Water Resour Manage (2013) 27:1–23
DOI 10.1007/s11269-012-0130-1

Error Correction Modelling of Wind Speed Through Hydro-Meteorological Parameters and Mesoscale Model: A Hybrid Approach

Asnor Muizan Ishak · Renji Remesan · Prashant K. Srivastava · Tanvir Islam · Dawei Han

Received: 8 May 2011 / Accepted: 27 August 2012 /
Published online: 21 November 2012

Abstract Accurate estimation of wind speed is essential for many hydrological applications. One way to generate wind velocity is from the fifth generation PENN/NCAR MM5 mesoscale model. However, there is a problem in using wind speed data in hydrological processes due to large errors obtained from the mesoscale model MM5. The theme of this article has been focused on hybridization of MM5 with four mathematical models (two regression models- the multiple linear regression (MLR) and the nonlinear regression (NLR), and two artificial intelligence models – the artificial neural network (ANN) and the support vector machines (SVMs)) in such a way so that the properly modelled schemes reduce the wind speed errors with the information from other MM5 derived hydro-meteorological parameters. The forward selection method was employed as an input variable selection procedure to examine the model generalization errors. The input variables of this statistical analysis include wind speed, temperature, relative humidity, pressure, solar radiation and rainfall from the MM5. The proposed conjunction structure was calibrated and validated at the Brue catchment, Southwest of England. The study results show that relatively simple models like MLR are useful tools for positively altering the wind speed time series obtaining from the MM5 model. The SVM based hybrid scheme could make a better robust modelling framework capable of capturing the non-linear nature than that of the ANN based scheme. Although the proposed hybrid schemes are applied on error correction modelling in this study, there are further scopes for application in a wide range of areas in conjunction with any higher end models.

A. M. Ishak (✉) · P. K. Srivastava · T. Islam · D. Han
WEMRC, Department of Civil Engineering, University of Bristol, Bristol BS81TR, UK
e-mail: asnormuizan.ishak@bristol.ac.uk

R. Remesan
Department of Geography, University of Hull, Cottingham Road, Hull HU6 7RX, UK

A. M. Ishak
Hydrology and Water Resources Division, Department of Irrigation and Drainage, Ministry of Natural Resources and Environment, KM 7, Jalan Ampang, 68000 Kuala Lumpur, Malaysia

Keywords MM5 dynamical downscaling · Linear and non-linear regression · Artificial neural network (ANN) · Support vector machine (SVM) · Numerical weather prediction (NWP) model · Input variable selection · Meteorology and hydrology

1 Introduction

The Pennsylvania State University–National Center for Atmospheric Research (PSU/NCAR) mesoscale modelling system 5 (MM5) is one of the sophisticated and widely accepted downscaling models in the hydro-meteorological field (Dudhia 1993; Ishak et al. 2010). Downscaled global assimilated weather data from the mesoscale downscaling model such as MM5 are a very useful source of information capable of making input data to many regional meteorological and hydrological models. For instance, the MM5 downscaled weather variables could effectively be used for reference evapotranspiration (ET_o) estimation, especially in ungauged catchments. It has been known that the ET_o is a main component in conventional water balance studies which has considerable significance on hydrological modelling and water resources management (Kashyap and Panda 2001; Chauhan and Shrivastava 2009) , where, wind speed is one of the major input weather variables influencing the estimation of ET_o (Allen et al. 1998). However, many studies have highlighted the modelling difficulty of wind speed using mesoscale models (Frank 1983; Zhong and Fast 2003). A recent study by Ishak et al. (2010) has demonstrated that the percentage error in wind speed is about 200–400 % in the MM5 downscaling study adopted at the Brue catchment in southwest England using the ERA-40 reanalysis data.

The advent of modern artificial intelligence (AI) technologies provides us with many useful approaches (e.g. artificial neural networks (ANN), support vector machines (SVMs) and many more) to tackle complex physical processes. ANNs are one of the very powerful mathematical tools and successfully used in hydrology for tackling many issues like river level forecasting, rainfall runoff modelling, rainfall estimation and forecasting, ground water modelling, reservoir inflow monitoring, water quality prediction and water resources management (Yang et al. 2001; Zhu et al. 2007; Islam et al. 2012a). ANNs are reliable tools to improve the estimation of hydrological and meteorological variables such as wind speed. Another new tool in hydrology from machine learning community field is called support vector machine (SVM) and recently has gained considerable attention in environmental science and related fields. One can find several applications of SVMs in literature like flood stage forecasting, statistical model of daily precipitation, runoff modeling and many more (Bray and Han 2004; Yu et al. 2006; Chen and Yu 2007; Wang et al. 2009). These artificial intelligence models (ANNs and SVMs) could be successfully used in conjunction with MM5 to tackle error correction issues in short-term wind speed prediction. A study by Salcedo-Sanz et al. (2009) has presented a hybrid system including weather forecast models (MM5) and artificial neural networks in a problem of short-term wind speed prediction. Another useful error correction approach is with regression models (linear or power type). Besides, the optimisation method (OP) and validation techniques are the common methods to identify the best model structure to tackle a specific modelling problem. Study by Efron (1986) has shown that validation with optimisation methods are a reliable and successful scheme for model selection of parameters.

Henceforth, in this paper, we recommend the hybridization of a mesoscale model with regression models (multiple linear regression (MLR), nonlinear (power type) regression model (NLR)), AI models like ANNs and SVMs to obtain an error correction system for the Brue catchment. The MM5 model dynamically downscales the ECMWF global data to

obtain meteorological variables including wind speed in the smaller area. Then the properly trained ANN and SVM models could process the wind speed data together with other variables in order to accurately predict the wind speed. Later the modelling capability is compared with that of relatively simpler regression models (linear and power type). The paper has the following structure. In Section 2, we detail the modelling system used, structure of the neural network and SVMs, study area and the data sets used. Section 3 details the results on the performance of the approaches. Section 4 gives some final remarks and conclusions of this work.

2 Materials and Methods

This section describes the materials and methods for wind speed correction, using models and different combinations of six meteorological variables obtained from the downscaled dataset. Figure 1 shows an outline of how the system works. It starts from a global ECMWF ERA 40-reanalysis data whose outputs are used as the boundary condition for the MM5 model. The MM5 derived variables such as wind speed, temperature, relative humidity, pressure, solar radiation and rainfall were used as input variables for the wind velocity correction approach. Four different models are considered for the correction viz. the ANN model, SVM model and two regression models (a linear model and a power form model). The details of the methodology are described in following sections.

2.1 Study Area

The Brue catchment is chosen as the study area which is located in the south-west of England, 51.075 °N and 2.58 °W, and drains an area of 135.2 sq km (Fig. 2). The

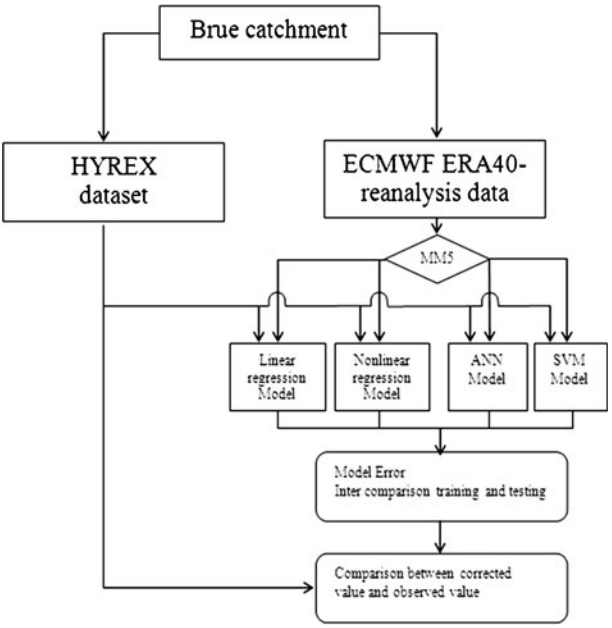

Fig. 1 Outline of the error correction system for downscaled wind speed data

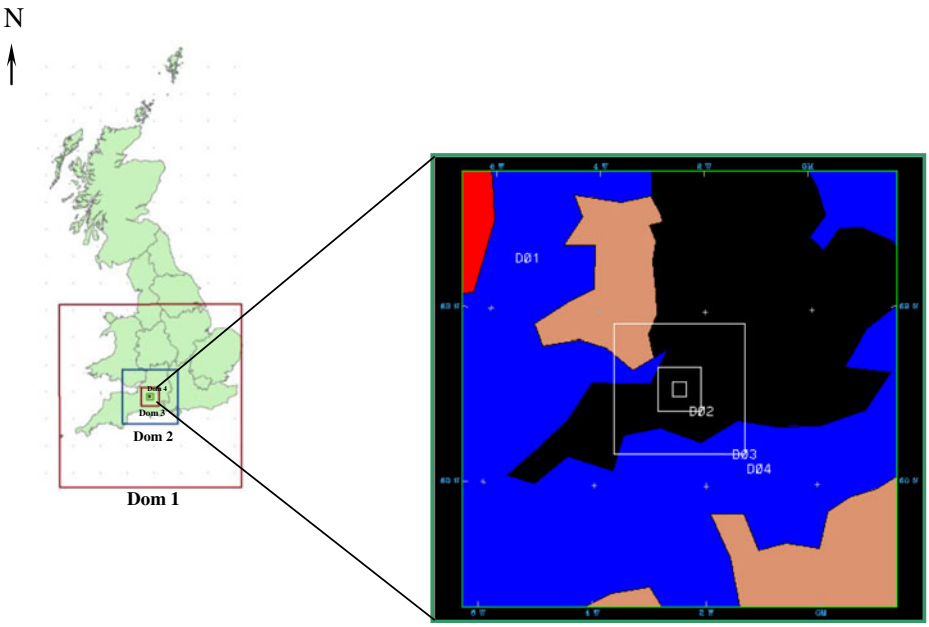

Fig. 2 The study area of the Brue catchment, Somerset, Southwest England

observation data for this study were obtained from the NERC (Natural Environment Research Council) funded HYREX project (Hydrological Radar Experiment). The ground observed data from the Brue catchment, provided by HYREX, are used for evaluating the downscaled wind speed data from the mesoscale regional model MM5. On top of that, this study makes use of the ERA-40 reanalysis global weather data in the years of 1995, 1996, and 1998. The resolutions of these data are 1°×1° in space and 6 h in time. Data from representative seasonal months like January, March, July and October were selected for the analysis representing winter, spring, summer and autumn seasons respectively. In addition, winter and spring seasons may also be considered as cold season while summer and autumn seasons may be considered as warm season (Islam et al. 2012b). Data from the first 2 years (1995 and 1996) have been used for the training and other 1 year (1998) for the testing purposes.

2.2 Data Analysis Techniques

2.2.1 MM5

The MM5 (Mesoscale Model 5) is the fifth generation PENN/NCAR mesoscale model descended from the model developed by Anthes in the 1970s at PSU. MM5 is a regional-scale primitive equation model that can be configured hydrostatically or non-hydrostatically (Grell et al. 1994). The MM5 model uses sigma coordinates in the vertical domain, and allows for 2-way interactive nesting of domains, with up to nine nested and interactive domains possible. MM5 is also equipped with four dimensional data assimilation capability and several new physics parameterisations which were not included in any of the previous releases of the modelling system viz. Betts-Miller, Kain-Fritsch and Fritsch-Chappell cumulus parameterizations, the Burk-Thompson planetary boundary layer scheme, two new

cloud microphysical schemes and the CCM2 radiation package (Warner et al. 1991; Mass and Kuo 1998; Chen and Dudhia 2001). In the MM5 setting, the set of parameterizations for the atmospheric model processes is Grell cumulus formation. We have made this selection based on pervious case study over the Brue catchment (Ishak et al. 2012). MRF parameterization for the planetary boundary layer has been chosen for the MM5 simulation. Technical detail of this scheme can be found in (Dudhia 1993). The adopted approach uses the PSU–NCAR mesoscale model (Dudhia 1993; Grell 1995) as a common test framework to host the output of wind speed for 4 months in each year of 1995, 1996 and 1998. The model was run with horizontal resolutions of 4 domains called Domain 1, 2, 3 and 4. As one can find in Fig. 2, Domains 1 to 4 have been structured with horizontal resolutions of 21×27 km, 19×9 km, 19×3 km, 19×1 km, respectively. The model was run using 23 vertical levels which are default in MM5. Figure 3 shows the hourly pattern of the MM5 derived wind speed and observed wind speed obtained from ground based HYREX project during periods 01 – 31 January, 01 – 31 March, 01 – 31 July and 01 – 31 October in 1995. The study has also used data from similar dates in 1996, and 1998

2.2.2 Model Selection

The general hypothesis of model selection is based on model complexity (here input space is considered as a measure of model complexity) and its influence during training and testing phases is shown in Fig. 4. A general hypothesis states that more complex models can simulate reality better than simpler models (i.e. less RMSE error), but they may fit to the data noise and will perform poorly in generalisation. On the other hand, simpler models are less influenced by the data noise, but may have poor training errors. The optimal model should be the one in between. In reality, both model structure and data noise would have an impact on the optimal input variables selection. Therefore, even for the same physical problem, different models may have different optimal input variable combinations.

2.2.3 Regression Models and Validation Method

This study has used two regression models viz. the linear and nonlinear power-form function models commonly used to describe a relationship between output (Y) with input of variables ($X_1, X_2, X_3, X_4, \ldots, X_n$) and with model parameters ($a_0, a_1, a_2, \ldots a_n$) (Thomas and Benson, 1970). These are reliable techniques and widely used in many estimation and forecasting problems. The linear and power form equations are given in Eqs. 1 and 2:

$$Y = a_o + X_1 a_1 + X_2 a_2 + X_3 a_3 + X_4 a_4 + X_5 a_5 + X_6 a_6 + \ldots\ldots X_n a_n + \varepsilon_o \tag{1}$$

$$Y = a_0 X_1^{a_1} X_2^{a_2} X_3^{a_3} X_4^{a_4} X_5^{a_5} X_6^{a_6} \ldots\ldots X_n^{a_n} + \varepsilon_o \tag{2}$$

where, $a_0, a_1, \ldots a_n$, are the model parameters, ε_0 is the error term, n is the number of data. In this study, Y is the observed wind speed while X_1 to X_n are input variables and these are from the MM5 outputs such as wind speed, surface temperature, surface pressure, solar radiation, rainfall and relative humidity. Optimization technique has been used to minimize the result estimated variables function, for instance $\min_x f(x)$ (Broyden 1970; Fletcher 1987). The Broyden-Fletcher-Goldfarb-Shanno (BFGS) Quasi-Newton gradient-based algorithms are the common methodology to solve this kind of unconstrained minimization problems. The unconstrained minimization case is applied due to the imposed conditions on the

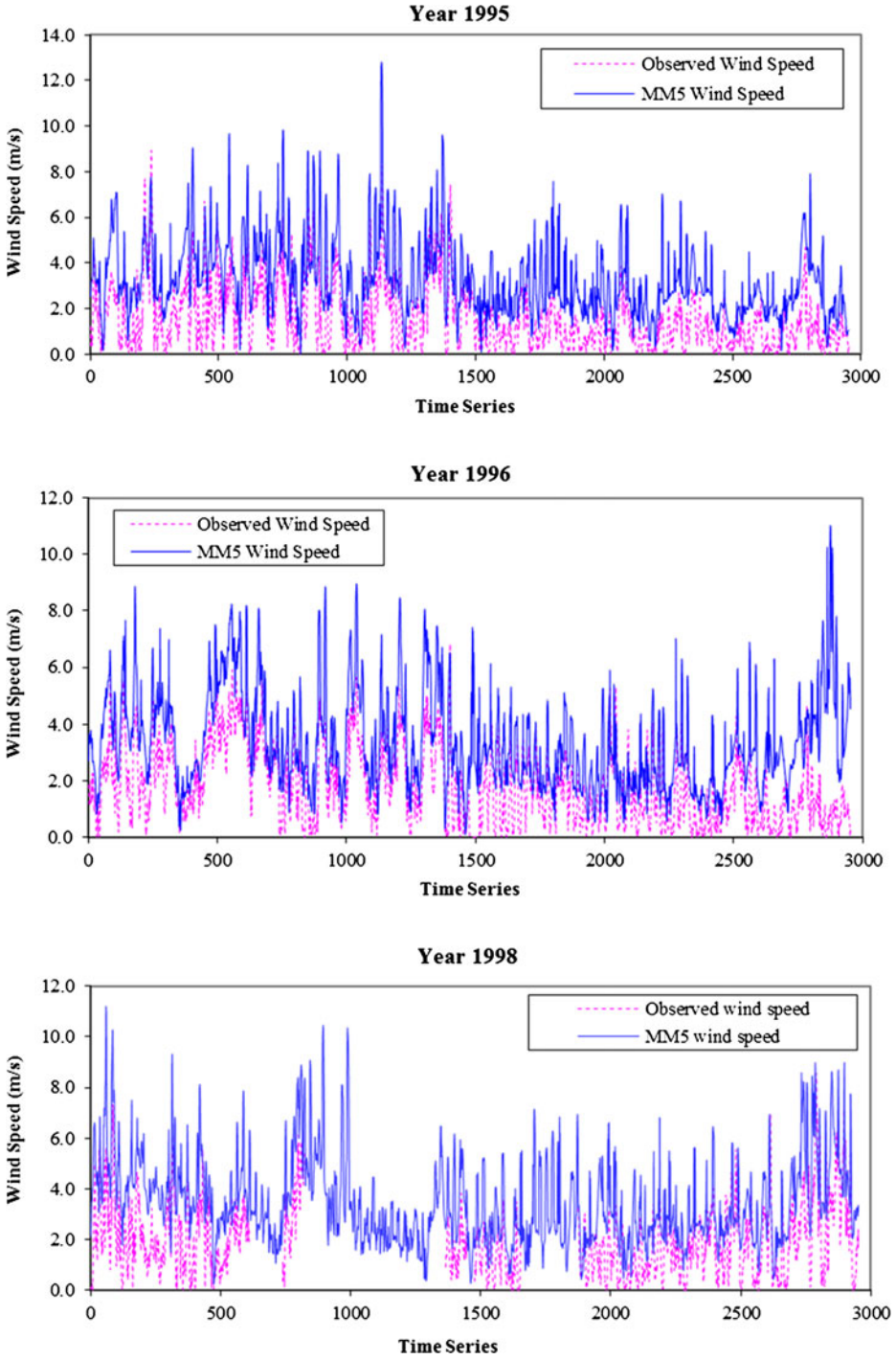

Fig. 3 Time series of hourly wind speed based on the observed ground data and MM5 derived for year 1995, 1996 and 1998

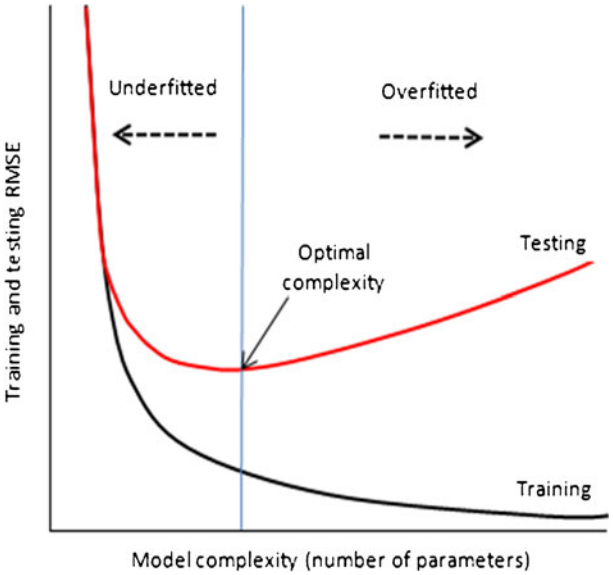

Fig. 4 Hypothesis showing effect of complexity during training and testing (Hastie et al., 2001)

independent variables X and it assumes that f is defined for all X. Therefore, the optimization uses an iteration process to find the most optimum value on this process. The value of a_0 (initial value) should be considered first, later the processes are carried out for next a_1, a_2, a_3, a_4,..., a_n. At the end of this, the process will succeed on estimation of those variables at the local minimum. Thus, this technique will end the process of analysis until it reaches the predefined number of iterations of k. The forward model selection method is applied to identify the suitable model. The concept is based on separation of the dataset into training and validation data sets (Cawley and Talbot 2003). In addition, the validation is considered as an estimator of model generalization error.

2.2.4 Artificial Neural Network (ANN)

The supervised learning is the most common learning approach used in ANNs, in which the input is presented to the network along with the desired output, and the weights are adjusted so that the network attempts to produce the desired output (Møller 1993). There are different learning algorithms and a popular algorithm is the back propagation algorithm. This study has adopted the artificial neural network with single hidden layer architecture as shown in Fig. 5. We have adopted a three layer network topology with six input in the first layer (layer A), six nodes units in the second layer (layer B or *hidden layers*) and a single node in the third layer (layer C or *output layer*). The 'trial and error' method was adopted to identify the number of hidden nodes (10, in this study). In the network, each input-to-node and node-to-node *connection* is modified by a *weight.* There is an extra input assumed in each node that is assumed to have a constant value of one. The weight that modifies this extra input is called the *bias*. Before performing training process, the weights and biases were initialized to appropriately scaled values. Appropriate normalisation of training data was essential to avoid saturating the activation function, hence our training data were normalised. The sigmoid activation function was employed in this study. The training of the network was

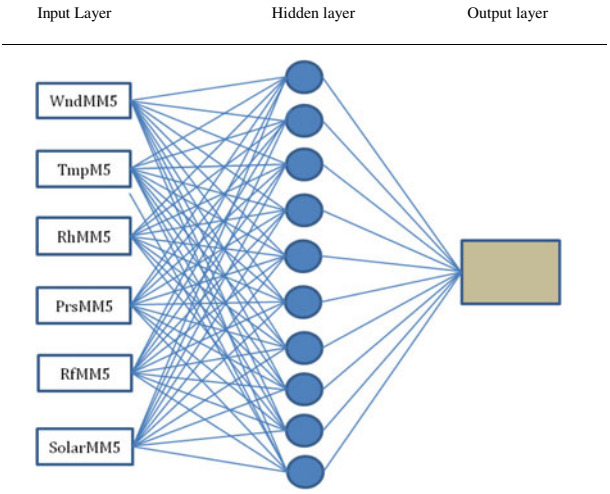

Fig. 5 The structure of a single hidden layer artificial neural network

carried out using the Levenberg–Marquardt algorithm. Various neuron number combinations at the hidden layer were tested for the ANN models to find the best number of the hidden layer nodes for modelling.

$$O_a = h_{hidden}\left(\sum_{p=1}^{P} i_{a,p} w_{a,p} + b_a\right) \tag{3}$$

where $h_{hidden}(x) = \frac{1}{1+e^{-x}}$

When the network runs, each hidden layer node makes a calculation as per Eq. 3 on its inputs and transfers the result (O) to the next layer of nodes. In the above equation, O_a is the output of the current hidden layer node a, P is either the number of nodes in the previous hidden layer or number of network inputs, $i_{a,p}$ is an input to node a from either the previous hidden layer p or network input p, $w_{a,p}$ is the weight modifying the connection from either node p to node a or from input p to node a, and b_a is the bias. The subscripts a, p, and n in the given equations in this section identify nodes in the current layer, the previous layer, and the next layer, respectively. The sigmoid activation function was employed in this research. In the above equation, h_{hidden} (x) is the sigmoid activation function of the node. Before performing training process, the weights and biases were initialized to appropriately scaled values. Appropriate normalisation of training data was essential to avoid saturating the activation function. For output layer, multi linear activation function was used. So the output layer nodes perform the calculation as follows

$$O_a = h_{output}\left(\sum_{p=1}^{P} i_{a,p} w_{a,p} + b_a\right) \tag{4}$$

where $h_{output}(x) = x$

where, O_a is the output of the output layer node unit a, P is the number of nodes in the previous hidden layer, $i_{a,p}$ is an input to node a from the previous hidden layer node p, $w_{a,p}$ is the weight modifying the connection from node p to node a, and b_a is the bias. h_{Output} (x) is a multi linear activation function.

Before performing modelling, the input data for ANN has been normalized within the range of −1 to 1. The shape of the sigmoid function plays an important role in ANN learning. The weight changes corresponding to a value near −1 or 1 are minimal (Rao and Rao 1996). The following normalise equation was used for normalization:

$$x_{norm} = \frac{x_o - \overline{x}}{x_{\max} - x_{\min}} \tag{5}$$

where x_{norm}=normalized value; x_0=original value; $\overline{x}$ = mean; x_{max}=maximum value; and x_{min}=minimum value.

2.2.5 Support Vector Machines (SVM)

The SVMs for regression were first introduced in (1998) by Vapnik which was developed at AT&T Bell Laboratory by Vapnik and co-workers in the early 1990s. Just like ANNs, SVM can be represented as two-layer networks (where the weights are non-linear in the first layer and linear in the second layer).

Mathematically, a basic function for the statistical learning process is

$$y = f(x) = \sum\nolimits_{i=1}^{M} \alpha_i \varphi_i(x) = w\varphi(x) \tag{6}$$

where the output is a linearly weighted sum of M. The nonlinear transformation is carried out by $\varphi(x)$.

The decision function of SVM is represented as

$$y = f(x) = \left\{ \sum\nolimits_{i=1}^{N} \alpha_i K(x_i, x) \right\} - b \tag{7}$$

where $\boldsymbol{K}$ is the kernel function, α_i and $\boldsymbol{b}$ are parameters, $\boldsymbol{N}$ is the number of training data, $\boldsymbol{x_i}$ are vectors used in training process and $\boldsymbol{x}$ is the independent vector. The parameters α_i and $\boldsymbol{b}$ are derived by maximising their objective function.

The least squares approach prescribes choosing the parameters (**w**, b) to minimise the sum of the squared deviations of the data, $\sum_{i=f1}^{l} (y_i - < w \cdot x > -b)^2$ (Cristianini and Shawe-Taylor 2000). To allow for some deviation ε, between the eventual targets y_i and the function$f(x) = < w \cdot x > +b$, modelling the data, the following constraints are applied: $y_i - \mathbf{w} \cdot \mathbf{x} - b < \varepsilon$ and $y_i - \mathbf{w} \cdot \mathbf{x} + b \leq \varepsilon$

This can be visualised as a band or a tube around the hypothesis function *f(x)* with points outside the tube regarded as training errors, otherwise called slack variables ξ_i. These slack variables are zero for points inside the tube and increase progressively for points outside the tube. This approach to regression is called ε -SV regression and it is the most common approach.

The task is now to minimise $\|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \left(\xi_i + \xi_i^*\right)$ subject to: $y_i - \mathbf{w} \cdot \mathbf{x} - b \leq \varepsilon + \xi_i$ and $(\mathbf{w} \cdot \mathbf{x} + b) - y_i \leq \varepsilon + \xi_i^*$

An alternative form of SVM is called ***n***SV regression. This model uses ν to control the number of support vectors. Given a set of data points, $\{(x_1,z_1),\ldots(x_l,z_l)\}$, such that $x_i \in R^n$ is an input vector and $z_i \in R^1$ the corresponding target, the form is:

$$\min_{w,b,\xi,\xi^*} \tfrac{1}{2} w^T w + C\left(v\varepsilon + \tfrac{1}{7} \sum_{i=1}^{l} (\xi_i + \sum_{i=1}^{l} \xi_i^* \right)$$

Subject to: $w^T\phi(x_i) + b - z_i \leq \varepsilon + \xi_i$ and $z_i - w^T\phi(x_i) - b \leq \varepsilon + \xi_i^*$ with ξ is the upper training bound and ξ_i^* the lower training bound.

The role of the kernel function simplifies the learning process by changing the representation of the data in the input space to a linear representation in a higher-dimensional space called a feature space. A suitable choice of the kernel allows the data to become separable in the feature space despite being non-separable in the original input space. Four standard kernels are usually used in classification and regression cases: linear, polynomial, radial basis and sigmoid.

$$\begin{array}{ll} \textit{Linear} & u' \times v \\ \textit{Polynomial} & \left(\gamma \times u' \times v + coef\right)^{\deg ree} \\ \textit{Radial basis} & e^{-\gamma \times |u-v|^2} \\ \textit{Sigmoid} & \tanh\left(\gamma \times u' \times v + coef\right) \end{array}$$

A number of support vector machine software packages are now available. The software used in this project was LIBSVM developed by Chih-Chung Chang and Chih-Jen, and supported by the National Science Council of Taiwan (Chang and Lin 2011). Figure 6 illustrates the SVM layout describing the processes carried out in this study. We have tried SVM modelling with different kernel functions and different SVR types (ν-SV regression and ε-SV regression). Note that, the results from the procedures set by (Bray and Han 2004), it was found that the ε-SV regression and linear kernel had better performance than the remaining models.

The deviation between the target value and the function describing the hypothesis found by the support vector machine is controlled by the ε parameter. ε Values were varied between ε=1 to ε=0.00001 (the default value is ε=0.01) whilst keeping all other parameters fixed at their default values. However, in this study the ε value was set as 1. If the data is of good quality, the distance between the two hyperplanes is narrowed down. If the data is noisy, it is preferable to have a smaller value of *C* which will not penalise the vectors. In this study, the cost value was chosen to be 10.

2.3 Statistical Parameters

In this study, we have compared the MM5 downscaled and error corrected values of wind speed with the HYREX land based observed data. Although there are many statistical

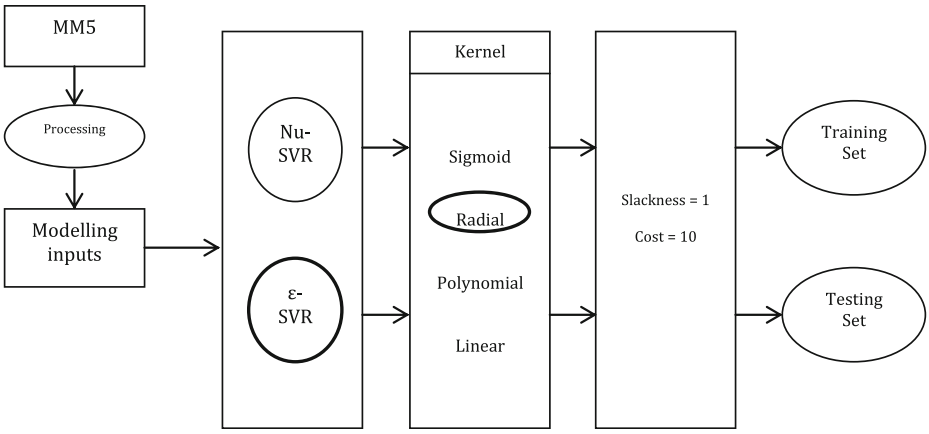

Fig. 6 The SVM based hybrid modelling scheme used in this study

indices available, the study has focused on two indices, root mean square error (RMSE) and mean bias error (MBE):

$$RMSE = \sqrt{\left(\frac{1}{n}\sum_{i-1}^{n}[y_i(i) - x_i(i)]^2\right)} \tag{8}$$

$$MBE = \frac{\sum_{i=1}^{n}(y_i - x_i)}{n} \tag{9}$$

where n is the number of observations; x_i=observed variable and y_i=estimated variable. The RMSE and MBE values are expressed as a percentage of the mean value of the observed data.

3 Results and Discussions

This section gives an overview of the four models performance on error correction methodology based on selected input variables.

3.1 Selection of Model Inputs

The models reported in this paper were developed to correct the wind speed obtained from MM5 [WndMM5(t)], using different hourly sets of data like MM5 derived air temperature [TmpMM5(t)], MM5 derived atmospheric pressure [PrsMM5(t)], MM5 derived relative humidity [RhMM5(t)], MM5 derived solar radiation [SolarMM5(t)], and MM5 derived rainfall [RfMM5(t)] with the observed HYREX wind velocity [WndOBS(t)] as the target data set. The study has used the MM5 outputs directly for modelling, without performing any bias correction for individual variables. It can be reasonably assumed that the correction models should be able to correct the biases in the input variables during the training process. A traditional approach to find the dominant inputs is cross correlation method. Normally in this approach, researches depend on linear cross-correlation analysis to determine the strength of the relationship between the input time series and the output time series (Haugh and Box 1977). The disadvantage associated with this method is its inability to capture any nonlinear dependence that may exist between the inputs and the output.

Table 1 shows the correlation coefficients of input data series with both training and testing data sets. The correlations are higher for the MM5 derived wind speed with the value of 0.69 during the training period while 0.70 during the testing period. The second higher correlation values are associated with the MM5 derived pressure values in both training and testing period with values of −0.30 and −0.40 respectively. The MM5 derived rainfall and surface temperature

Table 1 Correlation values between the observed wind speed (WndOBS) and input variables for training and testing results

	WndMM5	TmpMM5	RhMM5	PrsMM5	RfMM5	SolarMM5
1995 and 1996 (Training phase)						
WndOBS	0.6992	−0.1887	−0.1039	−0.3040	0.0650	0.1085
1998 (Testing phase)						
WndOBS	0.7029	0.0094	−0.1294	−0.4001	0.1267	0.1087

have shown weak correlation during the training and testing phases. Based on the correlation outputs, one could easily point out that the dominant inputs have a trend as follows- WndMM5 >PrsMM5>SolarMM5>RhMM5>TmpMM5>RfMM5. In addition, this study has also adopted forward selection approach to identify suitable input combinations for modelling. The adopted forward selection involves using a single dataset from the available input space for modelling and to identify the best input which gives optimised training and testing results. In the next step, this modelling is repeated with two inputs keeping the best input fixed and varying other input series. The performance of the forward selection is evaluated based on the value of RMSE in each model. We have adopted this approach for all four models in this study. The best model input structure obtained are shown in Table 2 and the corresponding figures for linear regression, nonlinear regression, ANNs and SVMs are shown in Fig. 7a, b, c and d respectively.

Table 2 Model selection showing RMSE for nonlinear form function, multi linear form function, ANN model and SVM model

List of Variables	RMSE Training (m/s) 1995 and 1996	RMSE Testing (m/s) 1998
For Power form (Nonlinear model)		
WndMM5	0.9895	0.9901
Wnd+TmpMM5	0.9697	0.9705
Wnd+Tmp+RfMM5	0.9691	0.9702
Wnd+Tmp+Rf+PrsMM5	0.9683	0.9694
Wnd+Tmp+Rf+Prs+RhMM5	0.9681	0.9693
Wnd+Tmp+Rf+Prs+Rh+SolarMM5	0.9677	0.9692
For linear form (Multilinear model)		
WndMM5	0.9878	0.9883
Wnd+TmpMM5	0.9825	0.9832
Wnd+Tmp+RhMM5	0.9801	0.9809
Wnd+Tmp+Rh+SolarMM5	0.9779	0.9789
Wnd+Tmp+Rh+Solar+PrsMM5	0.9627	0.9778
Wnd+Tmp+Rh+Solar+Prs+RfMM5	0.9767	0.9792
	Normalised data	
For ANN model		
WndMM5	0.0982	0.1076
Wnd+SolarMM5	0.0954	0.1048
Wnd+Solar+RfMM5	0.0933	0.1034
Wnd+Solar+Rf+RhMM5	0.0931	0.1031
Wnd+Solar+Rf+Rh+TmpMM5	0.0930	0.1051
Wnd+Solar+Rf+Rh+Tmp+PrsMM5	0.0927	0.1065
	Scaled data	
For SVM model		
WndMM5	0.1112	0.1296
Wnd+SolarMM5	0.1093	0.1278
Wnd+Solar+RfMM5	0.1080	0.1258
Wnd+Solar+Rf+RhMM5	0.1078	0.1252
Wnd+Solar+Rf+Rh+TmpMM5	0.1075	0.1255
Wnd+Solar+Rf+Rh+Tmp+PrsMM5	0.1074	0.1389

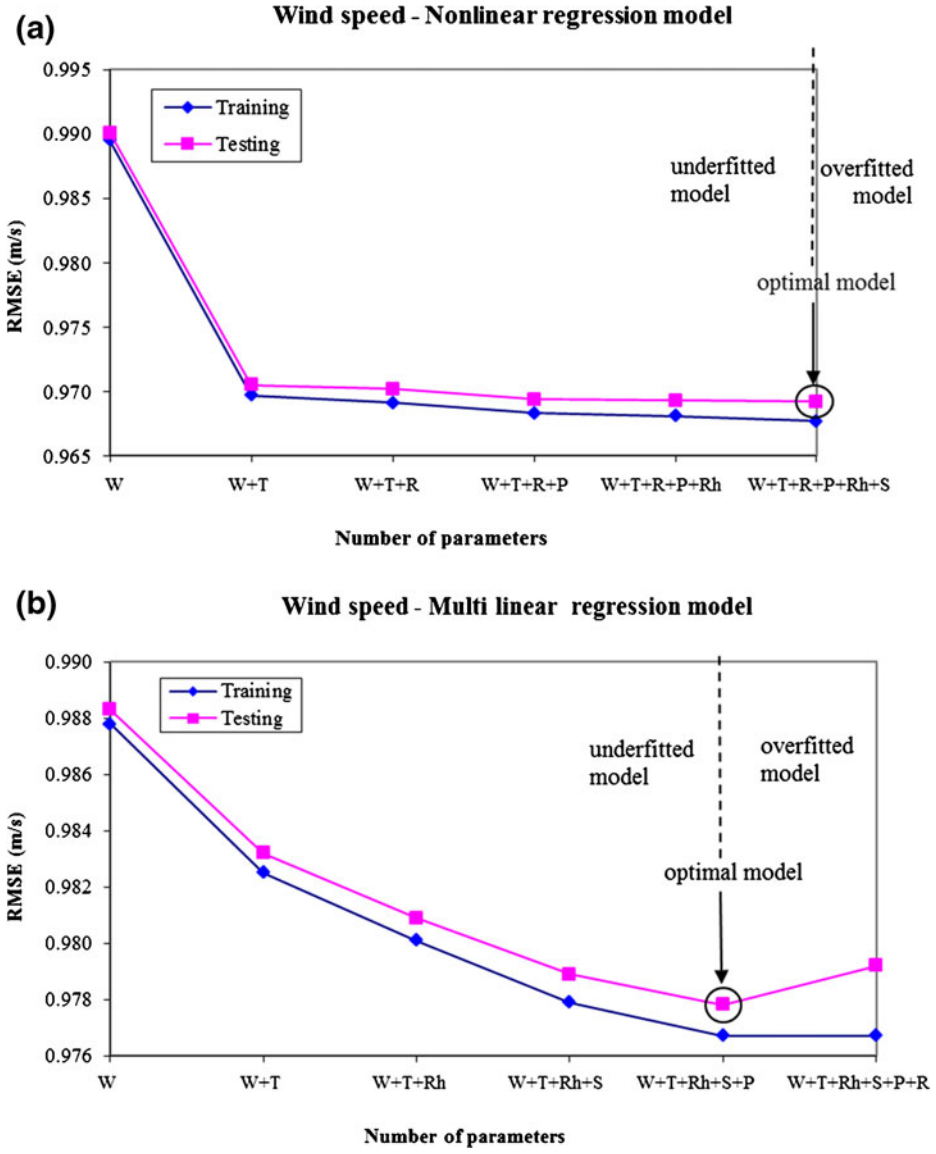

Fig. 7 Results showing the model selection method (**a**) nonlinear form model; (**b**) multi linear model; (**c**) ANN model and (**d**) SVM model

Various combinations based on six input variables were tested for all the models where the objective was to find the best combination with the least value of RMSE. For example, in Fig. 7a, for NLR, wind speed (W) performed the best within the six variables with the lowest of RMSE value. Meanwhile the lowest values of RMSE for two combinations of variables are wind speed and temperature (W+T); as for three best combinations are wind speed, temperature and rainfall (W+T+R); and so on. Similar descriptions can be applied for Fig. 7b, c and d.

In this regards, Fig. 7a describes the model selection for nonlinear power form function. It has shown that, the combination of the MM5 derived wind speed, surface temperature,

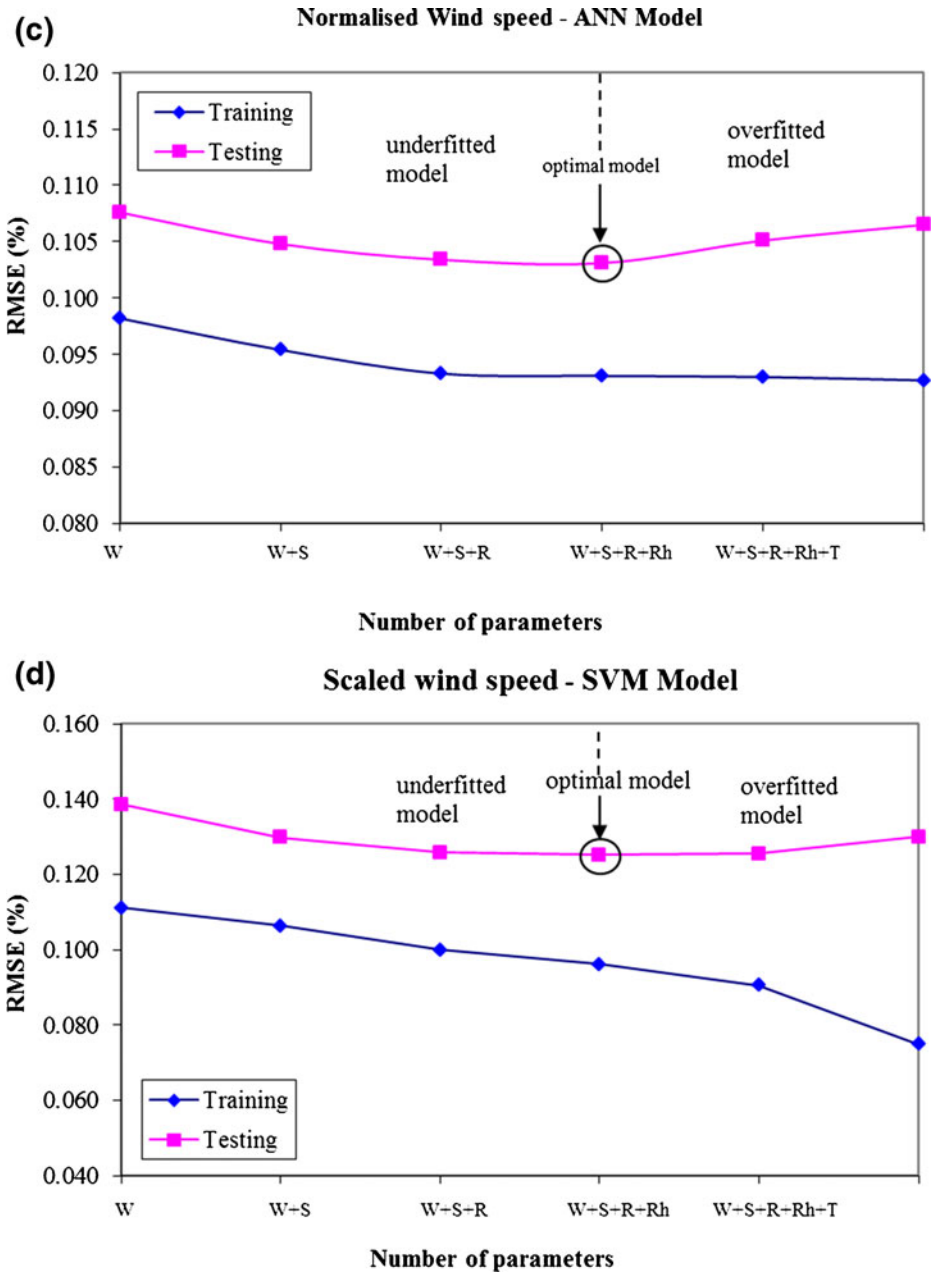

Fig. 7 (continued)

rainfall, atmospheric pressure, relative humidity and solar radiation can produce a better model with the least value of RMSE in the case of nonlinear regression models. The corresponding RMSE values can be found in Table 2. Whereas in the case of multi linear regression models, the best input combination of selection is identified as the MM5 derived wind speed, surface temperature, relative humidity, solar radiation and surface pressure.

When rainfall data were added to this combination, the RMSE value at the testing phase changed to a higher value, but the RMSE value during training phase remained unchanged (Table 2). In the case of ANN model, the best inputs are identified as a combination of the MM5 derived wind speed, solar radiation, rainfall and relative humidity. This combination has been identified considering the best RMSE values during the testing phase. For the SVM model, unlike ANN, a combination of wind speed, solar radiation, rainfall and relative humidity (W + S + R + Rh) is the best.

3.2 Application of Different Wind Velocity Error Correction Methods

This section describes the four models (multi linear regression, nonlinear regression, ANNs and SVMs) used for the MM5 derived wind velocity error correction at the Brue catchment.

3.2.1 Modelling with Multi Linear Regression (MLR) and Nonlinear Regression (NLR) Models

Before implementation of the multi linear and nonlinear regression models, it is important to standardize the input data with the target output ranging either for X_{max}, X_{mean} or X_{min}. The statistical details of the observed wind speed and other MM5 derived inputs are shown in Table 3. After standardization, the multi linear and nonlinear regression equations were modelled according to Eqs. 1 and 2 respectively. The study has followed the suggestions from the forward selection method, in which nonlinear regression model gave better results for the combination of [WndMM5, TmpMM5, RfMM5, PrsMM5, RhMM5, SolarMM5] with RMSE value of 0.967 m/s and 0.969 m/s during the training and testing phase respectively. The optimum nonlinear regression model is given in Eq. 10. This generalisation of the model is assessed based on its performance on the testing dataset as shown in Table 3 and Fig. 7a.

$$Y = 9.9537.WndMM5^{1.0303} \times \left(\frac{TmpMM5 + 10}{10}\right)^{-0.2715} \times (RfMM5 + 1)^{-0.1111} \times \left(\frac{\mathrm{Pr}\,sMM5}{100}\right)^{-1.1056} \times \left(\frac{RhMM5}{10}\right)^{-0.1327} \times \left(\frac{SolarMM5 + 1}{100}\right)^{-0.0093} \ldots\ldots\ldots \quad (10)$$

The performance of each model is indicated by the RMSE value on the training and testing (see Table 2). Generalization of the model is assessed based on its performance on the testing dataset. In the case of the multi linear regressive function model, [WndMM5, TmpMM5, RhMM5, SolarMM5, PrsMM5] input combination has shown better performance with RMSE value of 0.962 m/s and 0.978 m/s during training and testing periods respectively. The optimal multi linear regression model with five input variables and the corresponding parameters are shown in Eq. 11. In general, the negative terms in Eq. 11 indicates that those particular input parameters decrease while the output increases.

$$Y = 6.8047 + 0.5338.WndMM5 - 0.2460 \times \left(\frac{TmpMM5 + 10}{10}\right) - 0.1201 \times \left(\frac{RhMM5}{10}\right) - 0.0493 \times \left(\frac{SolarMM5 + 1}{100}\right) - 0.5203 \times \left(\frac{\mathrm{Pr}\,sMM5}{100}\right) \ldots\ldots\ldots \quad (11)$$

The values of RMSE and bias obtained after wind speed corrections based on MLR are given in Table 4 and 5 corresponding to the training and testing phases respectively. The time series plot after error correction with multiple linear regression model on training set is given in Fig. 8 (top).

Table 3 The information on statistical parameters for the observed wind speed (target output) and six input variables from the MM5 derived

	Year 1995 and 1996 Training Phase							Year 1998 Testing Phase						
	Wnd OBS (m/s)	WndMM5 (m/s)	TmpMM5 (°C)	RhMM5 (%)	PrsMM5 (mb)	RfMM5 (mm)	SolarMM5 (W/m2)	Wnd OBS (m/s)	WndMM5 (m/s)	TmpMM5 (°C)	RhMM5 (%)	PrsMM5 (mb)	RfMM5 (mm)	SolarMM5 (W/m^2)
Xmean	1.730	3.310	10.016	84.602	1014.434	0.050	110.083	1.995	3.575	10.438	86.040	1012.399	0.072	88.754
Xmax	8.935	12.816	28.973	100.000	1039.289	5.243	926.171	8.575	11.198	23.755	100.000	1035.346	3.781	927.963
Xmin	0.000	0.066	−8.547	27.296	978.118	0.000	0.000	0.000	0.238	−4.122	39.387	978.424	0.000	0.000

Table 4 Statistical indices showing performance of wind speed error correction models in training phase

	MM5		MLR (5VAR) (normalised data)		NLR (6VAR) (normalised data)		ANN (4VAR) (de-normalised)		SVM (1VAR) (unscaled)	
	(m/s)	%	(m/s)	%	(m/s)	%	(m/s)	%	(m/s)	%
1995 and 1996 data (training phase)										
Bias	1.5797	91.324	0.000	0.007	0.009	0.536	0.000	−0.003	0.005	0.311
RMSE	2.0020	115.733	0.962	55.607	0.967	55.928	0.898	51.890	0.963	55.658

The corresponding plots on testing data set are given in Fig. 8 (bottom). Before error correction, the MM5 derived wind velocity has shown higher values of bias and RMSE in comparison to the observed wind velocity for both selected training and testing sets. During 1995–1996 period (training period for the correction techniques) the MM5 simulated wind velocity has shown bias value of 1.58 m/s (91.3 %) and corresponding values for RMSE were 2.00 m/s (115.7 %). The MM5 simulation results during the year 1998 (testing period for the error correction models) have shown higher bias and RMSE values of 1.53 m/s (83.1 %) and 1.95 m/s (105.9 %) respectively. After the MLR modelling, bias values were considerably reduced to 0.007 % during the training period and −6.516 % (slight under estimation) during the testing period. The corresponding RMSE values are reduced to 55.6 % and 49.0 % during the training and testing periods respectively.

The results obtained from the nonlinear regression model are also given in Table 4 which emphasizes the close performance of the NLR model to that of the linear regression model. However for better visual understanding of the model accuracy, line plots between measured and NLR model error corrected wind velocity are shown in the Fig. 9 corresponding to training and testing phases.

3.2.2 Modelling with ANN Model

This study also makes an evaluation of the use of artificial neural network models to correct the distorted wind velocity time series data obtained from the MM5 simulation in the Brue catchment. The four member input structure (WndMM5, SolarMM5, RfMM5, RhMM5) is identified as the optimal, i.e. MM5 derived wind speed, solar radiation, rainfall and relative humidity. Just like the previous two models, the ANN model has used the data from years 1995 and 1996 (5,904 data points) for training and year 1998 data (2,032 data points) for testing. The time series plots of LM algorithm based ANN model results obtained in this error correction study during training and testing are given in the Fig. 10. The ANN model produced the wind speed with RMSE values of 0.898 m/s (51.9 %) during the training phase and 1.11 m/s

Table 5 Statistical indices showing performance of wind speed error correction models in testing phase

	MM5		MLR (5VAR)		NLR (6VAR)		ANN (4VAR)		SVM (4VAR)	
	(m/s)	%	(m/s)	%	(m/s)	%	(m/s)	%	(m/s)	%
1998 data (testing phase)										
Bias	1.5254	83.066	−0.130	−6.516	−0.155	−7.750	−0.079	−3.935	−0.015	−0.770
RMSE	1.9452	105.927	0.978	49.018	0.969	48.577	1.111	55.696	1.074	53.839

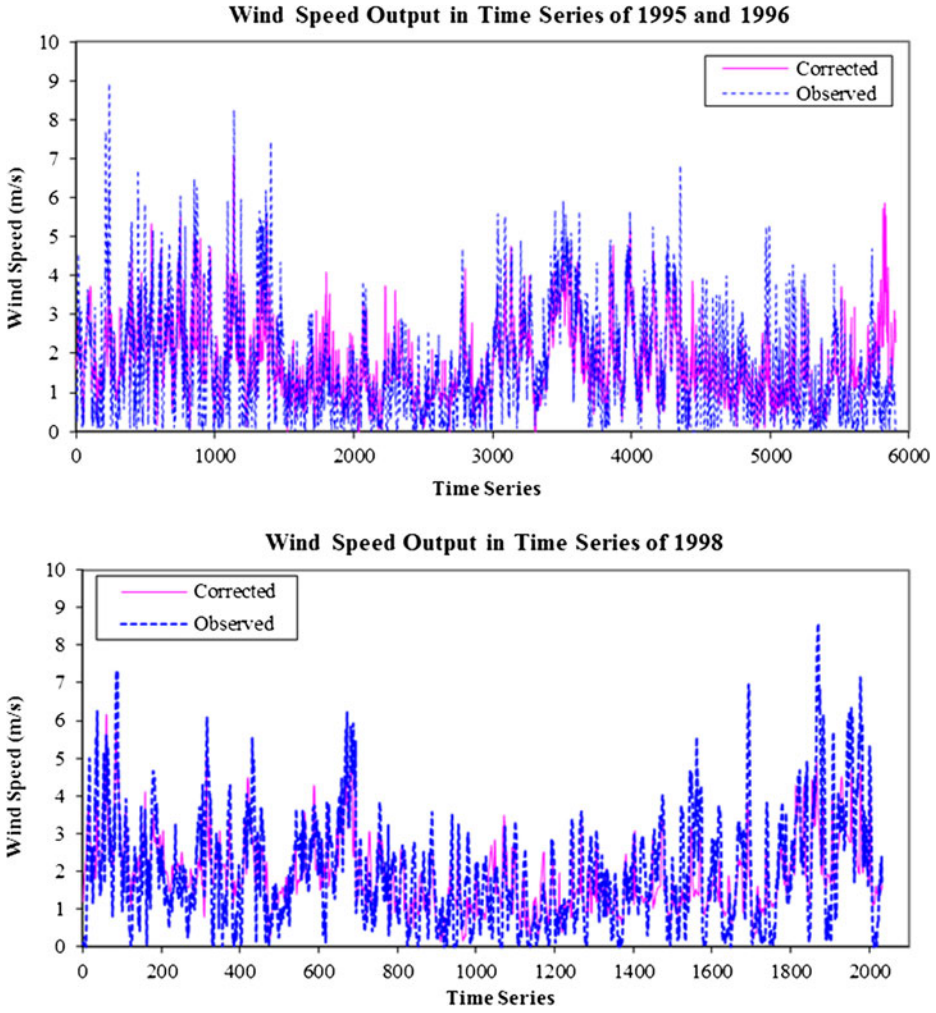

Fig. 8 The *line plots* of MLR corrected wind velocity and observed wind velocity at the Brue catchment during training phase (*top*); testing phase (*bottom*)

(55.70 %) during the testing phase. The bias values observed in the ANN model during the training phase were very close to zero, whereas during the testing phase, mean bias error (MBE) were observed as −0.079 m/s, which is −3.94 % of the mean observed HYREX wind speed during that study period. The analysis results in comparison with other models are given in Table 4 and Table 5 for the training and testing phases respectively. The bias values were 91.3 % for the MM5 simulation results during 1995–1996 in comparison to the mean observed wind velocity. The ANN modelling has considerably reduced this higher bias values to −0.003 % and a similar trend could also be observed in the training phase. Albeit the ANN gave better training results than the regression models, it slightly failed to show better skills in comparison with MLR and NLR function models during testing. The ANN model was trained using the same training data set and testing set as used for the regression models, so the reason for the disparity could be associated with inputs used for the model. The MLR model (5 input) gave better testing results than NLR model (6 input, the extra input is MM5 derived rainfall) and ANN

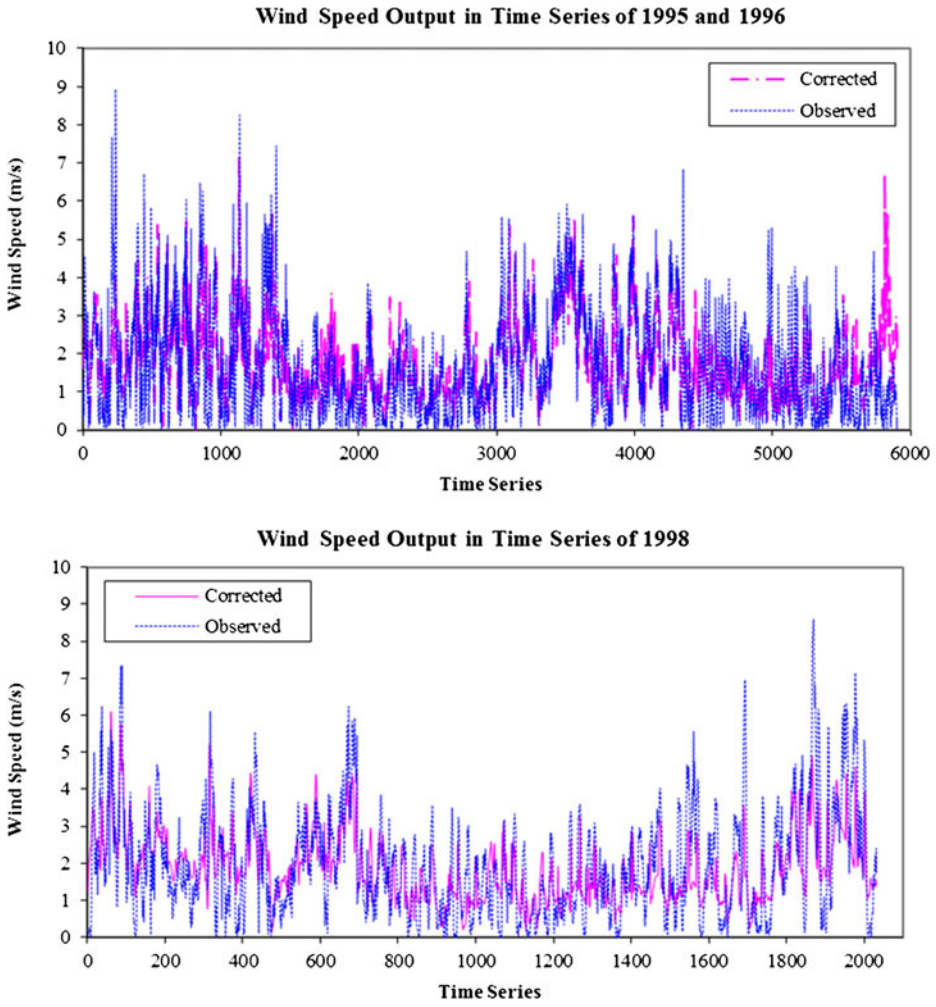

Fig. 9 The *linear plots* of NLR corrected wind velocity and observed wind velocity during training phase (*top*); testing phase (*bottom*)

model (4 input). The overall performance of the regression models in the testing phase indicates that such simple models are equally good as the ANN models to make reasonably good results in error correction modelling.

3.2.3 Modelling with SVMs

The study has explored error correction capability of support vector machines in wind velocity modelling on the data obtained from the MM5 simulation. The study has used four inputs for modelling as suggested by the model selection method. The statistical performance of support vector machine (SVM) technique with ε-SVR and linear kernel is presented in Tables 4 and 5 corresponding to the training and testing phases. It can be obviously seen from Table 5 that the SVM model approximates the measured values with the lowest value of bias during the testing phase than that of ANN, MLR and NLR models. In

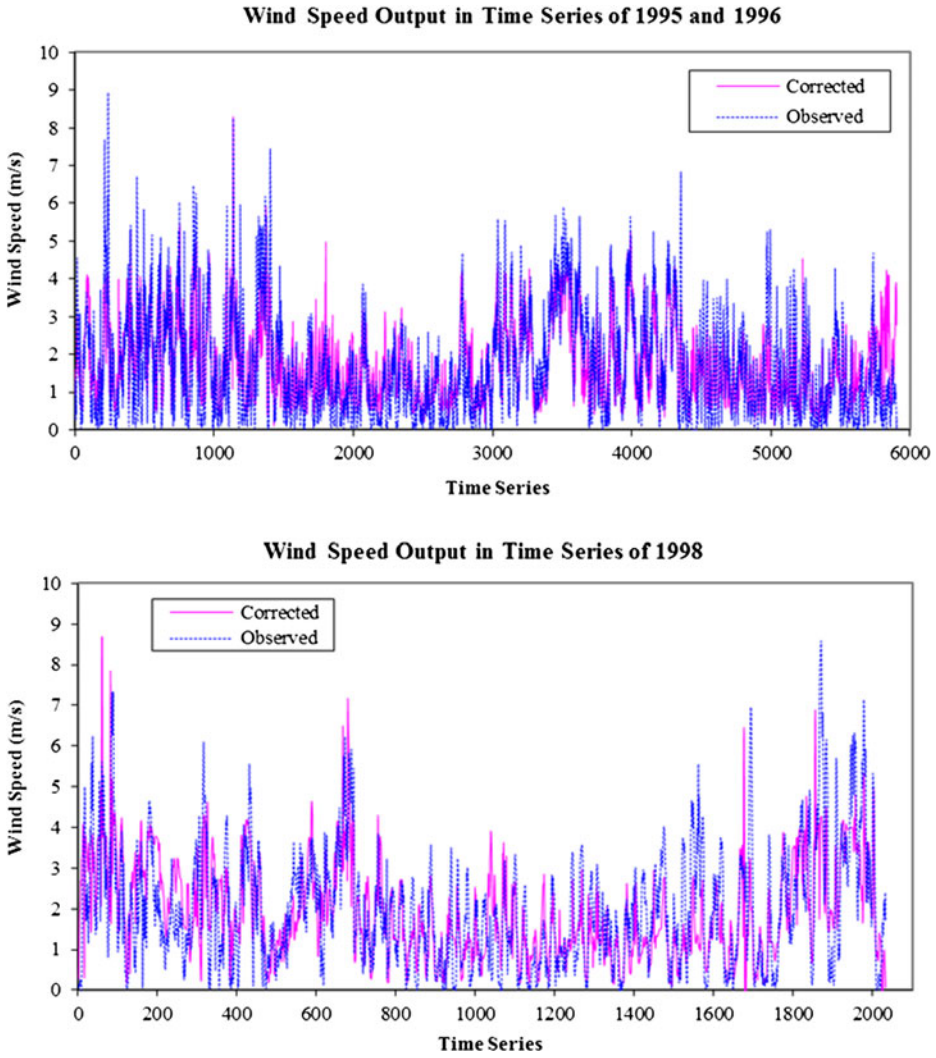

Fig. 10 The *line diagrams* of ANN corrected wind velocity and observed wind velocity at the Brue catchment during training phase (*top*); testing phase (*bottom*)

the training phase, the SVM model has shown better results in bias and RMSE as compared to the NLR model but weaker than ANN and MLR. The SVM model has better modelling results with RMSE value of 1.074 m/s (53.8 %) and mean bias error value of −0.015 m/s (−0.77 %) during the testing phase. The corresponding value during the training phase was 0.963 m/s (55.7 %) and 0.005 m/s (0.31 %) respectively. It was interesting to note that the performance of the multi linear regression in terms of RMSE was better than that of the SVM model during both the testing and training periods. The results have shown that the performance of MLR models is closer to that of ANNs and SVMs. But the significance of MLR is higher when we consider the modeller has to perform the tedious trial and error procedure to develop the optimal network architecture of ANNs/SVMs, while such a procedure is not required in developing simpler regression models. The observed and error

corrected wind speed values of the SVM model for the training and testing data are given in Fig. 11 (top and bottom respectively). In general, these results indicate that the error correction performance of the SVM model is better than that of the ANN model, because of its better predictability in the testing data set.

4 Conclusions

The main aim of this study was to develop a hybrid system with the MM5 model to modify the distorted wind speed data applied to the Brue catchment of the Southeast England. For

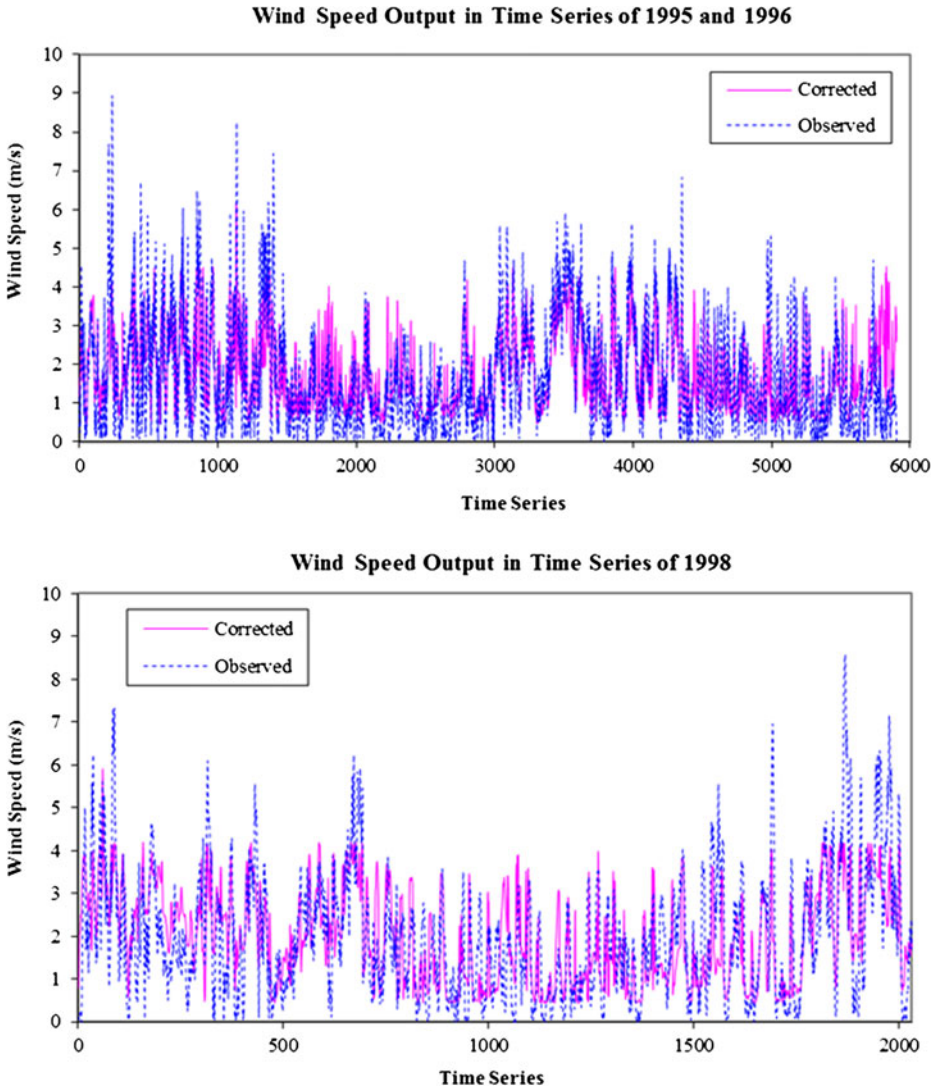

Fig. 11 The *line diagrams* of SVM corrected wind velocity and observed wind velocity at the Brue catchment during training phase (*top*); testing phase (*bottom*)

this purpose, two regression systems (with linear model (MLR) and nonlinear model (NLR)) and two AI systems (with ANNs and SVMs) were developed in conjugation with MM5, and their performances were inter-compared in error correction modelling. The input vector selections for these models are tricky part of this modelling scheme, which were performed through quantification of the statistical properties. Various outputs from the MM5 downscaling model were analysed and optimum input structure for each model was identified. The optimisation with the model input selection technique has identified the best input combinations for multi linear (MLR) model as [WndMM5, TmpMM5, RhMM5, SolarMM5, PrsMM5] while that of nonlinear form (NLR) model as [WndMM5, TmpMM5, RfMM5, PrsMM5, RhMM5, SolarMM5]. The AI models like ANNs and SVMs have shown better performance on four input variables [WndMM5, SolarMM5, RfMM5, RhMM5]. The exclusion of PrsMM5 could be managed with the presence of RhMM5. The inter-comparison of different hybrid schemes have shown that relatively simpler models like MLR have given reliable and close results to those of complex ANNs and SVMs during the testing phases. It is observed that the NLR model is capable of producing better statistical properties of the wind speed time series during the testing phase than those of ANNs but not SVMs. The SVM based scheme was observed as more robust than ANNs and regression models on unseen data sets, though its statistical values during the training phases were weaker. However if we consider difficulties in trial and error procedures associated with ANNs and SVMs, the regression based models may hold an upper hand. The improved performance of regression models may be because of a higher number of inputs in the model structure. In addition, this improvement is also influence from the well performed results during training and testing phases. This study highly depended on the model input selection approach; however, more studies using the same input series may be required to reinforce this conclusion. Error correction studies of this kind have useful implications in hydrology and earth sciences, especially in ungauged catchments as the inputs used for modelling can be directly obtained from the MM5 simulation models. One weakness of the models is their inability to capture small part of the training data near the end with significant overestimation. The overestimation is mainly due to the overestimation by the MM5 model near the end. The correction models are able to learn and correct the majority part of the training data, but unable to learn the part which departs from the majority patterns. In general, the results of the study are highly encouraging and suggest that all four models can provide reasonably reliable results using the MM5 derived variables as inputs.

Acknowledgements This research is funded by the Public Services Department of Malaysian Government. We also acknowledge the support from the Irrigation and Drainage Department, Malaysian Government. Many thanks are expressed to the anonymous reviewers for their valuable comments.

References

Allen RG, Pereira LS, Raes D, Smith M (1998) Crop evapotranspiration - Guidelines for computing crop water requirements - FAO Irrigation and drainage paper 56. Food and Agriculture Organization of the United Nations, Rome

Bray M, Han D (2004) Identification of support vector machines for runoff modelling. J Hydroinf 6(4):265–280

Broyden CG (1970) The convergence of a class of double-rank minimization algorithms 1. General considerations. IMA J Appl Math 6(1):76–90

Cawley GC, Talbot NLC (2003) Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recogn 36(11):2585–2592

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27

Chauhan S, Shrivastava R (2009) Performance evaluation of reference evapotranspiration estimation using climate based methods and artificial neural networks. Water resour manag 23(5):825–837
Chen F, Dudhia J (2001) Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: model implementation and sensitivity. Mon Weather Rev 129(4):569–585
Chen ST, Yu PS (2007) Real-time probabilistic forecasting of flood stages. J Hydrol 340(1–2):63–77
Cristianini N, Shawe-Taylor J (2000) An introduction to support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge
Dudhia J (1993) A nonhydrostatic version of the Penn State-NCAR mesoscale model: validation tests and simulation of an atlantic cyclone and cold front. Mon Weather Rev 121(5):1493–1513
Efron B (1986) How biased is the apparent error rate of a prediction rule? J Am Stat Assoc 81(394):461–470
Fletcher R (1987) Practical Methods of Optimization, second edition. John Wiley, Chichester
Frank WM (1983) The cumulus parameterization problem. Mon Weather Rev 111:1859–1871
Grell GA (1995) A description of the fifth-generation Penn State/NCAR mesoscale model (MM5), NCAR Technical Note, Colorado and Pennsylvania, USA
Grell GA, Dudhia J, Stauffer DR, Mesoscale NCfAR, Dicision MM (1994) A description of the fifth-generation Penn State/NCAR mesoscale model (MM5), NCAR Technical Note, Colorado and Pennsylvania, USA
Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: Data mining, inference, and prediction. Springer Series in Statistics. Springer-Verlag, New York, p 533
Haugh LD, Box GEP (1977) Identification of dynamic regression (distributed lag) models connecting two time series. J Am Stat Assoc 72(357):121–130
Ishak AM, Bray M, Remesan R, Han D (2010) Estimating reference evapotranspiration using numerical weather modelling. Hydrol Processes 24(24):3490–3509
Ishak AM, Bray M, Remesan R, Han D (2012) Seasonal evaluation of rainfall estimation by four cumulus parameterization schemes and their sensitivity analysis. Hydrol Processes 26(7):1062–1078
Islam T, Rico-Ramirez MA, Han D, Srivastava PK (2012a) Artificial intelligence techniques for clutter identification with polarimetric radar signatures. Atmos Res 109–110:95–113
Islam T, Rico-Ramirez MA, Thurai M, Han D (2012b) Characteristics of raindrop spectra as normalized gamma distribution from a Joss–Waldvogel disdrometer. Atmos Res 108:57–73
Kashyap PS, Panda R (2001) Evaluation of evapotranspiration estimation methods and development of crop-coefficients for potato crop in a sub-humid region. Agric Water Manag 50(1):9–25
Mass CF, Kuo YH (1998) Regional real-time numerical weather prediction: current status and future potential. Bull Am Meteorol Soc 79(2):253–264
Møller MF (1993) A scaled conjugate gradient algorithm for fast supervised learning. Neural netw 6(4):525–533
Rao V, Rao H (1996) C++ Neural networks and fuzzy logic, BPB. New Delhi, India, pp 380–381.
Salcedo-Sanz S, Pérez-Bellido ÁM, Ortiz-García EG, Portilla-Figueras A, Prieto L, Paredes D (2009) Hybridizing the fifth generation mesoscale model with artificial neural networks for short-term wind speed prediction. Renew Energy 34(6):1451–1457
Thomas DM, Benson MA (1970) Generalization of streamflow characteristics from drainage-basin characteristics. U. S. Geol. Surv. Water Supply Pap. 1975:55
Vapnik V (1998) The support vector method of function estimation. Nonlinear Model Adv Black-Box Tech 55:86
Wang WC, Chau KW, Cheng CT, Qiu L (2009) A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. J Hydrol 374(3):294–306
Warner TT, Kibler DF, Steinhart RL (1991) Separate and coupled testing of meteorological and hydrological forecast models for the Susquehanna river basin in Pennsylvania. J Appl Meteorol 30:1521–1533
Yang K, Huang G, Tamai N (2001) A hybrid model for estimating global solar radiation. Sol Energy 70 (1):13–22
Yu PS, Chen ST, Chang IF (2006) Support vector regression for real-time flood stage forecasting. J Hydrol 328(3):704–716
Zhong S, Fast J (2003) An evaluation of the MM5, RAMS, and Meso-Eta models at subkilometer resolution using VTMX field campaign data in the Salt Lake Valley. Mon Weather Rev 131(7):1301–1322
Zhu YM, Lu X, Zhou Y (2007) Suspended sediment flux modeling with artificial neural network: an example of the Longchuanjiang river in the Upper Yangtze Catchment, China. Geomorphology 84(1):111–125