Hindawi
Geofluids
Volume 2018, Article ID 8328167, 13 pages
https://doi.org/10.1155/2018/8328167

Research Article

Machine Learning Models for Spring Discharge Forecasting

Francesco Granata, Michele Saroli, Giovanni de Marinis, and Rudy Gargano

Department of Civil and Mechanical Engineering, University of Cassino and Southern Lazio, via G. Di Biasio 43, 03043 Cassino (FR), Italy

Correspondence should be addressed to Francesco Granata; f.granata@unicas.it

Received 3 May 2018; Revised 4 August 2018; Accepted 30 September 2018; Published 4 December 2018

Academic Editor: John A. Mavrogenes

Nowadays, drought phenomena increasingly affect large areas of the globe; therefore, the need for a careful and rational management of water resources is becoming more pressing. Considering that most of the world's unfrozen freshwater reserves are stored in aquifers, the capability of prediction of spring discharges is a crucial issue. An approach based on water balance is often extremely complicated or ineffective. A promising alternative is represented by data-driven approaches. Recently, many hydraulic engineering problems have been addressed by means of advanced models derived from artificial intelligence studies. Three different machine learning algorithms were used for spring discharge forecasting in this comparative study: M5P regression tree, random forest, and support vector regression. The spring of Rasiglia Alzabove, Umbria, Central Italy, was selected as a case study. The machine learning models have proven to be able to provide very encouraging results. M5P provides good short-term predictions of monthly average flow rates (e.g., in predicting average discharge of the spring after 1 month, $R^2 = 0.991$, RAE = 14.97%, if a 4-month input is considered), while RF is able to provide accurate medium-term forecasts (e.g., in forecasting average discharge of the spring after 3 months, $R^2 = 0.964$, RAE = 43.12%, if a 4-month input is considered). As the time of forecasting advances, the models generally provide less accurate predictions. Moreover, the effectiveness of the models significantly depends on the duration of the period considered for input data. This duration should be close to the aquifer response time, approximately estimated by cross-correlation analysis.

1. Introduction

In recent years, long and frequent droughts have affected many countries in the world. These events require an ever more careful and rational management of water resources. Most of the globe's unfrozen freshwater reserves are stored in aquifers. Groundwater is generally a renewable resource that shows good quality and resilience to fluctuations. Thus, if properly managed, groundwater could ensure long-term supply in order to meet increasing water demand.

For this purpose, it is of crucial importance to be able to predict the flow rates provided by springs. These represent the transitions from groundwater to surface water and reflect the dynamics of the aquifer, with the whole flow system behind. Moreover, spring influences water bodies into which they discharge. The importance of springs in groundwater research is highlighted in some significant contributions [1, 2]. In-depth studies on springs started only after the concept of sustainability was introduced in the management of water resources [3].

A spring hydrograph is the consequence of several processes governing the transformation of precipitation in the spring recharge area into the single output discharge at the spring. A water balance states that the change rate in water stored in the feeding aquifer is balanced by the rate at which water flows into and out of the aquifer. A quantitative water balance generally has to take the following terms into account: precipitation, infiltration, surface runoff, evapotranspiration, groundwater recharge, soil moisture deficit, spring discharge, lateral inflow to the aquifer, leakage between the aquifer and the underlying aquitard, well pumpage from the aquifer, and change of the storage in the aquifer.

In many cases, the evaluation of the terms of the water balance is very complicated. The complexity of the problem arises from many factors: hydrologic, hydrographic, and hydrogeological features, geologic and geomorphologic characteristics, land use, land cover, water withdrawals, and climatic conditions.

Even more complicated would be to estimate future spring discharges by using a model based on the balance equations. Therefore, simplified approaches are frequently pursued for practical purposes.

Many authors have addressed the problem of correlating the spring discharges to the rainfall through different approaches. Zhang et al. [4] used a lumped-parameter model and least-squares method to simulate temporal variations of discharge from a limestone aquifer in Iowa, USA. Lambrakis et al. [5] applied nonlinear time series analysis and artificial neural network to the study of the regime and the possibility of short-term forecasting of the discharges of the karstic spring of Almyros in Iraklion, Greece. Hu et al. [6] developed a model for simulating spring discharges using ANN and applied the model to Niangziguan Springs, China. Fiorillo and Doglioni [7] presented a study on the relation between rainfall and the discharge from two karst springs in Southern Italy based on cross-correlation analyses. Fan et al. [8] proposed an assembled extreme value statistical model to investigate spring discharge depletion processes under extreme climate change and intense groundwater development. Diodato et al. [9] proposed a lumped climatological model for spring discharge estimation in which groundwater is represented by a single reservoir and spring discharge is described by a single valued function of storage in the reservoir.

Recently, many researchers have investigated the feasibility of addressing hydraulic engineering issues by means of advanced models derived from artificial intelligence studies. Regression Tree models, Ensemble methods, and support vector machines have been increasingly used in solving water engineering problems.

Dibike et al. [10] compared support vector machines (SVMs) and artificial neural networks (ANNs) in rainfall-runoff modelling. Ahmad et al. [11] employed SVMs to estimate soil moisture on the base of remote sensing data. Raghavendra and Deka [12] provided an extensive review of the SVM applications to hydrology problems. Granata et al. [13] carried out a comparative study between support vector regression (SVR) and Stormwater Management Model in the rainfall-runoff modelling in urban drainage. Najafzadeh et al. [14] used SVM and ANFIS for scour prediction in long contractions in waterways. Granata et al. [15] applied SVR and regression trees in order to predict wastewater quality indicators.

Tree models or ensemble methods were implemented to forecast flood events [16], to foresee mean annual flood [17], to predict scour depth due to waves [18], to evaluate sediment yield in rivers [19], to predict maximum scour depth around piers [20], to address wastewater hydraulics issues [21], to forecast local scour depth downstream of sluice gates [22], to evaluate sediment transport [23], and to evaluate the flow discharge in main channels and floodplains [24].

The aim of this study is to assess the ability of a machine-learning algorithm-based approach in predicting average monthly discharge of a spring, when the forecast horizon does not exceed a few months, if few years of monthly flow rate measurements and rainfall data are available. Therefore, regression tree, random forest, and support vector regression were used to build forecasting models and to perform a comparative study. The proposed approach was tested by means of experimental data obtained from the spring of Rasiglia Alzabove, Umbria, Central Italy. Time series data are available from the Regional Agency for Environmental Protection (http://www.arpa.umbria.it). Future monthly average discharges were correlated to the average discharges of the past months and to the cumulative rainfall of the same months.

2. Materials and Methods

2.1. Machine Learning Models: Regression Trees. A regression tree (RT) model (Figure 1) develops a decision tree in order to make predictions [25]. The target variables take real values. In a regression tree, each internal node represents one of the input variables, whereas each leaf corresponds to an assigned value of the target variable and a root node contains all data.

During the growth of a regression tree model, the input data domain is recursively divided into subdomains. The predictions are made in each of them by means of multivariable linear regression models. At the first step of the iterative algorithm, all data are allocated into two branches, considering all the possible split on every field. Subsequently, the development process of the regression tree continues by splitting each branch into smaller partitions, as the system expands. At each stage, the procedure identifies the subdivision in two distinct partitions that minimize the sum of the squared deviations from the mean. This sum can be considered a measure of the “impurity” at a node that is a quantification of the predictive capability of the node. The algorithm continues until the lowest impurity level is obtained or until a stopping rule is met. Usually, a stopping rule is related to the threshold for the minimum impurity variation provided by new splits, the minimum number of units in each node, or the maximum tree depth.

The algorithm here used is commonly known as M5P and is based on Quinlan’s M5 algorithm [26]. The impurity of each node is estimated by the least-squared deviation (LSD), $R(t)$, defined as

$$R(t) = \frac{1}{N(t)} \sum_{i \in t} (y_i - y_{\mathrm{m}}(t))^2, \quad (1)$$

in which $N(t)$ is the number of sample units in node t, y_i is the value of the target variable for the i – th unit, and y_{m} is the mean of the target variable in node t.

The split process at each node is carried out based on the following function of the LSD:

$$\phi(s_p, t) = R(t) - p_{\mathrm{L}} R(t_{\mathrm{L}}) - p_{\mathrm{R}} R(t_{\mathrm{R}}), \quad (2)$$

in which t_{L} and t_{R} are the left and right nodes generated by the split s_p, while p_{L} and p_{R} are the portions allocated in the

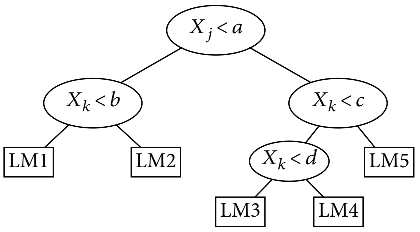


Figure 1: Typical architecture of a simple regression tree. LMs: linear models.

left and right child node. Finally, the subdivision s_p that maximizes the value of $\phi\,(s_p, t)$ is adopted.

A regression tree might suffer from overfitting when the model structure is fully developed. Overfitting occurs when a machine learning model has become too attuned to the data on which it was trained and therefore loses its applicability to any other dataset. Therefore, overfitting reduces the tree ability to make predictions, when the model is applied to novel data. To minimize this risk, a *pruning* process is generally carried out to reduce the size of the regression tree by removing the splits that do not bring significant improvements to the forecasting capability. Overfitting could also be prevented by extending the training dataset.

2.2. Random Forest. If the results of different regression trees are combined into a single prediction, an *ensemble method* is obtained. The simplest combination is a weighted average.

A random forest (RF) [27] is obtained by an ensemble of uncorrelated, simple regression trees (Figure 2). Each tree is built using a different bootstrap sample of the data, and each new training set is attained, with replacement, from the original training set. In addition, random forests differ in how the regression trees are built. In a standard tree, each node is divided by referring to the best split among all variables. A random forest, instead, is built by randomly choosing, at each node, a small group of input variables to split on. If the input variables are M, a number $m \ll M$ is identified such that at each node, m variables are randomly selected out of M and the best split on these m is used to subdivide the node. The value of m is kept constant during the expansion of the forest. Each tree is expanded as far as possible. No pruning is carried out.

The forest error rate is affected by two factors: the correlation between any two trees and the strength of each single tree. The forest error rate rises if the correlation increases. Moreover, the forest error rate decreases if the strength of the individual trees increases. If m is reduced, both the strength and the correlation decrease.

2.3. Support Vector Regression. Different is the approach of support vector machine algorithms [28]. They are supervised learning models that analyse data for classification or regression analysis [29]. If $\{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \subset X \times \mathrm{R}$ is a training dataset, where X is the space of the input patterns (e.g., $X = \mathrm{R}^n$), support vector regression (SVR) is aimed at identifying a function $f(x)$ that has a maximum ε deviation from the experimental target values y_i for all the training data. Smaller than ε errors can be tolerated, while the greater than ε errors are generally unacceptable (Figure 3). In addition, $f(x)$ should be as flat as possible.

Therefore, given a linear function in the form

$$f(x) = \langle w, x\rangle + b, \tag{3}$$

in which $w \in X$, $b \in \mathrm{R}$, and $\langle .\,, .\rangle$ is the dot product in X, the Euclidean norm $\|w\|^2$ has to be minimized, respecting the maximum deviation constraint. This condition leads to a convex optimization problem. However, in many cases, a certain error has to be tolerated. Therefore, slack variables ξ_l, ξ_l^* have to be introduced in the constraints of the optimization problem that can be formulated as

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right), \\ &\text{Subject to} \quad \begin{cases} y - \langle w, x_i\rangle - b \le \varepsilon + \xi_i^*, \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^*, \end{cases} \end{aligned} \tag{4}$$

The constant $C > 0$ affects both the flatness of f and the accepted deviations.

The optimization problem stated in (4) is generally solved in its dual formulation, by means of Lagrange multipliers:

$$\begin{aligned} L = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{l}\alpha_{\mathrm{i}}(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\ & - \sum_{i=1}^{l}\alpha_i^*\left(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b\right) - \sum_{i=1}^{l}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right), \end{aligned} \tag{5}$$

in which $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \ge 0$.

The partial derivatives of L with respect to the variables $(w, b, \xi_i\, \xi_i^*)$ must be equal to zero in correspondence with the optimal condition. It follows that the optimization problem in dual form can be stated as

$$\text{Maximize} \quad \begin{cases} -\dfrac{1}{2}\displaystyle\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)\langle x_i, x_j\rangle, \\ -\varepsilon\displaystyle\sum_{i=1}^{l}(a_i + a_i^*) + \sum_{i=l}^{l} y_i\left(a_i - a_j^*\right), \end{cases} \tag{6}$$

$$\text{Subject to} \quad \begin{cases} \displaystyle\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, \\ a_i, a_i^* \in [0, C]. \end{cases} \tag{7}$$

The evaluation of b is obtained imposing the Karush-Kuhn-Tucker conditions [30], according to which the product between constraints and dual variables must be equal to zero at the optimal condition.

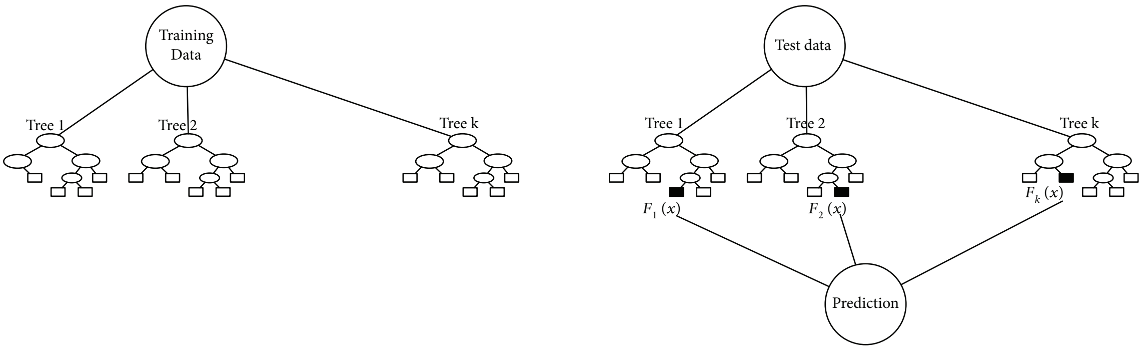


Figure 2: Regression tree ensemble: a typical random forest.

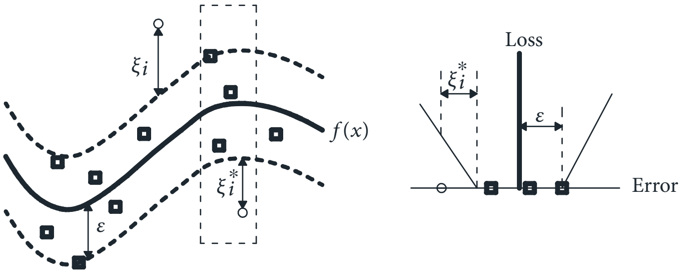


Figure 3: Example of support vector regression: smaller than ε errors are not relevant, while larger deviations are penalized in a linear fashion. "Loss" is the penalty for larger than ε deviations.

In order to make the SVR algorithm nonlinear, the training patterns x_i may be preliminarily processed by a function Φ: $X \longrightarrow F$, where F is some feature space. The SVR algorithm only depends on the dot products between the different patterns; therefore, it is possible to use a kernel $k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ instead of specifically using the map $\Phi(\cdot)$ (Figure 4).

It follows that condition (6) in the optimization problem can be replaced by

$$\text{Maximize} \quad \begin{cases} -\dfrac{1}{2}\displaystyle\sum_{i,j=1}^{l} (\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right) k\left(x_i, x_j\right), \\ -\varepsilon \displaystyle\sum_{i=1}^{l} (a_i + a_i^*) + \sum_{i=1}^{l} y_i \left(a_i - a_j^*\right), \end{cases} \tag{8}$$

while constraints expressed by (7) remain unchanged.

In the nonlinear case, the optimization problem requires finding the flattest function in the feature space, not in the input space. The expansion of f can be formulated as

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b. \tag{9}$$

In this research, a radial basis function (RBF) was selected as kernel. The RBF has the form

$$k\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \quad \gamma > 0. \tag{10}$$

In particular, in the case study, the parameters assume the following values: $C = 1$, $\varepsilon = 0.001$, and $y = 0.01$.

The described algorithms were implemented in a specific code written in MATLAB language. The search for the optimal structure of the models was conducted by means of a trial-and-error iteration procedure. The holdout method was used in the cross-validation process during training. It involves removing a part of the training data and using it to get predictions from the model trained on the rest of the data. The estimation error tells how the model is doing on unseen data or the validation set.

2.4. Case Study. The Menotre River Valley, in correspondence of the Rasiglia town (Umbria Region), is characterized by the Capo Vena spring (elevation of 670 m a.s.l. and average discharge of 700 l/s), Alzabove spring (elevation of 650 m a.s.l. and average discharge of 250 l/s), and the minor spring of Verchiano Aqueduct (elevation of 650 m a.s.l. and average discharge of 45 l/s) [31]. The Alzabove spring, partially collected by the Sella Valle Umbra Sud aqueduct (derived water discharge 125 l/s), represents an ideal site for the application of the proposed methodology, due to the available discharge time series and the favorable hydrogeological setting.

Carbonate deposits of the Umbria-Marche sequence characterize the area [31]. The stratigraphic and tectonic setting influences the groundwater circulation and the emergence point of the Alzabove spring. In wider terms, the Umbria-Marche series can be divided in three hydrogeological complexes consisting in the Corniola and Calcare Massiccio basal complex, the Maiolica complex, and the

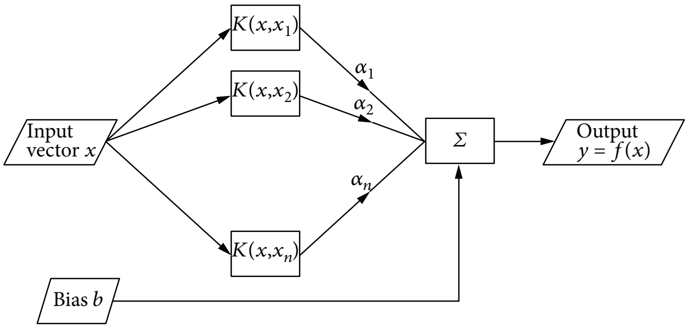


FIGURE 4: Typical architecture of the nonlinear SVR algorithm.

Scaglia Calcarea complex. These hydrogeological complexes with high permeability and high storing capacity constitute the main regional aquifers.

The complexes of "Marne del Sentino-Rosso Ammonitico-Marne ad Aptici Fm.," "Marne a Fucoidi," and "Scaglia Cinerea-Marnoso Arenacea" have an aquitard function and divide the groundwater circulation of the calcareous complexes. Groundwater of the Alzabove spring takes origin from the contact between Maiolica Fm. (lower Cretaceous) and Marne a Fucoidi Fm. (middle Cretaceous), via a contact of 500 m length, along the right side of Menotre River (Figure 5(a)). The contact between Maiolica Fm. and Marne a Fucoidi Fm. is located along the western limb of the Toricello Mt. anticlinal (Figures 5(a) and 5(b)) belonging to a large fold system. The anticline fold consists in layers of the Maiolica Fm. and constitutes the natural reservoir which feeds the Alzabove spring. Marne a Fucoidi Fm. represents the permeability limit of the aquifer [31–33] (Figures 5(a) and 5(b)).

In this context, the Alzabove spring can be considered an overflow spring (Figure 5(b)), collected via a trench drain, located along the Toricello Mt. and at the contact between Maiolica Fm. and Marne a Fucoidi Fm. From a geochemical point of view, the Alzabove spring is a bicarbonate-calcium water with a TDS of 300 mg/l, a Mg/Ca ratio of 0.12, and isotopic values of ∂34S equal to +7.5‰. Values are different from the water of the Capo Vena spring that has bicarbonate-sulphate-calcium characteristics with a TDS of 500 mg/l, Mg/Ca ratio of 0.25, and isotope value of ∂34S equal to +13‰ [33].

The chemical analysis proves that the Alzabove spring is characterized by low contents of Mg, typical of Maiolica Fm. On the contrary, groundwater circulation of the Capo Vena spring is affected by dolomitic horizons with sulphates. Therefore, the groundwater path of the Alzabove spring is shallower and less influenced by karst phenomena and tectonics, if compared with the Capo Vena reservoir which comprehends the basal portion of the Umbria-Marche sequence (lower Lias).

The identification of the hydrogeological model of the Alzabove spring confirms that it constitutes a useful case for the proposed algorithm training.

2.5. Effectiveness of the Models. The effectiveness of forecasting algorithms was evaluated by the following criteria: the coefficient of determination R^2, the mean absolute error (MAE), the root mean squared error (RMSE), and the relative absolute error (RAE).

The coefficient of determination, R^2, is a measure of the accuracy of the model, because it assesses how well the model replicates observed outcomes and how well it predicts future outcomes. R^2 is defined as

$$R^2 = \left(1 - \frac{\sum_{i=1}^{m}(f_i - y_i)^2}{\sum_{i=1}^{m}(y_a - y_i)^2}\right), \quad (11)$$

where m is the total number of observed data, f_i is the predicted value for data point i, y_i is the measured value for data point i, and y_a is the averaged value of the experimental data.

The mean absolute error measures how much the predictions are close to the observed values. It is evaluated by

$$\mathrm{MAE} = \frac{\sum_{i=1}^{m}|f_i - y_i|}{m}. \quad (12)$$

RMSE is the sample standard deviation of the differences between experimental and predicted values. It is given by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{m}(f_i - y_i)^2}{m}}. \quad (13)$$

Finally, RAE normalizes the total absolute error dividing it by the total absolute error of the simple predictor. Its definition is

$$\mathrm{RAE} = \frac{\sum_{i=1}^{m}|f_i - y_i|}{\sum_{i=1}^{m}|y_a - y_i|}. \quad (14)$$

3. Results and Discussion

The input data of the different models are the past monthly average flow rates, $Q_{-1}, Q_{-2}, \ldots, Q_{-k}$, and the cumulative monthly rainfall on the aquifer basin, $P_{-1}, P_{-2}, \ldots, P_{-k}$. A time series of 60 months (Figure 6) was used in the calculations; 80% of the available data was used for the algorithm training. The proposed approach is effective even with data from a single rain gauge; thus, rainfall data were taken from the Norcia rain gauge (Figure 5(c)). They optimally represent

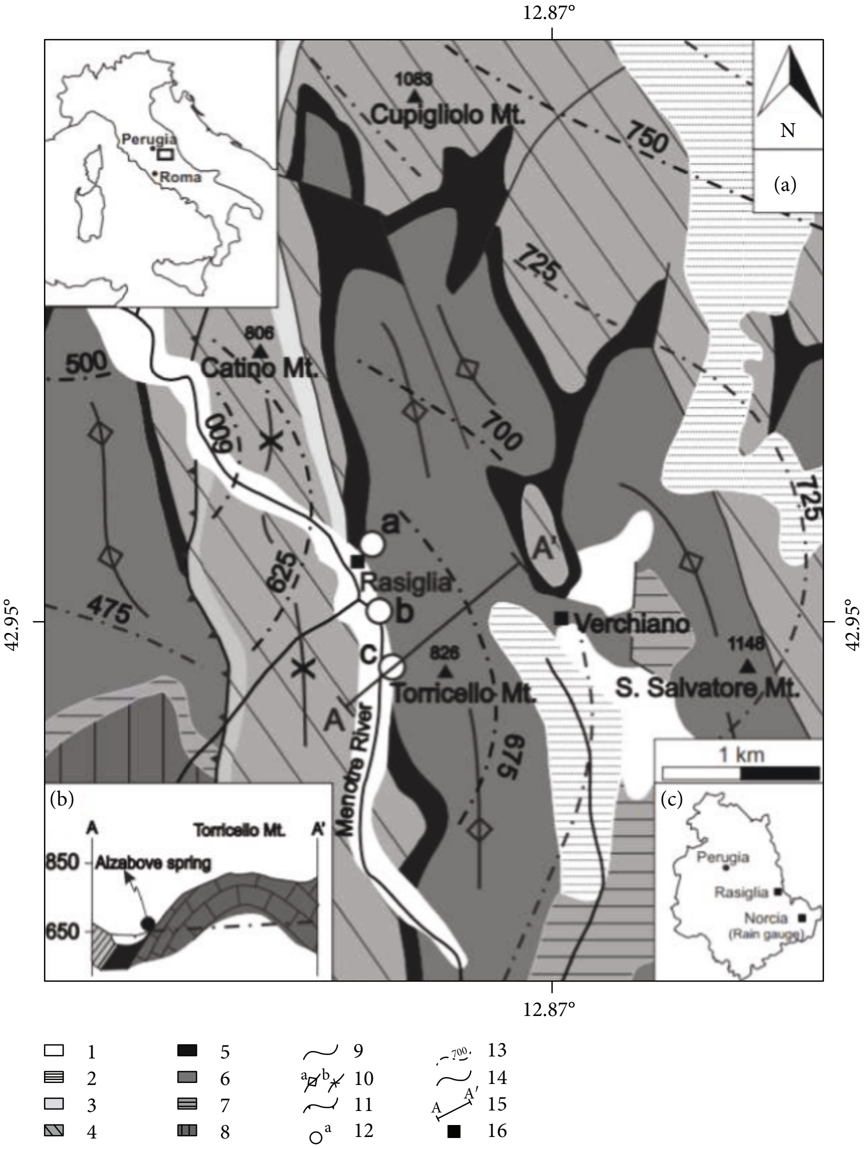


Figure 5: Geological and hydrogeological map (a), section of the Alzabove spring area (b), and rain gauge location (c). Key to the legend: (1) the talus and alluvial deposit complex (Olocene-Pleistocene) has high permeability which constitutes the local aquifer; (2) the lacustrine deposit complex (Olocene-Pliocene) has an aquitard function; (3) terrigenous complexes (marls, scaly clays, and sandstones) (Miocene) have an aquitard function; (4) the Scaglia calcarea complex (Scaglia Rossa and Scaglia Bianca Fm.) (Eocene-Cretaceous) has high permeability and high storing capacity which constitute the regional aquifer; (5) the Marne a fucoidi complex (Lower Cretaceous) has an aquitard function; (6) Maiolica complex (Lower Cretaceous-Jurassic) has high permeability and high storing capacity which constitute the regional aquifer; (7) the calcareous siliceous marly complex (Marne del Sentino-Rosso Ammonitico-Marne ad Aptici Fm.) (Upper Jurassic) has an aquitard function; (8) the Corniola-Calcare Massiccio basal complex (Lower Jurassic) has high permeability and high storing capacity which constitute the regional aquifer; (9) fault; (10) folds: (a) anticline and (b) syncline; (11) thrust; (12) springs: (a) Capo Vena, (b) Verchiano, and (c) Alzabove; (13) groundwater level: the numbers indicate the water above sea level; (14) river; (15) section trace; (16) village.

the cumulative rainfall on the basin. In addition, preliminary analyses showed that data from Norcia pluviometry are characterized by the highest values of cross-correlation with the flow rates of the Alzabove springs with respect to the other rain gauges located in the recharge basin.

The available time series is not very long, but this is not a limitation as one of the primary objectives of this study is to evaluate the predictive capabilities of the considered models when a few years of experimental observations are available.

In Figure 7 the cross-correlation between cumulative monthly rainfall and average monthly discharge is reported. Obviously, the positive values of the lag have no real interest. The analysis of the cross-correlogram shows that the cross-correlation takes a maximum value for a lag value of about 4 months. The hydrogeological characteristics of several

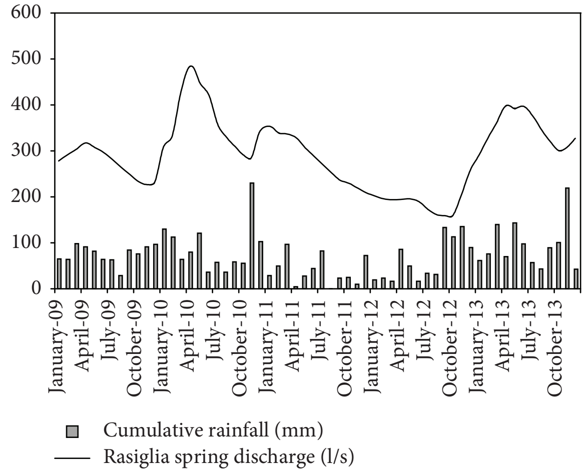


Figure 6: Time series of rainfall and spring discharge.

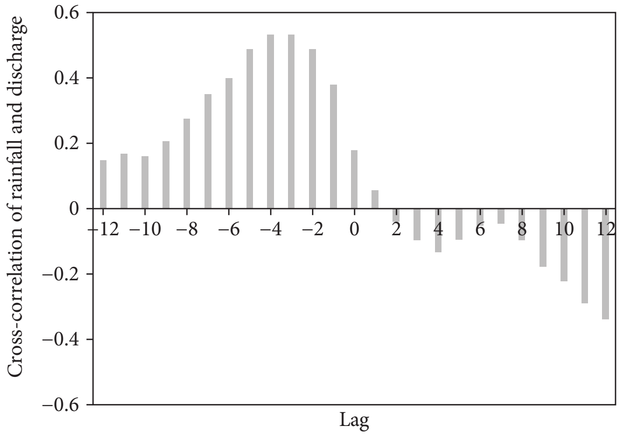


Figure 7: Cross-correlogram of cumulative monthly rainfall and average monthly discharge.

aquifers in Central Italy would suggest a higher lag. Therefore, data on the previous 4, 6, and 8 months were alternatively considered, in order to assess the influence of the input period duration on the model's response.

Preliminary analysis showed that better results are obtained if the input vector has the same number of flow rate and cumulative rainfall data. Thus, each vector of the input matrix to the three different models is composed as follows:

$$[Q_{-1}, Q_{-2}, \ldots, Q_{-k}, P_{-1}, P_{-2}, \ldots, P_{-k}], \quad (15)$$

where $k = 4, 6, 8$.

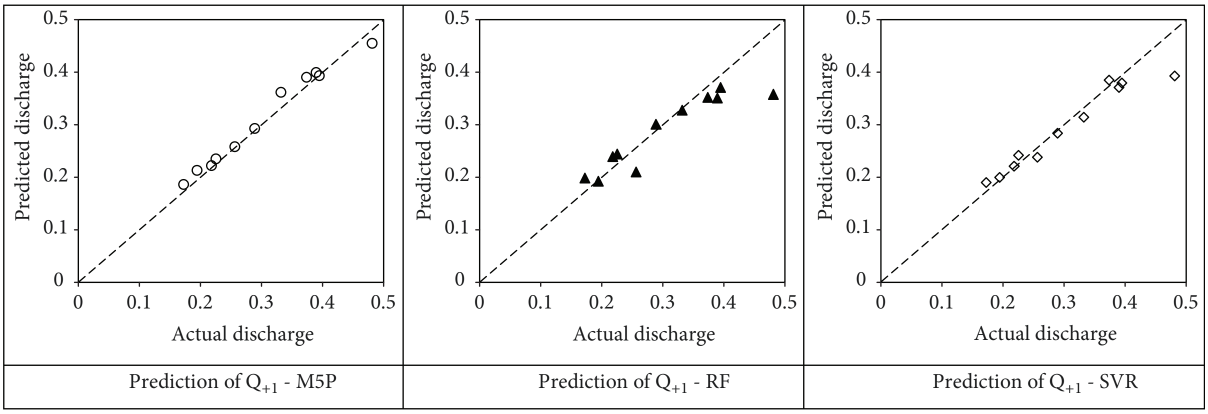
Different models were built to predict the monthly average discharge of the spring after 1 month, Q_{+1}, and the average discharge after 2 months, Q_{+2}, after 3 months, Q_{+3}, and after 4 months, Q_{+4}.

Table 1 and Figures 8–19 show a comparison of the results provided by the different models.

With regard to the average discharge of the following month, Q_{+1}, all models show adequate predictive capabilities (Table 1) when the input data takes into account the previous 4 months (Figure 8) or the previous 6 months (Figure 9). In these cases, the M5P algorithm, followed by the SVR algorithm, provides the best results. The RF algorithm leads to not fully satisfactory results, particularly for an input based on the previous 4 months, for which RAE = 37.29%. M5P provides slightly better results with an input based

Table 1: Comparative analysis of M5P, RF, and SVR by means of R^2, MAE, RMSE, and RAE.

	Model	R^2	MAE (m^3/s)	RMSE (m^3/s)	RAE
Q_{+1}					
4-month input	M5P	0.991	0.0124	0.0156	14.97%
	RF	0.926	0.0309	0.0446	37.29%
	SVR	0.97	0.0196	0.0299	23.67%
6-month input	M5P	0.987	0.013	0.018	15.67%
	RF	0.963	0.0261	0.035	31.50%
	SVR	0.976	0.0191	0.0291	22.97%
8-month input	M5P	0.889	0.0214	0.0312	41.24%
	RF	0.823	0.0297	0.0377	57.26%
	SVR	0.86	0.0275	0.0348	52.96%
Q_{+2}					
4-month input	M5P	0.962	0.0272	0.0309	32.50%
	RF	0.972	0.0322	0.0391	38.53%
	SVR	0.933	0.03	0.0402	35.91%
6-month input	M5P	0.976	0.0207	0.026	24.73%
	RF	0.972	0.0333	0.0389	39.81%
	SVR	0.95	0.028	0.0369	33.55%
8-month input	M5P	0.675	0.0484	0.0623	75.87%
	RF	0.834	0.0398	0.0491	62.24%
	SVR	0.84	0.0381	0.0489	59.61%
Q_{+3}					
4-month input	M5P	0.859	0.0487	0.0544	56.43%
	RF	0.964	0.0373	0.0435	43.12%
	SVR	0.791	0.0507	0.0637	58.69%
6-month input	M5P	0.921	0.0349	0.0405	40.40%
	RF	0.96	0.0388	0.0475	44.88%
	SVR	0.838	0.0464	0.0589	53.65%
8-month input	M5P	0.586	0.048	0.0591	84.85%
	RF	0.855	0.0409	0.0496	72.37%
	SVR	0.389	0.0638	0.0682	112.87%
Q_{+4}					
4-month input	M5P	0.755	0.0544	0.0612	71.83%
	RF	0.936	0.0359	0.0393	47.41%
	SVR	0.731	0.0528	0.0621	69.83%
6-month input	M5P	0.831	0.0449	0.051	59.39%
	RF	0.945	0.0373	0.0441	49.25%
	SVR	0.857	0.0415	0.05	54.88%
8-month input	M5P	0.709	0.0379	0.0475	70.49%
	RF	0.934	0.0277	0.0351	51.44%
	SVR	0.797	0.0314	0.0406	58.39%

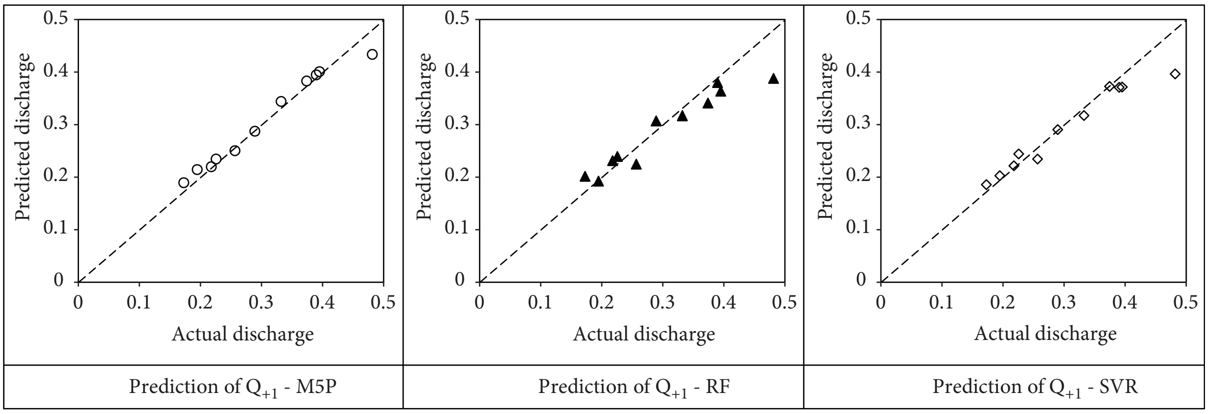


Prediction of Q_{+1} - M5P

Prediction of Q_{+1} - RF

Prediction of Q_{+1} - SVR

FIGURE 8: Comparison between predicted and observed discharges (m^3/s), 4-month input.

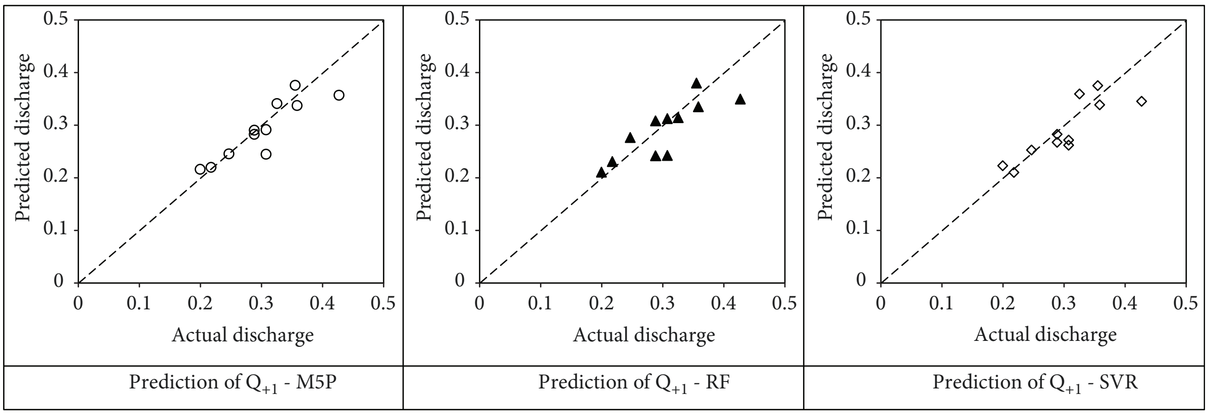


Prediction of Q_{+1} - M5P

Prediction of Q_{+1} - RF

Prediction of Q_{+1} - SVR

FIGURE 9: Comparison between predicted and observed discharges (m^3/s), 6-month input.



Prediction of Q_{+1} - M5P

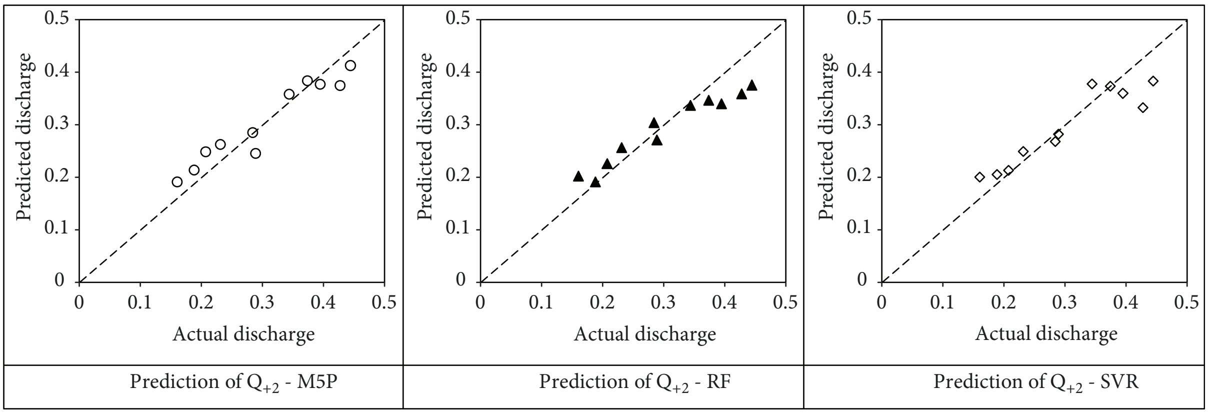
Prediction of Q_{+1} - RF

Prediction of Q_{+1} - SVR

FIGURE 10: Comparison between predicted and observed discharges (m^3/s), 8-month input.

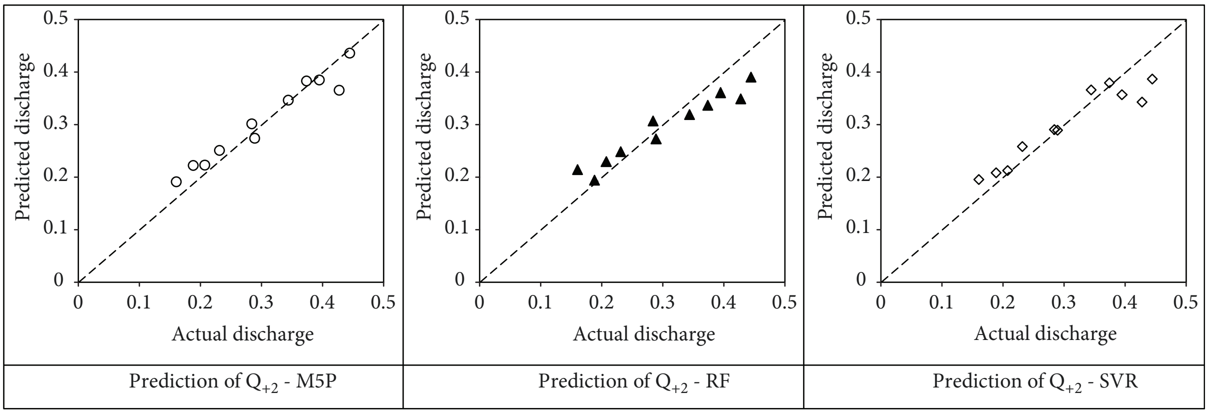
on the previous 4 months ($R^2 = 0.991$, MAE = 0.0124 m^3/s, RMSE = 0.0156 m^3/s, and RAE = 14.97%), while SVR performance is slightly better if an input based on the previous 6 months is considered ($R^2 = 0.976$, MAE = 0.0191 m^3/s, RMSE = 0.0291 m^3/s, and RAE = 22.97%). If, on the other hand, 8-month input is considered, the results of all models (Figure 10) get significantly worse (Table 1). Again, in this case M5P leads to the best results ($R^2 = 0.889$, MAE = 0.0214 m^3/s, RMSE = 0.0312 m^3/s, and RAE = 41.24%) while RF provides the least satisfactory ones ($R^2 = 0.823$, MAE = 0.0297 m^3/s, RMSE = 0.0377 m^3/s, and RAE = 57.26%).

Regarding the forecast of the average flow rate after two months, Q_{+2}, all models lead to reliable results (Table 1), both with 4-month input (Figure 11) and 6-month input (Figure 12). M5P and RF show the best performance in terms of coefficient of determination R^2, but RF predictions are characterized by higher values of MAE, RMSE, and RAE. The best predictions of Q_{+2} are provided



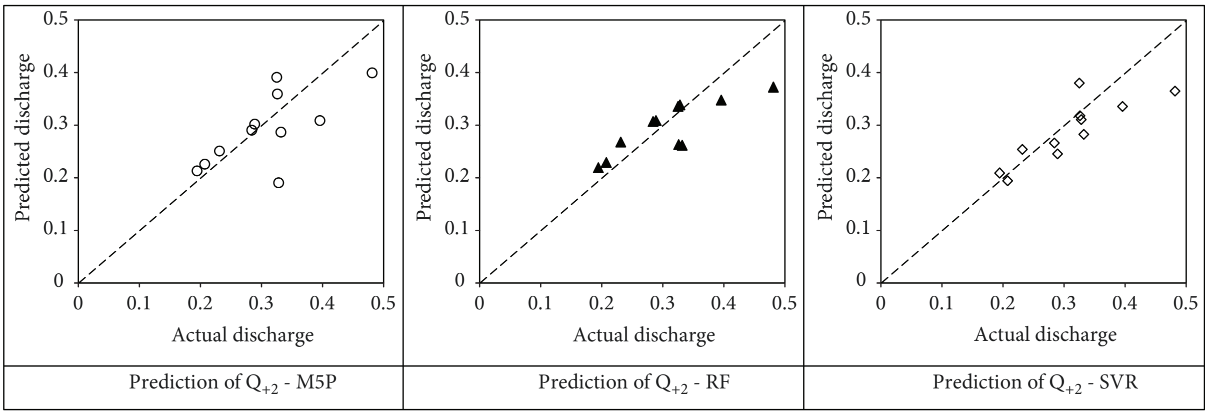
Prediction of Q_{+2} - M5P | Prediction of Q_{+2} - RF | Prediction of Q_{+2} - SVR

FIGURE 11: Comparison between predicted and observed discharges (m^3/s), 4-month input.



Prediction of Q_{+2} - M5P | Prediction of Q_{+2} - RF | Prediction of Q_{+2} - SVR

FIGURE 12: Comparison between predicted and observed discharges (m^3/s), 6-month input.



Prediction of Q_{+2} - M5P | Prediction of Q_{+2} - RF | Prediction of Q_{+2} - SVR

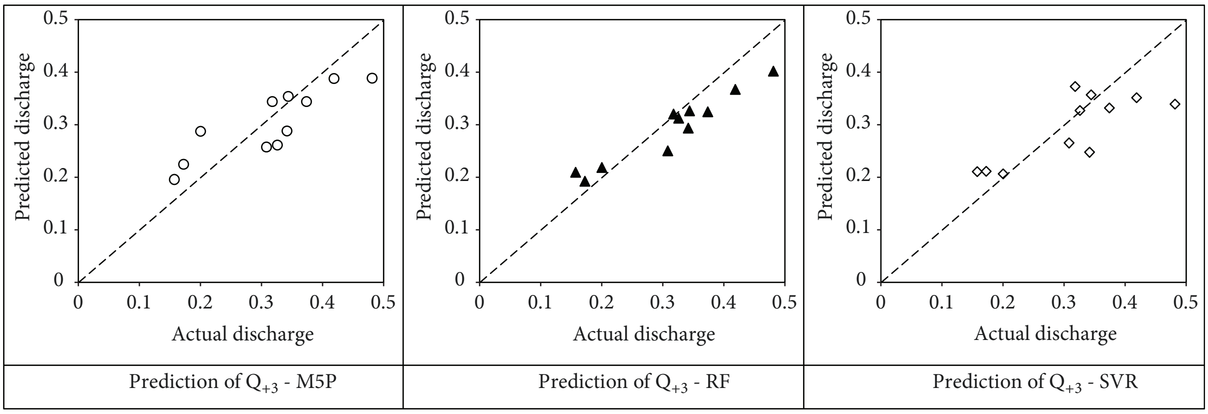
FIGURE 13: Comparison between predicted and observed discharges (m^3/s), 8-month input.

by M5P with a 6-month input ($R^2 = 0.976$, MAE = 0.0207 m^3/s, RMSE = 0.026 m^3/s, and RAE = 24.73%).

Considering an 8-month input (Figure 13), it can be noted that the performance of all models significantly declines (Table 1). M5P is characterized by the lowest R^2 and the highest errors, while performance of SVR and RF, which are very similar, is much better than that of M5P.

All models show the most significant errors for $Q > 0.35\,m^3/s$.

Analyzing the average flow rate after 3 months, Q_{+3}, referring to a 4-month input (Figure 14), it can be seen that RF provides the best results in terms of both R^2 and MAE, RMSE, and RAE, while M5P provides less accurate predictions and SVR leads to the worse results (Table 1). Considering instead a 6-month input (Figure 15), the accuracy of M5P improves significantly, with R^2 increasing from 0.859 to 0.921 and MAE decreasing from 0.0487 m^3/s to 0.0349 m^3/s, RMSE from 0.0544 m^3/s to 0.0405 m^3/s, and

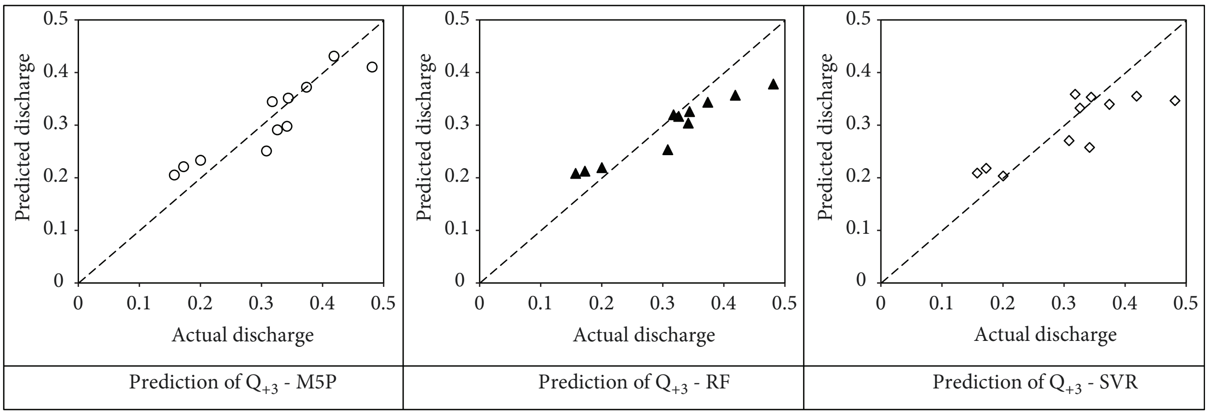


Prediction of Q_{+3} - M5P

Prediction of Q_{+3} - RF

Prediction of Q_{+3} - SVR

Figure 14: Comparison between predicted and observed discharges (m^3/s), 4-month input.

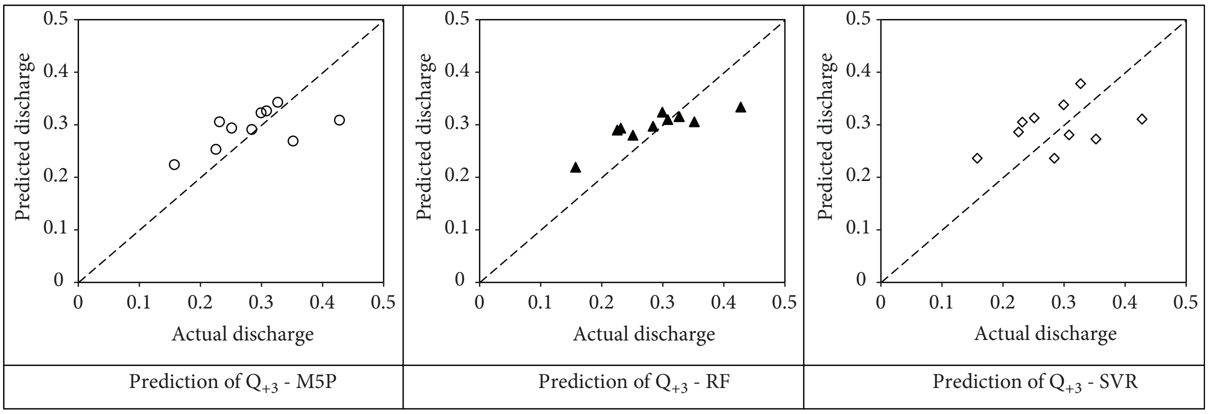


Prediction of Q_{+3} - M5P

Prediction of Q_{+3} - RF

Prediction of Q_{+3} - SVR

Figure 15: Comparison between predicted and observed discharges [m^3/s], 6-month input.



Prediction of Q_{+3} - M5P

Prediction of Q_{+3} - RF

Prediction of Q_{+3} - SVR

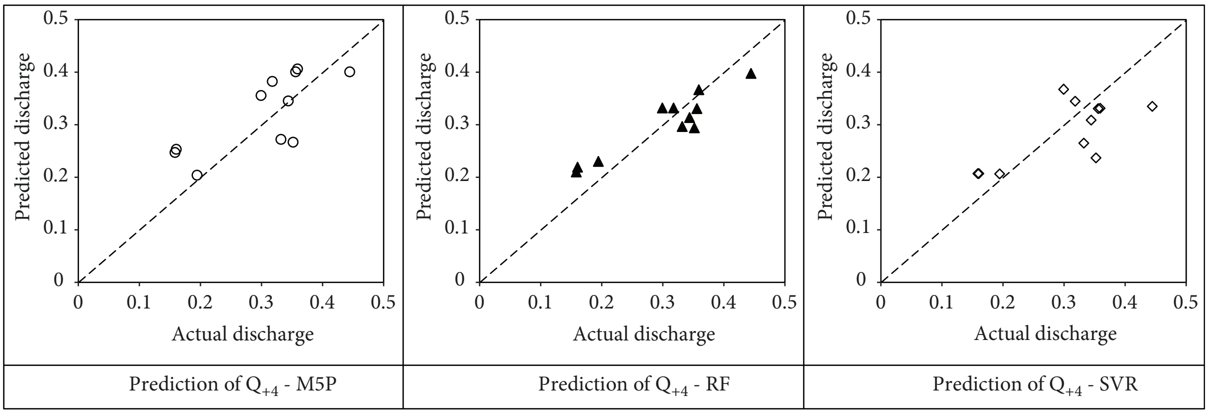
Figure 16: Comparison between predicted and observed discharges (m^3/s), 8-month input.

RAE from 56.43% to 40.40%. RF performance indices remain almost unchanged; therefore, RF continues to be characterized by the higher value of R^2, while error values are now higher than those of M5P. SVR provides less accurate predictions either with a 4-month input or with a 6-month input. In the first case, $R^2 = 0.791$, MAE = 0.0507 m^3/s, RMSE = 0.0637 m^3/s, and RAE = 58.69%; in the second case, $R^2 = 0.838$, MAE = 0.0464 m^3/s, RMSE = 0.0589 m^3/s, and RAE = 53.65%: even the accuracy of SVR predictions improves by switching from 4-month input to 6-month input.

If, on the other hand, an 8-month input is considered (Figure 16), once again the performance of forecast models significantly declines (Table 1). Only the RF algorithm is characterized by a fairly high value of R^2, equal to 0.855, while SVR provided results affected by excessive errors.

Again, all the models show the most significant errors for $Q > 0.35\ m^3/s$.

With regard to the average discharge after 4 months, Q_{+4}, the algorithm RF-based model provides the most accurate predictions for all input durations considered (Table 1).

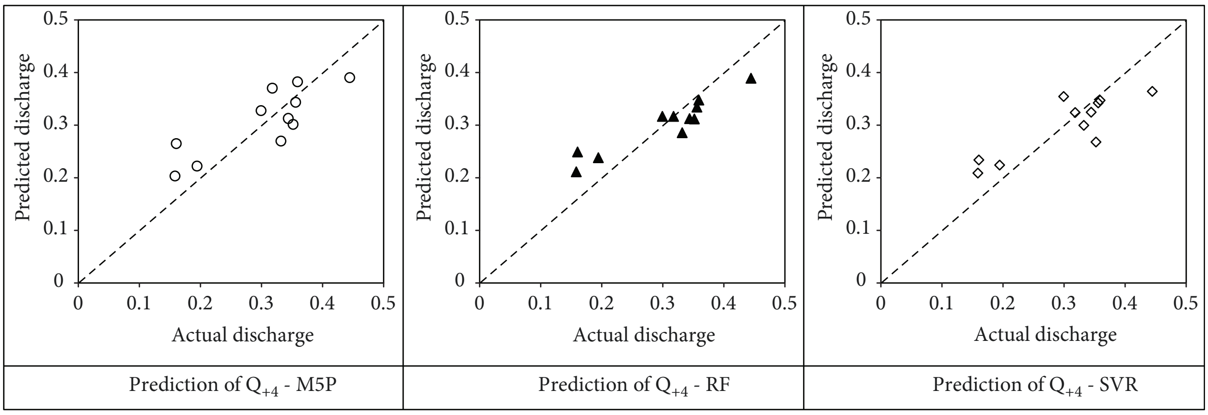


Prediction of Q_{+4} - M5P

Prediction of Q_{+4} - RF

Prediction of Q_{+4} - SVR

Figure 17: Comparison between predicted and observed discharges (m^3/s), 4-month input.

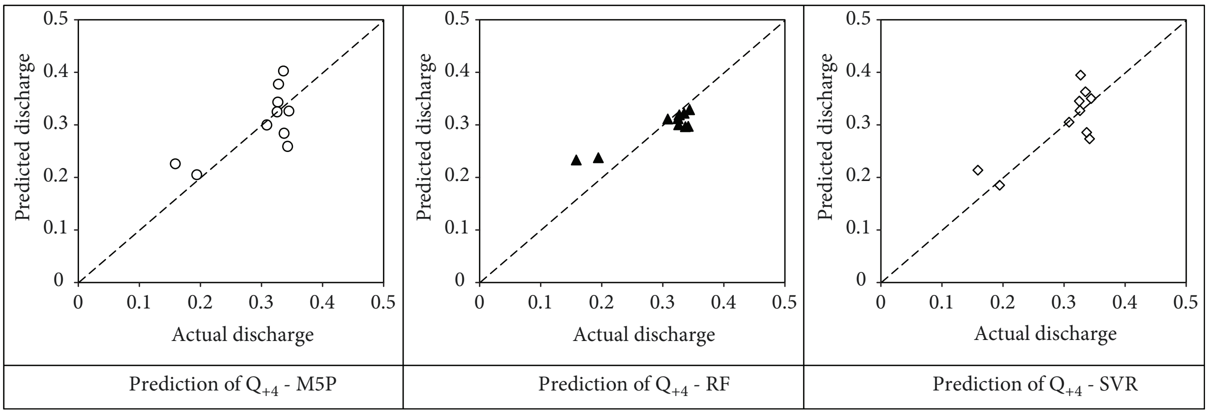


Prediction of Q_{+4} - M5P

Prediction of Q_{+4} - RF

Prediction of Q_{+4} - SVR

Figure 18: Comparison between predicted and observed discharges (m^3/s), 6-month input.



Prediction of Q_{+4} - M5P

Prediction of Q_{+4} - RF

Prediction of Q_{+4} - SVR

Figure 19: Comparison between predicted and observed discharges (m^3/s), 8-month input.

Considering a 4-month input (Figure 17), RF is characterized by $R^2 = 0.936$, MAE = 0.0359 m^3/s, RMSE = 0.0393 m^3/s, and RAE = 47.41%. For a 6-month input (Figure 18), instead, $R^2 = 0.945$, MAE = 0.0373 m^3/s, RMSE = 0.0441 m^3/s, and RAE = 49.25%: the performance of the model with two different inputs are very close. Better than expected is the RF performance for an 8-month input: $R^2 = 0.934$, MAE = 0.0277 m^3/s, RMSE = 0.0351 m^3/s, and RAE = 51.44%. Both in the case of 4-month input and 6-month input are performances of M5P and SVR very similar. With reference to the 8-month input (Figure 19), however, SVR provides more accurate predictions than M5P (Table 1).

The error of the models is fairly evenly distributed over the entire flow rate range considered for testing.

It can also be noted that as the forecasting horizon advances, the M5P model provides less accurate predictions. Similar is the behavior of SVR. The accuracy of the RF model, instead, is less reduced as the monthly timeframe advances. While M5P provides very good short-term forecasts, RF is able to provide fairly accurate predictions of average flow rates that will be available after some months.

All the considered models tend to be less effective if the input data relate to an appreciably longer period than the actual aquifer response to rainfall on the basin; therefore, it is appropriate to estimate this time by means of a cross-correlation analysis.

4. Conclusions

The capability of forecasting spring discharges is essential for a careful management and an accurate planning of water resources. In many cases, a prediction of the flow rate that will be available in the future on the basis of the basin water balance is very complicated or impossible. Machine learning models represent a very interesting alternative. These models can be built on the basis only of past discharges and cumulative rainfall.

Three different machine learning algorithms were used and compared in this study: M5P, random forest, and support vector regression. The spring of Rasiglia Alzabove, Umbria, Central Italy, was chosen as a case study.

The considered models have proven to be able to provide encouraging results even if the available time series for training is rather limited. M5P provides very good short-term predictions of monthly average flow rates, while RF is able to provide accurate medium-term forecasts.

As the time of forecasting advances, the models generally lead to less accurate predictions. Moreover, the effectiveness of the models significantly depends on the duration of the period considered for input data. This time should be approximately estimated by means of a cross-correlation analysis, in order to evaluate the actual aquifer response time.

Notations

b:	Bias in SVR algorithm
C:	Cost of error in SVR algorithm
N:	Number of sample units in the generic node
P_{-k}:	Cumulative monthly rainfall of k months before
p_{L}:	Portion of units assigned to the left child node
p_{R}:	Portion of units assigned to the right child node
Q_{+k}:	Predicted monthly average flow rate after k months
Q_{-k}:	Monthly average flow rate of k months before
R:	Least-squared deviation, within variance for the generic node
s_p:	Generic split during tree model growing
t:	Generic node of a tree model
t_{L}:	Left node generated by the generic split s_p
t_{R}:	Right node generated by split s_p
w:	Variable in SVR algorithm
y_i:	Value of the target variable for the i-th unit in the generic node
y_{m}:	Mean value of the target variable in the generic node
Φ:	Function used to make the SVR algorithm nonlinear
α_i:	Variable in SVR algorithm
α_i^*:	Variable in SVR algorithm
ε:	Deviation parameter in SVR algorithm
ϕ:	Function of the least-squared deviation
η_i:	Variable in SVR algorithm
η_i^*:	Variable in SVR algorithm
ξ_l:	Slack variable in SVR algorithm
ξ_l^*:	Slack variable in SVR algorithm.

Data Availability

The data used to support the findings of this study were provided by Umbria regional agency for environmental protection. They are freely available online (http://www.arpa.umbria.it/articoli/portata-delle-sorgenti-000).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Tóth, "Groundwater discharge: a common generator of diverse geologic and morphologic phenomena," *Hydrological Sciences Journal*, vol. 16, no. 1, pp. 7–24, 1971.

[2] J. Tóth, "Groundwater as a geologic agent: an overview of the causes, processes, and manifestations," *Hydrogeology Journal*, vol. 7, no. 1, pp. 1–14, 1999.

[3] N. Kresic and Z. Stevanovic, *Groundwater Hydrology of Springs*, Elsevier, 2010.

[4] Y.-K. Zhang, E.-W. Bai, R. Libra, R. Rowden, and H. Liu, "Simulation of spring discharge from a limestone aquifer in Iowa, USA," *Hydrogeology Journal*, vol. 4, no. 4, pp. 41–54, 1996.

[5] N. Lambrakis, A. S. Andreou, P. Polydoropoulos, E. Georgopoulos, and T. Bountis, "Nonlinear analysis and forecasting of a brackish karstic spring," *Water Resources Research*, vol. 36, no. 4, pp. 875–884, 2000.

[6] C. Hu, Y. Hao, T. C. J. Yeh, B. Pang, and Z. Wu, "Simulation of spring flows from a karst aquifer with an artificial neural network," *Hydrological Processes*, vol. 22, no. 5, pp. 596–604, 2008.

[7] F. Fiorillo and A. Doglioni, "The relation between karst spring discharge and rainfall by cross-correlation analysis (Campania, Southern Italy)," *Hydrogeology Journal*, vol. 18, no. 8, pp. 1881–1895, 2010.

[8] Y. Fan, X. Huo, Y. Hao et al., "An assembled extreme value statistical model of karst spring discharge," *Journal of Hydrology*, vol. 504, pp. 57–68, 2013.

[9] N. Diodato, L. Guerriero, F. Fiorillo et al., "Predicting monthly spring discharges using a simple statistical model," *Water Resources Management*, vol. 28, no. 4, pp. 969–978, 2014.

[10] Y. B. Dibike, S. Velickov, D. Solomatine, and M. B. Abbott, "Model induction with support vector machines: introduction and applications," *Journal of Computing in Civil Engineering*, vol. 15, no. 3, pp. 208–216, 2001.

[11] S. Ahmad, A. Kalra, and H. Stephen, "Estimating soil moisture using remote sensing data: a machine learning approach," *Advances in Water Resources*, vol. 33, no. 1, pp. 69–80, 2010.

[12] N. S. Raghavendra and P. C. Deka, "Support vector machine applications in the field of hydrology: a review," *Applied Soft Computing*, vol. 19, pp. 372–386, 2014.

[13] F. Granata, R. Gargano, and G. de Marinis, "Support vector regression for rainfall-runoff modeling in urban drainage: a comparison with the EPA's storm water management model," *Water*, vol. 8, no. 3, p. 69, 2016.

[14] M. Najafzadeh, A. Etemad-Shahidi, and S. Y. Lim, “Scour prediction in long contractions using ANFIS and SVM,” *Ocean Engineering*, vol. 111, pp. 128–135, 2016.

[15] F. Granata, S. Papirio, G. Esposito, R. Gargano, and G. de Marinis, “Machine learning algorithms for the forecasting of wastewater quality indicators,” *Water*, vol. 9, no. 2, p. 105, 2017.

[16] D. P. Solomatine and Y. Xue, “M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China,” *Journal of Hydrologic Engineering*, vol. 9, no. 6, pp. 491–501, 2004.

[17] K. K. Singh, M. Pal, and V. P. Singh, “Estimation of mean annual flood in Indian catchments using backpropagation neural network and M5 model tree,” *Water Resources Management*, vol. 24, no. 10, pp. 2007–2019, 2010.

[18] A. Etemad-Shahidi and N. Ghaemi, “Model tree approach for prediction of pile groups scour due to waves,” *Ocean Engineering*, vol. 38, no. 13, pp. 1522–1527, 2011.

[19] M. K. Goyal, “Modeling of sediment yield prediction using M5 model tree algorithm and wavelet regression,” *Water Resources Management*, vol. 28, no. 7, pp. 1991–2003, 2014.

[20] M. Najafzadeh, M. Rezaie Balf, and E. Rashedi, “Prediction of maximum scour depth around piers with debris accumulation using EPR, MT, and GEP models,” *Journal of Hydroinformatics*, vol. 18, no. 5, pp. 867–884, 2016.

[21] F. Granata and G. de Marinis, “Machine learning methods for wastewater hydraulics,” *Flow Measurement and Instrumentation*, vol. 57, pp. 1–9, 2017.

[22] M. Najafzadeh, A. Tafarojnoruz, and S. Y. Lim, “Prediction of local scour depth downstream of sluice gates using data-driven models,” *ISH Journal of Hydraulic Engineering*, vol. 23, no. 2, pp. 195–202, 2017.

[23] M. Najafzadeh, D. B. Laucelli, and A. Zahiri, “Application of model tree and evolutionary polynomial regression for evaluation of sediment transport in pipes,” *KSCE Journal of Civil Engineering*, vol. 21, no. 5, pp. 1956–1963, 2017.

[24] A. Zahiri and M. Najafzadeh, “Optimized expressions to evaluate the flow discharge in main channels and floodplains using evolutionary computing and model classification,” *International Journal of River Basin Management*, vol. 16, no. 1, pp. 123–132, 2018.

[25] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, CRC Press, 1984.

[26] J. R. Quinlan, “Learning with continuous classes,” *Machine Learning*, vol. 92, pp. 343–348, 1992.

[27] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

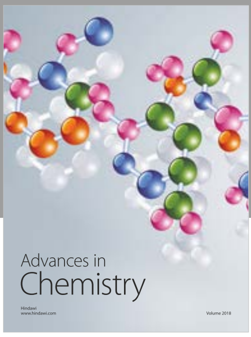
[29] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

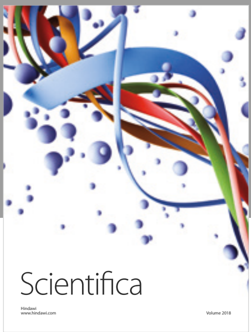
[30] H. W. Kuhn and A. W. Tucker, “Nonlinear programming,” in *Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics, 1951*, pp. 481–492, University of California Press, 1951.

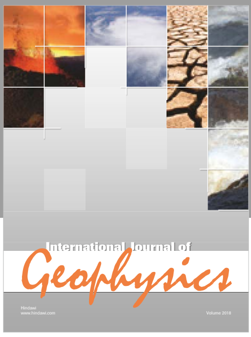
[31] C. Boni, D. Cascone, L. Mastrorillo, and C. Tarragoni, *Carta idrogeologica delle dorsali interne umbro-marchigiane. Foglio 1 e 2. Scala 1:50.000 con note illustrative. Pubblicazione N° 2865 CNR – Gruppo Nazionale Difesa Catastrofi Idrogeologiche*, CNR-GNDCI (ITALY), ROMA, 2005.

[32] C. Boni and L. Mastrorillo, “Rilevamento idrogeologico dei M. di Foligno,” in *Atti del Convegno “Ricerca e protezione delle risorse idriche sotterranee delle aree montuose”. Brescia, Ottobre 1991. Quaderni di sintesi, 42, 247 – 268*, A cura dell’Azienda Servizi Municipali di Brescia, 1993.

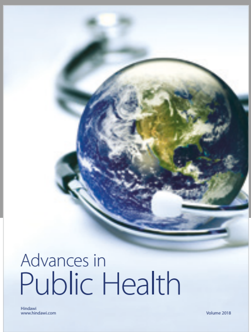
[33] P. Conversini and G. S. Tazioli, “Studio idrogeologico delle sorgenti di Rasiglia, Umbria orientale,” *Quaderni di Geologia Applicata*, vol. 2, pp. 117–126, 1994, (in Italian).

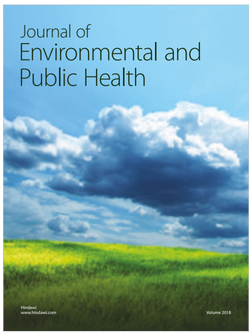

Advances in
Chemistry

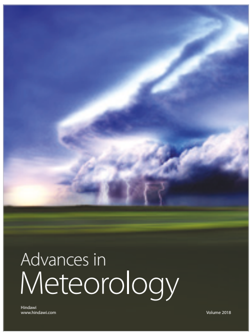

Scientifica

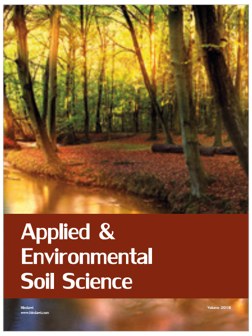

International Journal of
Geophysics

The Scientific
World Journal


Advances in
Public Health

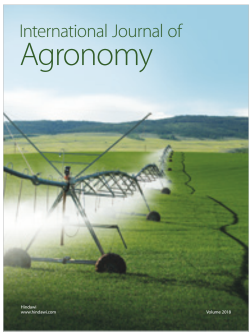

Journal of
Environmental and
Public Health


Advances in
Meteorology


Applied &
Environmental
Soil Science

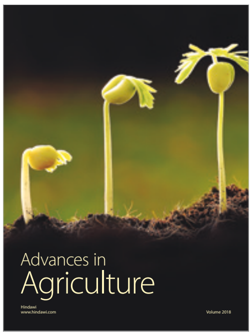
Journal of
Geological Research

International Journal of
Forestry Research

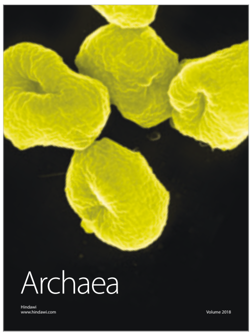

International Journal of
Agronomy

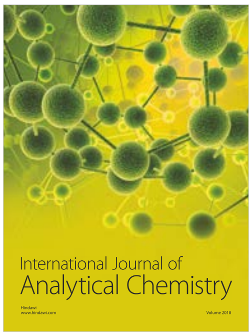
International Journal of
Microbiology

International Journal of
Ecology


Advances in
Agriculture

International Journal of
Biodiversity


Archaea


International Journal of
Analytical Chemistry

Journal of
Chemistry

Journal of
Marine Biology