Error correction modelling for high resolution dataset

Asnor Muizan Ishak[a,b], Renji Remesan[b] and Dawei Han[a]

[a] Water and Environmental Management Research Centre (WEMRC), Department of Civil Engineering, Faculty of Engineering, University of Bristol, BS8 1TR Bristol, United Kingdom.

[b] Water Resources Management and Hydrology Division, Department of Irrigation and Drainage Malaysia, KM 7, Jalan Ampang, Ampang, 68000 Kuala Lumpur, Malaysia.

[C] Research Fellow in Modelling Hydrological Impacts of Climate Change, Environmental Science and Technology Department, Cranfield, United Kingdom.

drasnorjps@gmail.com

Abstract

The hybridisation of a fifth-generation mesoscale model (MM5) comprising four statistical models—multiple linear regression (MLR), nonlinear regression (NLR), artificial neural network (ANN) and support vector machine (SVM)—facilitates the improvement in accuracy of downscaling weather variables derived from ECMWF ERA40-reanalysis data. This study will explore the best combination of variable selection by using forward-based selection. All the analysis is based on hourly time interval data. This paper explores in detail the regression models and the hybrid approach which uses the outputs from MM5 for improvements in statistical modelling. This study is also a first-time attempt to decide the best combination of weather variables for improvements in dynamical downscaling output. Through this study, new corrected values for each weather variable can be developed based on the best model to produce the corrected values of weather variables in the testing phase for hydrological modelling purposes.

Keywords: SVM, ANN, Statistical models, MM5, Error correction, Weather parameters

1. Introduction

It is indubitable that hydrological/weather modelling accuracy relies significantly on accurate hydro-meteorological data (Chowdhury and Ward, 2004; Coulibaly, 2003; Werner et al., 2005). The main input parameters of many hydrological models are evapotranspiration and rainfall (Ishak et al., 2010). However, accurate estimation of the above parameters also depends on the wind speed, surface temperature, relative humidity, and solar radiation. Hence, these other weather parameters are important in the hydrological cycle, particularly for hydrological modelling and forecasting (Chahine, 1992; Fowler et al., 2007; Heuvelmans et al., 2006).

The most common use of weather parameters is the estimation of reference evapotranspiration (ET_o) (Meza, 2005). As a weather parameter, ET_o can be computed from weather data (Xu et al., 2006). Since water is abundantly available at reference evapotranspiration surfaces and soil factors do not affect ET_o (Allen et al., 1998), estimation is derived mainly by calculating the wind speed, surface temperature, relative humidity, solar radiation and surface pressure (Allen et al., 1998; Sentelhas et al., 2010). The Penman-Monteith equation has been proven as the best for estimating ET_o and has been published in the FAO-56 report (Allen et al., 1998).

Another approach to reference ETo estimation is through downscaled weather variables from numerical weather prediction models (Ishak et al., 2010). However, downscaled weather variables tend to be erroneous and thus unreliable for ET_o estimation. For instance, Ishak showed that the error in the essential weather parameter of wind speed for ET_o estimation was about 200-400% through mesoscale model (MM5) downscaling. Other input parameters showing errors were very high air temperature (<10%), relative humidity (5–21%) and net radiation (4–23%). The only exception was for atmospheric pressure, which was accurately derived with less than 0.2% error. Thus, with the exception of atmospheric pressure, the other four weather variables must be improved before estimating evapotranspiration using MM5 downscaling with reliable accuracy (Ishak et al., 2010, 2011).

This paper demonstrates the improvement of three weather variables (surface temperature, relative humidity and solar radiation) and rainfall by using four types of empirical mathematical models. These mathematical models—multiple linear regression, nonlinear regression, artificial neural network and support vector machine—are proven tools for improving weather variables (Ghosh and Mujumdar, 2008; Heuvelmans et al., 2006). They have been successfully used in many hydrological problems involving river level forecasting, rainfall-runoff modelling, rainfall forecasting, ground water modelling, water quality prediction, and water resources management and operation (Ghosh and Mujumdar, 2008; Khalil et al., 2005; Oommen et al., 2007; Tripathi et al., 2006; Yoon et al., 2010; Young, 2002).

In this paper, the development of a hybrid system using mathematical models for the improvement of key hydrological and weather variables is explored. The paper is structured as follows: the description of the study area and observed data is provided in Section 2. Section 3 summarises the MM5 modelling and downscaling set-up. In Section 4, the four empirical mathematical models are applied to error correction of the four weather variables. The selection of input variables to the empirical models is illustrated in Section 5, while the error correction results using the models are elaborated in Section 6. The paper ends with a conclusion in Section 7.

2 Study Area and Observation Data

The Brue catchment in the United Kingdom was chosen as the study area, as shown in Figure 1. It is located in southwest England, 51.075 °N and 2.58 °W, and drains an area of 135.2 km^2. It is a predominantly rural catchment of modest relief with spring-fed headwaters rising in the Mendip Hills and Salisbury Plain. The observation data for this study were obtained from the Hydrological Radar Experiment (HYREX) project, which was funded by the Natural Environment Research Council (NERC). An automatic weather station (AWS) and automatic soil water station (ASWS), located in the catchment during the HYREX project, provided records of net radiation, wind speed, wet and dry bulb temperatures, atmospheric pressure and rainfall at hourly intervals. The rain gauge network consists of 49 Casells 0.2 mm tipping bucket-type rain gauges. The observation data were downloaded from the British Atmospheric Data Centre (BADC). The ground observed data from the Brue catchment, also provided by HYREX, were used for evaluating the downscaled weather variables produced by the mesoscale regional model MM5. The hybrid prediction system presented in this paper has been applied to data sets from 1995, 1996 and 1998 with the four weather variables. This paper discusses the results of the selected years based on four types of mathematical modelling for error correction of downscaled weather variables and verifies them with the observation data set from the Brue catchment.

Figure 1. Study area of the Brue catchment, Somerset, Southwest England

3 Mesoscale Modelling 5

The approach adopted for this study uses the PSU–NCAR mesoscale model version 5 (MM5) (Chen and Dudhia, 2001; Grell, 1995) as a common test framework to host the output of the weather variables for surface temperature, surface pressure, relative humidity, solar radiation and rainfall. The MM5-derived weather variables were extracted from four months of the year for 1995, 1996 and 1998. The model was simulated for the four seasons with representative months of the seasons [viz. winter (January), spring (April), summer (July) and autumn (October) seasons]. This study used ERA-40 reanalysis weather data, which is provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) through its data website. The resolutions of the data are 1° x 1° in space and 6 hours in time. The model was run with horizontal resolutions of four domains called Domain 1, 2, 3 and 4. Domains 1 to 4 have been structured with horizontal resolutions of 21 x 27km, 19 x 9km, 19 x 3km, 19 x 1km, respectively (see Figure 2). The model was run using a vertical level of 23, a default level for MM5. The set of parameters selected for the atmospheric model was the Grell cumulus formation based on a previous case study (Ishak et al., 2011) within the Brue catchment. The MRF (what is MRF?) parameters for the planetary boundary layer has been chosen for(?) MM5 simulation. Further details regarding these schemes may be found in the seminal study by Chen and Dudhia, 2001). The pattern of global model output of the four weather variables from MM5 with respect to the observed ground data from the HYREX project for years 1995, 1996 and 1998 has been systematically discussed in a previous study by Ishak et al., 2010. The performances of the six MM5-derived weather variables based on Domain 4 (inner domain) are shown in **Table 1**. The output of the MM5 model run is then taken as the input data for the statistical models, which are multiple linear regression, nonlinear regression, artificial neural network and support vector machine.

Figure 2. MM5 domains in the Brue catchment size of Domain 1 to Domain 4

Table 1. MM5 Downscaling performance statistics for 1995, 1996 and 1998 in the Brue catchment

4 Empirical Models for Error Correction

4.1 Regression Models

This study uses two regression models, namely, the multiple linear and nonlinear power-form function models, which are common mathematical methods to describe a relationship of output (Y) with input of variables (X_1, X_2, X_3, X_4, X_n) and model parameters ($a_0, a_1, a_2, \ldots a_n$) (Thomas and Benson, 1969). These methods are reliable and widely used in many estimation and forecasting problems. The linear and power-form equations are given as in Eq. (1) and (2) below:

$$Y = a_o + X_1 a_1 + X_2 a_2 + X_3 a_3 + X_4 a_4 + X_5 a_5 + X_6 a_6 + \ldots\ldots X_n a_n + \varepsilon_o \quad (1)$$

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

where, $a_0, a_1, \ldots a_n$, are the model parameters, ε_0 is the multiplicative error term, and n is the number of the data series. In this study, Y is the observed value of each weather variable while X_1 to X_6 are input of variables from MM5 output such as wind speed, surface temperature, surface pressure, solar radiation, rainfall and relative humidity.

Optimisation techniques are applied to minimise the result-estimated variable function, for instance, $\min_x f(x)$ (Broyden, 1970; Fletcher, 1970; Jones and Thornton, 2000; Shao, 1993). One such technique is the Broyden-Fletcher-Goldfarb-Shanno (BFGS), a Quasi-Newton gradient-based algorithm that is commonly used to solve unconstrained minimisation cases. Such cases occur when there are imposed conditions on the independent variables *X,* and it is assumed that *f* is defined for all *X*s. The BFGS uses an iteration process to find the optimal values for this function. The value of a_0 (initial value) is considered first, then the process is carried out for the rest of the values a_1, a_2, a_3, a_4, a_n. Eventually the iteration process successfully estimates the variables at the local minimum. The analysis ends when a predefined number of iterations, *k,* is reached. This optimisation technique is used throughout this study to determine the most optimal values of parameters for the function.

4.2 Artificial Neural Network

This study employs an artificial neural network (ANN) with a single hidden layer architecture, as shown in Figure 3. The network topology in this study has six nodes in the first layer (layer A) and ten nodes in the second layer (layer B), which are called *hidden layers*. The 'trial and error' method was adopted to identify the number of hidden nodes. In the third layer (layer C), which is called the *output layer,* there is one node. The network has six network inputs and one network output. There is an extra input assumed in each node that is considered to have a constant value of one. The weight that modifies this extra input is called the *bias*. The error correction was done by implementing a feed forward back propagation ANN with multiple-layer perceptron (MLP) architecture and was trained using the Levenberg–Marquardt (LM) optimisation technique. Before performing the training process, the weights and biases were initialized to appropriately scaled values. The sigmoid activation function was employed in this research. It has been observed that a feed forward ANN with six hidden neurons and learning rate of 0.08 gives better training and testing results.

Figure 3. The structure of a singlelayer artificial neural network

4.3 Support Vector Machines

Although the support vector machine (SVM) is a relatively new machine learning approach compared to the ANN, its network architecture and parameterizations are well documented in literature. Thus, the mathematical formulations of this model will not be repeated in this paper. For implementation of the SVM modelling, the software LIBSVM was used in this project, which was developed by Chih-Chung Chang and Chih-Jen and supported by the National Science Council of Taiwan (Chang and Lin, 2001). Figure 4 illustrates the SVM layout describing the processes carried out in this project. For this study, SVM modelling was carried out with different kernel functions and different SVR types (ν-SV regression and ε-SV regression). The ε value is set as 1 as suggested by previous studies on the Brue catchment (Remesan et al., 2009). The cost of error assigns a penalty for the number of vectors falling between the two hyperplanes. If the data quality is good, the distance between the two hyperplanes is narrowed down. If the data quality is noisy, it is preferable to have a smaller value of *C*, which will not penalise the vectors. However, in this study, the nu-SV regression for modelling was used as this kind of regressor with nonlinear kernel functions (radial basis function, RBF) performs better than the linear function as recommended and explored by Bray and Han, 2004 and Han et al., 2007. The error correction analysis was performed after fixing the parameters to default values (degree in kernel function is set as 3, coef0 in kernel function is set as zero, and the cache memory size is set as 40MB, tolerance of termination criterion is set as a default value of 0.001). Identification of the cost parameter (C) and slack parameter (ε) is very important for better performance of SVM. Different ranges of trial and error iterations were trained to find out the setting according to the least root mean square error (RMSE) values. This resulted in the selection of the cost parameter of 0.6 - 0.8 and slack parameter of 0.002 – 0.005, which yielded the best results.

Figure 4. The SVM-based hybrid modelling scheme used in this case study

5 Selection of Model Inputs

The four empirical models in this case study were developed to correct the downscaled weather variables obtained from MM5 (surface temperature, relative humidity, solar radiation, and rainfall). The target output of each model (MLR, NLR, ANN and SVM) was the observed ground value (HYREX weather data set) of weather variables in the Brue catchment. In this section, the results obtained from the cross correlation and leave-one-out cross-validation (LOOCV) identifying the best input combination for the corresponding model is described.

5.1 Cross-Correlation Method

The traditional approach to find dominant input series is the cross-correlation method. Table 2 shows the correlation coefficient of input data series for both training and testing target data sets of four downscaled weather variables taking ground data as a reference. It is to be noted that the training data sets comprising 5,904 data points are taken from years 1995 and 1996, while the testing data sets consisting of 2,032 data points are from the year 1998. From Table 2, it can be observed that MM5-derived surface temperature shows high correlation with a value of 0.955 during the training period and a value of 0.943 during the testing period. The second highest correlation values are associated with MM5-derived relative humidity values for both training and testing periods showing values of -0.547 and -0.520 respectively. MM5-derived wind

speed and rainfall have shown relatively weak correlation in the testing phase, while the same is observed for MM5-derived rainfall and pressure in the training phase. If the testing results based on the correlation outputs are considered as a reference, a trend of dominant inputs can be observed: TmpMM5 >RhMM5 >SolarMM5 >PrsMM5 >RfMM5 >WndMM5. Similarly, if the correlation statistics of other three attributes from the table are considered, the trends of the input variables will be as follows: RhMM5 >SolarMM5 >WndMM5 >TmpMM5 >RfMM5 >PrsMM5 for relative humidity, SolarMM5 >RhMM5 >TmpMM5 >WndMM5 >PrsMM5 >RfMM5 for solar radiation, and RfMM5 >PrsMM5 >WndMM5 >RhMM5 >SolarMM5 >TmpMM5 for rainfall.

Table 2. Correlation values between observed each weather variables and input of variables for 1995 to 1996 and 1998

5.2. Leave-One-Out Cross-Validation method

The leave-one-out cross-validation (LOOCV) method involves using a single data set from the available input space for modelling and identifying the best input for better training and testing results. This modelling is then repeated with two inputs, keeping the best input fixed and varying other input series, and so on. The performance of LOOCV is evaluated based on the value of RMSE in each model. For this case study, this modelling approach was adopted for all four models. The best model input structures obtained from LOOCV for all four weather variables are shown in Table 3. The corresponding graph pattern of the LOOCV results with the NLR, MLR, ANN and SVM models are illustrated in Figures 5 to 8 for surface temperature, relative humidity, solar radiation, and rainfall parameters respectively. For example, the graph pattern of surface temperature in Figure 5 describes the LOOCV-based model selection for NLR, MLR, ANN and SVM. It shows the best combination of input variables after running the four models with forward selection, optimisation and LOOCV. Various combinations based on the six input variables were tested for all models, where the objective was to find the best combination with least value of RMSE. For instance, for the NLR model, surface temperature (T) performed best among the four variables with the lowest RMSE value. Meanwhile, the lowest values of RMSE for two combinations of variables are surface temperature and wind speed (T+W); the three best combinations are surface temperature, wind speed and relative humidity (T+W+Rh); and so on. Similar descriptions can be applied for MLR, ANN and SVM.

Table 3. Model selection based on LOOCV method showing RMSE for NLR, MLR, ANN and SVM models

Figure 5. Results showing the LOOCV method of surface temperature by using NLR, MLR, ANN and SVM models

Considering all combinations, it can be seen from Figure 5 that for the NLR model the best combination of inputs were MM5-derived surface temperature, wind speed, relative humidity and solar radiation, yielding the least value of RMSE. The corresponding RMSE values can be found in Table 3 for surface temperature under the NLR model. For MLR models, the best input combination of selection is identified as MM5-derived surface temperature and surface pressure, while for ANN models, the best inputs are identified as a combination of MM5-derived surface temperature, wind speed, relative humidity, solar radiation, rainfall and surface pressure. Similarly, for SVM models, a combination

of surface temperature, wind speed, relative humidity, solar radiation, rainfall and surface pressure (T+W+Rh+S+R+P) is best. The same can be said for the other three variables of relative humidity, solar radiation and catchment average rainfall in different modelling scenarios. The best parameter combinations representing optimal model are illustrated in Figures 6 to 8 for relative humidity, solar radiation and rainfall respectively. It is worth mentioning that in cases of conflict between training and testing observations, the combination considering the lowest RMSE values during the testing phase was identified. The RMSE values are listed in Table 3, which also highlights the best corresponding parameter combinations.

Figure 6. Results showing the LOOCV method of relative humidity using NLR, MLR, ANN and SVM models

Figure 7. Results showing the LOOCV method of solar radiation using NLR, MLR, ANN and SVM models

Figure 8. Results showing the LOOCV method of rainfall using NLR, MLR, ANN and SVM models

6 Application of Different Error Correction Methods

This section describes the detailed results of the MM5-derived surface temperature, relative humidity, solar radiation and rainfall error correction modelling using the four methods (MLR, NLR, ANNs and SVMs) for the Brue catchment in southwest England. In this study, the downscaled MM5 and error-corrected values of weather variables is compared with the HYREX land-based observed data. The study has focused on two major indices as the performance criteria: RMSE and mean bias error (MBE). The RMSE and MBE values are mainly expressed as percentages of the mean value of observed data.

6.1 Modelling with MLR and NLR Models

Before implementing MLR and NLR models, it is important to standardise the input data with the target output for X_{max}, X_{mean} or X_{min}. For solar radiation and rainfall, normalised data was used for analysing MLR and NLR models. This is because negative values were too large in the solar radiation data, and there were many zero values in the rainfall data. After standardising temperature and relative humidity and normalising solar radiation and rainfall, MLR and NLR equations were then modelled according to Eq. (1) and Eq. (2) respectively. Generalisation of the four variables is made based on performance on the testing data set.

Surface temperature

In the case of error correction modelling for surface temperature, the NLR model gave better results for the combination of TmpMM5, WndMM5, RhMM5 and SolarMM5 with RMSE values of 1.902 m/s and 1.952 m/s (based on Table 3) during training and testing phase respectively. The optimum NLR model is given in Eq. (12).

$$Y = 0.575x(TmpMM5 + 0.583)x(WndMM5 + 5)^{0.129}\, x(RhMM5/10)^{0.030}\, x\left(\frac{SolarMM5}{100} + 5\right)^{0.067} \quad \text{.....(12)}$$

In MLR function models, the TmpMM5 and PrsMM5 input combination have shown better performance with RMSE values of 1.616 m/s and 1.925 m/s during training and testing periods respectively. The optimal MLR model with two input variables and corresponding parameters is shown in the Eq 13.

$$Y = -11.733 + ((TmpMM5 + 1)x0.856) + \left(\frac{\Pr sMM5}{100}\right)x1.229 \quad \text{.....(13)}$$

The values of RMSE and bias obtained after surface temperature corrections based on MLR and NLR during the training and testing phases can be found in Table 4. For better visual understanding of the model accuracy, scatter plots of the surface temperature training set before and after error correction with MLR model are depicted against observed data in Figure 9. The corresponding scatter plots of the testing data set are given in Figure 10. Similarly, scatter plots showing the difference between measured and NLR model error corrected surface temperature during both training and testing phase are shown in Figure 11. Before error correction, the MM5-derived surface temperature showed higher values of bias and RMSE in comparison with observed surface temperature for both training and testing sets. During the training period (1995-1996), the MM5-simulated surface temperature showed a bias value of 0.443^{o}C (4.510%) and corresponding RMSE values of 3.055^{o}C (31.074%) (see Table 1). MM5 simulation results during the testing period (1998) have shown higher bias and RMSE values of -0.310^{o}C (-3.154%) and 2.811^{o}C (28.588%) respectively. After error correction modelling, the NLR model performed better with less value of bias compared to the MLR model (refer to Table 4). The values of bias from NLR output reduced considerably to 0.261% during the training period and -2.492% (slightly underestimated) during the testing period. Likewise, the corresponding RMSE values also reduced to 16.186% and 19.722% during training and testing periods respectively.

Table 4. Statistical indices showing performance of surface temperature error correction models during the training and testing phases

Figure 9. Scatter plots of four weather variables error correction on training data set: MM5-derived output (left) after MLR modelling output (right)

Figure 10. Scatter plots of four weather variables error correction modelling on testing data set: MM5-derived output (left) after MLR modelling output (right)

Figure 11. Scatter plots of NLR modelling for four weather variables error correction modelling during training phase (left) and testing phase (right)

MLR and NLR modelling output results for the other three variables, relative humidity, solar radiation and catchment average rainfall are shown in Figures 9 to 11. The results are based on the best combinations of variables, as depicted in Figures 6 to 8 and Table 2. From the statistics of these three variables shown in Tables 5 to 7, it can be seen how such modelling approaches reduce the error from the raw MM5 downscaled performance. It is to be noted that the data inputs of the rainfall analysis improvement with mathematical models are based on normalised data. However, the values of RMSE and MBE were de-normalised for MLR and NLR. The reason for de-normalisation of this output was for comparative results. The established equations of MLR and NLR for other three corrected weather variables are as follows:

Relative humidity

NLR

$$Y = 139.999x(RhMM5)^{0.664}\,x((WndMM5x6) + 25)^{-0.111}\,x((RfMM5x10) + 20)^{0.076}\,x((TmpMM5x3) + 50)^{0.035}\,x\left(\frac{\Pr sMM5}{10}\right)^{-0.740}$$

.....(14)

MLR

$$Y = 87.341 + (RhMM5x0.721) + (((TmpMM5x3) + 50)x0.056) + (((RfMM5x10) + 20)x0.193) + \left(\left(\left(\frac{SolaMM5}{10}\right) + 10\right)x - 0.002\right) + (((WndMM5x6) + 25)x - 0.146) + \left(\left(\frac{\Pr sMM5}{10}\right)x - 0.654\right)$$

.....(15)

Solar radiation

NLR

$$Y = (SolarMM5 + 0.00071)x(RfMM5 + 0.826)$$ .....(16)

MLR

$$Y = 0.000106 + (SolarMM5x0.638) + (RfMM5x - 0.023) + (RhMM5x - 0.217) + (TmpMM5x0.022) + (WndMM5x0.011) + (\Pr sMM5x0.013)$$ ....(17)

Catchment average rainfall

NLR

$$Y = (RfMM5 + 0.179)x(WndMM5 + 0.557)x(RhMM5 + 0.596)x(\Pr sMM5 + 0.051)x(SolarMM5 + -0.639)x(TmpMM5 + 1.217)$$

.....(18)

MLR

$$Y = 0.00003 + (RfMM5x0.261) + (\Pr sMM5x - 0.044) + (RhMM5x0.021) + (WndMM5x0.020) + (TmpMM5x0.017) + (SolarMM5x0.004)$$

.....(19)

6.2 Modelling with the ANN Model

As with the previous two models, the ANN model uses data from years 1995 and 1996 for training and year 1998 data for testing. Scatter plots of the results from the LM (what is LM?) algorithm-based ANN model during training and testing are given in Figure 12. The corresponding training and testing statistics for surface temperature, relative humidity, solar radiation and rainfall are shown in Tables 4 to 7.

Surface temperature (°C) (Is there a reason for putting units of measurement in the subtitles?)

For surface temperature, the best combination was achieved with six inputs: TmpMM5, WndMM5, RhMM5, SolarMM5, RfMM5 and PrsMM5 (see Table 4). The inputs are MM5-derived surface temperature, wind speed, relative humidity, solar radiation, rainfall and pressure, which have been used for error correction based on LOOCV. The ANN model anticipated a surface temperature with RMSE values of 1.571 °C (15.869%) during the training phase and 1.802 °C (18.207%) during the testing phase. The bias value observed during the training phase is -0.003 °C (-0.032%), whereas the mean bias error (MBE) during the testing phase is observed as 0.371 °C, which is 3.751% of the mean observed value. From Table 1, it can be observed that the bias value was initially 4.510% for the MM5 simulation results during the training phase. The ANN modelling has significantly reduced this bias value to -0.032% during training and to 3.751% during the testing phase. Although the ANN produced better training results than those of the regression models, it was less effective in showing better skills in terms of numerical values compared to MLR and NLR function models during testing. All the four models (MLR, NLR, ANN and SVM) were trained and tested using the same data set. Thus, the reason of the disparity could be associated with inputs used for the models.

Figure 12. Scatter plots of ANN model results for four weather variables error correction modelling during training (left) and testing (right)

Relative humidity (%)

It can be observed from Table 5 that ANN model performed better than all other models in the training phase. The best combination of five inputs are RhMM5, WndMM5, PrsMM5, TmpMM5 and RfMM5. As shown in Table 1, the bias value of the MM5 simulation results is 4.326% when compared against the MBE of the mean observed relative humidity. The bias is significantly diminished through ANN modelling. The bias value observed in the ANN model during the training phase is close to zero, while the MBE during the testing phase is -0.824%, which is -0.984% of mean observed value.

Table 5. Statistical indices showing performance of relative humidity error correction models in the training and testing phases

Solar radiation (W/m^2)

Solar radiation shows a different trend from the other three variables. As indicated earlier, the best performing models during the training and testing periods are ANN and MLR respectively. Solar radiation behaves differently due to the large amount of negative values in the solar radiation data set, and therefore, only MLR can analyse these negative data set, especially during evening and night time. The statistical performance of error-corrected solar radiation is shown in Table 6 and Figure 12. Table 1 shows the value of bias from MM5-derived solar radiation, which is 4.040 W/m^2 (3.749%) for the training phase and -0.289 W/m^2 (-0.268%) for the testing phase. In contrast, the value of bias using the ANN model during the training phase is reduced to -0.006 W/m^2 (-0.005%), while MBE during the testing phase is observed as 30.983 W/m^2, which is 28.577% of mean observed value. (Add a phrase describing the RMSE), the value of RMSE from the ANN model is 72.960 W/m^2 (67.248%) during the training phase and 110.470 W/m^2 (101.822%) during the testing phase.

Table 6. Statistical indices showing performance of solar radiation error correction models in training and testing phases

Catchment average rainfall (unit of measurement?)

Results indicate that among all modelling cases, ANN constructed better models for catchment average rainfall as compared to MLR, NLR and SVM models in the training phase (see Table 7). In the ANN model, it was found that three input variables provided the most optimal combination of this model, and these are RfMM5, PrsMM5 and RhMM5. Improvements through this method can be observed in the series of tables 1 through 7 representing the statistics before and after error correction modelling with the ANN model. Although this is not the best method for catchment average rainfall improvement from downscaled MM5 as the errors are still very high, there is some reduction of errors, with approximately $\pm$ 20% in the training period while $\pm$ 10% in the testing period.

Table 7. Statistical indices showing performance of rainfall error correction models in the training and testing phases

6.3 Modelling with SVMs

SVMs use the LOOCV method for input modelling. The statistical performance of the SVM with nu-SVR and RBF kernel is presented in the series of tables 4 through 7 for surface temperature, relative humidity, solar radiation and rainfall error correction modelling in the training and testing phases. Scatter plots show the observed and error-corrected data of the four weather parameters (see Figure 13).

Surface temperature (^oC)

SVM showed satisfactory results for surface temperature. Results for error correction achieved higher accuracy compared to those of the other models during the testing period (Table 4). Similarly, the SVM was found to be the second best followed by the ANN model in the training period. The values of MBE and RMSE are 0.239% and 16.136% respectively during the training phase, while -1.172% and 17.5% during the testing phase.

Figure 13. Scatter plots of SVM model results for four weather variables of error correction modelling during training (left) and testing (right)

Relative humidity (%)

The evaluation of relative humidity error correction modelling using SVM is given in Table 5. The improvements for this parameter from MM5 downscaling shows similar output patterns as those of the other weather variables. In terms of performance, the SVM constructed better models as compared to the MLR, NLR and ANN in the testing phase. The best model results are obtained with five input combinations, which are RhMM5, WndMM5, PrsMM5, TmpMM5 and RfMM5. Improvements of the modelled relative humidity can be seen in Figure 13. There is significant improvement in error reduction from MM5 downscaling data, as shown in Table 1 and Table 5. The value of bias for relative humidity from MM5 simulation results is -3.622 % (4.326%) during training phase (see Table 1). In comparison, the value of bias in the SVM model for the same phase is 0.009% (0.010%), while the value of bias during testing phase is -0.119%, which is -0.142% of mean observed value. In this regard, the SVM model shows least value of RMSE of 7.427% (8.870%) during the training phase and 8.373% (10.0%) during the testing phase.

Solar radiation (W/m^2)

Statistical error results of solar radiation output, illustrated in Table 6, shows that error correction from the ANN model performed better than all other models in the training phase. Conversely, SVM performed better than ANN and NLR models in the testing phase. The SVM output shows that the three inputs that gave the best combinations are SolarMM5, RhMM5 and TmpMM5. A scatter plot of observed and corrected solar radiation for the training and testing data is depicted in Figure 13. Compared against the mean observed solar radiation, the MM5 simulation results gave a biased value of 4.040 W/m^2 (3.749%) in the training phase (see Table 1). The bias value observed in the SVM model during the training phase is -4.489 W/m^2 (-4.138%), while the MBE during the testing phase is 1.126 W/m^2, which is 1.037% of mean observed value. The SVM model analysed the solar radiation with RMSE values of 83.24 W/m^2 (76.725%) during the training phase and 96.877 W/m^2 (89.293%) during the testing phase. Similarly with the result of catchment average rainfall where even this four mathematical model did not performed quite well but at least the value of RMSE performed improvement ranging from ± 30 during training and $\pm 20\%$ during testing. (Why is catchment average rainfall being discussed here?)

Catchment average rainfall (mm)

Table 7 clearly shows that the SVM model gave the second lowest value of bias during testing phase than the ANN and NLR models. However, in the training phase, SVM model showed better bias and RMSE than the NLR model, but weaker than ANN and MLR. The SVM model produced better modelling results with RMSE value of 0.352 mm (392.139%) and MBE value of -0.035 mm (-39.563%) during the testing phase. The corresponding value during the training phase are 0.279 mm (311.022%) and -0.022 mm (-24.584%) respectively. The performance of SVM in terms of MBE and RMSE values was weaker than that of ANN model during the training and testing periods. In general, all the values of RMSE for both training and testing phases shows variation as 290% < RMSE < 400%. This implies that the four models of error correction performed unsatisfactorily in producing better simulation of the MM5-derived rainfall.

7 Conclusion

Although hydrological modelling has been extensively studied, there has been little study on some hydro-meteorological data, specifically variables such as wind speed, surface temperature, surface pressure, relative humidity, solar radiation and rainfall. These data nonetheless are very important for flood forecasting and water resources assessment, especially in ungauged catchments. One of the reasons for such little study is due to a general lack of hydro-meterological data, and although this is a common problem for hydrological and meteorological modellers worldwide, it is particularly so in developing countries. However, with the advancement of computer and telecommunication technology, it is now possible to collect huge quantities of data. The optimisation and LOOCV presented in this case study may offer a solution for hydrologists and meteorologist to estimate any hydro-meteorological variables in the future. Using a MM5 framework to downscale key parameters, bias from the initial data selection is further reduced through error correction models.

In this study, it has been demonstrated that statistical performance produces better result after correction on MM5-derived surface temperature, relative humidity, solar radiation and rainfall using four types of models including the two, three, four, five and six input variables of forward selection. Although cross-validation involves forward and backward selection, in this case study we have focused on forward selection due to little input variables for the error correction methodology. The result shows that cross-validation is useful for multiple linear, nonlinear, ANN and SVM structures systems as decreases in value of bias and RMSE become evident for the corrected four weather variables. It was also found that cross-validation (k-1) will need more time to process huge data sets, for instance data sets that have above 5,000 data points. Cross-validation with optimisation technique has the potential in helping hydrologists and meteorologists to decide the optimum input data combination for their models.

(NEED TO ASK DR. ASNO) Results from the case study showed varying performance levels among the parameters after error correction. The parameter showing the least significant improvement was rainfall. MM5-derived rainfall performance improved about 30% in terms of RMSE value. Although the most challenging work in this research was improving estimation of rainfall, ANN modelling showed a decrease in RMSE value (before correction is 375.4% while after correction is 345.691% during the training period). The second least-performing parameter is MM5-derived solar radiation where the values of RMSE before and after correction are 103.580% and 89.293% respectively during the training period. The RMSE value is based on SVM modelling output. The next parameter, MM5-derived surface temperature, saw a decrease of RMSE value from 12.470% to 10.0%, before and after correction respectively, according to SVM modelling

output. The most efficient error correction modelling in this case study is MM5-derived surface temperature??. The modelled surface temperature saw a decrease of RMSE values from 28.588% to 17.5%. This variable, which derived from MM5, performed with more than 50% improvement by decreasing the value of RMSE.

In summary, this case study demonstrates the improvement of MM5-derived weather parameters using six inputs, which is a novel approach that considers more input variables for error correction modelling instead of only one. Clearly, this marks the beginning of error correction through mathematical modelling, and there is the need for more study for further data quality improvement. This case study has not provided a final answer to error correction techniques for hydrological and meteorological modelling, but instead provides an impetus for the academic and practitioner communities to expand on this problem and explore the mathematical tools presented in this paper such that wide gaps of knowledge could be bridged by engaging in a wide range of trials using this approach.

8. REFERENCES

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. FAO, Rome 300, 6541.

Bray, M., Han, D., 2004. Identification of support vector machines for runoff modelling. Journal of Hydroinformatics 6, 265-280.

Broyden, C., 1970. The convergence of a class of double-rank minimisation algorithms 1. general considerations. IMA Journal of Applied Mathematics 6, 76.

Chahine, M.T., 1992. The hydrological cycle and its influence on climate. Nature 359, 373-380.

Chang, C.C., Lin, C.J., 2001. LIBSVM: a library for support vector machines.

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. Monthly Weather Review 129, 569-585.

Chowdhury, M., Ward, N., 2004. Hydro meteorological variability in the greater Ganges–Brahmaputra–Meghna basins. International journal of climatology 24, 1495-1508.

Coulibaly, P., 2003. Impact of meteorological predictions on real time spring flow forecasting. Hydrological processes 17, 3791-3801.

Fletcher, R., 1970. A new approach to variable metric algorithms. The Computer Journal 13, 317.

Fowler, H., Blenkinsop, S., Tebaldi, C., 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. International journal of climatology 27, 1547-1578.

Ghosh, S., Mujumdar, P., 2008. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. Advances in water resources 31, 132-146.

Grell, G.A., 1995. A Description of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5), NCAR/TN-398+ STR. NCAR TECHNICAL NOTE.

Han, D., Kwong, T., Li, S., 2007. Uncertainties in real time flood forecasting with neural networks. Hydrological processes 21, 223-228.

Heuvelmans, G., Muys, B., Feyen, J., 2006. Regionalisation of the parameters of a hydrological model: Comparison of linear regression models with artificial neural nets. Journal of Hydrology 319, 245-265.

Ishak, A.M., Bray, M., Remesan, R., Han, D., 2010. Estimating reference evapotranspiration using numerical weather modelling. Hydrological processes.
Ishak, A.M., Bray, M., Remesan, R., Han, D., 2011. Seasonal evaluation of rainfall estimation by four cumulus parameterization schemes and their sensitivity analysis. Hydrological processes.
Jones, P.G., Thornton, P.K., 2000. MarkSim: Software to generate daily weather data for Latin America and Africa. Agronomy Journal 92, 445-453.
Khalil, A., Almasri, M.N., McKee, M., Kaluarachchi, J.J., 2005. Applicability of statistical learning algorithms in groundwater quality modeling. Water resources research 41, W05010.
Meza, F.J., 2005. Variability of reference evapotranspiration and water demands. Association to ENSO in the Maipo river basin, Chile. Global and Planetary Change 47, 212-220.
Oommen, T., Misra, D., Agarwal, A., Mishra, S.K., 2007. Analysis and application of support vector machine based simulation for runoff and sediment yield. American Society of Agricultural and Biological Engineers, St. Joseph, MI, ASABE paper.
Remesan, R., Shamim, M.A., Han, D., Mathew, J., 2009. Runoff prediction using an integrated hybrid modelling scheme. Journal of Hydrology 372, 48-60.
Sentelhas, P.C., Gillespie, T.J., Santos, E.A., 2010. Evaluation of FAO Penman-Monteith and alternative methods for estimating reference evapotranspiration with missing data in Southern Ontario, Canada. Agricultural Water Management 97, 635-644.
Shao, J., 1993. Linear model selection by cross-validation. Journal of the American Statistical Association, 486-494.
Thomas, D.M., Benson, M.A., 1969. Generalization of streamflow characteristics from drainage basin characteristics. GEOL SURV OPEN-FILE REP, 1969. 45 P, 16 FIG, 11 TAB, 18 REF.
Tripathi, S., Srinivas, V., Nanjundiah, R.S., 2006. Downscaling of precipitation for climate change scenarios: A support vector machine approach. Journal of Hydrology 330, 621-640.
Werner, M., Reggiani, P., Roo, A.D., Bates, P., Sprokkereef, E., 2005. Flood forecasting and warning at the river basin and at the European scale. Natural hazards 36, 25-42.
Yoon, H., Jun, S.C., Hyun, Y., Bae, G.O., Lee, K.K., 2010. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. Journal of Hydrology.
Young, P.C., 2002. Advances in real–time flood forecasting. Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 360, 1433.