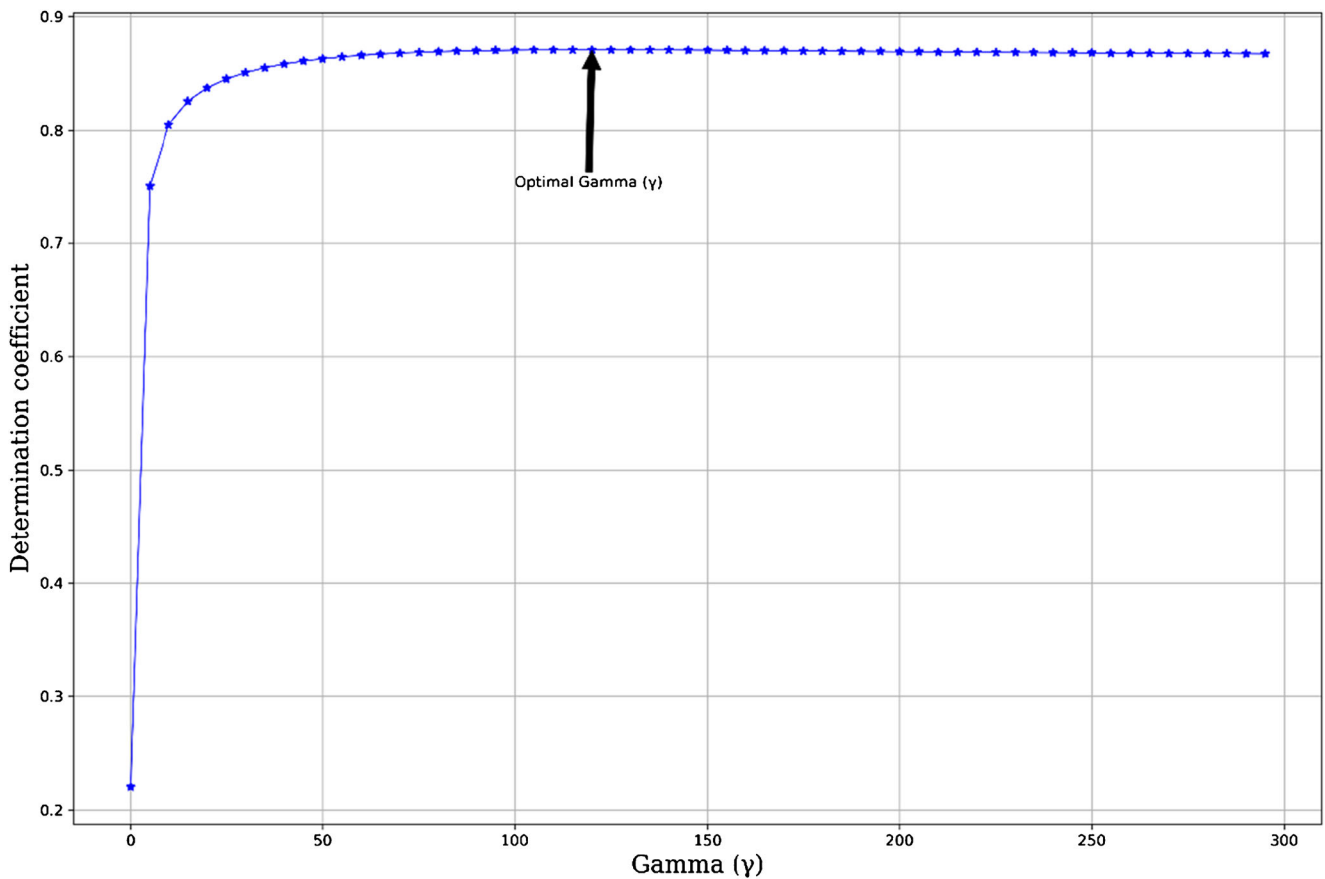

Fig. 12 The impact of γ on the determination coefficient

Table 4 Comparison of widely used kernel functions in terms of RMSE, MAE, and DC indexes

Indexes	Kernel functions				
	Linear	Polynomial	RBF (Gaussian)	Laplacian	Sigmoid
RMSE	4.3628	3.8474	3.6473	3.7030	4.0128
MAE	3.6770	2.7751	2.0394	2.0850	3.0070
DC	0.7016	0.7337	0.8613	0.8570	0.7216

flow recorded data points of this scenario. Furthermore, the combination {DS1, DS2, DS3} for training and DS4 for validation shows a weak correlation between river flow recorded data since it gives the worst MAE value (2.0394). The curves of both observed and predicted river flow using LSSVR-PSO are given in Figure 13. This figure reinforces what it was said earlier about the optimality of the scenario {DS1, DS3, DS4} for training and DS2 for validation; since it provides the best scores of the evaluation metrics compared to the other scenarios, the obtained similarity index DC is 0.8707; this value reflects an estimate of the strength of the relationship between the model and the response variable.

Discussion

In this study, a smart hydro-informatics model coupling LSSVR algorithm and PSO metaheuristic is proposed to predict monthly river flow time series in the Aghbalou catchment. The proposed method decays river flow time series into multiple components (12 lagged river flow values) and presents them to the model to forecast the future river flow time series. The built model is evaluated using three performance

evaluation indexes (RMSE, MAE, R^2). Monthly data of over 45 years were used to test the model; the found RSME, MAE, and R^2, respectively, are 2.9102, 1.5866, and 0.8707. In the light of this study, the most relevant results can be summarized as follows:

– Some authors believe that more accuracy of predictions is possible when big amounts of data are available; this contention was contested later by others like Makridakis et al. (1982). Whatever the right current, we provided a trade-off between the two ideas, and enough quantity of data has been provided so that the model was able to discover more information and patterns in the time series. The stability of data series is conventionally an impacting factor of the prediction accuracy. Consequently, many primary methods are made available to evaluate the stability; the autocorrelation function and quick statistical method are used in this study as relevant statistical techniques; they allow formally to look for any non-stationary caused by trends or seasonality effects in the time series; the log transform and the lag difference are used to clean the time series from non-stationarity effects; this preprocessing technique helps to select the appropriate extrapolation method. It is therefore important to notice that the regression-based model used in this work is mainly feasible with performances relatively similar, even if non-stationary is observed in the tested data, since preliminary processing is always performed.
– LSSVR-PSO minimizes the complexity of the prediction model; however, tested data may contain some aleatoric or epistemic uncertainty, and this is another influencing factor of forecasting error that needs to be addressed.
– Aghbalou catchment indicated lower values of RF_{RMSE} and RF_{MAE}; this result is justified by the lower mean

Table 5 LSSVR-PSO performance evaluation of multiple scenarios of 4-fold CV

Error indexes	Training datasets	Validation dataset	Performance validation of LSSVR-PSO (PVLSSVR)	Mean of PVLSSVR	Standard deviation of PVLSSVR
RF_{RMSE}	DS1 – DS2 – DS3	DS4	3.6473	3.2338	0.3123
	DS2 – DS3 – DS4	DS1	3.2693		
	DS3 – DS4 – DS1	DS2	2.9102		
	DS4 – DS1 – DS2	DS3	3.1084		
RF_{MAE}	DS1 – DS2 – DS3	DS4	2.0394	1.790375	0.1876
	DS2 – DS3 – DS4	DS1	1.5866		
	DS3 – DS4 – DS1	DS2	1.7447		
	DS4 – DS1 – DS2	DS3	1.7908		
RF_{DC}	DS1 – DS2 – DS3	DS4	0.8613	0.8608	0.00918
	DS2 – DS3 – DS4	DS1	0.8485		
	DS3 – DS4 – DS1	DS2	0.8707		
	DS4 – DS1 – DS2	DS3	0.8627		
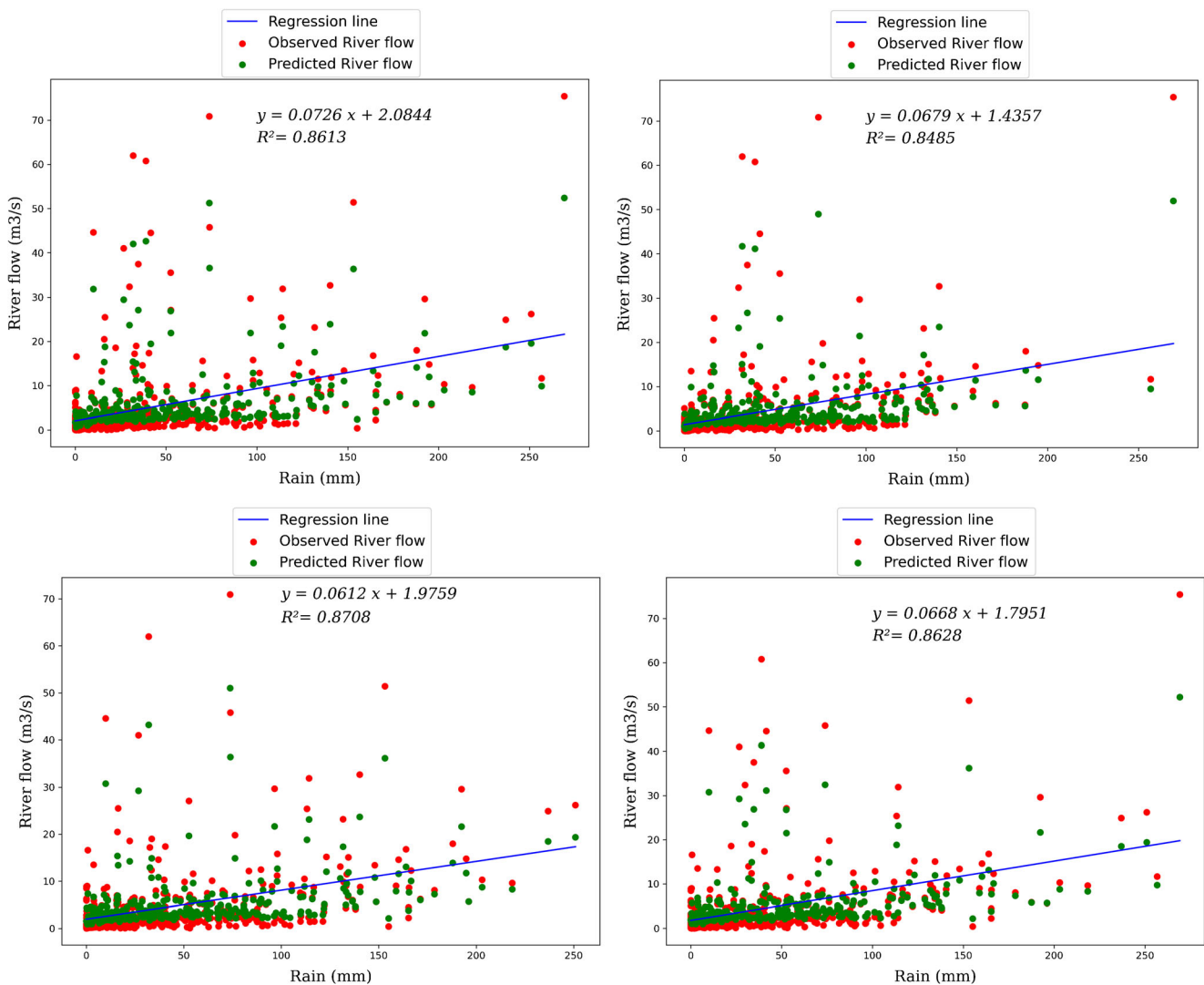
Fig. 13 Comparison between the observed and the predicted river flow values in 4 CV scenarios

value (4.82115 m³/s) of river flow in the dataset, as well as the ability of LSSVR to explore the hidden relationships in the time series.

– It was also found that the best accuracy value of river flow is obtained at a specific combination of sub-datasets ({DS1, DS3, DS4} for training and DS2 for validation); this optimal result leads us to report that not all the CV combinations can be used to accurately predict the river flow.

– The main reason behind the found quality of river flow prediction can be the robust generalization skills of LSSVR-PSO or may be the fine-tuning of the hyper-parameters of LSSVR-RBFKF notably gamma (γ) and sigma square (σ^2) using PSO metaheuristic. From the convergence point of view, PSO needs few generations to achieve its optimal hyper-parameters. The computational time needed

by PSO to achieve the convergence level is due to the communication between the populations during the optimization process; despite that, it is still reasonable since it runs on a machine with high computational performances.

– The key point of the success of LSSVR is its ability to capture more nonlinearity in data, so it can be used to predict the river flow more effectively using a specific subdivision of sub-datasets, and the accuracy is improved when PSO is used to achieve the best control of internal hyper-parameters.

FFWS accepts two types of data in the input: the water level measured in the Aghbalou river and the amount of rainfall recorded in the same region. Thus, two options to collect the required data are available, either we use the daily amount of rainfall recorded by

Table 6   Flood predefined thresholds

Catchment	Flows			
	Pre-alert situation		Alert situation	
	Water level on the coast (m)	Flow (m3/s)	Water level on the coast (m)	Flow (m3/s)
Aghbalou	13.58	125	14.69	500

internal meteorological services, these values are publically available on the platform of the Moroccan Department of National Meteorology (Meteo Maroc 2020), or we use our network of hydrology sensors based on water level gauges and rain sensors as shown in the "Material used in this study" section. The rainfall, the water level, and river flow collected data are used to forecast the hydraulic situation with high predictive power ($R^2 = 0.8707$). The forecasted river flow is later subject to the thresholding described in Table 6; these thresholds are defined by experience, in other words, based on flood historical events that normally occurred

between autumn and winter seasons and manual river flow measurements during each inundation event.

The thresholds described in Table 6 define two levels of alerts. When the forecasted river flow is less or equal to 125m³/s and the recorded water level in the Aghbalou catchment is also less than 13.58 m, the system understands that this is a pre-alert situation, so it delivers the pre-alert notification. However, if the predicted river flow value is greater or equal to 500 m³/s, and the recorded water level on the coast achieves 14.69 m, the system arises the warning of a probable flood. The flowchart of flood thresholding rules is described in Figure 14.

A review of existing flood forecasting methods

To comprehend and compare the performance of the system proposed in this study, it is advised to investigate the performance of already existing machine learning models used in hydrology and proposed for flood prediction; the next paragraph summarizes the most relevant ML models, notably

Fig. 14 Flowchart of flood thresholding rules
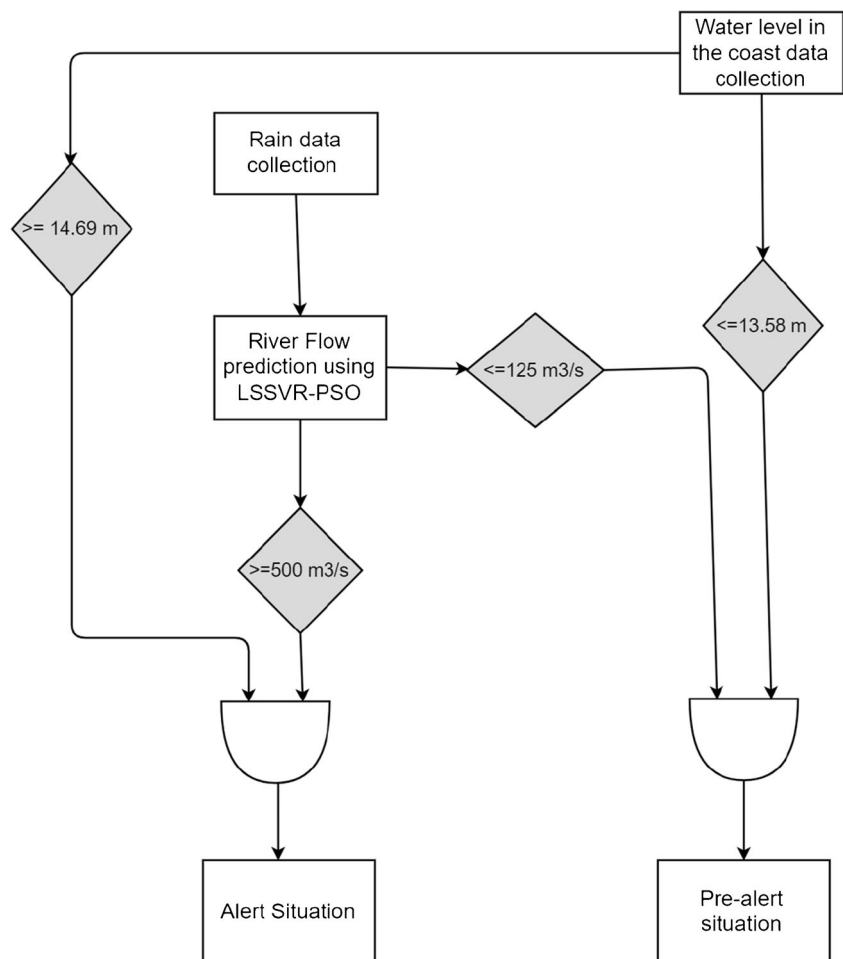
Fig. 15 ML methods for flood forecasting
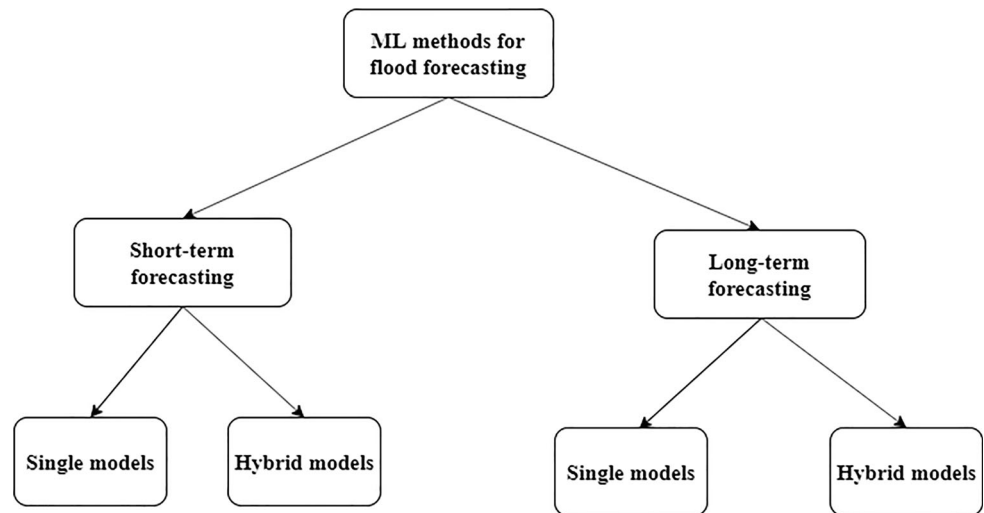
SVM, SVR, MLP, ANN, nonlinear autoregressive network with exogenous input (NARX), and others.

In literature, many studies (Mosavi et al. 2018) have classified flood prediction and water resource management methods into two big families: short-term and long-termmethods (Fig. 15). Each family consists of single and hybrid methods. The short-term machine learning methods are considered by Zhang et al. (2018) as highly important especially in an urban area since it helps to give more resistance and reduce damages in the more populated areas. However, long-term machine learning methods are significantly important for water resource management and to have more visibility about floods during periods considerably long (Choubin et al. 2016).

Starting from the summary presented in Table 7, the following comments can be performed: Kim et al. (2016) have demonstrated in their study the importance to select the appropriate dataset for better achievement using clustering analysis. Khosravi et al. (2018) confirmed the accuracy of ADT in flash flood position prediction compared to the other tested decision tree methods. Leahy et al. (2008) proposed an accurate optimization technique of the ANN model based on switching off the inter-neuron links before the training process and then adjusting the weights of the remaining connections using the classical backpropagation technique. To improve many metrics such as accuracy and time of training and reduce the complexity of the models, some references combine various ML models in a hybrid mode, Kim and Singh (2013) found that Kohonen self-organizing feature maps NN model (KSOFM-NNM) predict more accurately than multilayer perceptron NN model (MLPNNM) and generalized regression NN model (GRNNM) in the daily flood prediction. On the other hand, Tehrany et al. (2015) have reported that evaluation metrics can be improved using the SVM-FR ensemble method compared to the DT algorithm. Another proposition of the hybrid ML method is given by Hong (2008); the author showed that his hybrid model is a promising alternative to predict rainfall values.

Regarding long-term ML methods for flood forecasting, Deo and Sahin (2015) revealed that ANN represents a good data-driven tool to predict drought and its related properties. Another work (Lin, 2006) has shown the significant predictive capabilities of SVM compared to ARIMA and ANN in monthly river flow discharges prediction. Besides the aforementioned long-term methods known as single ML models, there exist other types of ML models called hybrid methods; this family was proposed in various works such as Li et al. (2009); the authors reported the good flood prediction skills of the modified NLPM-NN compared to the original NLPM-NN. Moreover, Zhu et al. (2016) reported that combining SVM, DWT, and EMD can greatly improve the accuracy of streamflow prediction.

Conclusion and perspectives

This study applied the LSSVR algorithm optimized using the PSO algorithm with a fine-tuning of penalty factor γ to build a monthly flood forecasting and warning system (FFWS) for the Moroccan atlas region (Aghbalou catchment).

The achieved quality of prediction (R-squared = 0.8707, root mean square error = 2.9102) explains the great ability of the regression model to describe the distribution of the observed data points, using RBF kernel and PSO algorithm to optimize LSSVR hyper-parameters. Three natural parameters are used to train the model, rainfall, the water level in the river coast, and previously recorded river flows. The proposed system combines sensors and a predictive learning algorithm (LSSVR-PSO) to build a civil security tool. Given the importance that is attached to the safety of lives and properties, this system among others can help a lot to anticipate flood disasters and critical damages that can occur.

Table 7 Shot-term and long-term flood prediction using ML methods

Reference	Flood prediction family	ML methods	Hydrological variables to forecast
Kim et al. (2016)	Short-term (hourly)	ANN	Water and after-runner surge level
Pereira Filho and dos Santos (2006)	Short-term (hourly)	ANN	Flash floods
Chang et al. (2014)	Short-term (hourly)	BPNN, NARX, Elman NN	The water level in the floodwater storage pond
Khosravi et al. (2018)	Short-term (hourly/daily)	NBT, ADT, REPT, LMT	Flash flood location forecasting
Leahy et al. (2008)	Short-term (hourly)	ANN	River level
Kim and Singh 2013	Short-term (daily)	MLPNNM, GRNNM, KSOFM-NNM	Flood prediction
Tehrany et al. (2015)	Short-term (real time)	SVM-FR, DT	Flood susceptibility assessment
Hong (2008)	Short-term (hourly)	RSVR optimized with CPSO	Rainfall depth values
Deo and Sahin (2015)	Long-term (monthly)	ANN	Drought forecasting based on Standardized Precipitation Evapotranspiration Index
Lin et al. (2006)	Long-term (monthly)	SVM, ARIMA, ANN	River flow discharges
Adnan et al. (2017)	Long-term (monthly)	MARS, M5RT, CMLR, HLGS	Flood prediction
Li et al. (2009)	Long-term (yearly)	A modified version of NLPM-NN	Catchment witnesses
Zhu et al. (2016)	Long-term (monthly)	SVM-DWT-EMD	Streamflow estimation

For future researches, an improved version of the model, combining time series preprocessing technique and data-driven approach, can be suggested to enhance the forecasting accuracy of the flooding situation. These preprocessing techniques can be addressed using more reliable statistical tests such as the Dickey-Fuller test. Also, the forecasting skills of the proposed system can be improved by considering more meteorological and geological variables notably valley morphology, slope and river gradients, sedimentology, vegetation cover, and soil characteristics. Another limitation related to the sources of data can also be compensated by increasing the density of hydrology sensors network installed in the basin, so more variability in the water level and the rainfall can be captured.

Abbreviations ADT, alternating decision trees; AIC, Akaike information criteria; ANN, artificial neural networks; ARIMA, autoregressive integrated moving average; BIC, Bayesian information criterion; BPNN, Backpropagation neural network; CMLR, conventional multiple linear regression; CPSO, chaotic particle swarm optimization; DC, determination coefficient; FFWS, flood forecasting and warning system; FR, frequency ratio; GRNNM, generalized regression NN model; HLGS, hybrid least square support vector regression-gravitational search; KF, kernel function; k-fold CV, K-fold cross-validation; KSOFM-NNM, Kohonen self-organizing feature maps NN model; LHLGS, log hybrid least square support vector regression-gravitational search; LMT, logistic model trees; LSSVM, least square support vector machine; LSSVR, least square support vector regression; M5RT, model 5 regression tree; MAE, mean absolute error; MARS, multivariate adaptive regression splines; MLP, multilayer perceptron; MLPNNM, multilayer perceptron NN model; NARX, nonlinear autoregressive network with exogenous input; NBT, Naïve Bayes trees; PSO, particle swarm optimization; QP, quadratic programming; RBFKF, radial basis function kernel function; REPT, reduced error pruning trees; RF, river flow; RMSE, root mean square error; RSVR, recurrent SVR; SARIMA, seasonal autoregressive integrated moving average; SVM, support vector machine; SVM-DWT-EMD, support vector machine-discrete wavelet transform-empirical mode decomposition; SVM-FR, support vector machine-frequency ratio; SVR, support vector regression

Declarations

References

Adnan RM, Yuan X, Kisi O, Anam R (2017) Improving accuracy of river flow forecasting using LSSVR with gravitational search algorithm. Adv Meteorol 2017:1687–9309. https://doi.org/10.1155/2017/2391621

Ahlert RC, Mehta BM (1981) Stochastic analyses and transfer functions for flows of the upper Delaware River. J Ecol Model 14:59–78. https://doi.org/10.1016/0304-3800(81)90014-4

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F (eds) 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8 1971, Budapest: Akadémiai Kiadó, pp. 267–281. Republished in: Kotz S, Johnson NL (eds) (1992), Breakthroughs in Statistics, I, Springer-Verlag, pp 610–624.

Chang F-J, Chen P-A, Lu Y-R, Huang E, Chang K-Y (2014) Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control. J Hydrol 517:836–846. https://doi.org/10.1016/j.jhydrol.2014.06.013

Choubin B, Khalighi SS, Malekian A (2016) Impacts of large-scale climate signals on seasonal rainfall in the Maharlu-Bakhtegan watershed. J Range Watershed Manag 69:51–63. https://doi.org/10.22059/jrwm.2016.61733

Dehghani M, Saghafian B, Farzin NS, Ashkan F, Roohollah N (2013) Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. Int J Climatol 34:1169–1180. https://doi.org/10.1002/joc.3754

Deo RC, Sahin M (2015) Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in Eastern Australia. Atmos Res 161:65–81. https://doi.org/10.1016/j.atmosres.2015.03.018

Eberhart R, Kennedy J (1995) A new optimizer using particle swarm theory. Paper presented at the Sixth International Symposium on Micro Machine and Human Science, Piscataway, NJ, USA. doi: https://doi.org/10.1109/MHS.1995.494215

El Khalki EM, Tramblay Y, El Mehdi SM, Bouvier C, Hanich L, Benrhanem M, Alaouri M (2018) Comparison of modeling approaches for flood forecasting in the High Atlas Mountains of Morocco. Arab J Geosci 11:410. https://doi.org/10.1007/s12517-018-3752-7

Gestel TV, Suykens JAK, Baestaens DE, Lambrechts A, Lanckriet G, Vandaele B, Moor BD, Vandewalle J (2001) Financial time series prediction using least squares support vector machines within the evidence framework. IEEE Trans Neural Netw 12:809–821. https://doi.org/10.1109/72.935093

Ghimire NS, Bhola (2017) Application of ARIMA model for river discharges analysis. J Nepal Phys Soc 4:27–32. https://doi.org/10.3126/jnphyssoc.v4i1.17333

Hanh PTM, Anh NV, Ba DT, Sthiannopkao S, Kim K-W (2010) Analysis of variation and relation of climate, hydrology and water quality in the lower Mekong River. Water Sci Technol 62:1587–1594. https://doi.org/10.2166/wst.2010.449

Hannan EJ, Quinn BG (1979) The determination of the order of an autoregression. J R Stat Soc 41:190–195. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x

Hong W-C (2008) Rainfall forecasting by technological machine learning models. Appl Math Comput 200:41–57. https://doi.org/10.1016/j.amc.2007.10.046

Kalteh AM (2013) Monthly river flow forecasting using artificial neural network and support vector regression models coupled with wavelet transform. Comput Geosci 54:1–8. https://doi.org/10.1016/j.cageo.2012.11.015

Katimon A, Shahid S, Mohsenipour M (2018) Modeling water quality and hydrological variables using ARIMA: a case study of Johor River, Malaysia. Sustain Water Resour Manag 4:991–998. https://doi.org/10.1007/s40899-017-0202-8

Khosravi K, Pham BT, Chapi K, Shirzadi A, Shahabi H, Revhaug I, Prakash I, Bui DT (2018) A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, Northern Iran. Sci Total Environ 627:744–755. https://doi.org/10.1016/j.scitotenv.2018.01.266

Kim S, Singh VP (2013) Flood forecasting using neural computing techniques and conceptual class segregation. J Am Water Resour Assoc 49:1421–1435. https://doi.org/10.1111/jawr.12093

Kim S, Matsumi Y, Pan S, Mase H (2016) A real-time forecast model using artificial neural network for after-runner storm surges on the Tottori Coast, Japan. Ocean Eng 122:44–53. https://doi.org/10.1016/j.oceaneng.2016.06.017

Kisi O, Parmar KS (2016) Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J Hydrol 534:104–112. https://doi.org/10.1016/j.jhydrol.2015.12.014

Landeras G, Ortiz-Barredo A, Lopez JJ (2009) Forecasting weekly evapotranspiration with ARIMA and artificial neural network models. J Irrig Drain Eng 135:323–334. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000008

Leahy P, Kiely G, Corcoran G (2008) Structural optimisation and input selection of an artificial neural network for river level prediction. J Hydrol 355:192–201. https://doi.org/10.1016/j.jhydrol.2008.03.017

Li C, Guo S, Zhang J (2009) Modified NLPM-ANN model and its application. J Hydrol 378:137–141. https://doi.org/10.1016/j.jhydrol.2009.09.017

Lin J-Y, Cheng C-T, Chau K-W (2006) Using support vector machines for long-term discharge prediction. Hydrol Sci J 51:599–612. https://doi.org/10.1623/hysj.51.4.599

Makridakis S, Andersen A, Carbone R, Fildes R, Hibon M, Lewandowski R, Newton J, Parzen E, Winkler R (1982) The accuracy of extrapolation (time-series) methods: results of a forecasting competition. J Forecast 1:111–153

Meteo Maroc (Weather Morocco). from http://www.meteomaroc.com/. Accessed 23 July 2020.

Mohan S, Arumugam N (1995) Forecasting weekly reference crop evapotranspiration series. Hydrol Sci J 40:689–702. https://doi.org/10.1080/02626669509491459

Mosavi A, Ozturk P, Chau K-W (2018) Flood prediction using machine learning models: literature review. Water 11:1536. https://doi.org/10.3390/w10111536

Nigam R, Bux S, Nigam S, Pardasani KR, Mittal SK, Haque R (2009) Time series modeling and forecast of river flow. Curr World Environ 4:79–87. https://doi.org/10.12944/CWE.4.1.11

Pereira Filho AJ, Dos Santos CC (2006) Modeling a densely urbanized watershed with an artificial neural network, weather radar and telemetric data. J Hydrol 317:31–48. https://doi.org/10.1016/j.jhydrol.2005.05.007

Rana I, Zhongmin L, Xiaohui Y, Ozgur K, Wattoo M, Li B (2019) Comparison of LSSVR, M5RT, NF-GP, and NF-SC models for predictions of hourly wind speed and wind power based on cross-validation. Energies 12:329. https://doi.org/10.3390/en12020329

Santhi C, Arnold JG, Williams JR, Dugas WA, Srinivasan R, Hauck LM (2001) Validation of the swat model on a large RWER basin with point and nonpoint sources. J Am Water Resour Assoc 37:1169–1188. https://doi.org/10.1111/j.1752-1688.2001.tb03630.x

Schwarz GE (1978) Estimating the dimension of a model. Ann Stat 6:461–464. https://doi.org/10.1214/aos/1176344136 MR 0468014

Shimi M, Najjarchi M, Khalili K, Hezavei E, Mirhoseyni SM (2020) Investigation of the accuracy of linear and nonlinear time series models in modeling and forecasting of pan evaporation in IRAN. Arab J Geosci 13. https://doi.org/10.1007/s12517-019-5031-7

Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9:293–300. https://doi.org/10.1023/A:1018628609742

Tehrany MS, Pradhan B, Jebur MN (2015) Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. Stoch Env Res Risk A 29:1149–1165. https://doi.org/10.1007/s00477-015-1021-9

Vapnik V, Lerner A (1963) Pattern recognition using generalized portrait method. Autom Remote Control 24:774–780

Yang S, Liu S, Li X, Zhong Y, He X, Wu C (2017) The short term forecasting of vaporation duct height (EDH) based on ARIMA model. Multimed Tools Appl 76:24903–24916. https://doi.org/10.1007/s11042-016-4143-2

Zhang J, Hou G, Ma B, Hua W (2018) Operating characteristic information extraction of flood discharge structure based on complete ensemble empirical mode decomposition with adaptive noise and permutation entropy. J Vib Control 24:5291–5301. https://doi.org/10.1177/1077546317750979

Zhu S, Zhou J, Ye L, Meng C (2016) Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of Yangtze River, China. Environ Earth Sci 75. https://doi.org/10.1007/s12665-016-5337-7