GEOLOGICAL SURVEY OF CANADA
OPEN FILE 6864

Support vector machine for the prediction of future trend of Athabasca River (Alberta) flow rate

Y. Liu

2017

Canada

GEOLOGICAL SURVEY OF CANADA
OPEN FILE 6864

Support vector machine for the prediction of future trend of Athabasca River (Alberta) flow rate

Y. Liu

SoftMirrors Ltd., 76 Hawkwood Road, NW, Calgary, Alberta

2017

© Her Majesty the Queen in Right of Canada, as represented by the Minister of Natural Resources, 2017

Information contained in this publication or product may be reproduced, in part or in whole, and by any means, for personal or public non-commercial purposes, without charge or further permission, unless otherwise specified.
You are asked to:
- exercise due diligence in ensuring the accuracy of the materials reproduced;
- indicate the complete title of the materials reproduced, and the name of the author organization; and
- indicate that the reproduction is a copy of an official work that is published by Natural Resources Canada (NRCan) and that the reproduction has not been produced in affiliation with, or with the endorsement of, NRCan.

Commercial reproduction and distribution is prohibited except with written permission from NRCan. For more information, contact NRCan at nrcan.copyrightdroitdauteur.rncan@canada.ca.

doi:10.4095/299739

This publication is available for free download through GEOSCAN (http://geoscan.nrcan.gc.ca/).

Recommended citation
Liu, Y., 2017. Support vector machine for the prediction of future trend of Athabasca River (Alberta) flow rate; Geological Survey of Canada, Open File 6864, 29 p. doi:10.4095/299739

Publications in this series have not been edited; they are released as submitted by the author.

Summary

River flow process usually was affected by a wide variety of factors and it is very difficult to make the trend prediction. Various mathematical techniques have been developed to tackle the prediction, but these techniques are less accurate compared with physically-based models. In this paper I presented the recurrent support vector machine methods to study a series of climate variables, such as temperature and precipitation, and then predict the future trend of flow river rate based on these climate variables. The Athabasca River data have been tested and the flow river rates in 100 years have been predicted.

Introduction

River flow rate prediction is an extremely complex problem because the flow process can be influenced by a wide variety of factors including the precipitation intensity and distribution, temperature, channel characteristics, watershed geology and topography, vegetation cover, human activities (e.g. land-use changes) and even, indirectly, the greenhouse gas releases [Hu 2001].

Attempts to make the predictions of river flow rate have been a relentless pursuit by hydrologists and water resources engineers. Broadly speaking, two different approaches can be used to predict the river flow rate: physically-based models and stochastic models. Physically-based models try to represent the physical processes observed in the real world. Typically, such models contain representations of surface runoff, subsurface flow, evapotranspiration, and channel flow, but the physically-based models can be far more complicated. In contrast, the stochastic models based on data are black box systems, using mathematical and statistical concepts to link a certain input (for instance precipitation) to the model output (for instance flow rate). The stochastic models have the advantages of being simple and reasonably accurate and many hydrologists favor the use of stochastic models.

Various mathematical techniques have been developed to tackle the stochastic models, such as regression, neural networks and system identification. Recently, a number of complex processes have been modeled with the aid of neural networks for the river flow rate prediction. All these techniques are often less accurate compared with physically-based models which need the extremely complex relationships between river flow and its influencing factors. However, the physical parameters and effort required for calibrating a physical model are tremendous such that prediction of river flow using a physical process model is not viable in many circumstances.

To introduce an explicit link between climate variables in forecasting of future water supply, we proposed a model based on Support Vector Machines in this work. Firstly, the Support Vector Machine recursively predicts the temperature and precipitation for 100 coming years. Then, the forecasts of the temperature and precipitation obtained can be used further to predict the river flow rate for 100 coming years using the Support Vector Machine.

Support Vector Machine for Regression

The Support Vector Machine (SVM) (Vapnik, 1998) has become widely established as one of the main stream approaches to pattern recognition and machine learning. It makes predictions in terms of a linear combination of kernel functions centred on a subset of the training data, known as support vectors.

Despite its widespread success, the SVM suffers from some of its weakness, notably the absence of probabilistic outputs. The relevance vector machine (Tipping, M. E., 2004) uses the same kernel functional form as SVM's, but generates predictive distributions instead of point predictions by introducing Bayesian statistics. Under the assumption of a Gaussian distribution of the data, a Bayesian-based Support Vector Machine Classification employs the same kernel functions and also can make predictions based on the Gaussian likelihood.

Under the assumption of regression, we are given a set of input variables, $X = \{x_i \; i = 1, \dots N\}$ together with corresponding targets, such as river flow rate, $Y = \{y_i \; i = 1, \dots N\}$

The goal is to predict the annual river flow rate target using existing observed data based on the regression through supervised learning of the Bayesian-based Support Vector Machine.

The proposed method makes the predictions based on the following linear combinations of the form:

$$\boldsymbol{Y}(x_j) = \sum_{n=1}^{N} \omega_n \, K(x_j, x_n) + \omega_0 \qquad (1)$$

Where $\{\omega_n \; n = 1, \dots N\}$ are the model weights, and $\{K(x_j, x_n\}$ denotes the kernel functions ($Schölkopf$ et al., 1999). The table 1 provides examples of kernel functions that have been extensively tested in the machine learning literature (Vapnik, 1998).

Table 1

Type of Kernel Function	Kernel Function
Polynomial of degree p	$K(u, v) = (u^T v + 1)^p$
Gaussian Radial Basis with width p	$K(u, v) = exp(-(u - v)(u - v)^T / p^2)$

To avoid notational clutter we will re-write equation (1) in matrix form:

$$Y = W\emptyset \qquad (2)$$

where $\{\emptyset\}$ denotes the design matrix of kernel functions, the vector of targets is given by$\{Y\}$, and $\{W\}$ indicates the vector containing the unknown weights.

In the Bayesian-based support vector machine, the likelihood function of the dataset can be written as:

$$p(t|w,\sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} exp\left\{-\frac{1}{2\sigma^2}\|t - W\emptyset\|^2\right\} \quad (3)$$

The posterior over the weights is then obtained from Bayes' rule:

$$p(w|t,\alpha,\sigma^2) = (2\pi)^{-\frac{(N+1)}{2}}|\sum|^{-1/2} exp\left\{-\frac{1}{2}((w-\mu)^T\sum^{-1}(w-\mu)\right\} \quad (4)$$

with

$$\sum = (\emptyset^T B\emptyset + A)^{-1} \quad (5)$$
$$\mu = \sum\emptyset^T Bt \quad (6)$$

where we defined $A = diag(\alpha_0, \alpha_1, \ldots, \alpha_N)$

By integrating the weights from (4), we obtain the marginal likelihood (Tipping, M. E., 2004):

$$p(t|\alpha,\sigma^2) = (2\pi)^{-\frac{(N+1)}{2}}|B^{-1} + \phi A^{-1}\phi^T|^{-1/2} exp\left\{-\frac{1}{2}t^T(B^{-1} + \phi A^{-1}\phi^T)^{-1}t\right\} \quad (7)$$

For the regression, we cannot experimentally explore the space of possible σ so we instead optimise directly via the iterative procedure to obtain the α_i and then calculate the weights using the following procedure.

1) Calculate $\alpha_i = \gamma_i/\mu_i^2$ where we defined $\gamma_i = 1 - \alpha_i\sum_{ii}$
2) Compute the variance matrix $\sum = (\emptyset^T B\emptyset + A)$
3) Calculate the residual between original and predicted results.
4) Repeat 1) to 3) until the residual does not be reduced or other convergence criteria are satisfied.

In practice, during the iteration, α_i could approach infinity and Eq. (4) become infinitely peaked at zero – implying that the corresponding kernel functions can be 'pruned'. Those α_i will be removed during the iteration.

Methods for Annual River Flow Rate Prediction

1) Recurrent Support Vector Machine to Forecast Annual Temperature and Precipitation

Support Vector Machines has been gaining popularity in regression and classification due to its excellent performance at the time of dealing with sparse inputs and also has been used as time series forecasters such as finance stock evolution and river flow rate predictions.

During river low-flow and high-flow periods the river flow rate is determined by a series of significant time series variables, such as the temperature and precipitation. During the annual river flow rate study, the annual flow rates (minimum, maximum, total and mean) of each year were calculated from the monthly observed dataset and the annual variables were estimated from the monthly observed dataset.

The temperature or precipitation based on the observed time series variables can be expressed as X: $\{x_i, i = 1, \dots N\}$

SVM Training Dataset
All the observed variables will be used to build the training dataset.

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & \dots & x_d \\ x_2 & x_3 & x_4 & x_5 & \dots & x_{d+1} \\ x_3 & x_4 & x_5 & x_6 & \dots & x_{d+2} \\ x_4 & x_5 & x_6 & x_7 & \dots & x_{d+3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_t & x_{t+1} & x_{t+2} & x_{t+3} & \dots & x_{N-1} \end{pmatrix} = \begin{pmatrix} f(X_1) \\ f(X_2) \\ f(X_3) \\ f(X_4) \\ \dots \\ f(X_t) \end{pmatrix} \quad (8)$$

$$\mathbf{Y} = \begin{pmatrix} x_{d+1} \\ x_{d+2} \\ x_{d+3} \\ x_{d+4} \\ \dots \\ x_N \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \dots \\ y_t \end{pmatrix} \quad (9)$$

Where **X** is training dataset and **Y** is target vectors. The d is the time lag, N is total observed samples and t = N-d. $\{X_i = \{x_{1,}x_{2,}x_{3,} \dots, x_d\}\ i = 1, \dots t\}$ is one of input training samples.

The following linear combinations will be applied to the training dataset to optimize and calculate the model weights $\{\omega_n\ n = 1, \dots N\}$

$$\boldsymbol{Y}(X_j) = \sum_{n=1}^{N} \omega_n\, K(X_j, X_n) + \omega_0 \quad (10)$$

When building the training dataset, having a sufficiently large time lag window is important for a time series predictor - if the window is too small then the system is being projected onto a space of insufficient dimension, in which the system cannot make reliable predictions. If a window is too large, it may also have problems: because all necessary information is populated in a subset of the window, the remaining fields will represent noise or contamination, in which the system will make wrong time-trend predictions. In Support Vector Machine training phase, if the lag time window is reasonable, the predicted values will be optimized; otherwise the equation will be ill-posed and cannot find the optimized results. Usually we need a series of test to determine whether the time lag window is reasonable or not.

SVM Predictions
To predict more than one-step ahead value of time series, support vector machine uses the predicted values as known data for the next ones. The recurrent SVM model can be constructed by first making one-step ahead prediction:

$$X_{t+1} = \{x_{t+1,}x_{t+2,}x_{t+3,} \dots, y_t\} \quad (11)$$

$$\boldsymbol{y_{t+1}} = \sum_{n=1}^{N} \omega_n\, K(X_{t+1}, X_n) + \omega_0 \quad (12)$$

where t denotes the numbers of input dataset and y_t denotes one-step ahead predicted value.

To predict the next value, the same prediction function is used:

$$X_{t+2} = \{x_{t+2,} x_{t+3,} x_{t+4,} \dots, y_{t+1}\} \quad (13)$$

$$\boldsymbol{y_{t+2}} = \sum_{n=1}^{N} \omega_n K(X_{t+2}, X_n) + \omega_0 \quad (14)$$

In this equation, the predicted value of y_{t+1} is used instead of the true value, which is unknown. Then, for the M-steps ahead prediction, y_{t+2} to y_{t+M} are predicted iteratively. When the predicted length M is larger than d, there are M-d real data to predict M^{th} value. When M exceeds d, all input variables are predicted values. The main problem of the recurrent support vector machine forecasting strategy is that there is a certain amount of error between the predicted value $\boldsymbol{y_{t+1}}$ and the true value. As the first predicted value is taken as input to obtain the second one, this error is propagated through the prediction function. The second predicted value has potentially twice more error: the difference between $\boldsymbol{y_{t+2}}$ and the true value plus the propagated error. With an increasing step ahead prediction, this accumulation can be important.

2) Support Vector Machine to Forecast Annual River Flow Rate

The river can be considered a complex, nonlinear system through which input variables, i.e. temperature, precipitation and Decadal Pacific Oscillation (DPO), are transformed by our support vector machine into output variables, i.e., river flow annual maximum rate, minimum rate, average rate and total rate.

Given a set of input variables, $\{XX = \{G_i\, S_i, P_i, D_i\}, i = 1, \dots N\}$ together with corresponding targets, such as river annual maximum flow rate and minimum flow rate:
$\{YY = \{F_i\, R_i, V_i, A_i\}, i = 1, \dots N\}$

The input observed variables can be expressed as following:

$$\text{XX} = \begin{pmatrix} G_1 & S_1 & P_1 & D_1 \\ G_2 & S_2 & P_2 & D_2 \\ G_3 & S_3 & P_3 & D_3 \\ G_4 & S_4 & P_4 & D_4 \\ \dots & \dots & \dots & \dots \\ G_N & S_N & P_N & D_N \end{pmatrix} = \begin{pmatrix} f(X_1) \\ f(X_2) \\ f(X_3) \\ f(X_4) \\ \dots \\ f(X_N) \end{pmatrix} \quad (15)$$

$\{G_i = \{g_1, g_2, \dots, g_d\}, i = 1, \dots N\}$ is training vectors of maximum annual temperatures and d is the lag time.
$\{S_i = \{s_1, s_2, \dots, s_d\},, i = 1, \dots N\}$ is training vector of minimum annual temperatures
$\{P_i = \{p_1, p_2, \dots, p_d\},, i = 1, \dots N\}$ is training vector of annual precipitation
$\{D_i = \{p_1, p_2, \dots, p_d\},, i = 1, \dots N\}$ is training vector of annual Decadal Pacific Oscillation (DPO)

The output observed variables can be expresses as:

$$YY = \begin{pmatrix} F_1 & R_1 & V_1 & A_1 \\ F_2 & R_2 & V_2 & A_2 \\ F_3 & R_3 & V_3 & A_3 \\ F_4 & R_4 & V_4 & A_4 \\ \cdots & \cdots & \cdots & \cdots \\ F_N & R_N & V_N & A_N \end{pmatrix} \quad (16)$$

$\{F_i, i = 1, \ldots N\}$ is training target of maximum annual river flow rate
$\{R_i, i = 1, \ldots N\}$ is training target of minimum annual river flow rate
$\{V_i, i = 1, \ldots N\}$ is training target of total annual river flow rate
$\{A_i, i = 1, \ldots N\}$ is training target of average annual river flow rate

The goal is to predict the annual river flow rate target using above existing observed data pairs (XX, YY) based on Bayesian regression through supervised learning of the Support Vector Machines.

During the training phase, the proposed method makes the trainings based on the following linear combinations of the form using the observed data pairs (XX, YY):

$$\boldsymbol{YY}(X_j) = \sum_{n=1}^{N} \omega_n K(X_j, X_n) + \omega_0 \quad (17)$$

During the prediction phase, the proposed method makes the river flow rate predictions based on the same linear combinations as the training phase:

$$\boldsymbol{YY}(Y_m) = \sum_{n=1}^{N} \omega_n K(Y_m, X_n) + \omega_0 \quad (18)$$

In this equation, the long-term predicted value of Y_m is unknown, which is forecasted iteratively using recurrent support vector machine. M is the predicted length and means the M-steps ahead prediction.

The main problem of the support vector machine forecasting strategy is that the recurrent predicted variables instead of the true variables which are not known, such as temperature and precipitation, are taken as input values to make the predictions of the river flow rate. Because there is a certain amount of error between the recurrent predicted value $\boldsymbol{y_{t+1}}$ and the true value and also the first recurrent predicted value is taken as input to obtain the second one, this error is propagated through the prediction function. So the accumulation error can be important for the long-term support vector machine. However, the support vector machine has been trained with all available observed input variables, such as temperature and precipitation, and output targets (such as flow rate), with the purpose of long-term river flow rate prediction which seems to be a better approach.

Methods for Daily Minimum River Flow Rate Prediction

1) Support Vector Machine to Forecast Daily Minimum Flow Rate

The river can be considered a complex, nonlinear system through which input variables, i.e. temperature, precipitation and Decadal Pacific Oscillation (DPO), are transformed by our

support vector machine into output variables, i.e., river flow annual maximum rate, minimum rate, average rate and total rate. During the daily minimum flow rate study, the daily minimum flow rate of each year was calculated from the daily dataset and other annual variables were estimated from the monthly dataset or calculated from the daily observed dataset if the dataset are available.

Given a set of input variables, $\{XX = \{G_i\, S_i, P_i, D_i\}, i = 1, \dots N\}$ together with corresponding targets, such as daily minimum flow rate: $\{YY = \{F_i\ \}, i = 1, \dots N\}$

The input observed variables can be expressed as following:

$$XX = \begin{pmatrix} G_1 & S_1 & P_1 & D_1 \\ G_2 & S_2 & P_2 & D_2 \\ G_3 & S_3 & P_3 & D_3 \\ G_4 & S_4 & P_4 & D_4 \\ \dots & \dots & \dots & \dots \\ G_N & S_N & P_N & D_N \end{pmatrix} \quad (15)$$

$\{G_i = \{g_1, g_2, \dots, g_d\}, i = 1, \dots N\}$ are training vectors of maximum annual temperatures and d is the lag time.
$\{S_i = \{s_1, s_2, \dots, s_d\},, i = 1, \dots N\}$ are training vectors of minimum annual temperatures
$\{P_i = \{p_1, p_2, \dots, p_d\},, i = 1, \dots N\}$ are training vectors of annual precipitation
$\{D_i = \{p_1, p_2, \dots, p_d\},, i = 1, \dots N\}$ are training vectors of annual Decadal Pacific Oscillation (DPO)

The output observed variables can be expresses as:

$$YY = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ \dots \\ F_N \end{pmatrix} \quad (16)$$

$\{F_i, i = 1, \dots N\}$ is training target of daily minimum river flow rate

The goal is to predict the daily minimum river flow rate using above existing observed data pairs (XX, YY) based on Bayesian regression through supervised learning of the Support Vector Machines.

During the training phase, the proposed method makes the trainings based on the following linear combinations of the form using the observed data pairs (XX, YY):

$$\boldsymbol{YY}(X_j) = \sum_{n=1}^{N} \omega_n\, K(X_j, X_n) + \omega_0 \quad (17)$$

During the prediction phase, the proposed method makes the river flow rate predictions based on the same linear combinations as the training phase:

$$\boldsymbol{YY}(Y_m) = \sum_{n=1}^{N} \omega_n \, K(Y_m, X_n) + \omega_0 \qquad (18)$$

In this equation, the long-term predicted value of Y_m is unknown, which is forecasted iteratively using recurrent support vector machine. M is the predicted length and means the M-steps ahead prediction. X is the temperature and precipitation which were obtained through the recurrent support vector machine described from section 3.

The main problem of the support vector machine forecasting strategy is that the recurrent predicted variables instead of the true variables which are not known, such as temperature and precipitation, are taken as input values to make the predictions of the river flow rate. Because there is a certain amount of error between the recurrent predicted value $\boldsymbol{y_{t+1}}$ and the true value and also the first recurrent predicted value is taken as input to obtain the second one, this error is propagated through the prediction function. So the accumulation error can be important for the long-term support vector machine. However, the support vector machine has been trained with all available observed input variables, such as temperature and precipitation, and output targets (such as flow rate), with the purpose of long-term river flow rate prediction which seems to be a better approach.

2) Recurrent Support Vector Machine to Forecast Daily Minimum Flow Rate

If the climate variables, such as temperature and precipitation are not available, we present a recurrent support vector machine to forecast the daily minimum flow rate. The daily minimum flow rate can be expressed as target input pairs (X, Y)$\{x_i, i = 1, \ldots d; x_{d+1}\}$. The training and prediction phase are the same as recurrent support vector machine to make the prediction of temperature and precipitation. The flow rate time series variables can be expressed:

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & \cdots & x_d \\ x_2 & x_3 & x_4 & x_5 & \cdots & x_{d+1} \\ x_3 & x_4 & x_5 & x_6 & \cdots & x_{d+2} \\ x_4 & x_5 & x_6 & x_7 & \cdots & x_{d+3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_t & x_{t+1} & x_{t+2} & x_{t+3} & \cdots & x_{N-1} \end{pmatrix} \qquad (19)$$

$$\mathbf{Y} = \begin{pmatrix} x_{d+1} \\ x_{d+2} \\ x_{d+3} \\ x_{d+4} \\ \cdots \\ x_N \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \cdots \\ y_t \end{pmatrix} \qquad (20)$$

Where **X** is the observed flow rate dataset from 1959 to 2008 and **Y** is target flow rate vectors. The d is the time lag, N is total observed samples and t = N-d.

The recurrent support vector machine uses the predicted flow rate values as known data for the next ones and the predicted model can be constructed by first making one-step ahead prediction:

$$X_{t+1} = \{x_{t+1,} x_{t+2,} x_{t+3,} \dots, y_t\} \quad (21)$$

$$\boldsymbol{y_{t+1}} = \sum_{n=1}^{N} \omega_n K(X_{t+1}, X_n) + \omega_0 \quad (22)$$

where t denotes the numbers of input dataset and y_t denotes one-step ahead predicted value. To predict the next value, the same prediction function is used:

$$X_{t+2} = \{x_{t+2,} x_{t+3,} x_{t+4,} \dots, y_{t+1}\} \quad (23)$$

$$\boldsymbol{y_{t+2}} = \sum_{n=1}^{N} \omega_n K(X_{t+2}, X_n) + \omega_0 \quad (24)$$

In this equation, the predicted flow rate value of y_{t+1} is used instead of the true value, which is unknown. Then, for the M-steps ahead flow rate prediction, y_{t+2} to $y_{t+M}$are predicted iteratively. As the first predicted flow rate value is taken as input to obtain the second flow rate one, this error is propagated through the prediction function. The second predicted flow rate has potentially twice more error: the difference between $\boldsymbol{y_{t+2}}$and the true value plus the propagated error. With an increasing step ahead prediction, this accumulation can be important.

Compared with flow rate prediction with climate variables, such as temperature and precipitation, the recurrent model is simple and only depends on the flow rate. Although the flow rate prediction based on the climate variables has an advantage which can consider the climate variables contributions, the recurrent predicted variables instead of the true variables which are not known, such as temperature and precipitation, are taken as input values to make the predictions of the river flow rate. So the accumulation error of climate variables can be important for the long-term support vector machine.

3) Validation for Daily Minimum Flow Rate Training

During the training phase of flow rate prediction, it is difficult to determine the kernel function parameters. The validation phase is one of powerful tool. Because there are limited flow rate dataset, we divide the annual flow rate dataset into two parts, one for training and another for validation. During our study, the validation dataset from one year to seven years has been tested.

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & \dots & x_d \\ x_2 & x_3 & x_4 & x_5 & \dots & x_{d+1} \\ x_3 & x_4 & x_5 & x_6 & \dots & x_{d+2} \\ x_4 & x_5 & x_6 & x_7 & \dots & x_{d+3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{t-v} & x_{t+1-v} & x_{t+2-v} & x_{t+3-v} & \dots & x_{N-1-v} \end{pmatrix} \quad (25)$$

$$\mathbf{Y} = \begin{pmatrix} x_{d+1} \\ x_{d+2} \\ x_{d+3} \\ x_{d+4} \\ \dots \\ x_{N-v} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \dots \\ y_{t-v} \end{pmatrix} \quad (26)$$

Where **X** is the observed flow rate dataset or observed climate variables from 1959 to 2008-v and **Y** is target flow rate vectors. The d is the time lag, N is total observed samples, t = N-d and v is validation years.

The support vector machine makes the river flow rate predictions based on the observed pairs (**X, Y**). The error is the difference between predicated values and validation values.

$$\boldsymbol{Y}(y_m) = \sum_{n=1}^{N} \omega_n \, K(Y_m, X_n) + \omega_0 \qquad (27)$$

$$\boldsymbol{error} = \sum_{n=1}^{V} ||\boldsymbol{Y}(y_n) - \boldsymbol{Y}(o_n)|| \qquad (28)$$

Because the validation value are true values and are available for comparison with the predicted values, so we can update the kernel function parameters so as there are minimum errors between the true value and predicted values. The best parameters with minimum errors will be applied to support vector machine to make predictions of future trend of flow rate.

The Athabasca River Example

The Athabasca River stretches from the Columbia Ice Fields near the Alberta-British Columbia border to its mouth in Lake Athabasca, at the northeastern corner of Alberta (Figure 1). Its length is estimated to be 1400 km, making it the third longest undammed river in North America, behind the Yukon and Mackenzie rivers, and slightly longer than the Fraser River. Over its length the Athabasca River drops about 800 m, with two-thirds of this drop occurring in the first 450 km [Schindler, 2007].

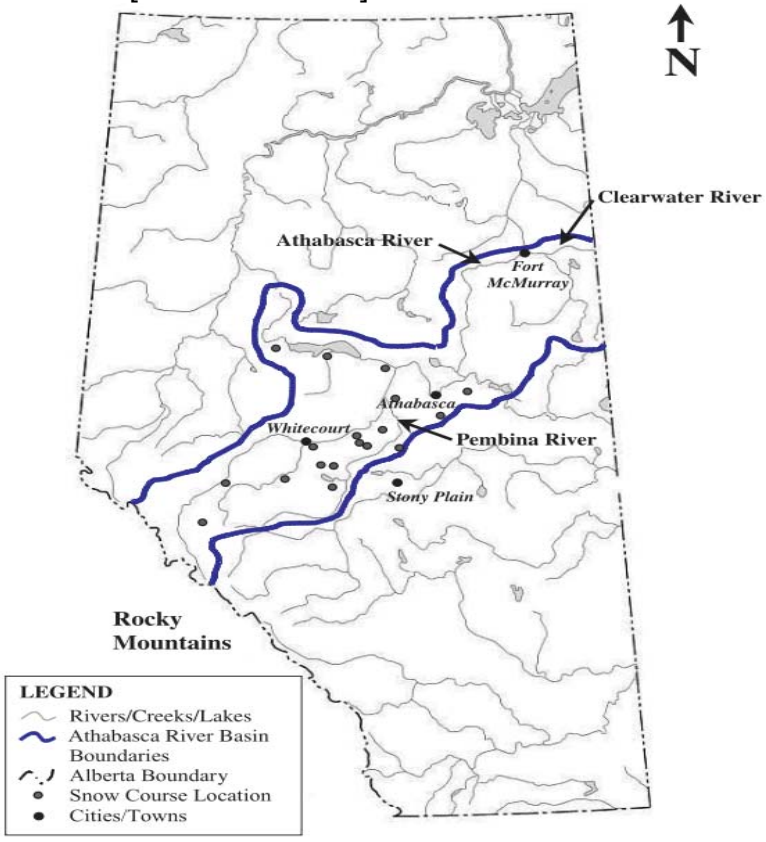

Figure 1: Bas Map of Athabasca River

Alberta's oil sands are not only the world's largest capital project but now represent 60 percent of the world's investable oil reserves. But in order to produce the oil sands, the intensive water requirements, combined with climate change, may threaten the water security of two northern territories, 300,000 aboriginal people and Canada's largest watershed: The Mackenzie River Basin (2007, University of Alberta). A 2006 Alberta report (Investing in Our Future) noted that "over the long term the Athabasca River may not have sufficient flows to meet the mining operations and maintain streams flows". To address these critical issues, we propose the Bayesian-based support vector machine method to make the prediction of future trend of the daily minimum flow rate using the historical observations of river flow rate and climate time series of temperature and precipitation.

Long-term flow monitoring records are not available for the study. A short period of records, such as river flow rate, temperature and precipitation, from 1959 to 2008, were collected at Clean Water, Edson Creek, Slave Lake, White Court and Fort McMurray stations. At Fort McMurray station, the highest daily minimum flow recorded during the above period was 211 cms in 1997. The lowest recorded flow was 75 cms in 2001. During a succession of dry years from 1997 to 2003, flows were less than 100 cms for almost four months in winter.
The minimum flows have been declined since 1959 (Figure 2). This is of concern whether the water flow in future can meet the needs of all the oil sands operations.

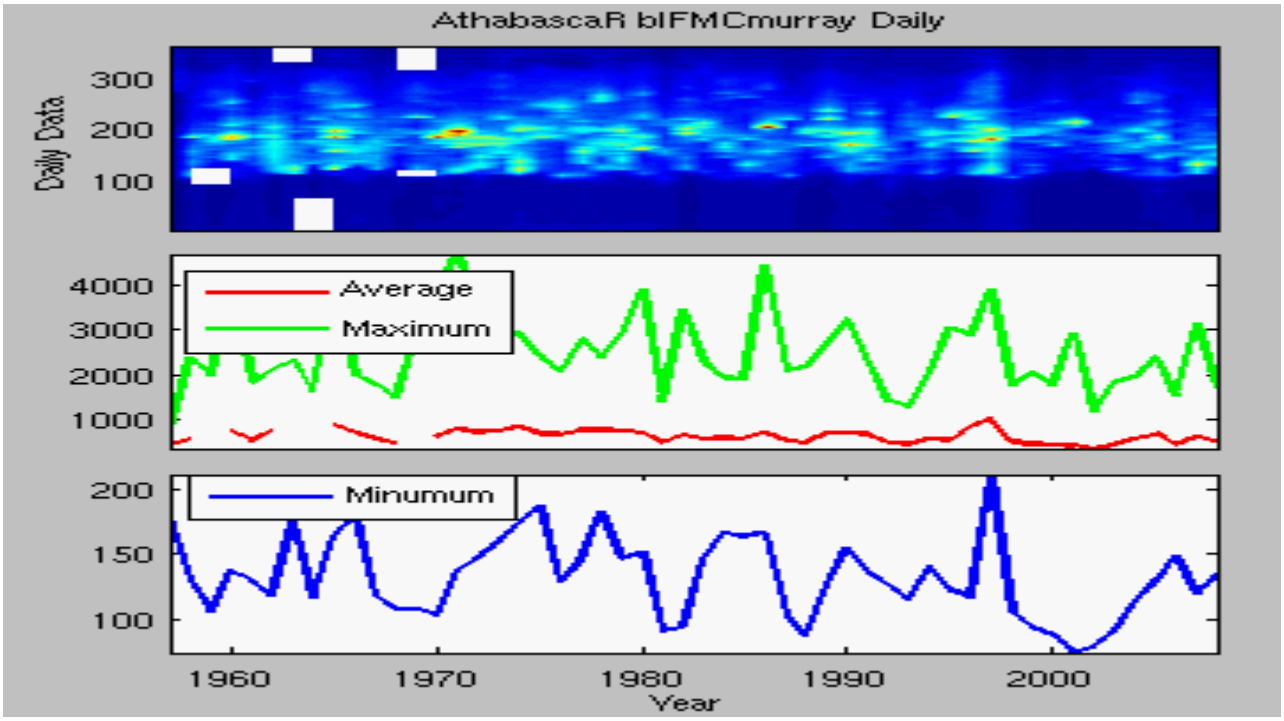

Figure 2: (Top) Daily flow rate of Fort McMurray station from 1959 to 2008. (Middle) the daily average and maximum flow rate. (Bottom) the daily minimum flow rate. The lowest recorded flow rate was 75 cms in 2001.

It is well known that the river flow rate prediction is an extremely complex problem and also is difficult for long term predictable because the flow process can be influenced by a wide variety of factors including the precipitation intensity and distribution, climate warning, drought, channel characteristics, watershed geology and topography, vegetation cover,

human activities (e.g. land-use changes) and even, indirectly, the greenhouse gas releases [Hu 2001].

To introduce an explicit link between climate variables in forecasting of future water supply of the Athabasca River, we proposed a model based on Support Vector Machines in this work. Firstly, the Support Vector Machine recursively predicts the temperature and precipitation for 100 coming years. Then, the forecasts of the temperature and precipitation obtained can be used further to predict the river flow rate for 100 coming years using the Support Vector Machine.

Athabasca Annual Average River Flow Rate Prediction

1) Data Used for Annual River Flow Rate Prediction

In this work, we use the Bayesian-based support vector machine method to make a long-term prediction for the river flow rate and time trends around Fort McMurray station in the Athabasca River. Generally, the river flow rate is affected by different climate change scenarios, such as temperature and precipitation. However, long-term flow rate, temperature and precipitation are not available for this study and a short period of observed records were collected at Fort McMurray, Clean Water, Edson Creek, Slave Lake and White Court stations from 1943 to 2007. The following figures show the historical time series of annual river flow, temperature changes and precipitation density at several stations located upstream of Fort McMurray. During the training phase of annual flow rate prediction, the validation phase is not applied, but we will apply the validation phase at daily minimum river flow rate study.

Figure 3 is Fort McMurray annual temperature time series trend. The upper window is monthly temperature display. The red color indicates high temperature and blue is low temperature. The Y axis is month and X is year. The middle is annual average of daily maximum temperature and bottom is annual sum of the monthly average.

Figure 4 is Fort McMurray annual minimum temperature time series trend. The upper window is monthly temperature display. The red color indicates the high and blue is low. The Y axis is month and X is year. The middle is annual average of daily minimum temperature and the bottom is annual sum of the monthly average.

Figure 5 is Fort McMurray annual precipitation time series trend. The upper window is monthly precipitation display. The red color indicates high precipitation and blue is low precipitation. The Y axis is month and X is year. The middle is annual average precipitation and the bottom is annual total precipitation.

Figure 6, 7, 8, 9 and Figure 10 are Athabasca annual precipitation display, Edson Creek annual maximum temperature time trend, Slave Lake annual maximum temperature time trend, White Court annual precipitation time trend and DPO time series index trend

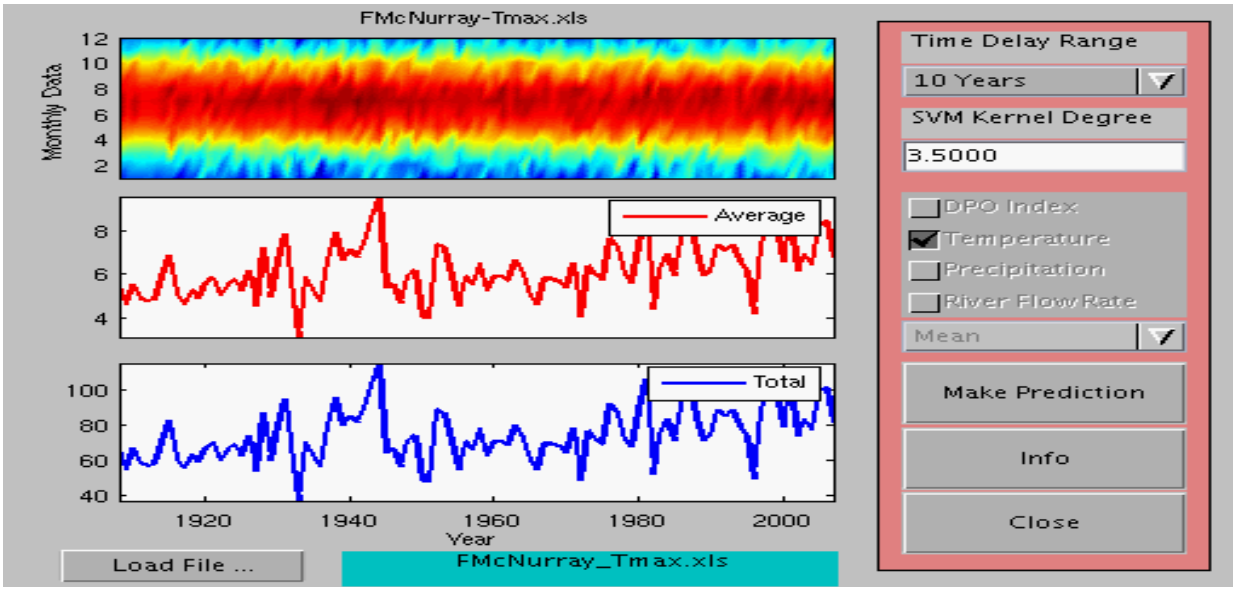

Figure 3: Fort McMurray annual maximum temperature time series trend

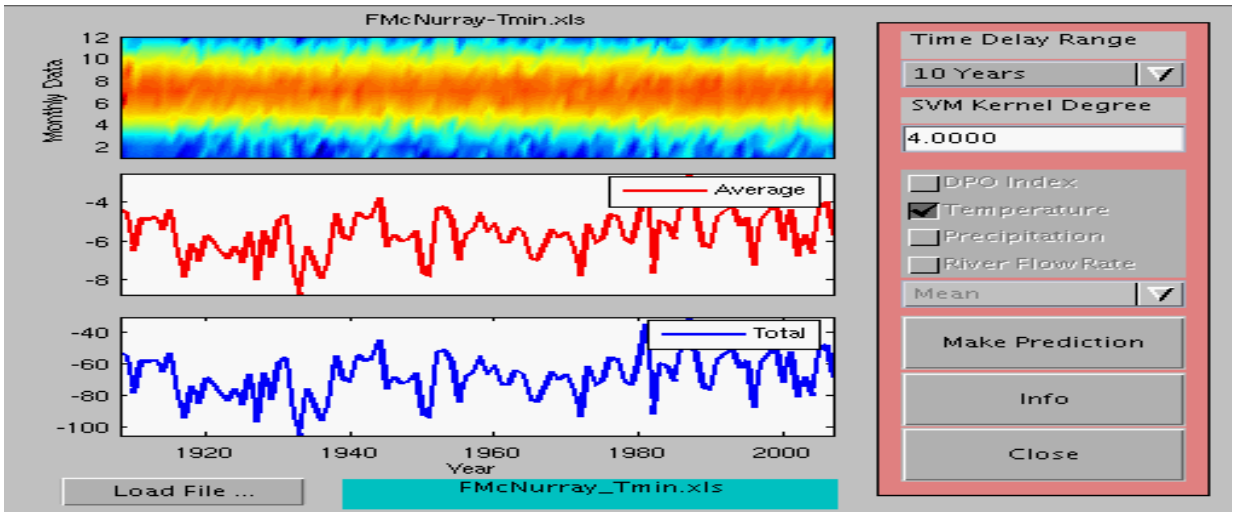

Figure 4: Fort McMurray annual minimum temperature time series trend

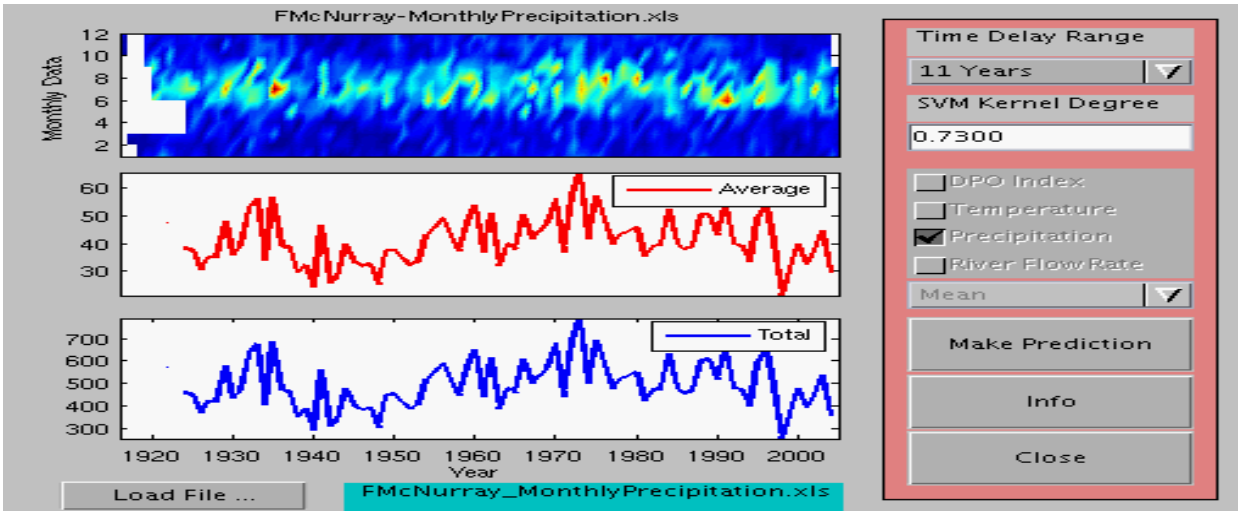

Figure 5: Fort McMurray annual precipitation time series trend

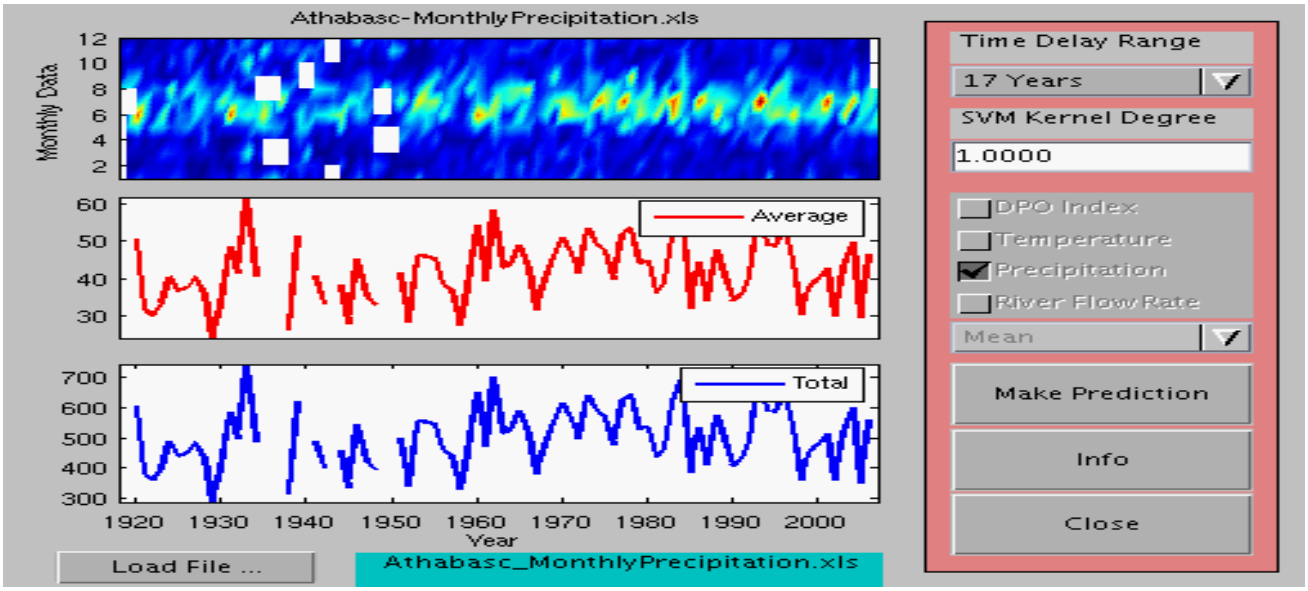

Figure 6: Athabasca annual precipitation time series trend

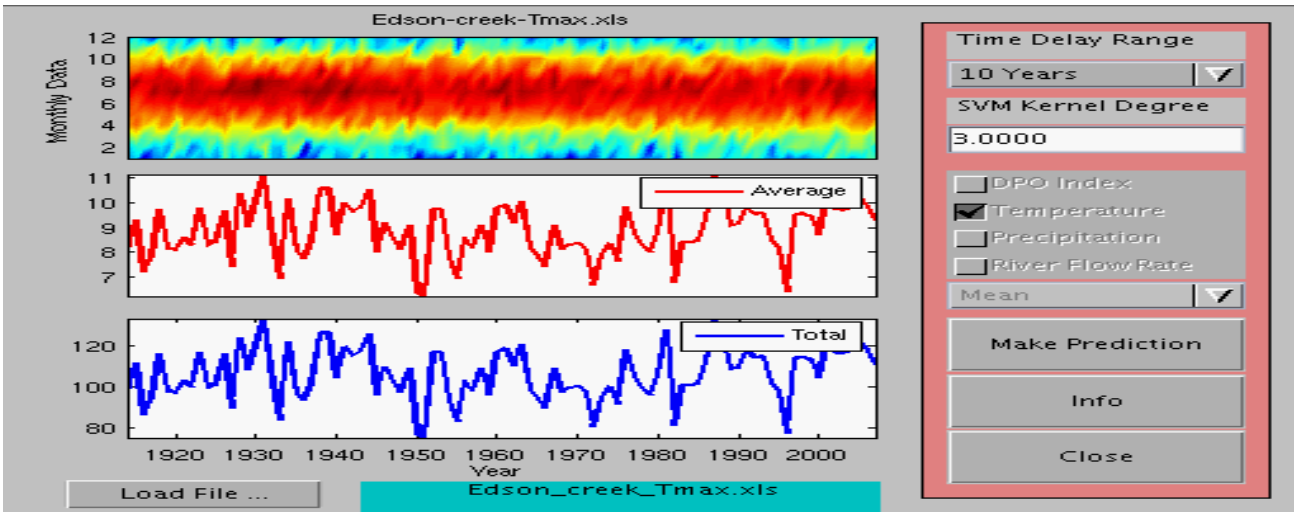

Figure 7: Edson Creek annual maximum temperature time series trend

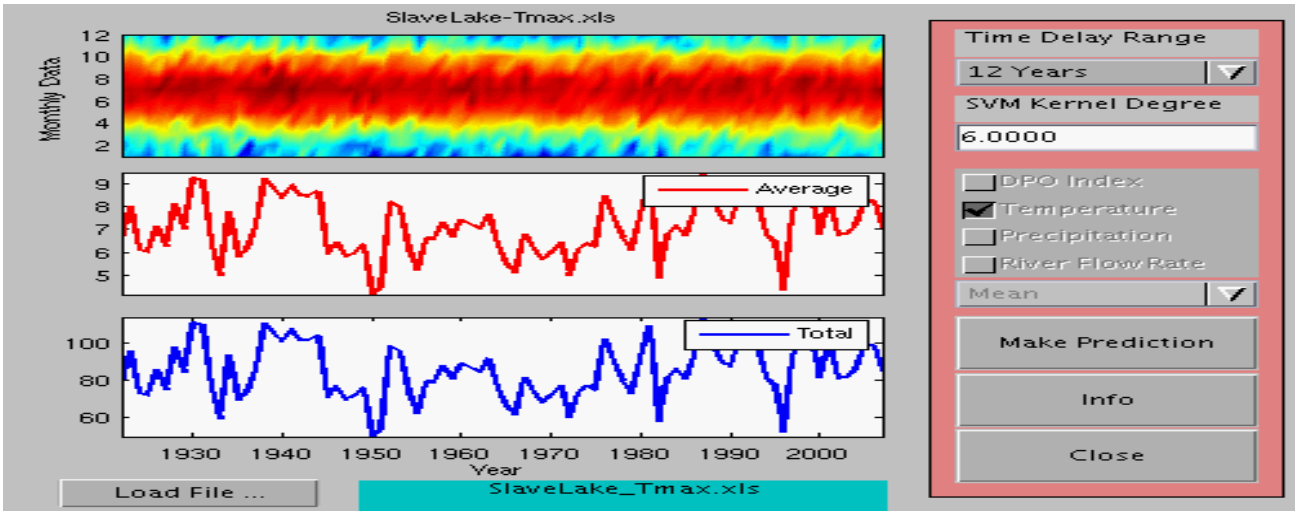

Figure 8: Slave Lake annual temperature time series trend

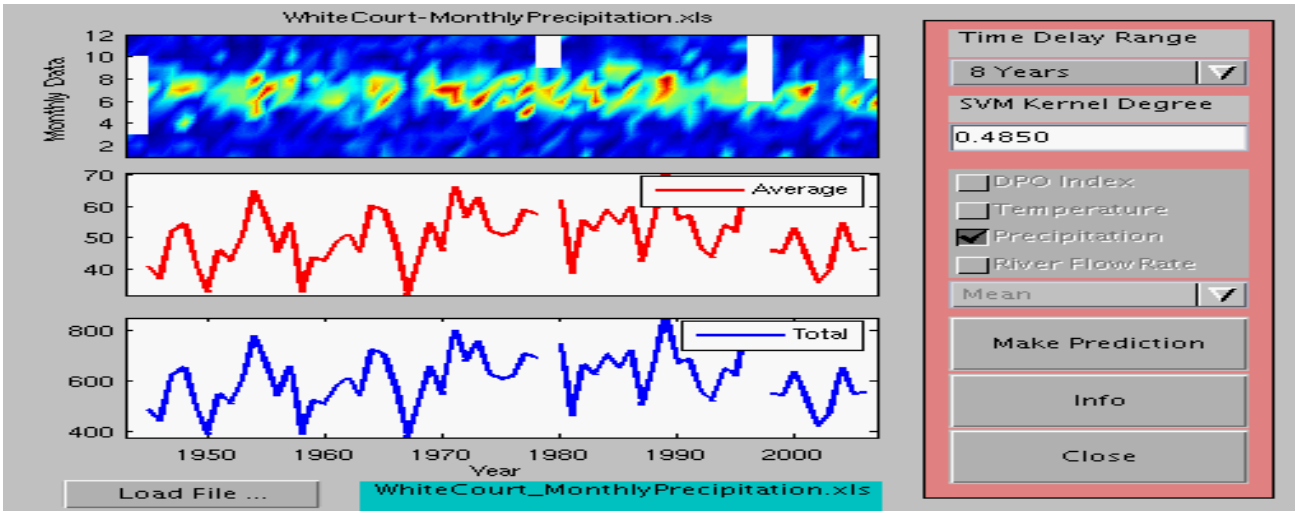

Figure 9: White Court annual precipitation time series trend

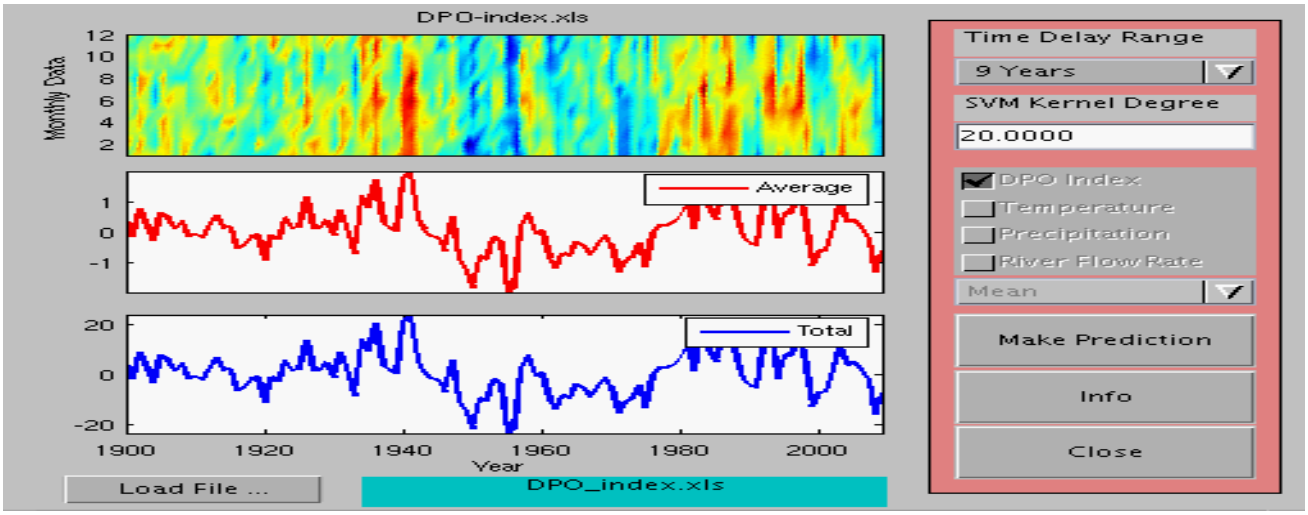

Figure 10: DPO annual index time series trend

2) Recurrent SVM Annual Temperature and Precipitation Predictions

The recurrent SVM models are trained using these observed temperature and precipitation data and the hyperparameter weights $\{\omega_n\ n = 1, \ldots N\}$ were optimised directly via the iterative procedure. The M-steps ahead predictions of the observed data are predicted iteratively based on the same linear combinations as the training's. The followings are the recurrent SVM long-term prediction results.

Figure 11 is Fort McMurray annual maximum temperature long-term trend prediction. The lag time is 10 years and the kernel function of SVM is Gaussian Radial Basis with width 3.5. These parameters need a series of test to determine based on the experience and reasonable trends.

Figure 12 is Fort McMurray annual minimum temperature long-term trend prediction. The lag time is 10 years and the kernel function of SVM is Gaussian Radial Basis with width 4.

Figure 13 is Fort McMurray annual precipitation long-term trend prediction. The lag time is 11 years and the kernel function of SVM is Gaussian Radial Basis with width 0.73.

Figure 14 is Athabasca annual precipitation long-term trend prediction. The lag time is 17 years and the kernel function of SVM is Gaussian Radial Basis with width 1.0.

Figure 15 is Edson Creek annual maximum temperature long-term trend prediction. The lag time is 10 years and the kernel function of SVM is Gaussian Radial Basis with width 3.0.

Figure 16 is Slave Lake annual maximum temperature long-term trend prediction. The lag time is 12 years and the kernel function of SVM is Gaussian Radial Basis with width 6.0.

Figure 17 is White Court annual precipitation long-term trend prediction. The lag time is 8 years and the kernel function of SVM is Gaussian Radial Basis with width 0.485.

Figure 18 is DPO annual long-term trend prediction. The lag time is 11 years and the kernel function of SVM is Gaussian Radial Basis with width 20.

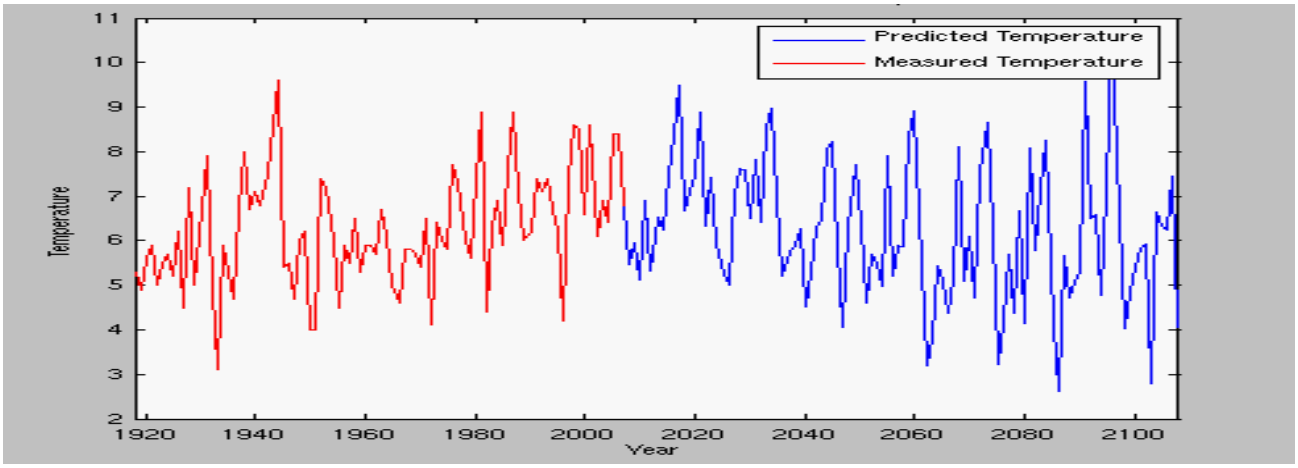

Figure 11: Fort McMurray annual maximum temperature long-term trend prediction. The left part (red line) is the observed records and the right part (blue) is the predicted annual maximum temperature values

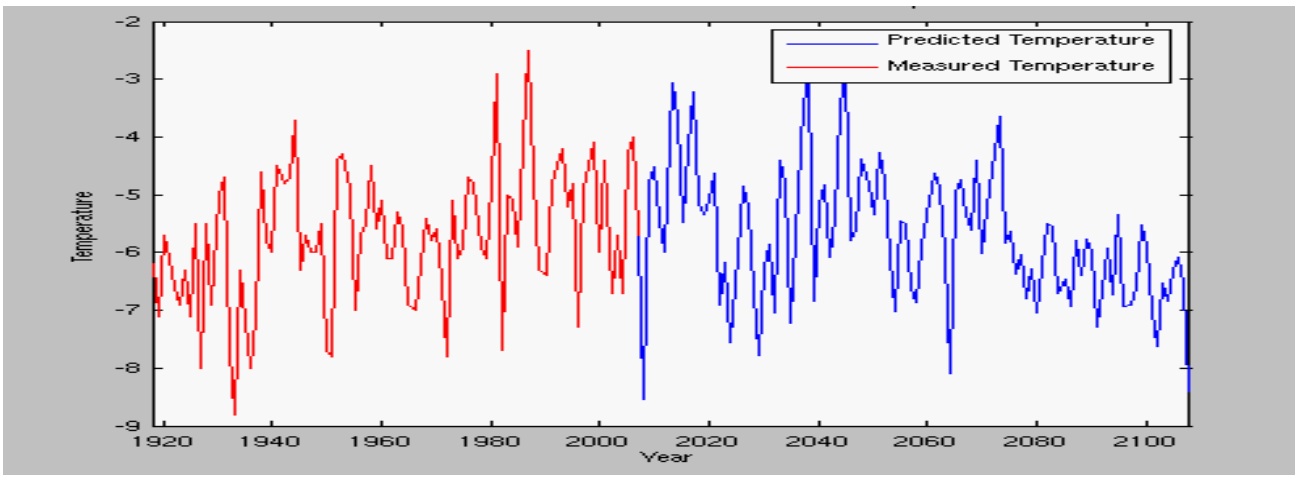

Figure 12: Fort McMurray annual minimum temperature long-term trend prediction. The left part (red line) is the observed records and the right part (blue) is the predicted annual minimum temperature values

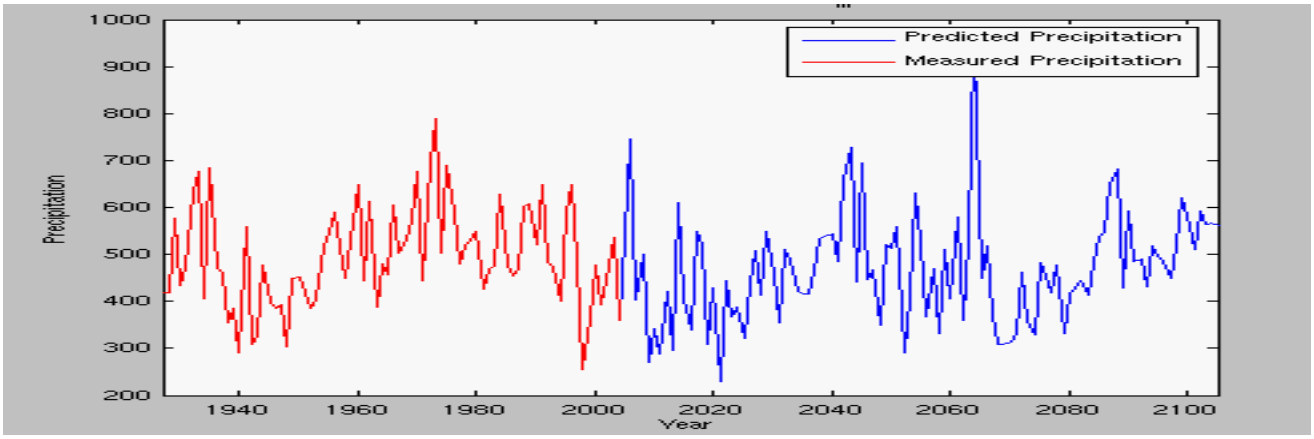

Figure 13: Fort McMurray annual precipitation long-term trend prediction. The left part (red line) is the observed annual precipitation records and the right part (blue) is the predicted annual precipitation values

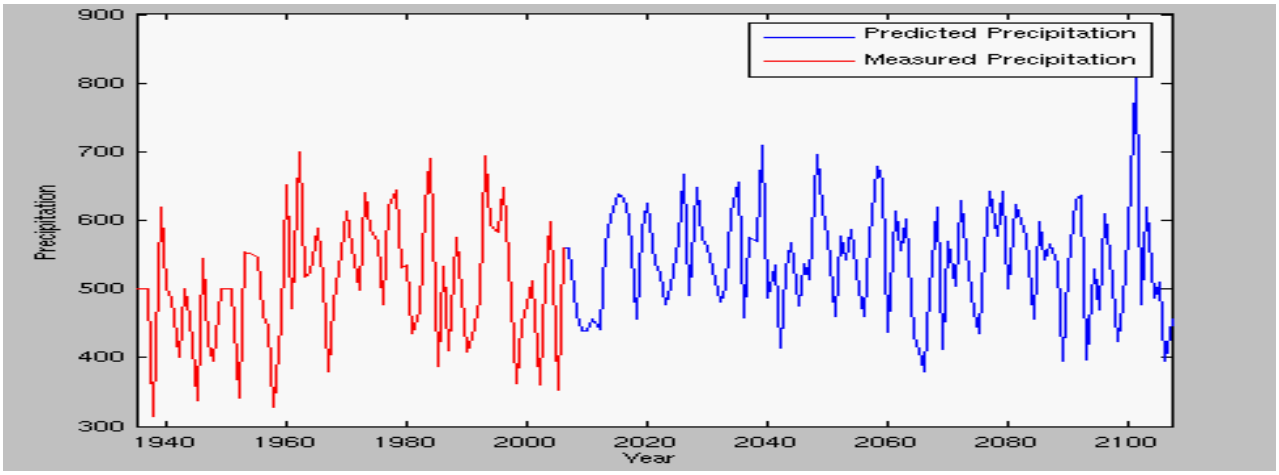

Figure 14: Athabasca annual precipitation long-term trend prediction. The left part (red line) is the observed annual precipitation records and the right part (blue) is the predicted annual precipitation values

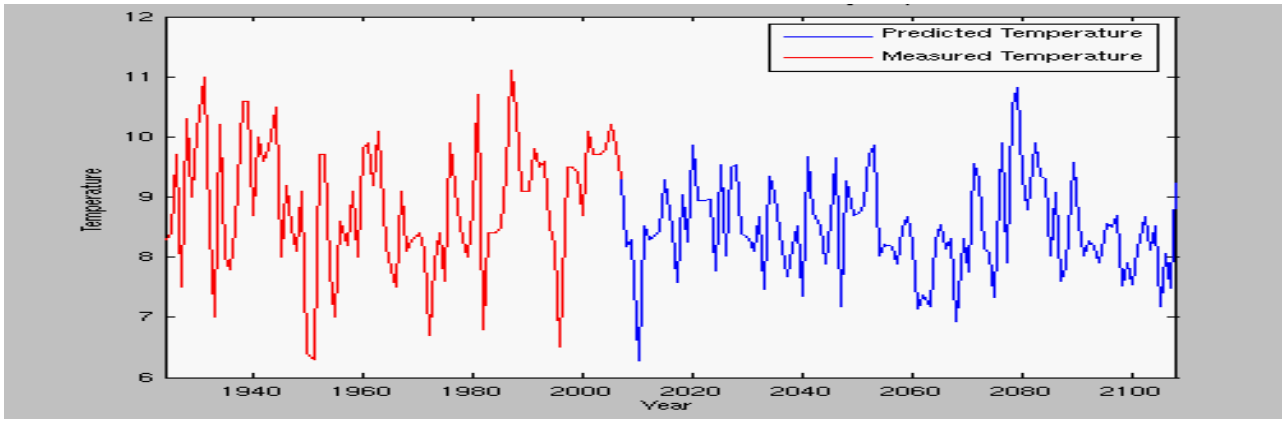

Figure 15: Edson Creek annual maximum temperature long-term trend prediction. The left part (red line) is the observed annual maximum temperature records and the right part (blue) is the predicted annual maximum temperature values

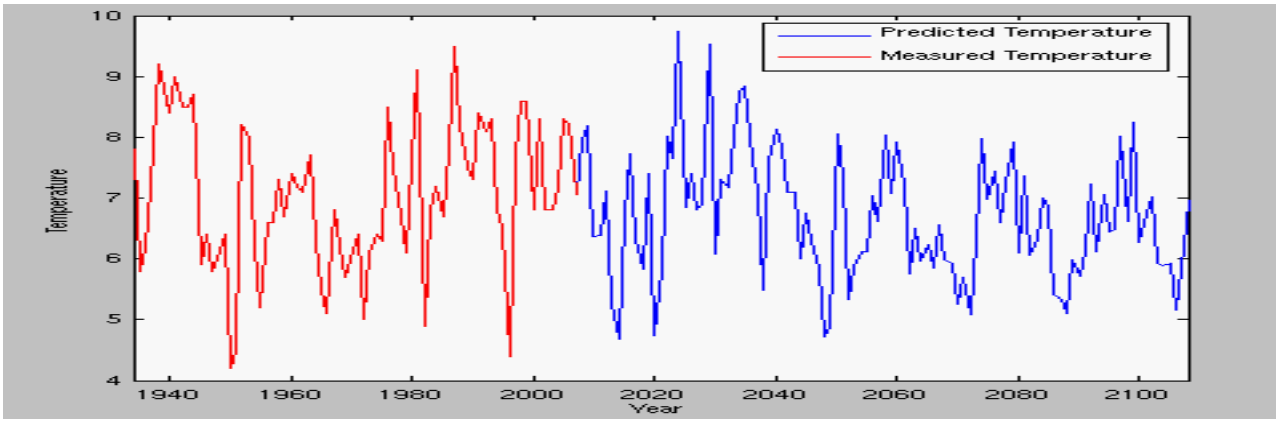

Figure 16: Slave Lake annual maximum temperature long-term trend prediction. The left part (red line) is the observed annual maximum temperature records and the right part (blue) is the predicted annual maximum temperature values

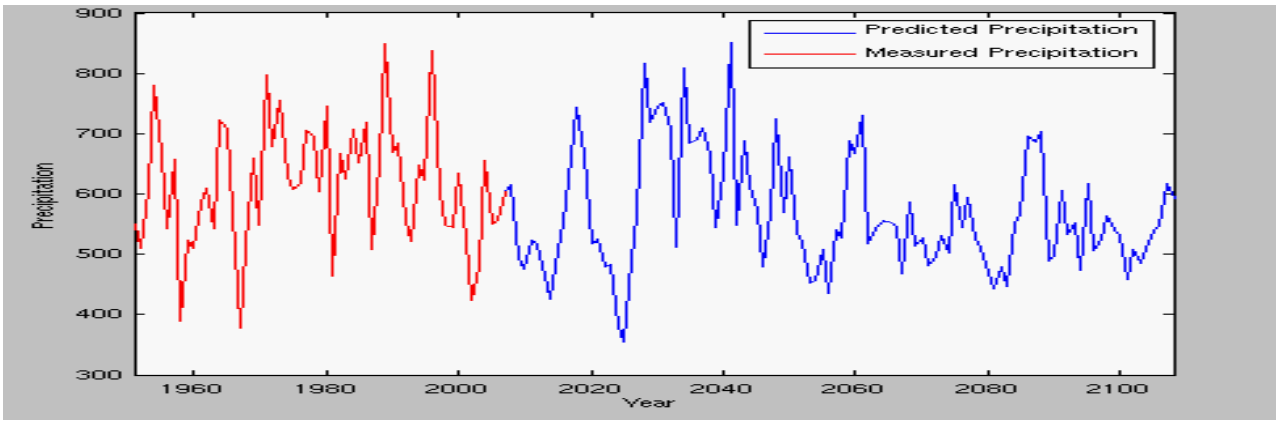

Figure 17: White Court annual precipitation long-term trend prediction. The left part (red line) is the observed annual precipitation records and the right part (blue) is the predicted annual precipitation values

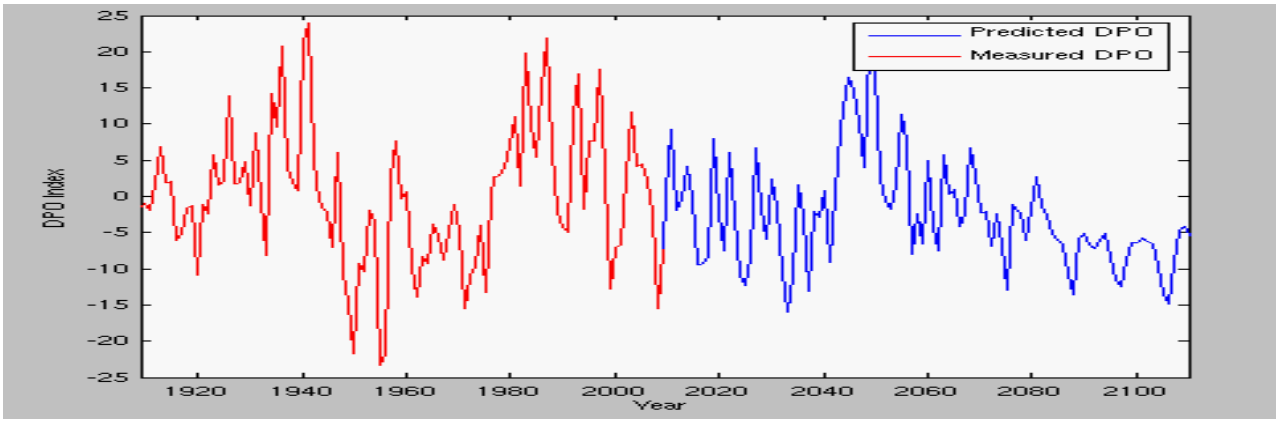

Figure 18: Decadal Pacific Oscillation (DPO) annual long-term trend prediction. The left part (red line) is the observed annual DPO records and the right part (blue) is annual the predicted DPO index values

3) Annual River Flow Rate Predictions

For the prediction of future river flow rates, climate variables, such as the temperature and precipitation are used. For the Athabasca River, temperature and precipitation data were obtained from several stations. These stations located upstream of Fort McMurray, covers an area of about thousands square kilometers. In this area, the temperature, precipitation and flow rate were recorded continuously from 1920s to 2008. The annual maximum, minimum, total and average temperature, precipitation and flow rate can be calculated from these records. All these observed records were shown from Figure 3 to Figure 10.

The support vector machine attempts to establish an empirical (statistical) relationship between annual flow rate and climate variables, such as temperature and precipitation from the available observed records. The training variables include maximum temperature of Fort McMurray station, minimum temperature of Fort McMurray, precipitation of Fort McMurray, precipitation of Athabasca, maximum temperature of Edson Creek, maximum temperature of Slave Lake and precipitation of White court. All these observed will be integrated into a training dataset with the lag time 9 years, which is estimated via a number of tests. The tests of varying the window lag length shows that if the lag time is reasonable, the predicted values will be optimized; otherwise the equation will be ill-posed and cannot find an optimization result.

The support vector machine makes predictions on long-term river flow rate using the recurrent SVM predicted temperature, precipitation and DPO data if available.

Figure 19 is a long-term annual maximum flow rate predicted trend. The input variables are all available observed temperature, precipitation and DPO. The lag time is 9 years. The target input is annual flow rate observed at Fort McMurray station. The SVM kernel function is Gaussian Radial Basis with degree 3.

Figure 20 is a long-term annual minimum flow rate predicted trend. The input variables are all available observed temperature, precipitation and DPO. The lag time is 9 years. The target input is annual flow rate observed at Fort McMurray station. The SVM kernel function is Gaussian Radial Basis with degree 3.

Figure 21 is a long-term annual total flow rate predicted trend. The input variables are all available observed temperature, precipitation and DPO. The lag time is 9 years. The target input is annual flow rate observed at Fort McMurray station. The SVM kernel function is Gaussian Radial Basis with degree 3.

Figure 22 is a long-term annual average flow rate predicted trend. The input variables are all available observed temperature, precipitation and DPO. The lag time is 9 years. The target input is annual flow rate observed at Fort McMurray station. The SVM kernel function is Gaussian Radial Basis with degree 3.

Figure 23 is a long-term annual maximum flow rate predicted trend. The input variables are all available observed temperature and precipitation. The lag time is 9 years. The target input is annual flow rate observed at Fort McMurray station. The SVM kernel function is Gaussian Radial Basis with degree 3.

Figure 24 is a long-term annual minimum flow rate predicted trend. The input variables are all available observed temperature and precipitation. The lag time is 9 years. The target input is annual flow rate observed at Fort McMurray station. The SVM kernel function is Gaussian Radial Basis with degree 3.

Figure 25 is a long-term annual total flow rate predicted trend. The input variables are all available observed temperature and precipitation. The lag time is 9 years. The target input is annual flow rate observed at Fort McMurray station. The SVM kernel function is Gaussian Radial Basis with degree 3.

Figure 26 is a long-term annual average flow rate predicted trend. The input variables are all available observed temperature and precipitation. The lag time is 9 years. The target input is annual flow rate observed at Fort McMurray station. The SVM kernel function is Gaussian Radial Basis with degree 3.

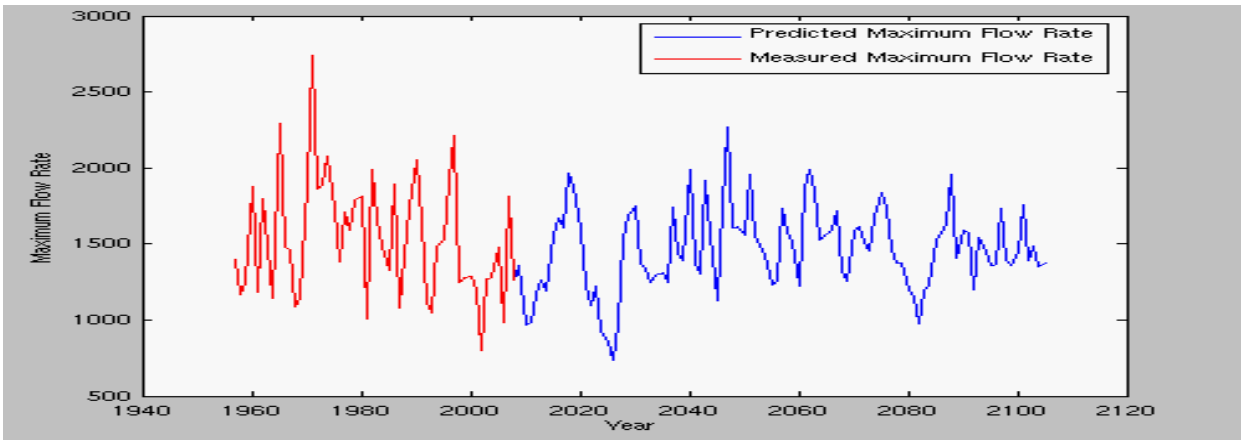

Figure 19: The predicted annual maximum flow rate trend based on temperature, precipitation and DPO. The left part (red line) is the observed annual maximum flow rate records and the right part (blue) is the predicted annual maximum flow rate

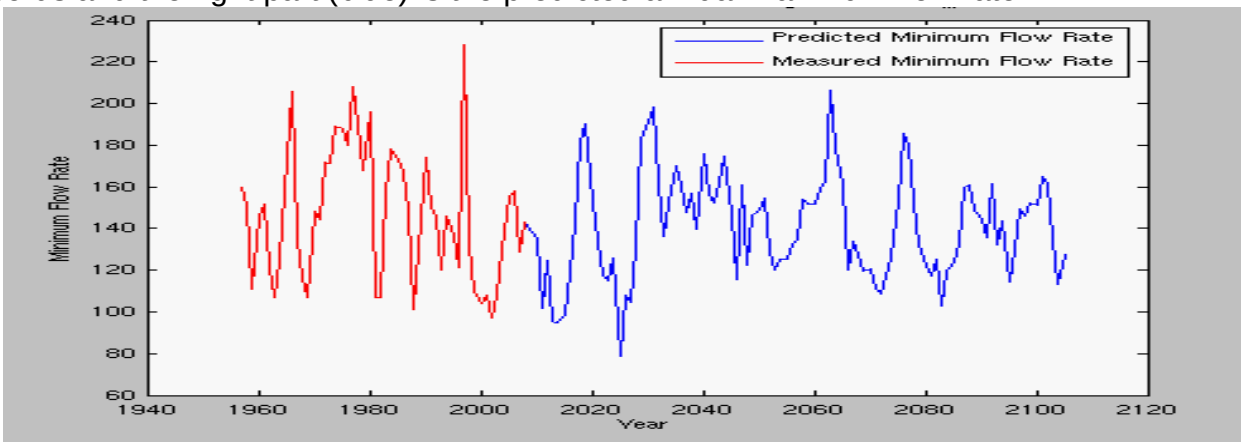

Figure 20: The predicted annual minimum flow rate trend based on temperature, precipitation and DPO. The left part (red line) is the observed annual minimum flow rate records and the right part (blue) is the predicted annual minimum flow rate

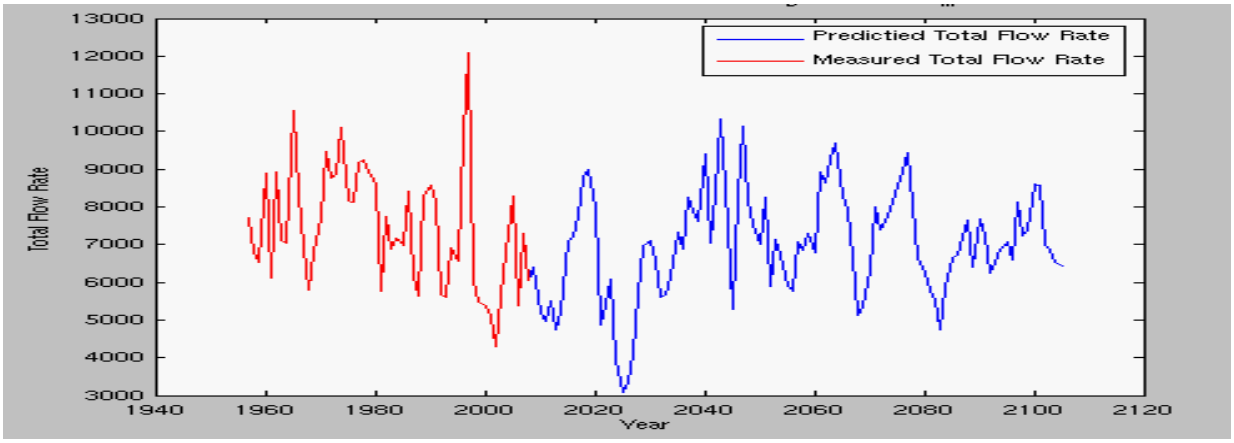

Figure 21: The predicted annual total flow rate trend based on temperature, precipitation and DPO. The left part (red line) is the observed annual total flow rate records and the right part (blue) is the predicted annual total flow rate

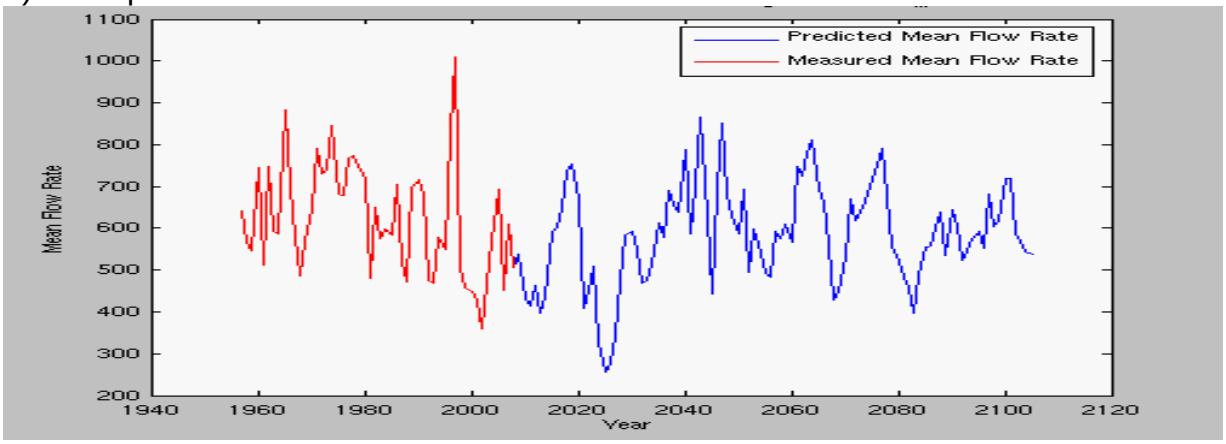

Figure 22: The predicted annual average flow rate trend based on temperature, precipitation and DPO. The left part (red line) is the observed annual average flow rate records and the right part (blue) is the predicted annual average flow rate

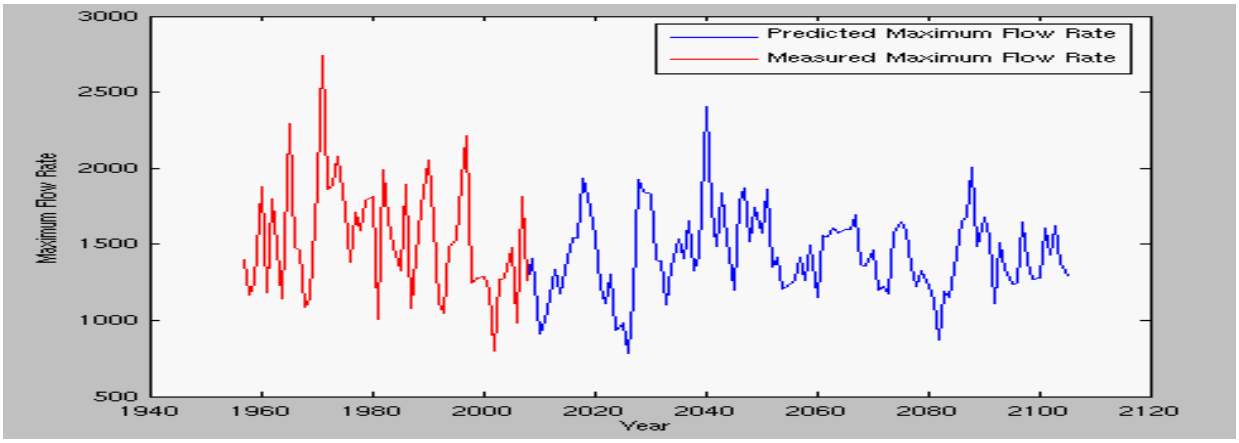

Figure 23: The predicted annual maximum flow rate trend based on temperature, and precipitation. The left part (red line) is the observed annual maximum flow rate records and the right part (blue) is the predicted annual maximum flow rate

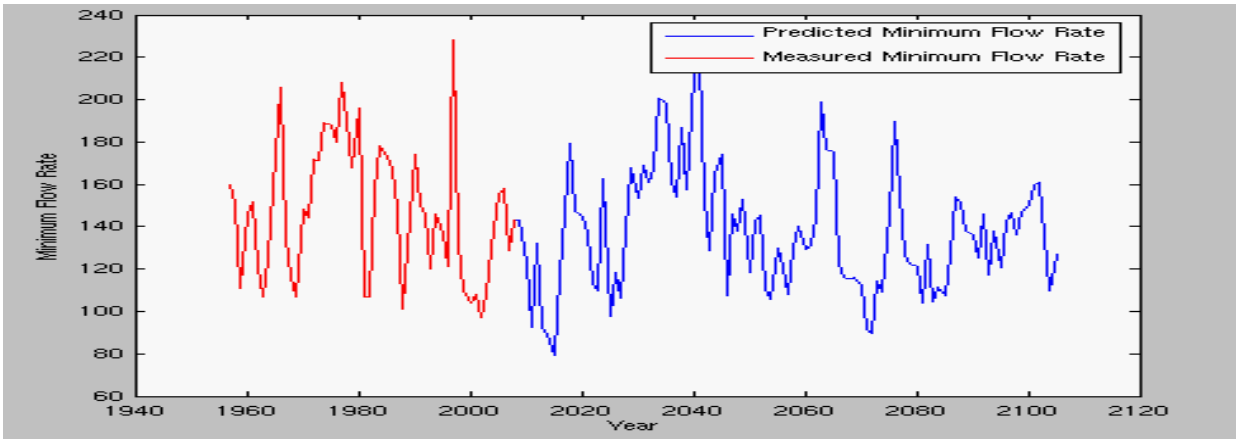

Figure 24: The predicted annual minimum flow rate trend based on temperature, and precipitation. The left part (red line) is the observed annual minimum flow rate records and the right part (blue) is the predicted annual minimum flow rate

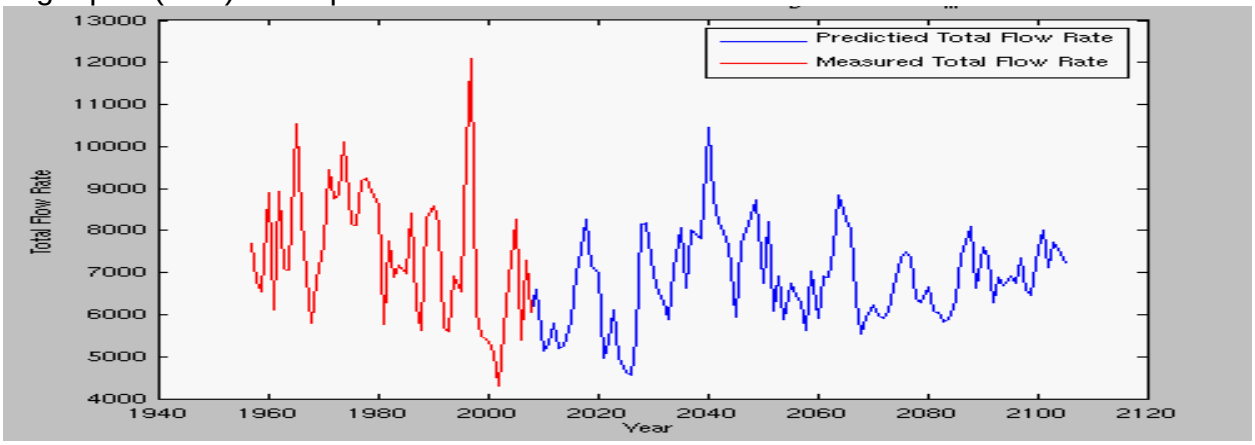

Figure 25: The predicted annual total flow rate trend based on temperature, and precipitation. The left part (red line) is the observed annual total flow rate records and the right part (blue) is the predicted annual total flow rate

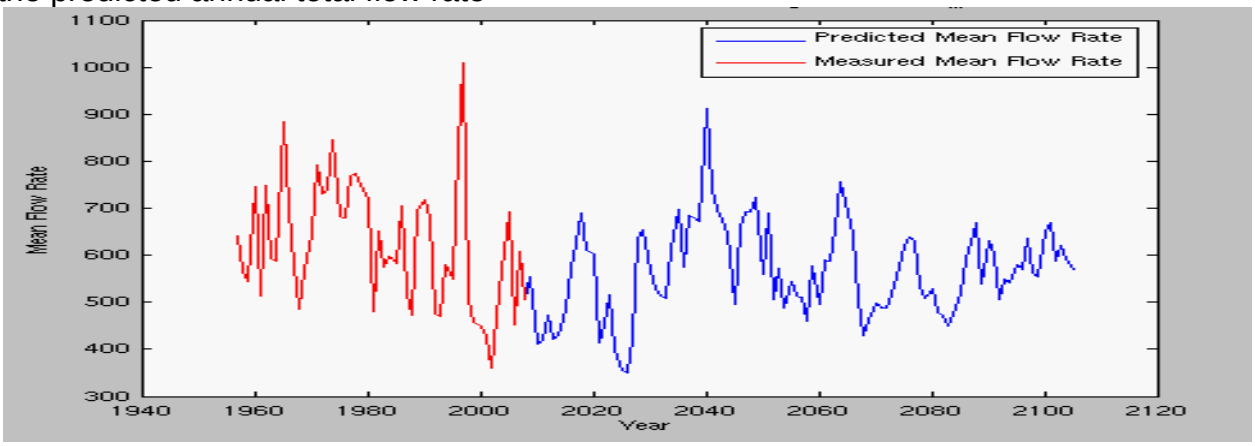

Figure 26: The predicted annual average flow rate trend based on temperature, and precipitation. The left part (red line) is the observed annual average flow rate records and the right part (blue) is the predicted annual average flow rate

Athabasca River Minimum Flow Rate Prediction

1) Climate Data Used for Daily River Flow Rate Prediction

In this work, we use the Bayesian-based support vector machine method to make a long-term prediction for the river flow rate and time trends around Fort McMurray station in the Athabasca River. Generally, the river flow rate is affected by different climate change scenarios, such as temperature and precipitation. However, long-term flow rate, temperature and precipitation are not available for this study and a short period of observed records were collected at Fort McMurray, Clean Water, Edson Creek, Slave Lake and White Court stations from 1959 to 2008. At Fort McMurray station, the highest daily minimum flow recorded during the above period was 211 cms in 1997. The lowest recorded flow was 75 cms in 2001. During a succession of dry years from 1997 to 2003, flows were less than 100 cms for almost four months in winter. The minimum flows have been declined since 1959 (Figure 2). In order to make the predictions of flow rate at Fort McMurray station, the following variables at several stations located upstream of Fort McMurray have been selected:

1) Average of daily maximum temperature, daily minimum temperature and precipitation at Fort McMurray station.
2) Athabasca annual precipitation
3) Edson Creek annual maximum temperature time trend
4) Slave Lake annual maximum temperature time trend
5) White Court annual precipitation time trend
6) DPO time series index trend

Because the above variables are only available from 1959 to 2008, we applied the recurrent support vector machine to predict the future trend of these variables from 2009 to 2100. Figure 27 is the result of recurrent support vector machine. The left parts are the observed dataset and the right parts are the future trends of the prediction of recurrent support vector machine.

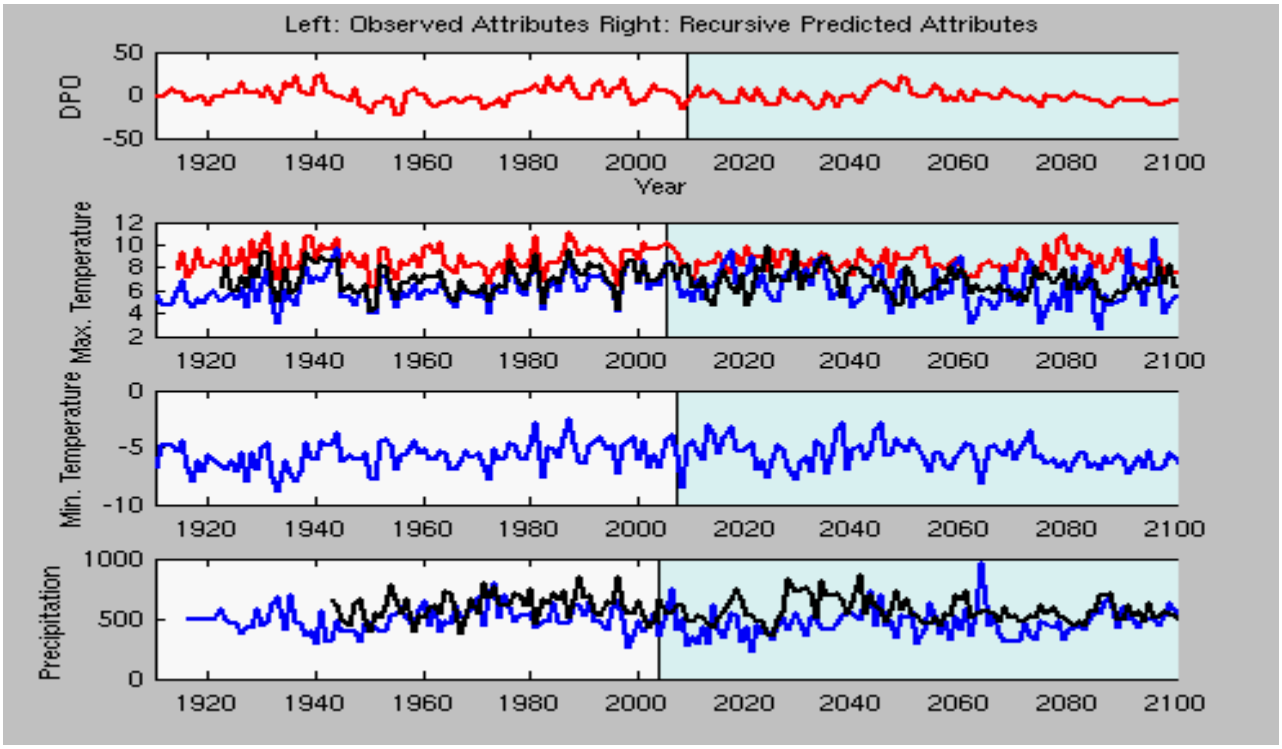

Figure 27: Left: the observed climate variables and Right: the future trends of predictions of recurrent support vector machines.

2) The Future Trends of SVM Daily Prediction of Flow Rate Using Climate Data

For the prediction of future river flow rates, climate variables, such as the temperature and precipitation are used. For the Athabasca River, temperature and precipitation data were obtained from several stations. These stations located upstream of Fort McMurray, covers an area of about thousands square kilometers. In this area, the temperature, precipitation and flow rate were recorded continuously from 1920s to 2008. The annual maximum, minimum, total and average temperature, precipitation and flow rate can be calculated from these records. All these observed records and future trends of the climate variables were shown from Figure 28.

The support vector machine attempts to establish an empirical (statistical) relationship between annual flow rate and climate variables, such as temperature and precipitation from the available observed records. The training variables include maximum temperature of Fort McMurray station, minimum temperature of Fort McMurray, precipitation of Fort McMurray, precipitation of Athabasca, maximum temperature of Edson Creek, maximum temperature of Slave Lake and precipitation of White court. All these observed will be integrated into a training dataset with the lag time 9 years, which is estimated via a number of tests using the different validation years and lag window. The test of varying the window lag length shows that if the lag time is reasonable, the predicted values will be optimized; otherwise the equation will be ill-posed and cannot find an optimization result. After determining the lag window, the different validation years (0-7 years) have been tested to achieve the minimum error between validation and predicted values through updating the parameters of kernel function of support vector machine.

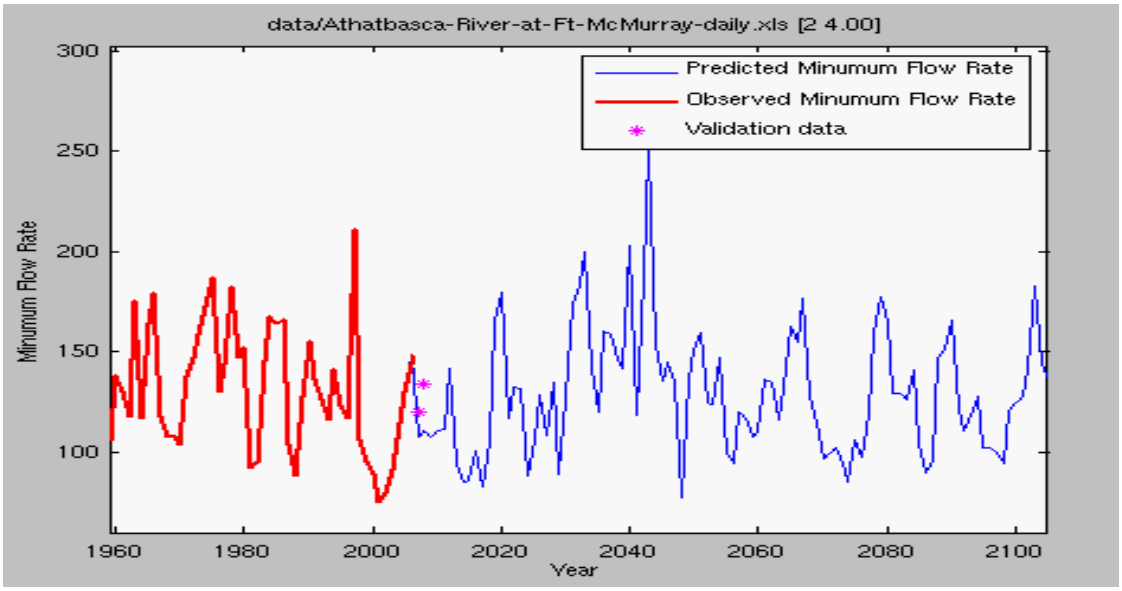

Figure 28: Left : the observed flow rate (red), Right: the predicted minimum flow rate using climate variables (blue). The star points are validation data (magenta)

In order to determine the contributions of climate variables, we make a different combination of climate variables, such as temperature, precipitation and DPO with different validation years. Figure 29 is the future trend of predictions of flow rate with all three temperature, precipitation and DPO variables. The left part is observed flow rate; the right is future trend of flow rate. The star points with green color are the validation dataset. The top is the future trend of the flow rate with zero year (red) and one year validation. The upper middle is the future trend of flow rate with two and three-year validation. The lower middle is the future trend of flow rate with four and five-year validation. The bottom is the future trend of flow rate with six and seven-year validation.

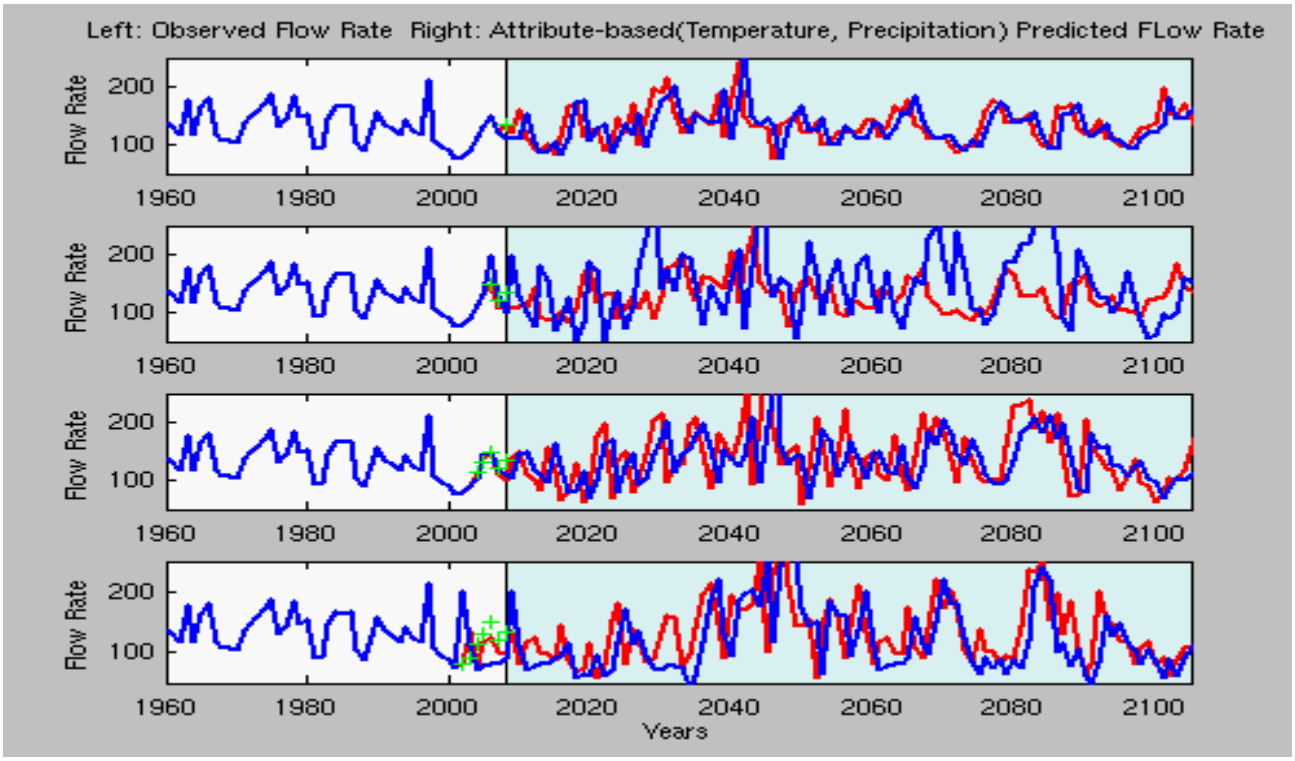

Figure 29: Left: Observed flow rate, Right: future trend of predictions of flow rate based on two climate variables (Temperature and Precipitation) with the different validations. The upper is zero (red) and one year (blue) validation, the bottom is six (red) and seven-year (blue) validation

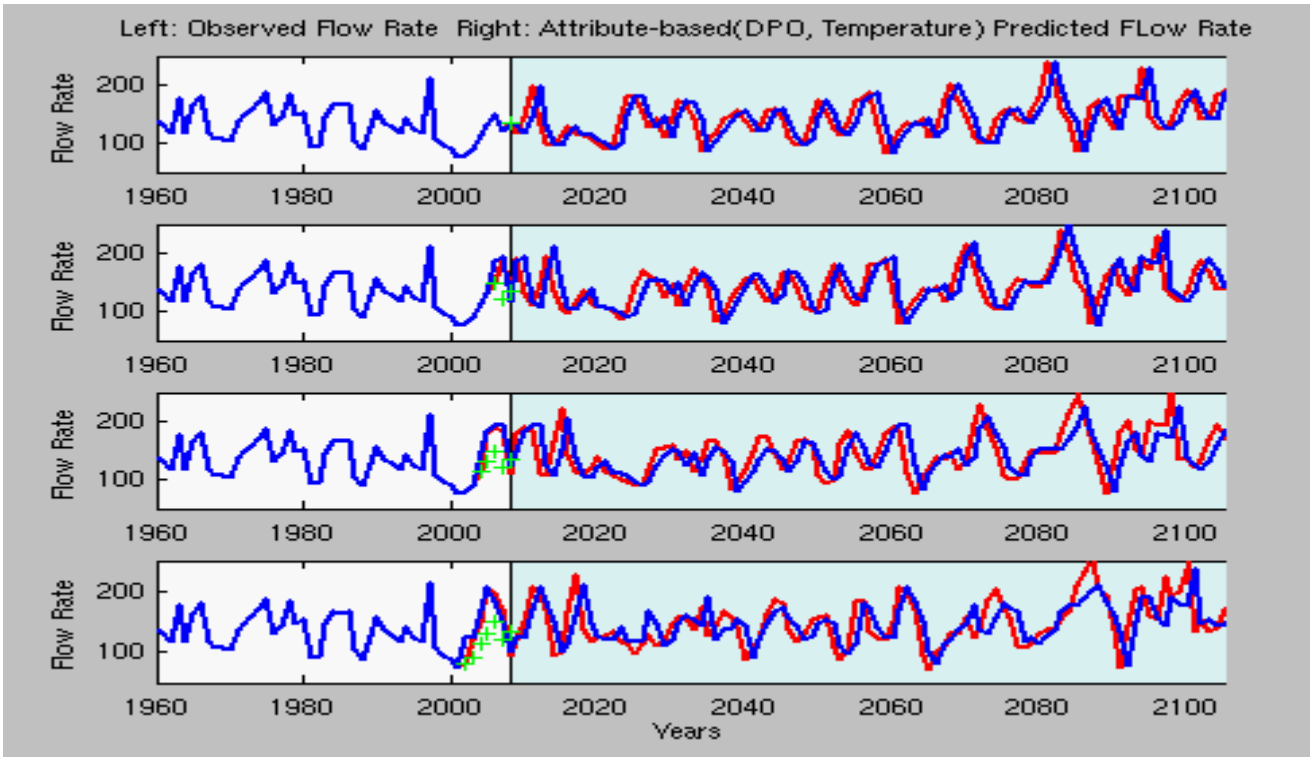

Figure 30: Left: Observed flow rate, Right: future trend of predictions of flow rate based on two climate variables (DPO and Temperature) with the different validations. The upper is zero (red) and one year (blue) validation, the bottom is six (red) and seven-year (blue) validation

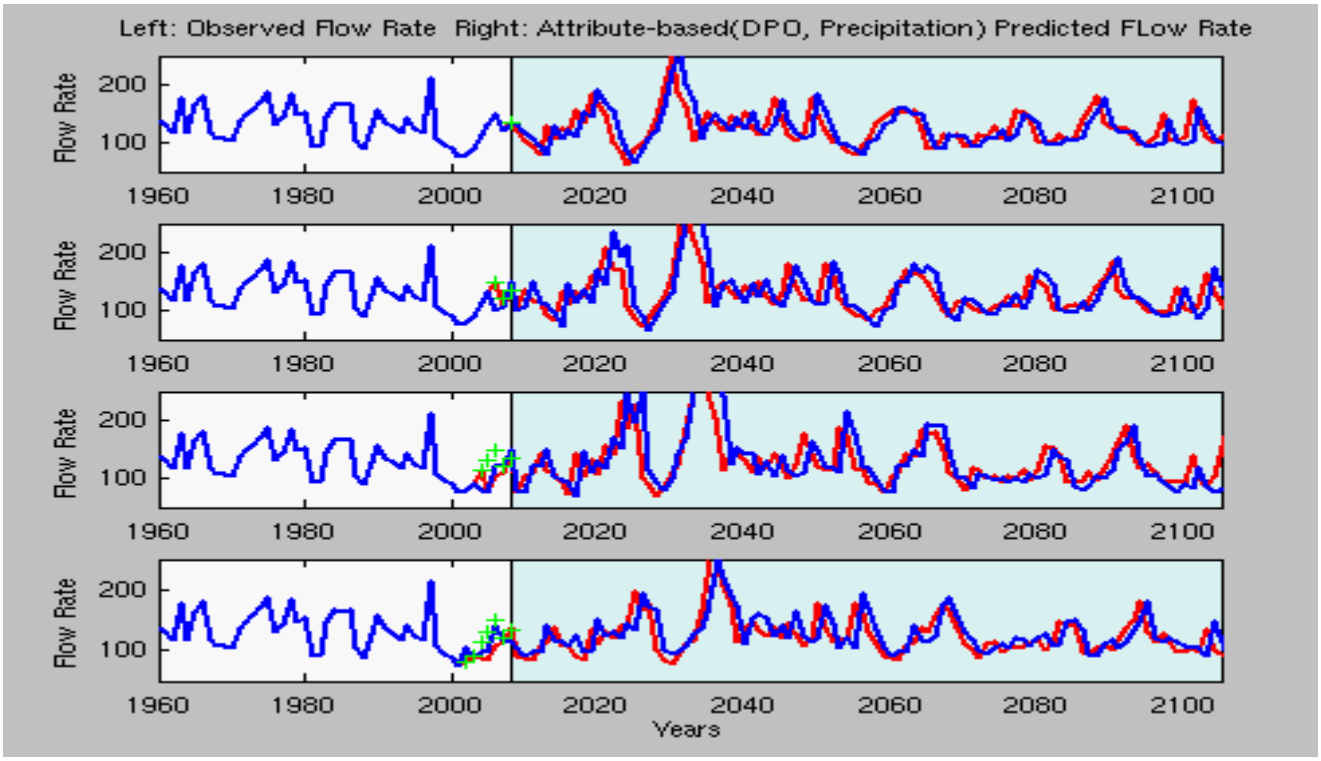

Figure 31: Left: Observed flow rate, Right: future trend of predictions of flow rate based on two climate variables (DPO and Precipitation) with the different validations. The upper is zero (red) and one year (blue) validation, the bottom is six (red) and seven-year (blue) validation

3) The Future Trends of Recurrent SVM Daily Prediction of Flow Rate

The recurrent flow rate SVM models are trained using these observed flow rate data from 1959 to 2008 and the hyperparameter weights $\{\omega_n\ n = 1, \ldots N\}$ were optimised directly via the iterative procedure. The M-steps ahead predictions of the observed flow rate data are predicted iteratively based on the same linear combinations as the training's. In order to best make predictions of future trend, the validation was used to achieve the minimum errors between the validation data and predicted values and then determine reasonable parameters for recurrent support vector machines. The followings are the recurrent SVM long-term prediction results.

Figure 32 is the future trend of recurrent predictions of flow rate. The left part is observed flow rate; the right is future trend of flow rate. The star points with green color are the validation dataset. The top is the future trend of the flow rate with zero year (red) and one year validation. The upper middle is the future trend of flow rate with two and three-year validation. The lower middle is the future trend of flow rate with four and five-year validation. The bottom is the future trend of flow rate with six and seven-year validation.

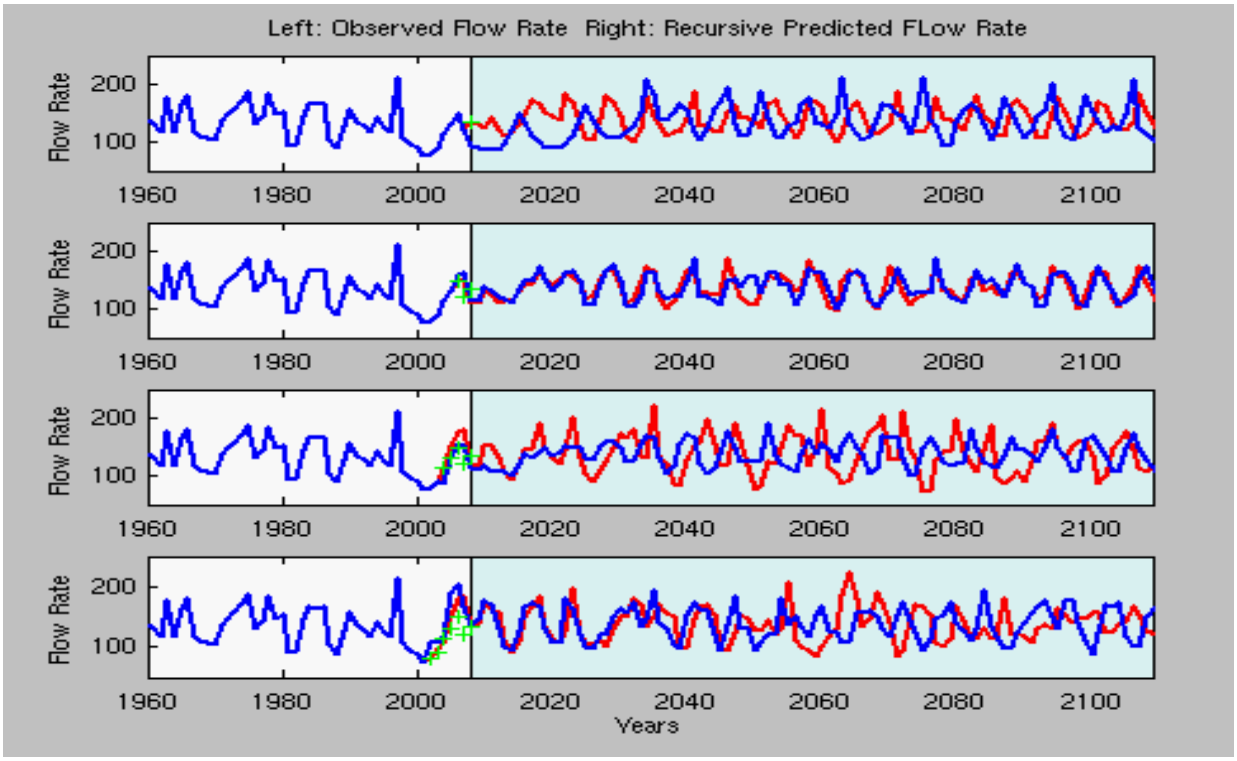

Figure 32: Left: Observed flow rate, Right: future trend of predictions of flow rate with the different validations. The upper is zero (red) and one year (blue) validation, the bottom is six (red) and seven-year (blue) validation

Discussion and Conclusions

This study employs the Bayesian Support Vector Machine to make predictions of future trend of daily river flow rate. In order to have an explicit link between the climate change and river flow rate, climate variables, such as temperature, precipitation and decadal pacific oscillation index are used as input for prediction of future river flow. The different combined climates and different validation tool have been used to better select the lag time and parameters in order to achieve the minimum errors between the validation true values and predicted values.

If the climate variables are not available, the recurrent support vector machine will be applied to make the prediction of future trend of flow rate.

The future river minimum flow rates of the Athabasica River near Fort McMurray stations, of the next 60 years are forecasted used the methods developed in this study. The predicted future trends of the river flow rate depend on the climate variables. If there are more geological experience and knowledge it seems the method can make reasonable prediction of future trend flow rate. The proposed Bayesian Support Vector Machine seems to be a useful tool for predicting future trends of time series.

More accurate prediction from Bayesian Support Vector Machine for the river flow rate depends on its learning kernel functions, kernel function's width (degree) and training

dataset. The validation tool has solved part of issues, but there remain some future problems to be solved. An important problem is how to determine the lag window and design a proper partition mechanism to build the input-target training pairs. In addition, more studies should be done with the selection of the climate variables in order to get the optimized and reasonable results.

References

Mackay, D. J. C., 1992, Bayesian interpolation, Neural Computation, 4(3), 415-447

Schölkopf, B., Burges, C.J.C., and Smola. A. J., 1999, Advances in Kernel Methods: Support Vector Learning, MIT Press.

Tipping M.E., 2001, Sparse Bayesian Learning and the Relevance Vector Machine, Journal of Machine Learning Research 1, 211-244.

Tipping, M. E. (2004). Bayesian inference: An introduction to principles and practice in machine learning. In O. Bousquet, U. von Luxburg, and G. Rätsch (Eds.), *Advanced Lectures on Machine Learning*, pp. 41–62. Springer.

Vladimir N Vapnik. Statistical Learning Theory, Wiley, New York, 1998.

Campolo, M. etc. 1995, Forecasting river flow rate during low-flow periods using neural networks, Water Resources Research, Vol. 35, No.11, P3547-3553, Nov., 1995

Hu, T.S. etc, River flow time series prediction with a range-dependent neural network, Hydrological Science, 46(5) October, 2001

Ozg¨ur KISI, Daily River Flow Forecasting Using Artificial Neural Networks and Auto-Regressive Models, Turkish J. Eng. Env. Sci. 29 (2005) , 9 - 20.

Schindler, D.W. Donahue W.F. and Thompson, John P., Future Water Flows and Human Withdrawals in the Athabasca River*, University of Alberta, May, 2007*

LI, Boyang, HU, Jinglu and HIRASAWA, Kotaro, Financial Time Series Prediction Using a Support Vector Regression Network, *2008 International Joint Conference on Neural Networks (IJCNN 2008)*

Iffat A. Gheyas, Leslie S. Smith, A Neural Network Approach to Time Series Forecasting, Proceedings of the World Congress on Engineering 2009 Vol. II WCE 2009, July 1 - 3, 2009, London, U.K.

Amaury Lendasse, etc, Fast Bootstrap applied to LS-SVM for Long Term Prediction of Time Series, IJCNN'2004 proceedings – International Joint Conference on Neural Networks Budapest (Hungary), 25-29 July 2004, IEEE, pp. 705-710

Liu, Y. and Sacchi, M. D., Propagation of borehole derived properties via a Support Vector Machine (SVM), CSEG Recorder, Canada, December 2003, 54-58

Radke, D., Investing in our future: Responding to the Rapid Growth of Oil Sands Development, Dec. 29, 2006, P.112-113