INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **30**: 1256–1268 (2010)
Published online 12 May 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1954

RMetS
Royal Meteorological Society

A modified support vector machine based prediction model on streamflow at the Shihmen Reservoir, Taiwan

Pei-Hao Li,[a,b]* Hyun-Han Kwon,[b] Liqiang Sun,[c] Upmanu Lall[b,c] and Jehng-Jung Kao[a]

[a] *Institute of Environmental Engineering, National Chiao Tung University, Hsinchu, 30090, Taiwan, ROC*
[b] *Department of Earth and Environmental Engineering, Columbia University, NY 10027, USA*
[c] *International Research Institute for Climate and Society, Columbia University, NY 10964, USA*

ABSTRACT: The uncertainty of the availability of water resources during the boreal winter has led to significant economic losses in recent years in Taiwan. A modified support vector machine (SVM) based prediction framework is thus proposed to improve the predictability of the inflow to Shihmen reservoir in December and January, using climate data from the prior period. Highly correlated climate precursors are first identified and adopted to predict water availability in North Taiwan. A genetic algorithm based parameter determination procedure is implemented to the SVM parameters to learn the non-linear pattern underlying climate systems more flexibly. Bagging is then applied to construct various SVM models to reduce the variance in the prediction by the median of forecasts from the constructed models. The enhanced prediction ability of the proposed modified SVM-based model with respect to a bagged multiple linear regression (MLR), simple SVM, and simple MLR model is also demonstrated. The results show that the proposed modified SVM-based model outperforms the prediction ability of the other models in all of the adopted evaluation scores. Copyright © 2009 Royal Meteorological Society

KEY WORDS forecast; streamflow; climate; support vector machine; bagging

Received 21 January 2008; Revised 13 March 2009; Accepted 18 April 2009

1. Introduction

In Taiwan, water resource availability is greatly influenced by the variation of East Asia Monsoon activities which contribute to well marked wet and dry seasons (Chang, 2004). The unanticipated fluctuations of water availability in the spring growing season have led the water authority to expend a lot of its budget towards compensation of the losses caused by its failure to deliver water allocated for irrigation (Shu, 2003). Forecasts of streamflow in the winter season are desirable to mitigate such possible negative impacts, through better allocation of water for domestic and irrigation use.

In recent years, the unusual behaviour of East Asian Monsoon has led to occurrence of extensive drought/flood disasters in East Asia. As mentioned in Chang (2004), these extreme events include the 2-month long persistent excessively heavy rainfall/floods over the Yangtze-Huaihe River Basins during the 1991 Meiyu season (Ding, 1993) and the prolonged unprecedented heavy rainfall/floods over the Yangtze River Basins during the 1998 Meiyu season. Both events caused the loss of numerous human lives and billions of Chinese Yuan. During the spring and the summer seasons of 2002, Northern Taiwan experienced the most severe drought in several decades. Faced with two nearly empty major water-supply reservoirs, emergency water restrictions, such as suspending nonessential water uses and cutting irrigation supplies, were implemented to save water for domestic consumption (Shiau and Lee, 2005).

In North Taiwan, which is the social and technical centre, the mean annual rainfall is 2,934mm, but distributed unevenly, with 62% coming during the wet season, May through October, and 38% during the dry season, November through April. Therefore, reservoirs are widely installed in this area to retain excess water for dry seasons. However, they do not have the capacity to deal with inter-annual drought. Public water supply takes priority over irrigation and water restrictions are imposed to mitigate the impact of drought (Huang and Chou, 2005). Critical decisions for a water release strategy need to be made before allocating water for the irrigation scheduled for two seasons in this area. The first season starts in late January (spring growing season) and the second in mid-July (summer growing season). A lack of streamflow predictors, however, often leads to the termination of irrigation in the middle of the spring growing season. About twenty million dollars was paid to farmers for crop yield losses due to the termination of irrigation during the severe drought of 2002 (Shu, 2003).

Climate-based forecasts of streamflow have been shown to improve the reliability of water supply (Kim and Palmer, 1997; Hamlet *et al*., 2002; Westphal *et al*., 2003). There are many studies that have used climate

* Correspondence to: Pei-Hao Li, Institute of Environmental Engineering, National Chiao Tung University, Hsinchu, 30090, Taiwan, ROC.
E-mail: peihaoli@gmail.com

signals such as El Niño-Southern Oscillation (ENSO) as reasonable predictors to forecast seasonal streamflow and rainfall (Chiew *et al*. 1998; Liu *et al*., 1998; Hamlet and Lettenmaier, 1999; Piechota *et al*., 1999; Fowler and Kilsby, 2002; Harshburger *et al*., 2002; Eldaw *et al*., 2003; Souza Filho and Lall, 2003; Karamouz and Zahraie, 2004; Xu *et al*., 2007). However, for the complex weather systems in East Asia, predictability of the streamflow has received limited attention and no previous studies have been focused on Taiwan where catchments area is relatively small and hence predictability is not expected to be high at seasonal to inter-annual time scales.

The streamflow in Taiwan has been shown to be mainly because of localized rainfall over Taiwan (Yu *et al*., 2006) with limited contributions from groundwater. There are many studies on the climatic mechanisms responsible for the rainfall in East Asia during the winter monsoon season (Yang *et al*., 2002; Wu *et al*., 2003; Chang, 2004), but studies focused on the precipitation in Taiwan during boreal winter, which is crucial for the available water resource during the driest season, are few. Chen *et al*. (1983) investigated the relationship between the diurnal cycle of local circulation and precipitation of North Taiwan. Chen and Chen (2003) addressed the general characteristics of the rainfall and the evolution of the mean circulation patterns for all seasons. However, these studies only considered local or regional circulation around Taiwan, and predictability using global or regional climate indicators was not assessed. In this study, we directly focus on the prediction of the December–January flow into the Shihmen reservoir in North Taiwan as illustrated in Figure 1, recognizing that the streamflow is a measure of the spatially averaged rainfall over the catchment during the two months, given the predominantly overland flow response to rainfall and the runoff travel time, which is about 1–4 h (Chang and Chang, 2006) and is substantially less than the 2 month window.

Although climate information is potentially valuable in improving hydrologic prediction in support of water resources management, there are still some challenges in developing such predictions:

- Deterministic climate-hydrologic model-based forecasts seem to have rather limited success in many instances.
- Given the acceptance of the idea that the climate is changing, there is concern that the statistical models built for forecasting streamflow using climate indicators are unlikely to work as conditions change.
- Hydrologic records at places of interest are usually short and building a reliable statistical model, including predictor identification, poses a challenge.
- The relationships between streamflow or local precipitation and climate indicators of the sort identified in this study are often expected to be nonlinear – both from the exploratory data analysis and from an examination of the governing equations of atmospheric dynamics as they apply to convection, moisture advection, and precipitation dynamics.

Collectively, these constitute a formidable challenge, particularly if assessing the uncertainty of forecasts is also a goal. A more robust prediction framework should thus be proposed to mitigate the difficulties.

Support vector machines (SVMs) are an advanced machine learning technique based on structural risk minimization (SRM) which minimizes expected error of a learning model and reduces the problem of overfitting (Yu *et al*., 2006). SVM was developed in the early 1990s (Boser *et al*., 1992; Vapnik, 1998) and has been successfully applied in many hydrologic studies, such as runoff prediction problems (Dibike *et al*., 2001; Asefa *et al*., 2006; Yu and Liong, 2007), flood forecasting problems (Liong and Sivapragasam, 2002; Yu *et al*., 2006), groundwater monitoring network design (Asefa *et al*., 2004), and lake water level prediction (Asefa *et al*., 2005; Khan and Coulibaly, 2006). However, SVM has not hitherto been applied in the climate based streamflow prediction problem. Ensemble modelling approaches, in addition, are widely adopted techniques in climatic modelling related studies to reduce the variance of predictive uncertainty (Colman and Davey, 2003; Raftery *et al*., 2005; Chowdhury and Sharma, 2009). In view of the relative short hydrologic records available in this study, we therefore enhance the usual application of SVM with bagging (Breiman, 1996), which is a technique for reducing the variability in statistical models through the averaging of candidate models formulated with the same data set, to improve the predictive performance of the constructed models. We also adopted a genetic algorithm for SVM parameterization.

In the following sections, proper predictors for the streamflow forecast and the possible mechanisms are first selected and investigated. A SVM based prediction model is proposed to learn and predict the streamflow at Shihmen reservoir with the selected predictors. Genetic algorithm (GA) based parameter determination and bagging technique are also introduced to improve the performance

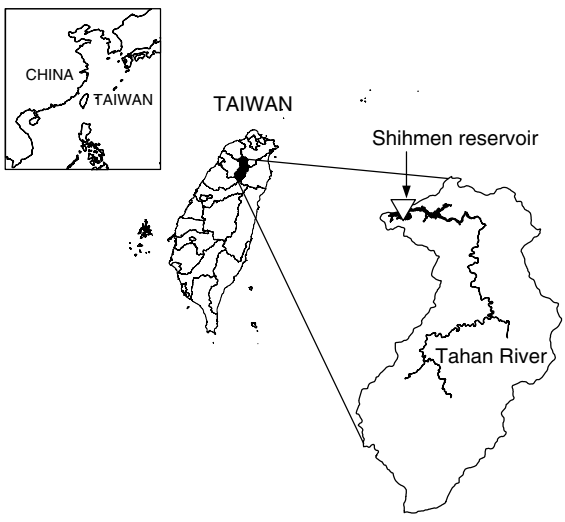

Figure 1. The Shihmen reservoir basin and Tahan River in North Taiwan.

of the constructed SVM model. Consequently, the predicted results of SVM model are compared with those forecasted by bagged multiple linear regression (MLR), simple SVM, and simple MLR based prediction model to present the superiority of the proposed model and to explore the predictability of streamflow in this area.

2. Study area and data

The Shihmen reservoir built in 1964 is located on the upstream reaches of the Tahan River (Figure 1) and is one of the largest water reservoirs in Taiwan. The Tahan River is 126 km long with a drainage area of 1163 km^2 and an average slope of 1/37. The Shihmen reservoir has effective capacity of 234 million cubic meters (MCM) and its catchment area is 754 km^2. The total demand supplied by Shihmen reservoir is 1,163.59 MCM/yr in 2001. About 529.09 MCM/yr of the supply is used for irrigation over the Taoyuan area and public water supply takes 634.50 MCM/yr for the people living in nearby Taipei city (Huang *et al.*, 2002). The streamflow data were recorded at the Shihmen reservoir gauge station from 1964. As illustrated in Figure 2, the annual cycle of monthly streamflow presents obvious wet and dry seasons in a year. The variation of monthly streamflow also varies in different seasons. As shown in Figure 3, the time series of 2-month averaged streamflow at Shihmen reservoir in December and January (Dec–Jan) was used in this study (Huang and Chou, 2005). Larger fluctuations are observed in recent years. The predictors considered for the December–January flow were past streamflow, from October and November (Oct–Nov) and a suite of climate predictors. Monthly climate data including global sea surface temperature (SST), sea level pressure (SLP), and outgoing longwave radiation (OLR) were extracted from NCDC extended reconstructed analysis, the NCEP reanalysis, and the interpolated OLR dataset, respectively. SST and SLP datasets were collected from KNMI web site (http://climexp.knmi.nl/) and the OLR dataset from the data archive of NOAA/ESRL Physical Sciences Division (http://www.cdc.noaa.gov/PublicData/). The SST, SLP and OLR datasets are available from 1880, 1948, and 1974, respectively. Only the common data from 1974 to 2005 were used in this study.

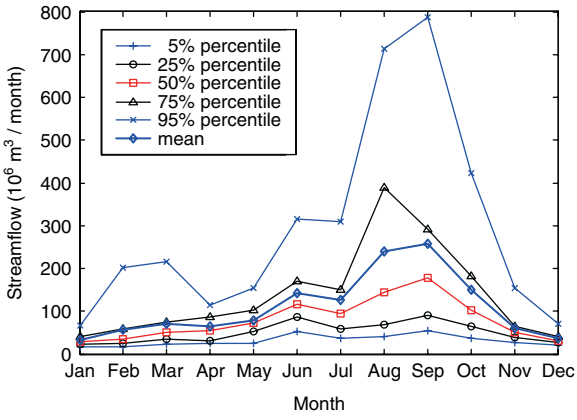

Figure 2. Annual cycle of percentiles and mean of monthly streamflow at Shihmen reservoir in North Taiwan. This figure is available in colour online at www.interscience.wiley.com/ijoc

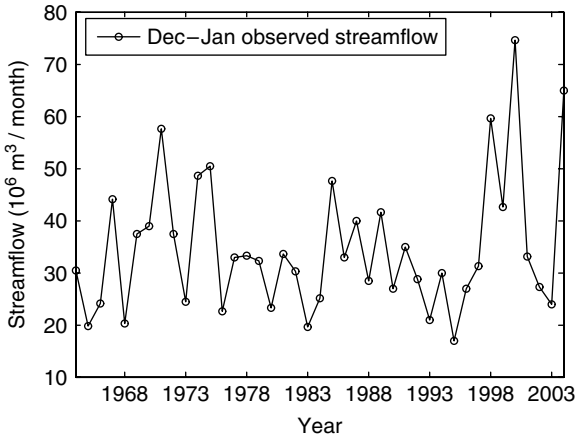

Figure 3. The 1964–2004 time series of 10-day averaged streamflow in Dec–Jan at Shihmen reservoir in North Taiwan.

3. Methodology

3.1. Climatic predictors

First, potential predictors for the Dec–Jan inflow into the Shihmen reservoir are screened by using linear correlation as a criterion. Only a limited number of potential predictors, such as some regions with a statistically significant correlation with Dec–Jan streamflow, have been found for the SST, SLP, and OLR fields, as illustrated in Figure 4. The averages of the respective fields over these areas, as indicated by the boxes in the figure and also identified in Table I, are then selected as climatic predictors. These predictors are selected because they are persistently correlated to the Dec–Jan streamflow and the mechanism contributing to the streamflow can be explained in the manner of global circulation reasonably. Although the selected SST predictor is from

Table I. Correlation coefficient between Dec–Jan streamflow and the selected climatic predictors.

Climatic predictors	Zone selected	Lag time (month)	Correlation coefficient with Dec–Jan streamflow
Oct–Nov streamflow	–	2	0.648
SST	25°N–35°N; 140°E – 160°E	6	0.516
SLP	15°N–30°N; 115°E – 125°E	2	−0.460
OLR	−10°N to −25°N; 135°E–160°E	2	−0.610

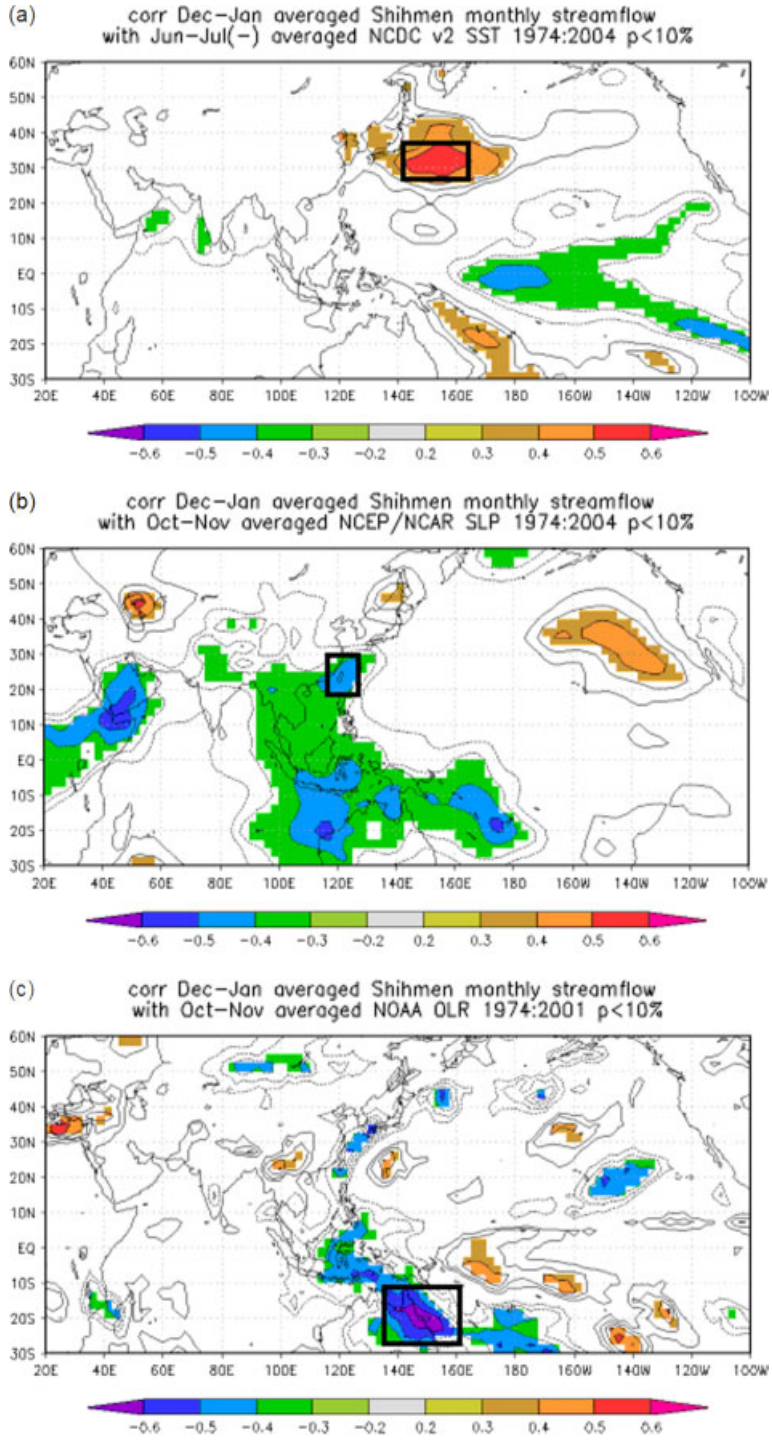

Figure 4. Correlation analysis between Dec–Jan streamflow and (a) Jun–Jul SST, (b) Oct–Nov SLP, and (c) Oct–Nov OLR. Climatic predictors at given locations are indicated by boxes. This figure is available in colour online at www.interscience.wiley.com/ijoc

6 months prior to Dec–Jan, the SST around the same area persistently retains a high correlation with the streamflow in Dec–Jan at Shihmen reservoir throughout the following six months. As for the case of numerous potential predictors, the approaches for partial dependence, such as the partial mutual information criterion (Sharma, 2000), are recommended to select proper predictors that offer new information.

During the East Asian winter monsoon (EAWM) season, the East Asian region is dominated by the Siberian high and the Aleutian low over the Eurasian continents and the Northern Pacific at middle and high latitudes. Northeasterly winds bring cold and dry air across Taiwan from the continent (Xue *et al.*, 2005; Wang, 2006). As shown in Figure 4(a), the region of the selected SST predictor near Japan is positively related with the streamflow at Shihmen reservoir. The higher SST over this region will contribute higher moisture near Japan area and together with the northerly monsoon winds, the moisture would be gradually drawn southward towards Taiwan

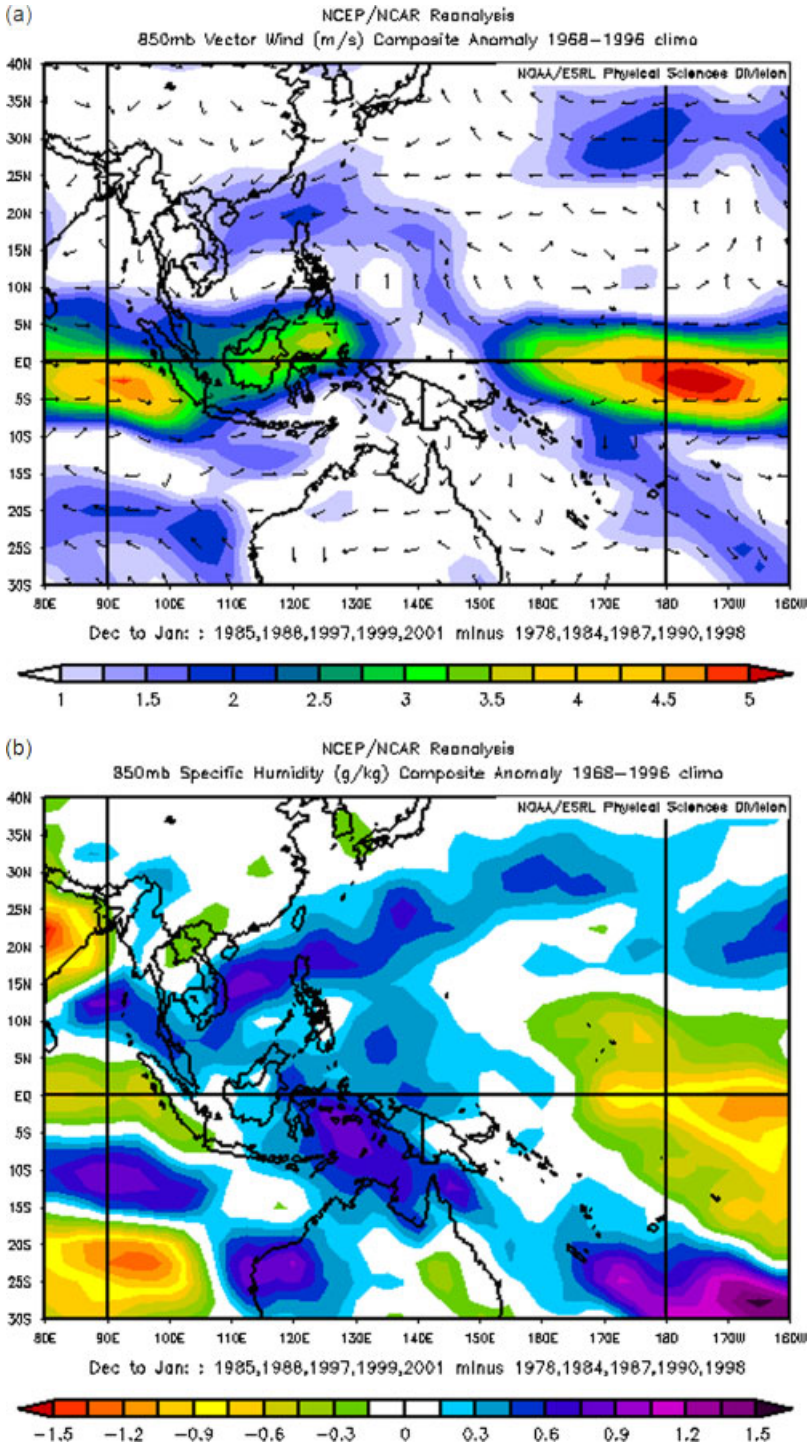

Figure 5. The differences between (a) anomalous wind vector and (b) anomalous specific humidity at 850 mb pressure level in 5 years with lowest SLP and those in five years with highest SLP over selected area near Taiwan in Dec–Jan. The area with larger speed difference of wind flows and difference of specific humidity are shaded in different colors in (a) with the direction indicated by arrows and in (b) respectively. This figure is available in colour online at www.interscience.wiley.com/ijoc

in Dec–Jan and could contribute to rainfall. On the other hand, as shown in Figure 4(b), the streamflow has a negative correlation with the sea level pressure near Taiwan. Clearly, a local high pressure would suppress convections and block fronts or migrating systems that may bring rainfall. The difference between the anomaly of wind vector in five years with lowest SLP and that in 5 years with the highest SLP is composited as illustrated in Figure 5(a). Easterly anomaly winds converged over Taiwan region is observed (Figure 5(a)) for the lower SLP over Taiwan area. This strengthens/increases the convection activities and lead to more precipitation over Taiwan region. Increase of water vapour at lower atmosphere (Figure 5(b)) also contributes to the increase of precipitation over Taiwan region. The difference of anomalous specific humidity at 850 mb pressure level between years

with lowest SLP and highest SLP over the Taiwan area is drawn as Figure 5(b) to show the moisture brought in.

OLR is a measure of cloudiness and strong convection. In this study, OLR over the ocean is used to identify areas where SSTs are sufficiently high in a sustained manner such that convection is sustained. Therefore, OLR in an ocean region that is a potential moisture source, or by virtue of its temperature influences circulation patterns, is a useful measure of SST levels exceeding a physically important threshold. According to the correlation analysis between the streamflow at Shihmen reservoir and OLR, a strong relationship exists between the convection in the Australian monsoon rainfall region and the streamflow at Shihmen reservoir as illustrated in Figure 4(c). The stronger the convection over Australia the higher the streamflow at Shihmen reservoir would be. To explore the possible mechanism contributing to the connection between the selected predictor and the Shihmen streamflow, anomalous wind flows and anomalous OLR in the 5 years with highest OLR in Australian monsoon area in Oct–Nov are composited as shown in Figure 6(a) and (b), respectively. For comparison, the same composite plots for anomalous wind flows and anomalous OLR in the five years with lowest OLR are plotted as in Figure 6(c) and (d), respectively. In the wettest years, strong easterly anomalous winds from the Pacific Ocean and strong westerly anomalous winds from Indian Ocean prevailed in Oct–Nov. These strong anomalous winds constructed a strong eastern Walker cell (EWC) and western Walker cell (WWC) to the Maritime Continent region and converged over that region resulting in strong convection and a negative OLR anomaly (Meehl and Arblaster, 2002). The equatorial Rossby wave propagates westwards along the equator accompanied by large variations in the rotational wind (Wheeler *et al.*, 2000) and the symmetric circulation cells of anomalous winds are therefore observed on either side of the equator (Allan, 1983; Wang, 2006), as also shown in Figure 6(a). The observed anticyclone over the Northern Hemisphere then advects the OLR anomaly northwards across the Taiwan area to the southern area of Japan in the wettest years, as shown in Figure 6(b). The moisture brought

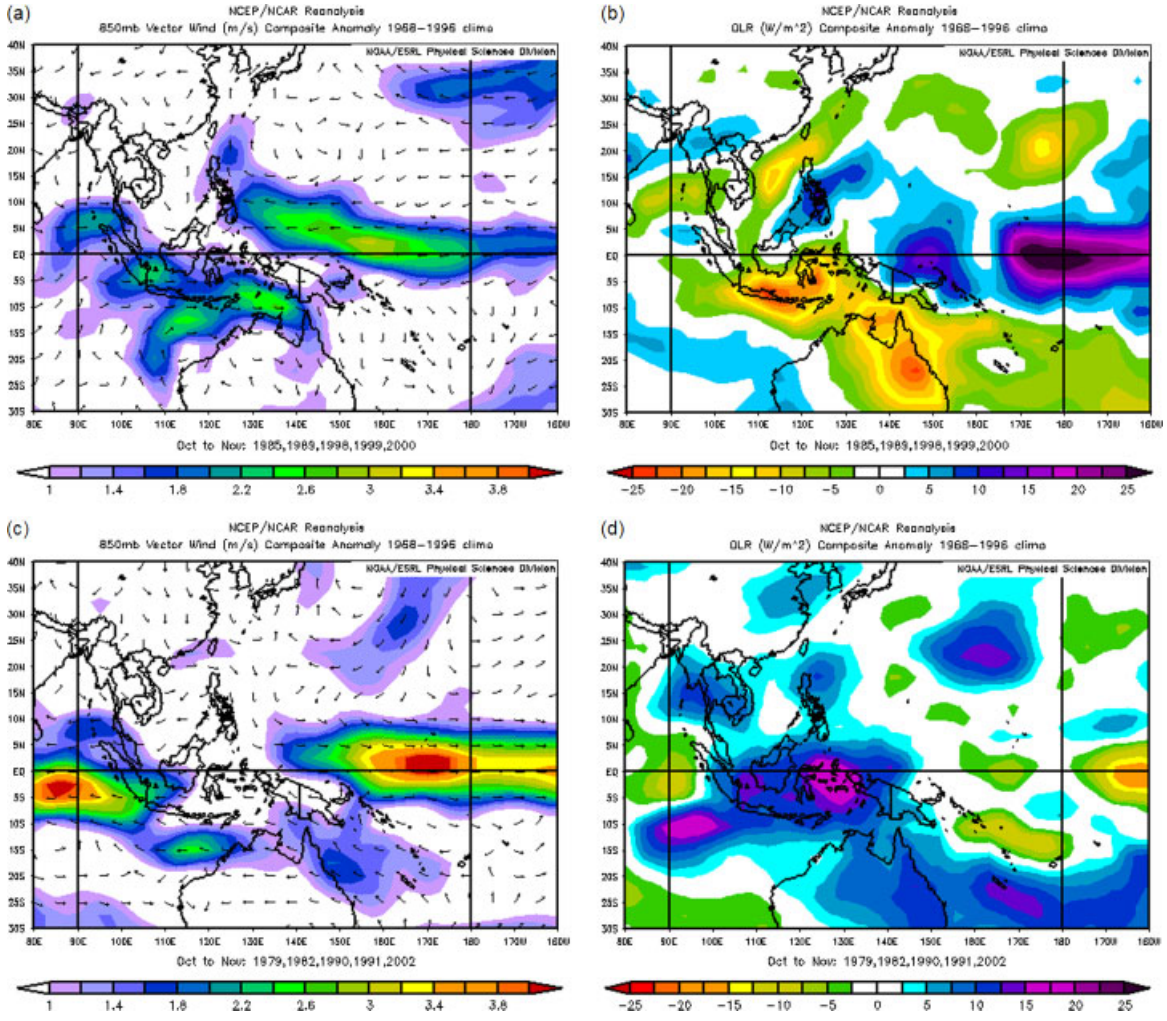

Figure 6. The composite of (a) anomalous wind flows at 850 mb pressure level and (b) anomalous OLR in five years with lowest OLR in Oct–Nov over selected Australian area, and the composite of (c) anomalous wind flows and (d) anomalous OLR in 5 years with highest OLR in Oct–Nov over selected Australian area. The area with stronger wind flows are shaded in different colors in (a) and (c) with the direction indicated by arrows. The degree of anomalous OLR is shaded in various colors in (b) and (d). This figure is available in colour online at www.interscience.wiley.com/ijoc

in increases the possibility of rainfall over the Taiwan area. Conversely, in the driest five years, because of the weaker EWC and WWC, neither the anomalous anticyclone nor anomalous OLR over the Maritime Continent region could be constructed (Meehl and Arblaster, 2002), as shown in Figure 6(c). Consequently, as illustrated in Figure 6(d), no anomalous moisture was brought into the Taiwan area to increase the possibility of rainfall events over Taiwan. Therefore, the OLR over the highly correlated region could be a reliable precursor and was selected as a predictor in this study.

3.2. Approach

In the practical situation, the water authority would use historical data to predict the streamflow for the next year and improve the accuracy of the developed model each year as more data become available. Consequently, the procedure we used and tested in this study corresponds to this sequential data acquisition and model re-calibration process as illustrated in Figure 7. Given T years of data to predict the inflows in year $T+1$, first a Monte Carlo procedure is used to build the forecast. A sample of random size $m < T$ is drawn randomly from the T years of available data. A SVM regression model for predicting the inflow is then built using these m years of data, with parameters estimated using a genetic algorithm (GA) and leave-one-out cross validation. This procedure is repeated 100 times, providing 100 candidate forecast models that cover the variability in sampling and parameter estimation with the fixed set of potential predictors. Each of these models is then applied to a forecast of the flow for the $T+1$ year and the median of these forecasts is used as the forecast. The entire forecast ensemble of 100 is also available to estimate the uncertainty in the forecast. The details of the SVM, the GA, and the resampling approach are presented next.

3.3. Support vector machines

The principal and methodology of SVM based regression technique is briefly provided below. Consider a general regression model that relates the predictors v_i to the streamflow y_i in year i:

$$y_i = f(v_i) + e \tag{1}$$

and denote by $\hat{f}(v)$ the estimate of the regression function $f(v)$. The function $f(v)$ can be any nonlinear function and in the SVM literature it is typically

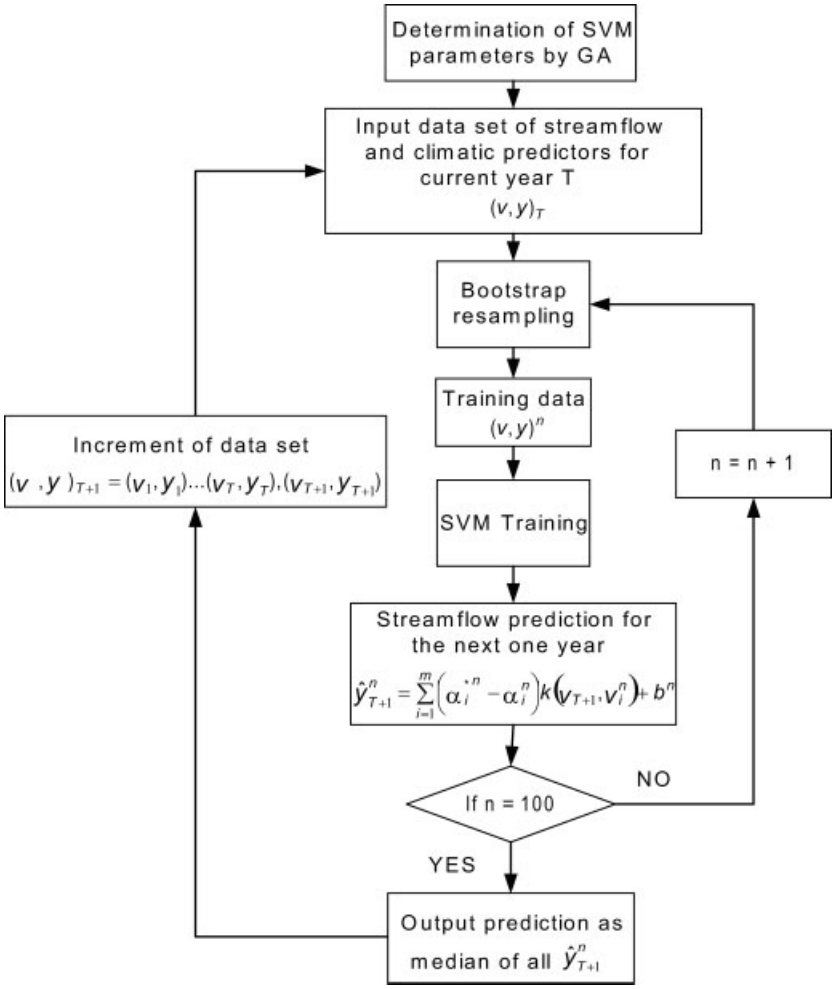

Figure 7. Proposed procedure for streamflow prediction, where T is the current year; n is the number of constructed models; the other symbols are the same as those defined in Section 3.3.

considered to be estimated through the sum of a set of kernel functions. Using these kernel functions, the input vector v_i is mapped into a new feature space in which linear regression is performed rather than non-linear regression. Given this mapping, the estimate of the regression model could be expressed as follows.

$$\hat{f}(v) = \langle w, v \rangle + b \tag{2}$$

where w represents the support vector weights (basis functions), angle brackets denote a dot product and b is a bias term, similar to an intercept in linear regression. The parameters of $\hat{f}(v)$ are estimated by minimizing the followingregularized risk function (Vapnik, 1998).

$$\text{Min} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \tag{3}$$

subject to

$$f(v) - \langle w, v \rangle - b \le \varepsilon + \xi_i \tag{4}$$

$$\langle w, v \rangle + b - f(v) \le \varepsilon + \xi_i^* \tag{5}$$

$$\xi_i^*, \xi_i \quad \ge 0 \tag{6}$$

Where ξ_i and ξ_i^* are slack variables that determine the degree to which state space samples with error more than ε be penalized; and ε is the degree to which one would like to tolerate errors in constructing the predictor $\hat{f}(v)$ (Figure 8). The above formulation is referred as the ε-insensitive approach.

The objective function given in Equations (3, 4, 5 and 6) minimizes the complexity (i.e., the magnitude of w) of the Shihmen streamflow estimator (i.e., the estimator will tend to be flat if no other considerations are imposed), leading to regularization of the solution, and penalizes errors in estimation that lie outside an ε tube (goodness of fit). In other words, for any (absolute) error smaller than ε, $\xi_i = \xi_i^* = 0$. The constant C>0 trades off the importance between the complexity of f and the amount to which deviations larger than ε are tolerated.

Usually, the optimization problem given in Equations (3, 4, 5 and 6) is solved in its dual form using Lagrange multipliers. Maximizing Equations (3, 4, 5 and 6) in its dual form, differentiating with respect to primal variables (w, b, ξ_i, ξ_i^*), and rearranging the following is obtained (Asefa *et al*., 2004, Appendix):

$$\begin{aligned}\text{Max} \quad W(\alpha^*, \alpha) = &-\varepsilon \sum_{i=1}^{N}(\alpha_i + \alpha_i^*) \\ &+ \sum_{i=1}^{N} Z_i(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*) \\ &\times (\alpha_j - \alpha_j^*)k(v_i, v_j)\end{aligned} \tag{7}$$

subject to

$$\sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0 \qquad 0 \le \alpha_i^*, \alpha_i \le C \tag{8}$$

to obtain

$$\hat{f}(v) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)k(v, v_i) + b \tag{9}$$

where α_i^* and α_i are Lagrange multipliers, $k(v, v_i)$ is a kernel that measures non-linear dependence between two state space realizations, and m is the number of selected state space points that are outside the ε tube and explain the underlying dynamic relationship. From the Kuhn–Tucker condition it follows that only for $|\hat{f}(v) - y_i| \ge \varepsilon$ the Lagrange multipliers may be nonzero. Therefore, for points inside the ε tube, α_i^* and α_i would vanish. The data points outside the ε tube are kept and are therefore called 'support vectors' to facilitate the learning process; hence the name SVMs.

By using different kernel functions for inner product evaluations, various types of non-linear models in the original space can be constructed. Radial-basis functions (RBF) are a reasonable choice of kernel functions with more flexibility and fewer parameters (Hua *et al*., 2007) than other choices. The RBF kernel function can be expressed as follows.

$$k(v, v_i) = \exp\left(\gamma^2 \|v, v_i\|\right) \tag{10}$$

where $\gamma$is user specified kernel parameter. This kernel is translation invariant, and can be written as Gaussian covariance kernel with unit variance:

$$k(v, v_i) = \sigma^2 \exp\left(\frac{\|v - v_i\|^2}{r^2}\right) = \exp\left(-\frac{h^2}{r^2}\right) \tag{11}$$

where $\sigma^2 = 1$, $r^2 = 1/\gamma^2$, and $h^2 = \|v - v_i\|^2$.

According to Thissen *et al*. (2003), the SVM model has the following characters: (1) a global optimal solution exists, which will be found (2) the result is a general solution avoiding overtraining (3) the solution is sparse and only a limited set of training points contribute to this solution and (4) non-linear solutions can be calculated efficiently because of the usage of inner products.

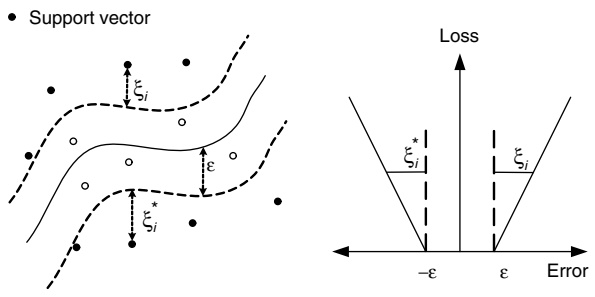

Figure 8. The ε -insensitive loss function.

3.4. Parameter determination

The macro level parameters for the SVM are the cost constant C, pointwise error tolerance or the radius of the insensitive tube ε, and the width r of the RBF kernel. In previous studies, the parameters have usually been determined by trial-and-error process which is less efficient and not easy to reach a better set of parameters promising the performance of SVM model. In this study, a GA was implemented to determine optimal parameters for the SVM model as illustrated in Figure 9. The GA (Goldberg, 1989) is a heuristic global optimization technique that imitates the natural selection of chromosomes to survive with better fitness in the environment, has been applied to several difficult problems, and shown to converge to near optimal solutions (Winston and Venkataramanan, 2003). The macro level parameters for the SVM are mutually dependent. The implemented GA assigns the individual chromosome as the combination of three real value variables for the SVM parameters. GA begins with randomly generated population of chromosomes. For each set of parameters (chromosome), leave-one-out cross validation is implemented to train SVM models by solving the optimization model defined in Equations 7–11 using the set of parameters with one data point dropped from training data sequentially. The fitness value of each chromosome is then assessed by the associated mean square error (MSE) of the predicted errors of the trained SVM models. The chromosomes with better fitness will tend to survive to the next generation and crossover with each other to generate new chromosomes in the new population. During the iteration between generations, mutation might also occur in individual chromosomes to increase the diversity of the population and to avoid being trapped in local optima. This search procedure keeps superior chromosomes and traverses through possible solution spaces to reach a near optimal solution without considering all possible solutions. During the iteration in the pre-specified number of generations, the fitness value of the best chromosome in each population can be improved continuously and the final best set of parameters (chromosome) can be retrieved by a search procedure.

3.5. Bagging

Bagging, or bootstrap aggregation, was originally proposed and applied in classification and regression trees (Breiman, 1996). This approach is on the basis of the

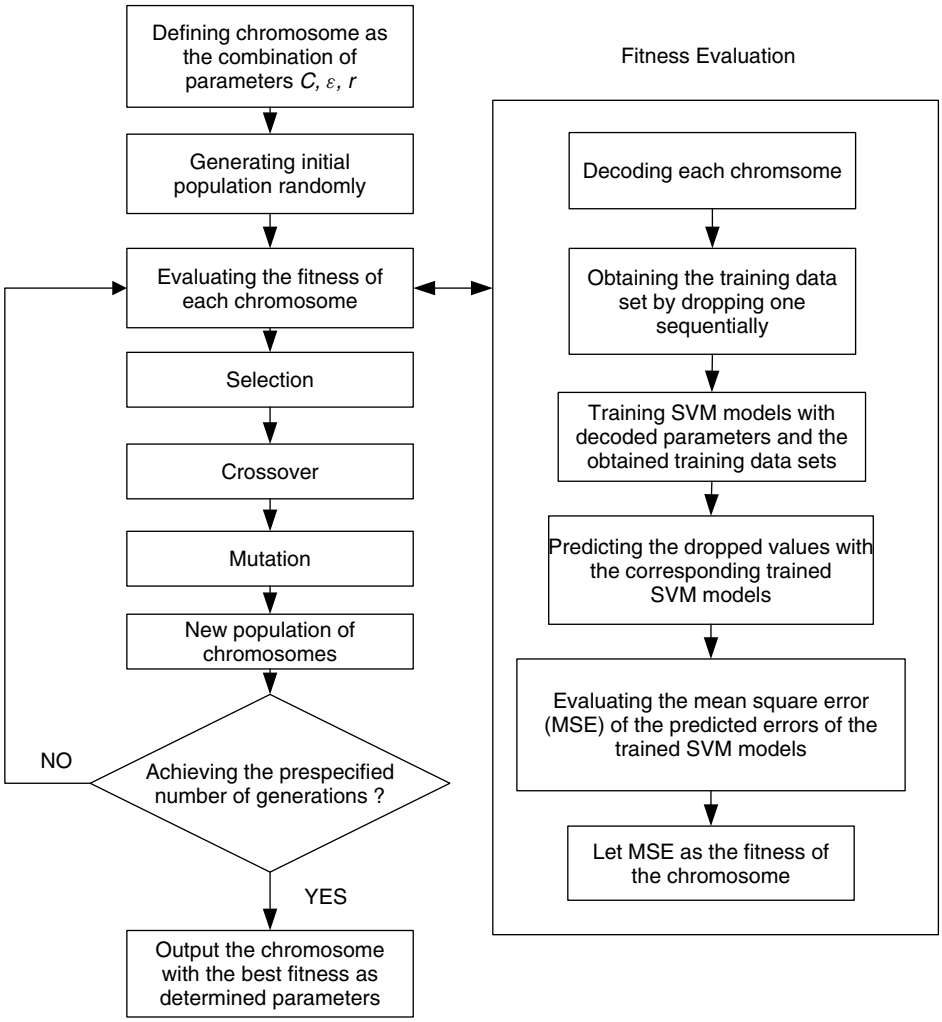

Figure 9. The procedure for parameter determination using genetic algorithm.

bootstrap statistical resampling technique (Efron and Tibshirani, 1993), to generate diverse training sets that are used to train the members composing an ensemble. It has been shown to improve the predictive performance of regression or classification trees (Gentle *et al*., 2004). In this study, bagging is applied to improve the predictability of the proposed SVM based prediction model. A prediction model is constructed with each bootstrapped sample. The resampling procedure is repeated 100 times. For each of these 100 samples, a SVM model is fit. Each of these SVM models is then used to predict the streamflow in the next year. The final forecast is the median of these 100 ensemble forecasts. This approach addresses the uncertainty in model estimation and also reduces the variance associated with the forecast based on just one set of parameters. In this study, for the relatively short hydrologic records, 100 times of bootstrap resampling were adopted to reduce the variation of the constructed prediction models. One could use a larger number of samples, particular as longer records are available. The reduction in variance as a function of the number of samples used can actually be assessed on a case by case basis. For the current application, use of 500 samples does not appreciably reduce the variance over the use of 100 samples given that the validation period is only 10 years long.

3.6. Multiple linear regression

A multiple linear regression (MLR) was also developed for streamflow forecast using the same predictors, to offer a comparison to the bagging-SVM forecast model. The MLR model is expressed as follows.

$$Y = b_0 + \sum_i b_i x_i \tag{12}$$

where Y is the Shihmen streamflow in Dec–Jan, the b_i are the regression coefficients that are estimated using the observed data, and the x_i are the regressors. The model was fit by regular least squares procedures.

4. Results and discussion

In this study, the SVM toolbox for MATLAB developed by University of Southampton (Gunn, 1997) was used to implement the proposed prediction procedure. The GA toolbox developed by MathWorks Inc. was used, and bagging was implemented directly in MATLAB.

A subset of the data (1974 to 1994) was used for model training and the rest of the data (1995 to 2004) was used to test the performance of the forecasts. The parameters of SVM model C, ε, and r are first determined by GA with the full record from 1974 to 2004. The "optimal values" were 211, 0.1, and 94 respectively. Note that these are structural parameters of the model that control how the fitting is done and are not the equivalents of regression coefficients or kernel parameters. Using these control parameters, SVM models are constructed by the proposed procedure to forecast the streamflow from 1995 to 2004. The prediction results from the proposed procedure are shown in Figure 10(a). A box-and-whiskers plot in which the bottom, middle, and upper line of the box present the first quartile (25%), median (50%), and the third quartile (75%) of the bootstrap predictions is provided for each year.

For comparison, the bagged MLR based prediction model with the same design and two other prediction models based on simple SVM and MLR approaches were also constructed. The prediction results for 1995 to 2004 are illustrated in Figures 10(b), 11(a) and (b). The relative performance of these four models in terms of correlation with the observed sequence and mean square error between predicted and observed values of the streamflow for the data withheld from model building during 1995 to 2004 is presented in Table II. The bagged SVM based prediction model outperformed bagged MLR, simple SVM, and simple MLR in terms of these measures. The worse performance of both bagged and simple MLR model suggests that there is some nonlinear relationship between climatic predictors and the Shihmen streamflow which could not be learned well in pure linear based prediction model. The performance of the bagged prediction models was better than that of simple ones in reducing the variance between the constructed models. The difference in the performance

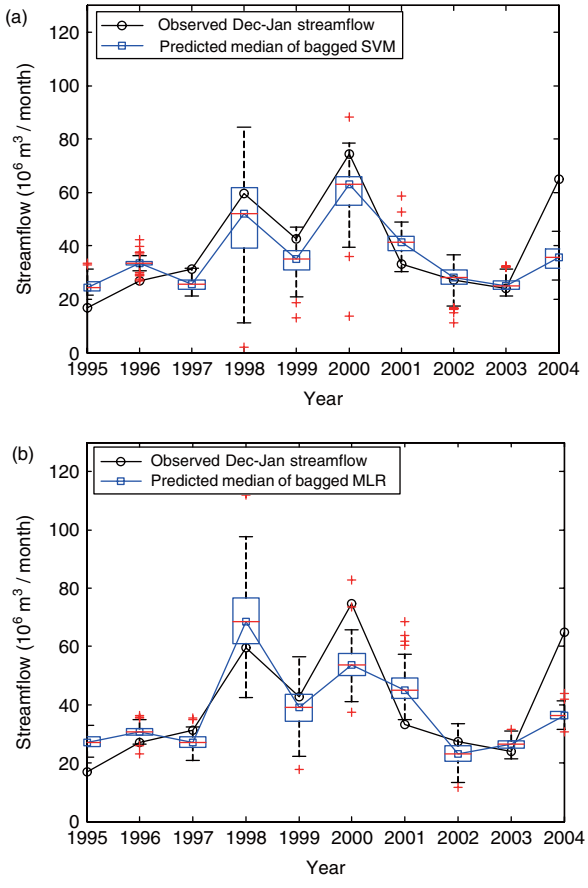

Figure 10. The predicted results of (a) SVM and (b) MLR models with 100 times bootstrap for each year from 1995 to 2004. This figure is available in colour online at www.interscience.wiley.com/ijoc

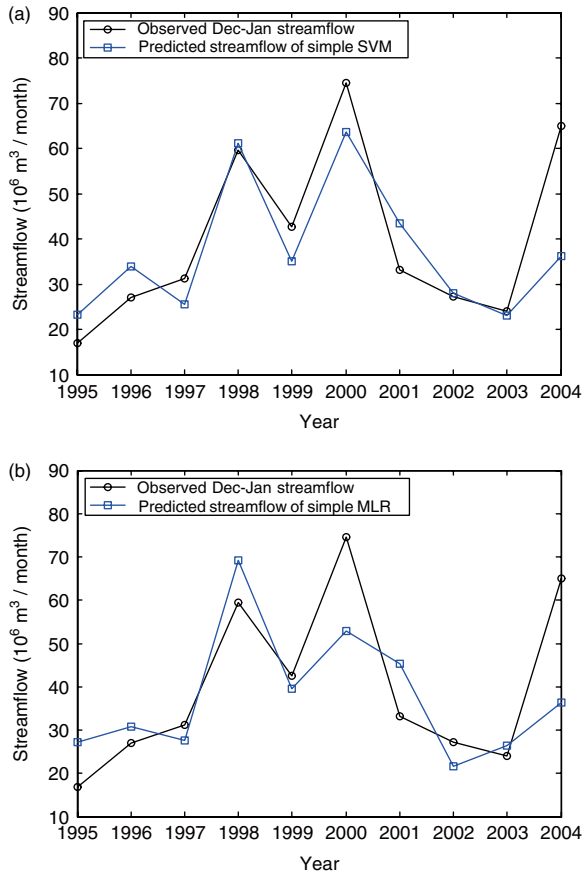

Figure 11. The predicted results of simple (a) SVM and (b) MLR models for each year from 1995 to 2004. This figure is available in colour online at www.interscience.wiley.com/ijoc

between the bagged and simple models is, however, much smaller than that between SVM and MLR-based ones. Therefore, the non-linear pattern underlying the system is more important than variance in the development of a reliable prediction model for the Shihmen streamflow.

The largest prediction errors occurred in 1998 for both the SVM and MLR models. As shown in Figure 3, the fluctuation of Shihmen streamflow seems to have increased in recent years. Although there were similar peaks in 1974 and 1975 during the study period, most of the streamflow data in the training period between 1976 and 1997 were relatively low. Therefore, the predictors identified in this study may not be able to adequately reflect the mechanisms responsible for the extreme wet conditions. Alternately, model fitting uncertainty for the small sample size for fitting the model is going to be high and this will be reflected in cases where prediction errors may be large. The forecasts for subsequent years with high flows are actually better once 1998 is included in the data set, suggesting that sampling of extremes is an issue in model building.

Table II. Performance of the prediction results of SVM and MLR models.

Prediction model	Cross validated correlation coefficient	Cross validated mean square error	R^2	RPSS[a] (%) *versus* climatology
Bagged SVM	0.83	130.2	0.62	31
Bagged MLR	0.73	164.7	0.52	18
Simple SVM	0.8	137.6	0.6	–
Simple MLR	0.72	169.8	0.5	–

[a] RPSS, rank probability skill score (Wilks, 1995).

A final alternative was to consider prediction of the Shihmen streamflow using the simulation of general circulation model (GCM) (Landman and Goddard, 2002) of the ocean-atmosphere system. To explore this alternative, a map presenting the correlation between the precipitation simulated by ECHAM 4.5 (Roeckner *et al.*, 1996) on the resolution of 2.8° × 2.8° lat/lon grid and the observed streamflow from 1974 to 2004 was also generated as in Figure 12. The ECHAM runs used were simulations using concurrent, observed data and hence represent an upper limit on potential predictability from the prior season. As shown in Figure 12, there was no simulated precipitation near Taiwan which was significantly related to the streamflow at the Shihmen reservoir. As mentioned in Yu *et al.* (2002), GCMs normally simulate the precipitation of the atmosphere on the basis of the conceptualization of physics of the atmospheric circulation, surface energy, and water fluxes. The grid size used in the GCM models is normally significantly larger than the size of the catchments in Taiwan. For the course resolution, the land–sea contrast and topography in the regional scale such as catchments in Taiwan cannot be properly represented in global models (Chu *et al.*, 2008) and this leads to the poor simulation results.

5. Conclusion

In Taiwan, the unanticipated fluctuations of water available in the spring growing season have led the water

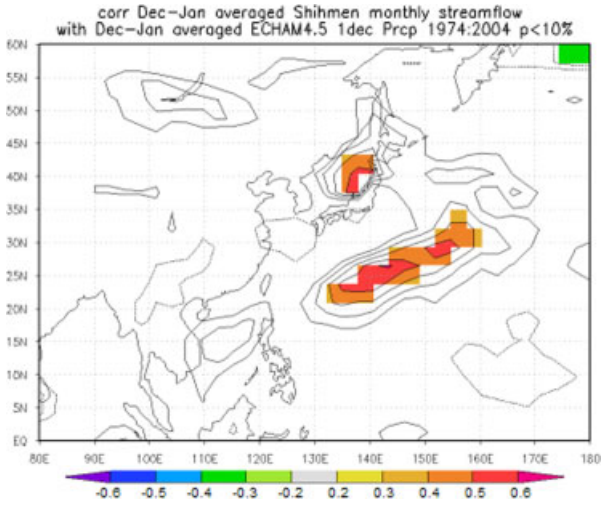

Figure 12. The correlation map between simulated precipitation by ECHAM and streamflow in Dec–Jan from 1974 to 2004. This figure is available in colour online at www.interscience.wiley.com/ijoc

authority to assign a lot of the budget to compensate for the losses caused by their failure to deliver water allocated for irrigation. A modified SVM-based prediction framework has therefore been proposed to improve the predictability of the inflow to Shihmen reservoir in December and January in order to mitigate such possible negative impacts, using the identified highly correlated climate precursors. A process of discussion about potential physical factors and exploratory data analysis (largely linear) was used to identify prospective predictors. Then, a non-linear regression approach that had been shown to be effective in identifying non-linear functional relations in sparse data, multivariate predictor situations was used. Bagging and GA were introduced to improve the uncertainty assessment and reliability of the scheme. The assessment of the performance of the scheme was done in a manner similar to what may happen in the real world. Model development based on a short record followed by model updating as more data became available. The potential of the strategy developed relative to the use of a linear regression approach with the same predictors and relative to the potential predictability from a GCM was demonstrated under these conditions. The proposed SVM-based prediction framework reasonably forecasts the Dec–Jan inflow to Shihmen reservoir, even with the relatively short hydrologic records, and outperforms the other alternatives for learning the non-linear pattern underlying the climate systems more robustly. Climate information was thus found to be potentially valuable for improved hydrologic prediction in support of water resource management in Taiwan and the non-linear pattern was more important in the development of reliable prediction models for the inflow into Shihment reservoir. The variance in the prediction from the constructed models by the proposed approach has also been reduced more efficiently. Even so, if the structural relationship between the predictors used and the streamflow were to change as climate changes or other predictors become more important, the performance of the scheme developed will be impacted. Also, the fitting uncertainty of the constructed model will be higher when the available record data are shorter.

In this study, only the predictability of the streamflow at Shihmen reservoir in Dec–Jan is explored. To facilitate the water authority to manage Shihmen reservoir more efficiently, the streamflow in the other seasons should also be predicted reasonably. Further work is needed to develop reliable prediction model for the streamflow in the other seasons and to investigate the reasonable climatic predictors and the corresponding mechanism underlying the global climate systems. The use of GCM's to explore and test the efficacy of these predictors is also needed. At this stage we have not had the resources to systematically analyse the precipitation response in Taiwan to systematic forcing of GCM's with anomalous conditions in each of the regions of prediction. Thus, our understanding of the physical mechanisms is incomplete. However, a utility of the statistical modelling approach as shown in this study is that it can stimulate systematic investigation of specific climate features that may be responsible for regional precipitation outcomes and thereby help overcome the perception that the GCM simply does not work for a certain region. It is also possible to use predictors from a GCM forecast run and preseason climate indicators in an SVM model. However, we have not yet explored this strategy for Taiwan. This initial effort was focused towards the development and quantitative testing of a tool that could be used by the water agency for medium range planning in a critical water supply season and for identifying some of the climate factors that need to be better understood through subsequent modelling.

Acknowledgements

The authors would like to thank National Science Council of the Republic of China for providing partial financial support of this work under the Graduate Students Study Abroad Program Grant NSC 095-2917-I-009-011. Valuable comments and suggestions provided by anonymous reviewers are also greatly appreciated. Without them, our manuscript could not have been significantly improved.

References

Allan RJ. 1983. Monsoon and teleconnection variability over Australasia during the Southern Hemisphere summers of 1973–77. *Monthly Weather Review* **111**(1): 113–142.

Asefa T, Kemblowski MW, Lall U, Urroz G. 2005. Support vector machines for nonlinear state space reconstruction: Application to the Great Salt Lake time series. *Water Resources Research* **41**: W12422. DOI: 10.1029/2004WR003785.

Asefa T, Kemblowski MW, McKee M, Khalil A. 2006. Multi-time scale stream flow predictions: The support vector machines approach. *Journal of Hydrology* **318**(1–4): 7–16.

Asefa T, Kemblowski MW, Urroz G, McKee M, Khalil A. 2004. Support vectors–based groundwater head observation networks design. *Water Resources Research* **40**: W11509. DOI: 10.1029/2004WR 003304.

Boser BE, Guyon I, Vapnik V. 1992. A training algorithm for optimal margin classifiers, *5th Annual ACM Workshop on Computational Learning Theory*. ACM, Pittsburgh: 144–152.

Breiman L. 1996. Bagging predictors. *Machine Learning* **24**(2): 123–140.

Chang CP. 2004. *East Asian Monsoon*. World Scientific: Singapore.

Chang FJ, Chang YT. 2006. Adaptive neuron-fuzzy inference system for prediction of water level in reservoir. *Advances in Water Resources* **29**: 1–10.

Chen CS, Chen YL. 2003. The rainfall characteristics of Taiwan. *Monthly Weather Review* **131**(7): 1323–1341.

Chen GTJ, Chen SY, Yan MH. 1983. The winter diurnal circulation and its influence on precipitation over the coastal area of Northern Taiwan. *Monthly Weather Review* **111**(11): 2269–2274.

Chiew FHS, Piechota TC, Dracup JA, McMahon TA. 1998. El Ni no–Southern Oscillation and Australian rainfall, streamflow and drought: Links and potential for forecasting. *Journal of Hydrology* **204**(1): 138–149.

Chowdhury S, Sharma A. 2009. Long range NINO3.4 predictions using pair wise dynamic combinations of multiple models. *Journal of Climate*. DOI: 10.1175/2008JCLI2210.1.

Chu JL, Kang H, Tam CY, Park CK, Chen CT. 2008. Seasonal forecast for local precipitation over northern Taiwan using statistical downscaling. *Journal of Geophysical Research* **113**: D12118. DOI: 10.1029/2007JD009424.

Colman AW, Davey MK. 2003. Statistical prediction of global sea surface temperature anomalies. *International Journal of Climatology* **23**: 1677–1697.

Dibike BY, Velickov S, Solomatine D, Abbot BM. 2001. Model induction with support vector machines: introduction and applications. *Journal of Computing in Civil Engineering* **15**(3): 208–216.

Ding YH. 1993. *A Study on the Excessively Heavy Rainfall Over the Yangtze-Huaihe River Basin in 1991 (in Chinese)*. China Meteorological Press: Vol. 2–4: 254.

Efron B, Tibshirani RJ. 1993. *An Introduction to the Bootstrap*. Chapman & Hall: New York.

Eldaw AK, Salas JD, Garcia LA. 2003. Long-range forecasting of the Nile River flows using climatic forcing. *Journal of Applied Meteorology* **42**(7): 890–904.

Fowler HJ, Kilsby CG. 2002. Rainfall and the North Atlantic Oscillation: a case study of climate variability in northern England. *International Journal of Climatology* **22**: 843–866.

Gentle JE, Härdle W, Mori Y. 2004. *Handbook of Computational Statistics*. Springer: Berlin.

Goldberg DE. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley: Boston.

Gunn SR. 1997. *Support Vector Machines for Classification and Regression*, Technical Report. Image Speech and Intelligent Systems Research Group, University of Southampton: Southampton.

Hamlet AF, Huppert D, Lettenmaier DP. 2002. Economic value of long-lead streamflow forecasts for Columbia River hydropower. *Journal of Water Resources Planning and Management* **128**(2): 91–101.

Hamlet AF, Lettenmaier DP. 1999. Columbia River stream-flow forecasting based on ENSO and PDO climate signals. *Journal of Water Resources Planning and Management* **125**(6): 333–341.

Harshburger B, Hengchun Y, Dzialoski J. 2002. Observational evidence of the influence of Pacific SSTs on winter precipitation and spring stream discharge in Idaho. *Journal of Hydrology* **264**(1): 157–169.

Hua XG, Ni YQ, Ko JM, Wong KY. 2007. Modeling of temperature–frequency correlation using combined principal component analysis and support vector regression technique. *Journal of Computing in Civil Engineering* **21**(2): 122–135.

Huang WC, Chou CC. 2005. Drought early warning system in reservoir operation: Theory and practice. *Water Resources Research* **41**: W11406. DOI: 10.1029/2004WR003830.

Huang WC, Yuan L, Lee C. 2002. Linking genetic algorithms with stochastic dynamic programming to the long-term operation of a multireservoir system. *Water Resources Research* **38**(12): 1304. DOI: 10.1029/2001WR001122.

Karamouz M, Zahraie B. 2004. Seasonal streamflow forecasting using snow budget and El Niño–Southern Oscillation climate signals: application to the Salt River Basin in Arizona. *Journal of Hydrological Engineering* **9**(6): 523–533.

Khan MS, Coulibaly P. 2006. Application of support vector machine in lake water level prediction. *Journal of Hydrologic Engineering* **11**(3): 199–205.

Kim YO, Palmer RN. 1997. Value of seasonal flow forecasts in bayesian stochastic programming. *Journal of Water Resources Planning and Management* **123**(6): 327–335.

Landman WA, Goddard L. 2002. Statistical recalibration of GCM forecasts over southern Africa using model output statistics. *Journal of Climate* **15**(15): 2038–2055.

Liong SY, Sivapragasam C. 2002. Flood stage forecasting with support vector machines. *Journal of the American Water Resources Association* **38**(1): 173–186.

Liu ZJ, Valdes JB, Entekhabi D. 1998. Merging and error analysis of regional hydrometeorologic anomaly forecasts conditioned on climate precursors. *Water Resources Research* **34**(8): 1959–1969.

Meehl GA, Arblaster JM. 2002. The tropospheric biennial oscillation and Asian-Australian monsoon rainfall. *Journal of Climate* **15**(7): 722–744.

Piechota TC, Dracup JA. 1999. Long-range streamflow forecasting using El Niño-southern oscillation indicators. *Journal of Hydrological Engineering* **4**(2): 144–151.

Raftery AE, Geneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian Model Averaging to calibrate forecast ensembles. *Monthly WeatherReview* **133**: 1155–1174.

Roeckner E, Arpe K, Bengtsson L, Christoph M, Claussen M, Dümenil L, Esch M, Giorgetta M, Schlese U, Schulzweida U. 1996. *The Atmospheric General Circulation Model ECHAM-4: Model Description and Simulation of Present-Day Climate*. Max-Planck-Institute for Meteorology report No. 218: 90.

Sharma A. 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1–A strategy for system predictor identification. *Journal of Hydrology* **239**(1–4): 232–239.

Shiau JT, Lee HC. 2005. Derivation of optimal hedging rules for a water-supply reservoir through comprising programming. *Water Resources Management* **19**: 111–132.

Shu YT. 2003. *Study on Transfer of Tou-Chien Creek's Agricultural Water in Drought Period* (in Chinese). Master Thesis. Chung Hua University: Hsinchu.

Souza Filho FA, Lall U. 2003. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: applications of a multivariate, semiparametric algorithm. *Water Resources Research* **39**(11): 1307–1325. DOI:10.1029/2002WR001373.

Thissen U, van Brakel R, de Weijer AP, Melssen WJ, Buydens LMC. 2003. Using support vector machines for time series prediction. *Chemometrics and Intelligent Laboratory Systems* **69**: 35–49.

Vapnik V. 1998. *Statistical Learning Theory*. John Wiley: Hoboken.

Wang B. 2006. *The Asian Monsoon*. Springer: New York.

Westphal KS, Vogel RM, Kirshen P, Chapra SC. 2003. Decision support system for adaptive water supply management. *Journal of Water Resources Planning and Management* **129**(3): 165–177.

Wheeler M, Kiladis GN, Webster PJ. 2000. Large-scale dynamical fields associated with convectively coupled equatorial waves. *Journal of Atmospheric Sciences* **57**(5): 613–640.

Wilks DS. 1995. Forecast verification, *Statistical Methods in the Atmosphere Sciences*. Academic Press: San Diego, CA; 233–283.

Winston WL, Venkataramanan M. 2003. *Introduction to Mathematical Programming*. Brooks/Cole: Pacific Grove.

Wu R, Hu ZZ, Kirtman BP. 2003. Evolution of ENSO-related rainfall anomalies in East Asia. *Journal of Climate* **16**(22): 3742–3758.

Xu K, Brown C, Kwon HH, Lall U, Zhang J, Hayashi S, Chen Z. 2007. Climate teleconnections to Yangtze river seasonal streamflow at the Three Gorges Dam, China. *International Journal of Climatology* **27**(6): 771–780.

Xue YK, Sun S, Lau JM, Ji J, Poccard I, Kang HS, Zhang R, Wu G, Zhang J, Schaake J, Jiao Y. 2005. Multiscale variability of the river runoff system in China and its link to precipitation and sea surface temperature. *Journal of Hydrometeorology* **6**(4): 550–570.

Yang S, Lau KM, Kim KM. 2002. Variations of the East Asian jet stream and Asian–Pacific–American winter climate anomalies. *Journal of Climate* **15**(3): 306–325.

Yu PS, Chen ST, Chang IF. 2006. Support vector regression for real-time flood stage forecasting. *Journal of Hydrology* **328**(3–4): 704–716.

Yu XY, Liong SY. 2007. Forecasting of hydrology time series with ridge regression in feature space. *Journal of Hydrology* **332**(3–4): 290–302.

Yu PS, Yang TC, Wu CK. 2002. Impact of climate change on water resources in southern Taiwan. *Journal of Hydrology* **260**: 161–175.

