Modular optimized data assimilation and support vector machine for hydrologic modeling

M. Mehrparvar and K. Asghari

ABSTRACT

Accurate and reliable simulation models are crucial for the operation and management of systems. Developing a simulation model to forecast future states of a system is generally followed by errors in prediction. Frequently, data-based models such as support vector machines (SVM) are used as forecasting techniques. This paper introduces a modular method which couples the machine learning technique of support vector regression (SVR) as a prediction model and a modified data assimilation (MDA) technique to partially correct the predicted values based on the observation data. To improve the performance and accuracy of the system output, the ensemble Kalman filter (EnKF) as a data assimilation procedure is implemented with an optimization procedure. As a case study, inflow quantities to Zayandehroud reservoir is considered as the state vector in the assimilation process to enhance the system output. Evaluation criteria such as root mean square error (RMSE) and R-squared criteria are implemented to evaluate the performance of the proposed model. The adjusted values of a hybrid model compared to the SVR model and standard DA indicate improved performance of the proposed model.

Key words | data assimilation, ensemble Kalman filter, optimization, streamflow prediction, support vector regression

M. Mehrparvar
K. Asghari (corresponding author)
Department of Civil Engineering,
Isfahan University of Technology,
Isfahan, Iran
E-mail: *kasghari@cc.iut.ac.ir*

INTRODUCTION

One of the main challenges in simulation models is reliability of their performance. Concerns about performance and accuracy of simulation models receive attention for reducing errors and ambiguities about models that produce uncertainty. Presenting approaches or methods to reduce errors and improve accuracy of models is the main concern of any data assimilation (DA) modeling to increase certainty and reliability of the systems. For example, uncertainty about the inflow discharge to a reservoir is the key factor in optimization of reservoir operation. The inflow usually consists of the runoff from upstream which is related to different parameters such as precipitation, temperature and water transfer from other basins. The accurate estimation of these parameters, which can be distributed spatially and temporally, has a significant effect on prediction of inflow. Deterministic optimization models which directly use historical data usually suffer from obtaining reliable results for future planning and management. To balance the uncertainty in the data, a simulation model is developed to predict future estimation while incorporating the information inherited into the observation data by using techniques such as DA.

For the last two decades, data-based methods have been developed and implemented in water resource system analysis as simulation models with acceptable performance. Artificial neural networks (ANN) and support vector machines (SVM) are among the frequently used data-based models capable of proper prediction and flexible operation. Introduction of an SVM model as a statistical data analysis technique was presented by Vapnik (1995) and initiated a wide variety of

doi: 10.2166/hydro.2018.009

research in the 1990s using implementations of SVM. The studies of rainfall-runoff modeling by Dibike *et al.* (2001); Behzad *et al.* (2009) and Jajarmizadeh *et al.* (2015), river inflow prediction by Asefa *et al.* (2006) and Lin *et al.* (2006) and controlling groundwater levels in aquifers by Asefa *et al.* (2004) are examples of these studies using SVM in water resources fields. Often, the results of these data-based modeling techniques require improvement due to confounded noise from both known and unknown sources.

DA has been used in many engineering applications to improve modeling accuracy. One of the most efficient and current sequential DA methods is the Kalman filter (KF) which was developed by Kalman (1960). This filtering process reads the variables of previous state or time step methods in sequential order and analyzes the forcing terms and observations of the current state to update the variables for the subsequent state.

DA can be a useful tool in hydrological modeling as it improves the model operation using background data in the numerical model. Refsgaard *et al.* (1983) developed a lumped rainfall-runoff model in state space form to show the implementation of KF where the input uncertainties are predominant for uncertainty of simulated runoff values. Wood & O'Connell (1985) studied the KF and extended Kalman filter (EKF) for rainfall-runoff modeling and applied it on state and parameter estimation of National Weather Service River Forecasting and Sacramento Soil Moisture model at the same time. Lee & Singh (1999) applied KF in the rainfall-runoff process of a tank model. The model parameters used as state vectors and uncertainty in rainfall-runoff process decreased by variation of parameters in time. Hartnack & Madsen (2001) combine ensemble Kalman filter (EnKF) in MIKE 11 as river modeling where the implemented model corrected the model perturbations on boundary conditions. In order to improve accuracy of SVM, Gill *et al.* (2007) and Liu *et al.* (2010) integrated SVM and EnKF to predict soil moisture in different soil layers. The results in both studies show better performance of SVM-EnKF in comparison with SVM without using EnKF. Dechant & Moradkhani (2011) proposed a framework to consider uncertainty in ensemble streamflow prediction (ESP) within the National Weather Service River Forecasting system by developing DA to improve ESP. Nasseri *et al.* (2011) coupled EKF and genetic programming to predict urban water demand. Ricci *et al.* (2011) applied the KF algorithm to update river water level observation for better flood forecasting. Significant improvements were achieved in the water level and discharge values in both analysis and forecast modes. The proposed DA modeling technique of Li *et al.* (2012) includes integration of EnKF and SVM to decrease data predicted of SVM via variable precision rough set theory. Dumedah & Coulibaly (2014) implemented EnKF and particle filter methods for a hydrological DA procedure. Estimation of measurement error is improved via integration of pareto-optimality into the DA techniques. Li *et al.* (2014) coupled EnKF and SVM for rainfall-runoff simulation and proved better accuracy of coupled models for real-time flood forecasting. Liu *et al.* (2016) evaluated the uncertainty of coupled SVM and EnKF to estimate soil moisture in the ground. Siswantoro *et al.* (2016) applied linear KF to improve the performance of neural network classification. Wang & Babovic (2016) developed a hybrid model to improve water level forecasting. The coupled model includes a data-driven model and KF as the data assimilation technique.

In this study, we considered modified DA (MDA) as a process tool which includes the combination of two robust techniques, namely SVM as a simulation forecasting model to estimate inflow discharge of Zayanderoud reservoir and then the EnKF was used to update and balance the SVM regression (SVR) estimates by available observed historical data. According to the studies by Babovic *et al.* (2000), Mancarella *et al.* (2008) and Sun *et al.* (2010), in analyzing the error propagation in different modeling techniques, the predicted residuals or errors of models can be used to correct prediction of simulation models. Within the same framework for error correction, a modified approach is proposed in the updating process by an optimization procedure. This optimization model makes sure that the error in the next step is less than or equal to the previous time error. In the following sections the theory of SVM and EnKF is briefly described and further sections explain the implementations, results and discussions of the case study.

METHODOLOGY

Using physical simulation models as prediction tools involves a variety of parameters which usually require calibration. An alternative approach would be a robust data-driven model based on analyzing the data about a system, and finding connections between the system state variables (input, internal and output variables) without explicit knowledge of the physical behavior of the system. In order to simulate the outlet streamflow of a sub-basin (reservoir inflow), historical data including rainfall, temperature and inflow discharge are required to develop a data-based model such as SVM. All data-based models need to be trained with collected data and then be tested for simulation. In the following study, EnKF as a data assimilation technique is implemented to improve the efficiency of an SVR model in each simulation step. The proposed approach is presented to improve the performance of integration of SVR and EnKF. The methodology of SVR, EnKF, the embedding approach and evaluation criteria are briefly described in the next sections.

Support vector machine

SVM is a mathematical model trained to determine the correlation between $x\epsilon R^d$ as d-dimensional input vector and the output vector $y\epsilon R$. SVM is divided into two classification and regression (SVR) methods in which SVR is more commonly used in water-related applications. Let vector (x_j, y_j), $j\epsilon\{1, 2, 3, \dots, N\}$ be the set of input and output observed data. The aim of a data-based model such as SVR generally is searching for a function $f(x)$ as an approximation of the value y_i with minimum risk, and is only based on the available independent and identically distributed data. In the SVR algorithm, the estimation function is determined by a small subset of training samples, namely support vectors. Also in this algorithm, a specific loss function called ε-insensitive loss is developed to create a sparseness property for SVR. It means instead of minimizing the empirical error over the training data, SVR minimizes a regulated risk function which states that for achieving the minimum risk, simultaneous control of the complexity of the model and the error owing to training data is essential (principle of the structural risk minimization theory). In a linear SVR shown in Equation (1), the input vector is mapped into the N-dimensional feature space by non-linear kernel transformation function $k_j(x)$, where $j = 1, 2, \dots, N$. α and b denote the weighted vector and bias term, respectively. The transformation of input x vector into the N-dimensional feature is shown in Figure 1.

$$f(x, \alpha) = \sum_{j=1}^{N} \alpha_j k_j(x) + b \tag{1}$$

The regularization is carried out by developing an optimization model as in Equation (2). The model is developed by minimizing the $(1/2)||\alpha||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$ as its objective function and denoting the errors without trespassing ε as constraints. The errors are defined by the difference between the estimates values of $f(x, \alpha)$ and the output vector y. Vapnik (1998) defined loss function to make the optimization model feasible. Thus, two slack variables ξ_i, ξ_i^* are defined to determine upper and lower errors over the ε value and parameter C indicates the loss function factor. The loss function is shown in Equation (3) which is also called ε-insensitive function:

$$\begin{aligned} &\text{Minimize } \frac{1}{2}||\alpha||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &\text{subject to:} \\ &f(x_i, \alpha) - y_i \le \varepsilon + \xi_i \\ &y_i - f(x_i, \alpha) \le \varepsilon + \xi_i^* \\ &\xi_i, \xi_i^* \ge 0\ ,\ \ i = 1, 2, \dots, n \end{aligned} \tag{2}$$

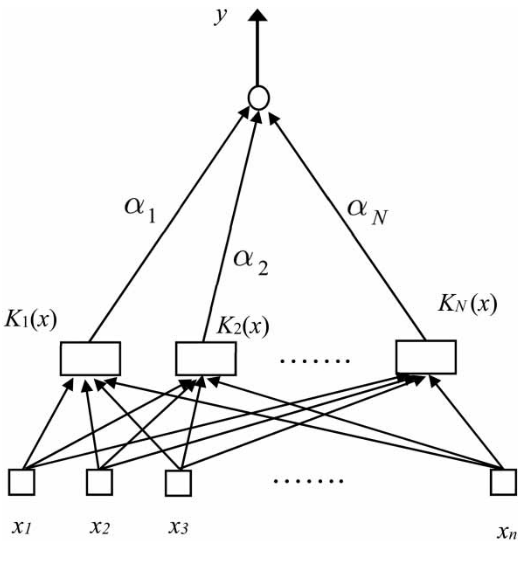


Figure 1 | SVR structure (Vapnik 1998).

$$L_\varepsilon(y, f(x, \alpha)) = |y - f(x, \alpha)|_\varepsilon = \begin{cases} 0 & \text{if } |y - f(x, \alpha)| \le \varepsilon \\ |y - f(x, \alpha)| \le \varepsilon & \text{otherwise} \end{cases} \tag{3}$$

we used the radial basis function (RBF) as a proper kernel function. This function has proven to be more efficient than other kernel functions based on studies reported on the use of RBF (Asefa *et al.* 2006; Behzad *et al.* 2009).

Kalman filter

The KF technique is one of the data assimilation methods rooted from the Monte-Carlo and Bayesian techniques. Generally, approximate Bayesian state estimation is considered in this method which states that the errors increase linearly and they are distributed normally. KF estimations are performed in two steps. In the updating step, the estimation of uncertain forecast state for time step t is adjusted and improved by a kernel function. By incorporating the observation of time step t in the kernel function, the forecast state will be modified and updated. In the forecasting step, the solution will propagate into the next time step $(t+1)$ by using the previous updated state variable.

EKF and EnKF are a variant of KF that can be used for the nonlinear filtering problem. Implementation of EKF is infeasible for large systems, such as hydrological systems, due to their calculation complexity. So EnKF, which considers a statistical approximation of EKF by sampling the forecast and analysis error, is substituted for EKF in this study.

Let us consider background vector X_t^b an ensemble of forecasts with size m which randomly sample the background of model errors at time t and the average of X_t^b which is shown by $\bar{x}_t^b$. Similarly, the observation vector is shown by Y_t:

$$X_t^b = (x_{1,t}^b, x_{2,t}^b, \ldots, x_{m,t}^b),$$
$$Y_t = (y_{1,t}, y_{2,t}, \ldots, y_{m,t}), \quad \bar{x}_t^b = \frac{1}{m}\sum_{i=1}^{m} x_{i,t}^b \tag{4}$$

EnKF uses an ensemble data that contains m datum in Equation (5) as an update step. In Equation (5), each ensemble member $(i = 1, 2, \ldots, m)$ is updated to the ith member of the X_t^a vector based on the observation vector and observation operator that is shown by $H(x)$. The H operator is briefly explained in Appendix A (available with the online version of this paper). This operator is determined by the type of problem and can be equal to the identity matrix:

$$x_{i,t}^a = x_{i,t}^b + K(y_{i,t} - H(x_{i,t}^b)) \tag{5}$$

$$K = \mathrm{P}^b\mathrm{H}^T(\mathrm{H}\mathrm{P}^b\mathrm{H}^T + \mathrm{R})^{-1} \tag{6}$$

where K is the Kalman gain matrix, P^b is background-error covariance, and R is observation-error covariance. P^b is created from ensemble data that are made from nonlinear forecasts and is determined by Equation (7). Background-error covariance is a function of deviation matrix $X_t^{'b}$ where its members are computed by $x_{i,t}^{'b} = x_{i,t}^b - \bar{x}_t^b$. Similarly, the perturbed observation members are determined by $y_{i,t}^t \sim N(0, \mathrm{R})$ which is normally distributed with an average equal to zero with covariance R:

$$\mathrm{P}^b = \frac{1}{m-1} X'^b X'^{b^T} \tag{7}$$

The components of Kalman gain matrix, which are the variance of perturbation observation operator matrix $\mathrm{HP^bH^T}$ and covariance of perturbation observation operator and background matrixes $\mathrm{P^bH^T}$, are computed by the following equations:

$$\overline{H(x_t^b)} = \frac{1}{m}\sum_{i=1}^{m} H(x_{i,t}^b) \tag{8}$$

$$\mathrm{HP^bH^T} = \frac{1}{m-1}\sum_{i=1}^{m} (H(x_{i,t}^b) - \overline{H(x_t^b)})(H(x_{i,t}^b) - \overline{H(x_t^b)})^T \tag{9}$$

$$\mathrm{P^bH^T} = \frac{1}{m-1}\sum_{i=1}^{m} (x_{i,t}^b - \overline{x_t^b})(H(x_{i,t}^b) - \overline{H(x_t^b)})^T \tag{10}$$

In the forecasting step, the model prediction transmits analysis or updated vector X^a from time t to $t+1$ by model operator M (x_t^a) which the prediction step follows by $x_{t+1}^b = M(x_t^a)$.

Model development

For developing the SVR model to predict or simulate streamflow, EnKF is implemented to update and optimize SVR predictions. Figure 2 shows the combined use of SVR and EnKF. In time step *t*, SVR predicts the background vector X_t^b by reading input data for three regions including rainfall (p_t^1, p_t^2, p_t^3), temperature (T_t^1, T_t^2, T_t^3) and inter-basin water transfer (I_t). The other input data is the inflow to the reservoir from the previous time step. Inflow to the reservoir from the previous time step is considered as a background vector in time step $t-1$ and is accounted as a state variable of EnKF. The output of SVR is the current inflow to the reservoir and is adjusted for the background vector in time step *t*. After prediction by the model forecast, the background data are updated to analyze the vector X_t^a by an assimilation procedure and then the updated data will be used as input for time step $t+1$ (Figure 2).

A modified method is implemented to enhance the X_{t-1}^a estimation. To improve the performance of SVR simulation, an error function is defined for each time step which must be minimized during each time step. To reach this goal, an optimization procedure for the forecast step is defined as follows:

$$\text{Minimize}\, Er_t = 1 - \frac{\min(X_t^b, y_t)}{\max(X_t^b, y_t)} \tag{11}$$

Subject to:

$$X_t^b = M(X_{t-1}^a + \varepsilon) \tag{12}$$

$$-1 \le \varepsilon \le 1 \tag{13}$$

$$Er_t \le Er_{t-1} \tag{14}$$

$$Er_{t-1} = 1 - \frac{\min(X_{t-1}^a, y_{t-1})}{\max(X_{t-1}^a, y_{t-1})} \tag{15}$$

where *M* is the SVR operator, X_{t-1}^a is updated data from a previous time, X_t^b is the resulted decision variable from simulation by SVR, y_t is the observation data and ε is the noise variable. Since the input and output data of SVR are normalized, the ε domain is limited to -1 and 1. By employing the optimization procedure, the best ε will be found to correct the input state variable of SVR. It must be noted that ε and X_t^b are the only variables of the optimization procedure. Er_t is an objective function of the optimization procedure within the DA process for time step *t* which defines the process of minimizing the error ratio as a function of model output and

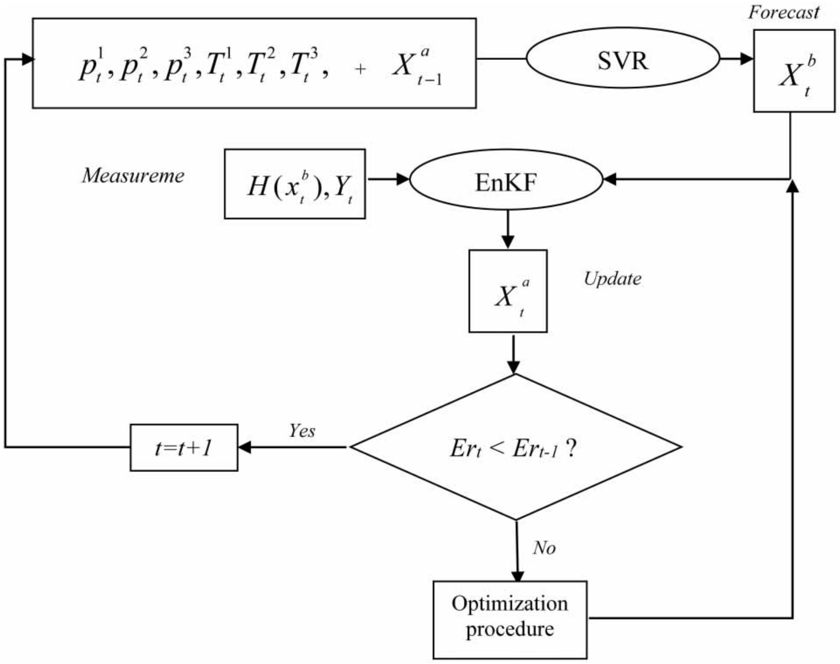


Figure 2 | Proposed model implementation flowchart.

observation data. The simulation output value, which is the state variable (X_t^b), can be lower or higher than y_t. The ratio of min(…)/max(…), which is a normalized error (similar to error measures used in any data-driven model), has a value between 0 and 1 to indicate the difference between observation and simulation values, therefore $1-Er_t$ demonstrates the closeness of these two values at time step t.

The proposed model is called modified data assimilation (MDA) which is an integration of the EnKF technique and an optimization procedure. This model will be implemented to predict the streamflow of a watershed which is an inflow to a downstream reservoir. MDA will be evaluated and compared with model prediction without using an EnKF technique (named forecast results), and will be compared with the model using an EnKF technique without any optimization (named DA results). The proposed procedure will address the uncertainty associated with parameters involved in hydrological modeling. A schematic diagram of different steps of this model is illustrated in Figure 2.

Evaluation criteria

To evaluate the operation of proposed models, methods such as root mean square error (RMSE), correlation coefficient (R) and R-squared are implemented to compute the errors or differences between observations and calculated data by model. In the following equations, n is the number of data points, and y_i and y_i^o indicate the model measurements and observational values, respectively:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i^o)^2}{n}} \tag{16}$$

$$R = \frac{\sum_{i=1}^{n} y_i y_i^o - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} y_i^o / n}{\sqrt{\left[\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n\right]} \sqrt{\left[\sum_{i=1}^{n} (y_i^o)^2 - \left(\sum_{i=1}^{n} y_i^o\right)^2 / n\right]}} \tag{17}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i^o - y_i)^2}{\sum_{i=1}^{n} \left(y_i^o - \frac{1}{n}\sum_{i=1}^{n} y_i^o\right)^2} \tag{18}$$

Study area and data

Zayanderoud reservoir, located in central Iran (Figure 3), was selected to assess the proposed models and to measure their performances. The catchment area is 3694 km^2 which is divided into three sub-basins. More than 23 rainfall and climate stations are located in the catchment area. ARCGIS analysis tools, such as kriging and zonal, are implemented to compute the average values of rainfall and climate data for three sub-basins. This procedure has been implemented for 30 years to develop an SVR model as a monthly prediction and simulation model. To forecast monthly inflow to the reservoir from October 2012 to September 2013, the SVR model uses 360 months of historical data for training and testing. Inputs include eight types of data such as rainfall (three zones), temperature (three zones), and inter-basin water transfer and monthly reservoir inflow at the outlet of the catchment from the previous month. The data sets were divided into two sets: a training set consisting of 25 years of data and validation set of five years of data.

RESULTS AND DISCUSSION

Twenty-five years of data were used in the training phase of SVR modeling, and in the testing phase, five years of data were used. Results from the developed SVR model are shown in Table 1. In this table, the optimized parameters and goodness-of-fit values for two training and testing phases are highlighted. To find proper values of SVR parameters, both trial and error and cross-validation approaches were used. The results of both methods, considering the two evaluation criteria, reveal several valuable features. Firstly, both the trial and error and cross-validation approaches are relatively well behaved. Secondly, all two criteria values show the same trend in training and testing phases. Finally, the generalization performance of SVR for three criteria is closely established.

The ensemble data, including 30 measured samples for the first month of assimilation, is considered. By using SVR as a forecasting model, the next month value will be predicted. The procedure of KF is separated into two forecasting and updating steps. For updating data, the

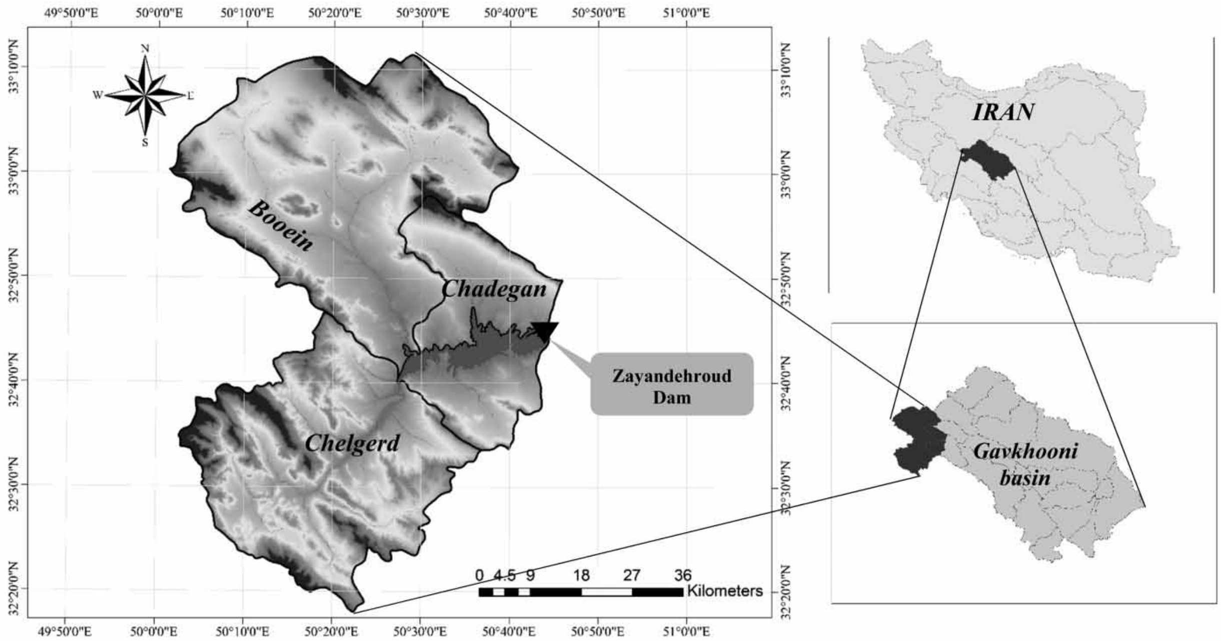


Figure 3 | Location of three sub-basins of study area.

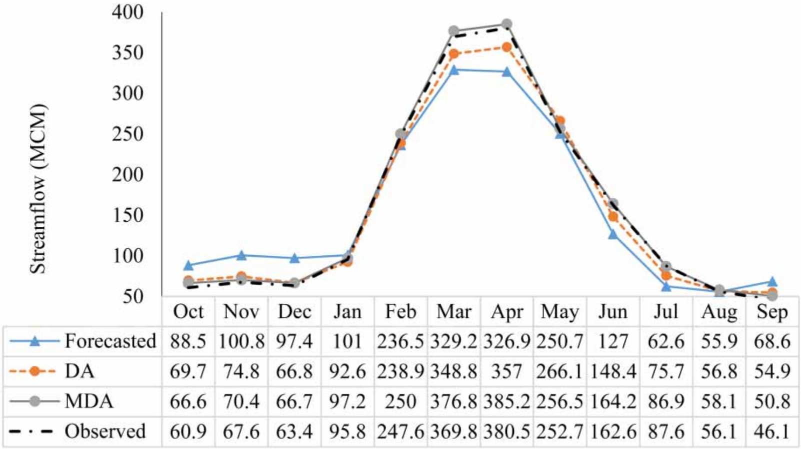
Table 1 | Parameters evaluation of SVR modeling

Trial and error	$C = 25$ $\varepsilon = 0.1$ $\gamma = 0.9$	
	RMSE (MCM[a])	R
Training	38.14	0.948
Testing	45.92	0.89
Cross-validation	$C = 15$ $\varepsilon = 0.9$ $\gamma = 0.7$	
	RMSE (MCM[a])	R
Training	44.87	0.94
Testing	49.2	0.9

[a]Million cubic meter.

forecasted results are assimilated using kernel updating of Equation (5) and incorporation of the EnKF technique. The results are the background data (streamflow) which is the output of SVR simulation corrected based on observation data. There is no guarantee that the error will not propagate for the next simulation time step. If X^a_{t-1} is the updated data, X^b_t will be predicted by SVR operator M, for the next step time. By implementing the optimization procedure, $X^b_t = M(X^a_{t-1})$ is substituted using Equation (12). ε variable is a noise value that is summed with X^a_{t-1} to force the model to simulate proper values in the frame of optimization. The results in Figure 4 compare the observed monthly average of streamflow values (state vector) with forecasted, DA and MDA solutions.

In the process of the error correction, the optimization procedure was incorporated into a conventional DA method to further reduce the error which was produced by the SVR. The analysis of error propagation is directed by the concept of error ratio in time step t (Er_t) and error ratio in time step $t-1$ to obtain the proper output which happens when Er_t is less than Er_{t-1}. To show the improved performance of the proposed method, Figure 5 shows the cumulative error ratio of the assimilation process for a period of five years (2003–2008). Monthly perturbation of forecasts and updates are shown by these error ratios which were the solution of the objective function of the optimization model. As seen in Figure 5, the error ratios show the differences between approximate values of forecasted, DA and MDA modeling with respect to the observational value. Monthly values refer to the average of the ensemble for each month. Referring to Figure 5, the trend of error propagation due to the MDA technique shows substantial improvement compared to SVR and DA, however it must be noticed that the error propagation cannot completely be eliminated. As indicated by Figure 6,



	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
▲ Forecasted	88.5	100.8	97.4	101	236.5	329.2	326.9	250.7	127	62.6	55.9	68.6
● DA	69.7	74.8	66.8	92.6	238.9	348.8	357	266.1	148.4	75.7	56.8	54.9
● MDA	66.6	70.4	66.7	97.2	250	376.8	385.2	256.5	164.2	86.9	58.1	50.8
Observed	60.9	67.6	63.4	95.8	247.6	369.8	380.5	252.7	162.6	87.6	56.1	46.1

Figure 4 | Comparisons of monthly average streamflow values (MCM).

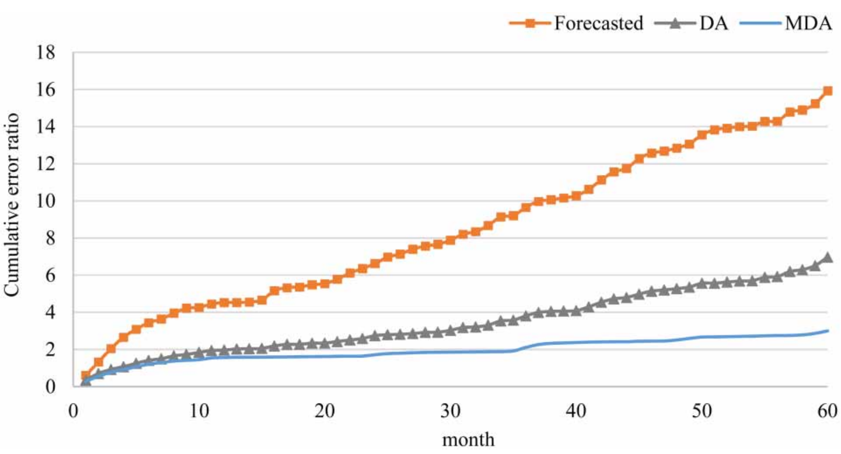


Figure 5 | Cumulative error assessment of assimilation process for selected five years.

the error ratio is generally reducing in 24 months of the simulation period. It is clear that in any non-physical modeling there is always a contingency of error variation in time series of predicted values, which is no exception in MDA techniques. To quantify the performance of the optimization procedure, Table 2 illustrates the variation of X_t^b, y_t, and Er_t for the first 12 months corresponding to data in Figure 6. The decreasing trend of Er_t from 0.27 to 0.02 with respect to state variable X_t^b and observation y_t variations explains the effectiveness of this procedure.

Figure 7 shows last-year monthly inflow computations from three modeling techniques of SVR, DA and MDA, compared to observed data. Clearly, the DA results are further corrected when the optimization procedure is integrated into the assimilation process. Although observation data for the last month is not considered in the modeling scheme, the trend of predicted values using the updated values of DA and MDA from the previous month closely follows the observation data.

To quantify the modeling performance, two error criteria were implemented to measure the significance of improvement resulting from the DA and MDA modeling techniques. In Table 3, the lower values of RMSE for the MDA technique stated improved performance of the proposed model for the planning horizon compared with the forecasted SVR or DA modeling process. Scattered diagrams of Figure 8 compare the MDA, DA and forecasted inflow for 30 years of data, which indicates a significant improvement

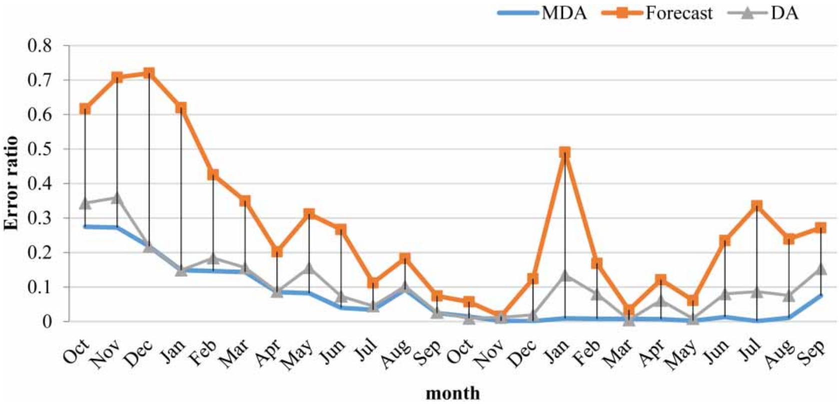


Figure 6 | Trend of error ratio reduction for two consecutive years.

Table 2 | Results of error ratio reduction with respect to variation of state variable X_t^b and observation y_t

***t* (month)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Er_t	0.27	0.27	0.22	0.15	0.15	0.14	0.09	0.08	0.04	0.03	0.09	0.02
X_t^b (MCM)	50.3	54.4	68.6	76.0	174.9	286.4	355.7	232.7	148.5	75.8	50.4	76.0
y_t (MCM)	36.5	39.6	53.6	64.7	149.3	245.4	325.3	213.5	142.6	78.5	45.7	74.2

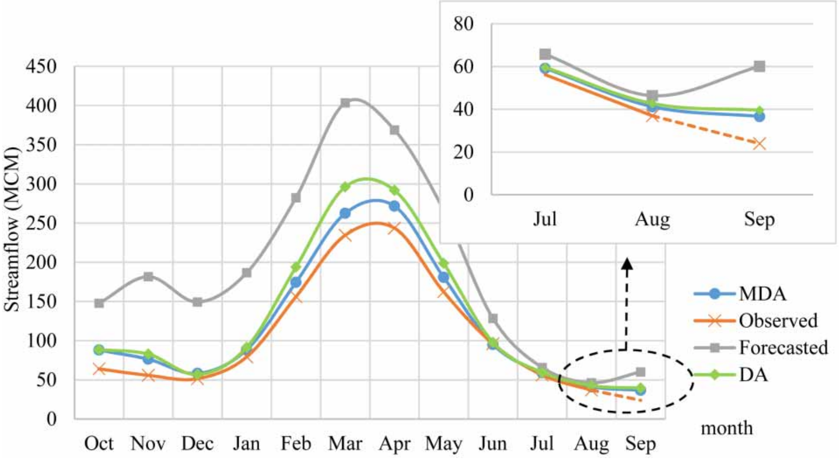


Figure 7 | Predicted and assimilated inflow compared with observed data.

performance of MDA. The correlation of 0.73 between observed inflow and SVR forecasted inflow is increased to 0.85 and 0.99 applying the DA and MDA techniques, respectively.

CONCLUSIONS

In this study, the combined use of a regression support vector machine (SVR) and modified ensemble Kalman

Table 3 | RMSE comparison of different models

	RMSE (MCM)		
Month	**SVR forecasted**	**DA**	**MDA**
October	52.66	16.55	13.45
November	70.12	14.03	10.36
December	62.97	5.65	4.54
January	55.24	7.11	4.23
February	71.15	24.79	9.35
March	87	39.27	18.11
April	99.95	46.47	14.57
May	75.35	25.97	8.2
June	43.55	15.11	3.32
July	30.25	15.06	10.94
August	16.3	8.21	6.96
September	34.21	12.69	10.48

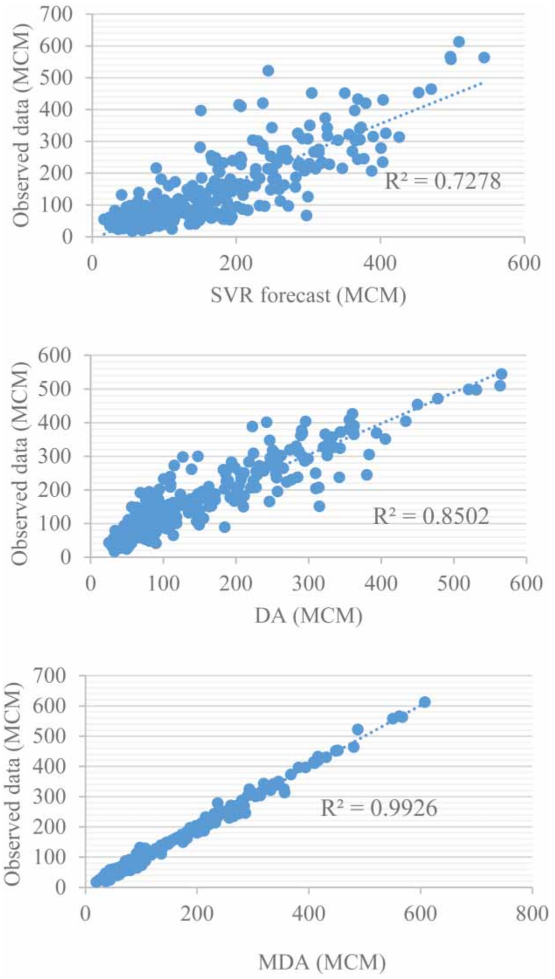


Figure 8 | Comparison of observed vs. computed inflow in different models.

filter (EnKF) as a data assimilation process is investigated. Considering the importance of developing a real-world model to predict water inflow into the reservoir, Zayanderoud reservoir located in central Iran is selected as the case study to evaluate the performance of the proposed model. The data-based SVR model is trained to forecast inflow with a correlation coefficient of 0.95 and 0.89 for training and testing, respectively. Eight hydrologic input data including rainfall (for three regions), temperature (for three regions), and inter-basin water and reservoir inflow from the previous month are considered to predict the current inflow. By considering 30 years' measurement samples for the first month of planning horizon, the EnKF is implemented to update (DA) and optimize (MDA) the prediction values for the next month and is repeated continuously for five years. The propagation of errors is controlled when applying the DA and is further reduced by MDA techniques. It is concluded that, when EnKF as a data assimilation technique is integrated with an optimization procedure, the modified assimilated routine is capable of reducing the computational error inherited in any simulation model. Evaluation criteria such as RMSE are used to compare the results of SVR (forecast without a DA technique), DA (using EnKF) and MDA (using modified data assimilation). The quantified results indicate a better performance of the MDA approach over SVR and DA with RMSE average values of 9.54, 58.23 and 19.24 MCM, respectively. Additionally, the performance of the proposed technique was measured by comparison of computed inflow with observed data using R-squared criteria. The R-squared value for the MDA model was improved to 0.992, comparable to 0.85 of DA model.

REFERENCES

Asefa, T., Kemblowski, M. W., Urroz, G., Mckee, M. & Khalil, A. 2004 Support vector–based ground water head observation networks design. *Water Resour. Res.* **40**, W11509.

Asefa, T., Kemblowski, M., Mckee, M. & Khalil, A. 2006 Multi-Time scale stream flow prediction: the support vector machines approach. *J. Hydrol.* **318** (1), 7–16.

Babovic, V., Keijzer, M. & Bundzel, M. 2000 From global to local modelling: a case study in error correction of deterministic models. In: *Proceedings of Fourth International Conference on Hydroinformatics*. Iowa City, USA.

Behzad, M., Asghari, K., Eazi, M. & Palhang, M. 2009 Generalization performance of support vector machines and neural networks in runoff modeling. *Expert Syst. Appl.* **36** (4), 7624–7629.

DeChant, C. M. & Moradkhani, H. 2011 Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. *Hydrol. Earth Syst. Sci.* **15** (11), 3399–3410.

Dibike, Y. B., Velickov, S., Solomatine, D. P. & Abbott, M. B. 2001 Model induction with support vector machines: introduction and application. *J. Comput. Civil Eng.* **15** (3), 208–216.

Dumedah, G. & Coulibaly, P. 2014 Integration of an evolutionary algorithm into the ensemble Kalman filter and the particle filter for hydrologic data assimilation. *J. Hydroinform.* **16** (1), 74–94.

Gill, M. K., Kemblowski, M. W. & McKee, M. 2007 Soil moisture data assimilation using support vector machines and ensemble Kalman filter. *J. Am. Water Resour. Assoc.* **43** (4), 1004–1015.

Hartnack, J. & Madsen, H. 2001 Data assimilation in river flow modeling. In: *4th DHI Software Conference*, 6–8 June. Helsingor, Denmark.

Jajarmizadeh, M., Lafdani, E. K., Harun, S. & Ahmadi, A. 2015 Application of SVM and SWAT models for monthly streamflow prediction, a case study in south of Iran. *KSCE J. Civil Eng.* **19** (1), 345–357.

Kalman, R. E. 1960 A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82** (1), 35–45.

Lee, Y. H. & Singh, V. P. 1999 Tank model using Kalman filter. *J. Hydrol. Eng.* **4** (4), 344–349.

Li, X. L., Du, Z. L., Jiao, L. X. & Shen, K. 2012 Data assimilation by coupling uncertain support vector machine with ensemble Kalman Filter. In: *IEEE International Conference on Machine Learning and Cybernetics*, 15–17 July. Xian, China.

Li, X. L., Lü, H., Horton, R., An, T. & Yu, Z. 2014 Real-time flood forecast using the coupling support vector machine and data assimilation method. *J. Hydroinform.* **16** (5), 973–988.

Lin, J.-Y., Cheng, C.-T. & Chau, K.-W. 2006 Using support vector machines for long-term discharge prediction. *Hydrol. Sci. J.* **51** (4), 599–612.

Liu, D., Yu, Z. B. & Lue, H. S. 2010 Data assimilation using support vector machines and ensemble Kalman filter for multi-layer soil moisture prediction. *Water Sci. Eng.* **3** (4), 361–377.

Liu, D., Mishra, A. K. & Yu, Z. 2016 Evaluating uncertainties in multi-layer soil moisture estimation with support vector machines and ensemble Kalman filtering. *J. Hydrol.* **538**, 243–255.

Mancarella, D., Babovic, V., Keijzer, M. & Simeone, V. 2008 Data assimilation of forecasted errors in hydrodynamic models using inter-model correlations. *Int. J. Num. Methods Fluids* **56** (6), 587–605.

Nasseri, M., Moeini, A. & Tabesh, M. 2011 Forecasting monthly urban water demand using extended Kalman filter and genetic programming. *Expert Syst. Appl.* **38** (6), 7387–7395.

Refsgaard, J. C., Rosbjerg, D. & Markussen, L. M. 1983 Application of the Kalman filter to real-time operation and to uncertainty analyses in hydrological modeling. *Scientific Procedures Applied to the Planning, Design and Management of Water Resources Systems*. IAHS Publication No. 147, pp. 273–282.

Ricci, S., Piacentini, A., Thual, O., Pape, E. L. & Jonville, G. 2011 Correction of upstream flow and hydraulic state with data assimilation in the context of flood forecasting. *Hydrol. Earth Syst. Sci.* **15** (11), 3555–3575.

Siswantoro, J., Prabuwono, A. S., Abdullah, A. & Idrus, B. 2016 A linear model based on Kalman filter for improving neural network classification performance. *Expert Syst. Appl.* **49**, 112–122.

Sun, Y., Babovic, V. & Chan, E. S. 2010 Multi-step-ahead model error prediction using time-delay neural networks combined with chaos theory. *J. Hydrol.* **395** (1), 109–116.

Vapnik, V. 1995 *The Nature of Statistical Learning Theory.* Springer, New York.

Vapnik, V. 1998 *Statistical Learning Theory.* John Wiley and Sons, Toronto.

Wang, X. & Babovic, V. 2016 Application of hybrid Kalman filter for improving water level forecast. *J. Hydroinform.* **18** (5), 773–790.

Wood, E. F. & O'Connell, P. E. 1985 Real-time forecasting. In: *Hydrological Forecasting*. M. G. Anderson and T. P. Burt (eds). John Wiley and Sons, New York, pp. 505–558.

First received 9 January 2017; accepted in revised form 11 January 2018. Available online 5 February 2018