The application of support vector regression (SVR) for stream flow prediction on the Amazon basin

Melise du Toit[1, 2], Josefine M. Wilms[2], Gideon J.F. Smit[1] and Willie Brink[1]

[1] *Department of Mathematical Sciences (Applied Mathematics), Stellenbosch University, South Africa*
[2] *Advanced Mathematical Modelling, Modelling and Digital Sciences, CSIR Stellenbosch, South Africa*

Long-term forecasting of river runoff is important for climate scientists and hydrologists. By analysing the processes of a river basin characterized by measurable variables, an empirical data-driven model can be constructed. The support vector regression technique is used in this study to analyse historical stream flow occurrences and predict stream flow values for the Amazon basin. Up to twelve month predictions are made and the coefficient of determination and root-mean-square error are used for accuracy assessment. Compared to previous studies, satisfactory results are obtained. Inclusion of environmental aspects such as precipitation and evaporation are suggested for more accurate predictions.

Keywords: Support vector machine, Support vector regression, Amazon basin, Stream flow prediction

1. Introduction

Research on model-generated river runoff is essential for climate scientists and hydrologists to predict and understand future changes in river runoff that may be associated with global climate change. The hydrologic cycle is closed by returning the correct amount of water to the river mouth with the appropriate timing and position (Miller *et al*., 1993). River engineers and scientists use these results for the study of various hydro-environmental aspects, such as the increasing international concern of riverine pollution problems and the growing flood levels of rivers (Falconer *et al*., 2005). Furthermore, sediment transport and salinity changes within the river basin can be examined and predicted (Falconer *et al*., 2005; Miller *et al*., 1993).

Numerous hydrological models have been implemented by researchers to analyse the behaviour of river basins and to model river flow in such basins by mapping the natural phenomena to a simulation program (Falconer *et al*., 2005). These models are known as physically based or process models, since they are based on the physical behaviour of the specific river basin system as well as the mathematical description of the river flow (Falconer *et al*., 2005; Solomatine and Ostfeld, 2008). A physically based model consists of a numerical process which involves the computation of an efficient and accurate solution to equations based on the physical laws obtained for the specific system. The accuracy of a process model is tested by comparing its results to past observations, and if a desired accuracy is obtained, such a model may be used to calculate and predict future changes in the particular system.

Even though various hydraulic and hydrologic process models have been constructed for river basin systems, limited knowledge of the required modelling processes in a system may result in an unreliable model. However, such a system may consist of a process characterized by measurable variables and contain a sufficient amount of concurrent input and output data associated with the particular process (Solomatine and Ostfeld, 2008). By analysing the relationship between the input and output data an empirical mathematical model, known as a data-driven model, can be constructed to model and predict future output variables (Solomatine and Ostfeld, 2008).

A detailed understanding of the physical processes and behaviour of a river basin system is therefore not required for the construction of a data-driven model. Instead, data-driven modelling involves a study of the relationship between the system's state variables (Solomatine *et al*., 2008). This may allow for the improvement of physically based models.

The objective of this study is the description and implementation of an empirically based (data-driven) model for river runoff. In particular, a supervised machine learning model known as support vector regression (SVR) will be considered. This model is used to analyse the stream flow history of gauging stations in a river basin in order to determine future stream flow. The Amazon River in South America is considered for the application of this data-driven

Fax: +27 21 888 2693 Telephone: +27 76 198 0472 E-mail: mdutoit2@csir.co.za

model and an attempt to accurately predict stream flow is made.

2. Instrumentation and Method

2.1. Study area and available data

Stream flow data for the Amazon basin have been obtained from the Observation Service for the geodynamical, hydrological and biogeochemical control of erosion/alteration and material transport (SO HYBAM). This association manages 20 gauging stations that are distributed in the Amazon. The stream flow records of three are considered for this study: the Obidos station in Rio Amazonas, the Manacapuru station in Rio Solimões, and the Lábrea station in Rio Purus, shown in Fig. 1.

Figure 1. Study area and location of the gauging stations.

2.2. Support vector regression: model formulation

In order to forecast an outcome $y(t+\Delta t)$ at an instant Δt from current time t, a regression method can be constructed. The purpose of such a method is to formulate a function $f(\boldsymbol{x})$ such that $f(\boldsymbol{x}) = y(t+\Delta t)$. The function f takes an input vector $\boldsymbol{x} = (x_1, x_2, \dots, x_m)$ of m known variables, including current and past data records $[y(t), y(t-1), \dots, y(t-q)]$, where $q \leq m$. The input vector may also consist of any other available numerical variables.

An extension of the support vector machine (SVM), formulated by Cortes and Vapnik (1995), is known as the support vector regression (SVR) technique. A thorough description on the construction of the SVR technique, its optimization parameters (C and ϵ) and its applications in the field of hydrology can be found in Raghavendra and Deka (2014). An important concept of the SVR method is that it attempts to find a simple function that can fit all the data while minimizing the sum of prediction errors above a predefined margin (Callegari *et al*., 2015).

For cases where the SVR model has to optimize nonlinear functions, the input vector $\boldsymbol{x}$ is mapped to a feature space where its relationship with y is linearized. This mapping function is known as a kernel function. A detailed discussion on kernel functions is given by Raghavendra and Deka (2014).

2.3. Model training and testing

The process of formulating a function $f(\boldsymbol{x})$ on a given subset of the available data (known as the *training set*) is known as training. During training, the model is tested by fitting it to a second sample set (known as the *validation set*). Finally, the trained SVR model is verified by an accuracy measurement on a third subset of the given samples, known as the *test set* (Solomatine and Ostfeld, 2008).

For the Obidos gauging station, monthly stream flow data from 1970 to 2000 are considered. Furthermore, data from 1973 to 2003 and from 1968 to 1998 are available for the Manacapuru and Lábrea stations, respectively. For each station, data for the first 15 years are used as training sets. The following 10 years' data constitutes the validation set and the remaining 5 years' data are used for testing.

2.4. Feature and kernel function selection

Each input vector $\boldsymbol{x}$ consists of 12 antecedent stream flow periods (months). The value of y represents the flow in the next period. One month predictions are made, where after forecasting is extended for up to 12 months. Evaluation is done by calculating the coefficient of determination (R^2) of the predicted and observed stream flow values. The purpose of R^2 is to give an estimation of how well observed models are replicated by the fitted model, based on the percentage of total variation of outcomes interpreted by the model. The R^2 percentage therefore represents the percentage of variation of predicted outcomes that are explained by the fitted model. Furthermore, the root-mean-square error (*RMSE*) percentage indicates residual variance between observed and forecasted outcomes, and will be used for evaluation in this study.

Linear, Polynomial (Poly) and Radial Basis (RBF) kernel functions are considered. These SVR formulations are expressed as follows:

Linear: $k(x_i, x_j) = x_i^T x_j$,

Poly: $k(x_i, x_j) = (\gamma x_i^T x_j + r)^d$, and

RBF: $k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right);\ \gamma > 0.$

The mapping of features x_i and x_j to the feature space is represented by $k(x_i, x_j)$. An outline of the kernel functions and their hyperparameters are given by Granata *et al*. (2016). A built-in module in the Python programming language, known as Optunity, is used to optimize the parameters of each kernel.

3. Results and Discussion

3.1. Optimal hyperparameters and kernel functions

Historical stream flow records of the respective stations are examined. For each station, the optimal hyperparameters of the considered kernel functions are calculated in order to determine the best generalized model for the given data. The R^2 value for each optimized model is listed in Tables 1 to 3. For the training and validation sets, every kernel function provides an R^2 greater than 0.9, indicating that at least 90% of the total variation of predicted outcomes are explained by the fitted models. The RBF and polynomial kernel functions provide the best results for each station. However, the RBF kernel is less complex in comparison to polynomial kernels, since it contains fewer parameters. Further investigation is therefore done by only considering the RBF kernel.

Table 1. Optimized kernel-specific hyperparameters and R^2 for one month predictions of river flow at the Obidos gauging station.

OBIDOS GAUGING STATION							
	Optimal Parameters					R^2	
Kernel	C	ϵ	γ	d	r	Training set	Validation set
RBF	641	0.0303	0.067	-	-	0.983	0.967
Linear	304	0.0262	-	-	-	0.976	0.965
Poly	381	0.0307	0.1	2	0.3	0.981	0.966

Table 2. Optimized kernel-specific hyperparameters and R^2 for one month predictions of river flow at the Manacapuru gauging station.

MANACAPURU GAUGING STATION							
	Optimal Parameters					R^2	
Kernel	C	ϵ	γ	d	r	Training set	Validation set
RBF	570	0.02	0.03	-	-	0.937	0.923
Linear	385	0.048	-	-	-	0.912	0.904
Poly	78	0.0056	0.1	3	0.5	0.942	0.925

Table 3. Optimized kernel-specific hyperparameters and R^2 for one month predictions of river flow at the Lábrea gauging station.

LABREA GAUGING STATION							
	Optimal Parameters					R^2	
Kernel	C	ϵ	γ	d	r	Training set	Validation set
RBF	255	0.05	1.7	-	-	0.985	0.965
Linear	84	0.015	-	-	-	0.956	0.951
Poly	675	0.0329	0.1	5	0.11	0.984	0.959

3.2. Extended stream flow forecasting

The optimized RBF models are applied to the testing data for forecasting. At an instant (month) t, twelve antecedent observed flow values $\boldsymbol{x} = [y(t), y(t-1), \ldots, y(t-11)]$ are used to predict flow $f(\boldsymbol{x})_{\{t+1\}}$ for month $t+1$. This is known as *one month forecasting*. Similarly, for *two month forecasting*, an input vector $\boldsymbol{x} = [f(\boldsymbol{x})_{\{t+1\}}, y(t), \ldots, y(t-10)]$ is used to predict stream flow for month $t+2$. Forecasting extending up to 12 months is done on the given test set of each station. The corresponding R^2 values and $RMSE$ percentages are determined and shown in Figs. 2 and 3, respectively.

For each gauging station the best results were obtained for one month forecasting. An R^2 of 0.973 is obtained for the Obidos station, whereas R^2 values of 0.94 and 0.95 are obtained for the Manacapuru and Lábrea stations, respectively. Furthermore, the *RMSE* percentages are obtained respectively as 5.06%, 6.49% and 21.38%. R^2 is a relative error of fit, whereas $RMSE$ is an absolute measure of fit. Since $RMSE$ is the square root of a variance, it can be explained as the standard deviation of the unexplained variance. This clarifies the larger $RMSE$ values obtained for the Lábrea station. Compared to stream flow forecasting studies done by Veiga *et al.* (2015), Lin *et al.* (2006) and Callegari *et al.* (2015), these results are quite satisfactory.

Extended forecasting produces less accurate results. However, it should be taken into account that predicted stream flow values were used to make future predictions. Also, stream flow is the only environmental/hydrological variable considered.

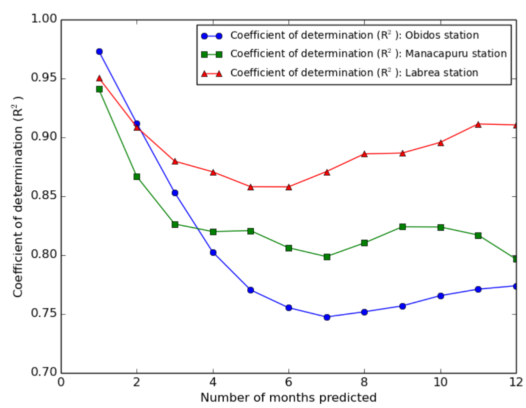

Figure 2. R^2 results for extended forecasting.

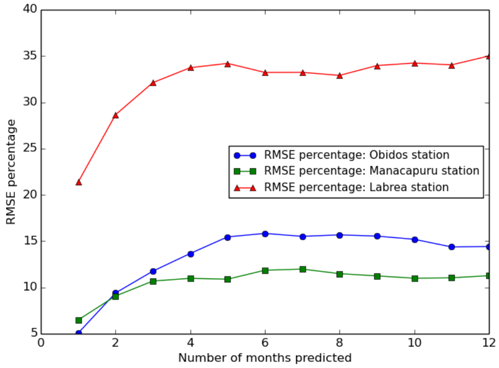

Figure 3. *RMSE* percentages for extended forecasting.

3.3. Illustrations of stream flow predictions

Figure 4 is an illustration of one, six and twelve month extended stream flow forecasting compared to observed stream flow. The worst predictions are made at the minimum and maximum stream flow occurrences, whereas good results are obtained for the upward and downward flow tendencies.

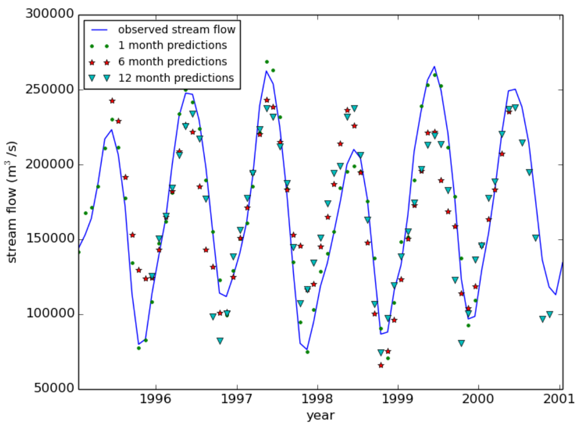

Figure 4: Stream flow discharge at the Obidos station for 1, 6, and 12 month predictions.

4. Conclusions

Research on long-term forecasting of river runoff predictions is important for climate scientists and hydrologists, since these results are used for the study of various hydro-environmental aspects. Numerous physically based hydrologic models have been implemented by researchers for this task, but due to limited knowledge of the necessary modelling processes in a river basin, inaccurate results have been obtained. Therefore, by analysing the processes of a river basin characterized by measurable variables, an empirical data-driven model can be constructed. The support vector regression (SVR) machine learning technique was used in this study to analyse historical stream flow occurrences in order to predict stream flow values. Predictions for up to twelve months were made and the coefficient of determination as well as the root-mean-square error were used as accuracy measurements. Satisfactory results were obtained and local stream flow data proved to be a trustworthy hydrological factor when predicting a specific river's stream flow. Even though the effects of precipitation may already be present in stream flow data, an understanding of the relationship between stream flow and precipitation may lead to a more accurate prediction of stream flow. Explicitly including precipitation and other environmental aspects such as temperature and evaporation when building an SVR model will therefore be addressed in further studies.

5. References

Callegari, M., Mazzoli, P., De Gregorio, L., Notarnicola, C., Pasolli, L., Petitta, M., Pistocchi, A. (2015). Seasonal river discharge forecasting using support vector regression: a case study in the Italian Alps. *Water*. 7: 2494-2515.

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning*. 20: 273-297.

Falconer, R., Lin, B., Harpin, R. (2005). Environmental modelling in river basin management. *International Journal of River Basin Management*. 3: 169-184.

Lin, J., Cheng, C., Chau, K. 2006. Using support vector machines for long-term discharge prediction. *Hydrological Sciences*. 51(4): 599-612.

Miller, J.R., Russel, G.L., Caliri, G. (1993). Continental-scale river flow in climate models. *Journal of Climate*. 7: 914-928.

Raghavendra, S., Deka, P.C. (2014). Support vector machine applications in the field of hydrology: a review. *Applied Soft Computing*. 19: 372-386.

Solomatine, D., Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics*. 10: 3-22.

Veiga, V.B., Hassan, Q.K., He, J. (2015). Development of flow forecasting models in the Bow River at Calgary, Alberta, Canada. *Water*. 7: 99-115.