Int. J. Appl. Math. Comput. Sci., 2012, Vol. 22, No. 4, 841–854
DOI: 10.2478/v10006-012-0062-1

A RAINFALL FORECASTING METHOD USING MACHINE LEARNING MODELS AND ITS APPLICATION TO THE FUKUOKA CITY CASE

S. MONIRA SUMI [*], M. FAISAL ZAMAN [*,**], HIDEO HIROSE [*]

[*] Department of Systems Design and Informatics
Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka, Japan
email: {sumi,zaman}@ume98.ces.kyutech.ac.jp, hirose@ces.kyutech.ac.jp

[**] School of Electronic Engineering
Dublin City University, Glasnevin, Dublin, Ireland
email: faisal.zaman@dcu.ie

In the present article, an attempt is made to derive optimal data-driven machine learning methods for forecasting an average daily and monthly rainfall of the Fukuoka city in Japan. This comparative study is conducted concentrating on three aspects: modelling inputs, modelling methods and pre-processing techniques. A comparison between linear correlation analysis and average mutual information is made to find an optimal input technique. For the modelling of the rainfall, a novel hybrid multi-model method is proposed and compared with its constituent models. The models include the artificial neural network, multivariate adaptive regression splines, the k-nearest neighbour, and radial basis support vector regression. Each of these methods is applied to model the daily and monthly rainfall, coupled with a pre-processing technique including moving average and principal component analysis. In the first stage of the hybrid method, sub-models from each of the above methods are constructed with different parameter settings. In the second stage, the sub-models are ranked with a variable selection technique and the higher ranked models are selected based on the leave-one-out cross-validation error. The forecasting of the hybrid model is performed by the weighted combination of the finally selected models.

Keywords: rainfall forecasting, machine learning, multi-model method, pre-processing, model ranking.

1. Introduction

Accurate forecasting of rainfall has been one of the most important issues in hydrological research because early warnings of severe weather can help prevent casualties and damages caused by natural disasters, if timely and accurately forecasted. To construct a predictive system for accurate rainfall, forecasting is one of the greatest challenges to researchers from diverse fields such as weather data mining (Yang *et al.*, 2007), environmental machine learning (Hong, 2008), operational hydrology (Li and Lai, 2004), and statistical forecasting (Pucheta *et al.*, 2009). A common question in these problems is how one can analyse the past and use future prediction. The parameters that are required to predict rainfall are enormously complex and subtle even for a short term period.

Physical processes in rainfall are generally composed of a number of sub-processes. A accurate modelling of rainfall by a single global model is sometimes not possible (Solomatine and Ostfeld, 2008). To overcome this difficulty, the concept of modular modelling and combining different models has attracted more attention recently in rainfall forecasting. In modular models, several sub-processes are first identified, and then separate models (also called local or expert models) are established for each of them (Solomatine and Ostfeld, 2008). So far, various modular models have been proposed, depending on soft or hard splitting of training data. Soft splitting means that the dataset can be overlapped, and the overall forecasting output is the weighted average of each local model (Shrestha and Solomatine, 2006; Wu *et al.*, 2008).

In the hard splitting, there is no overlap of data and the final forecasting output is derived explicitly from only one of the local models (Wu *et al.*, 2008). The approach of combining several models is also known as ensemble modelling. The basic idea behind the ensemble model is to build several different models for the same process and to integrate them together

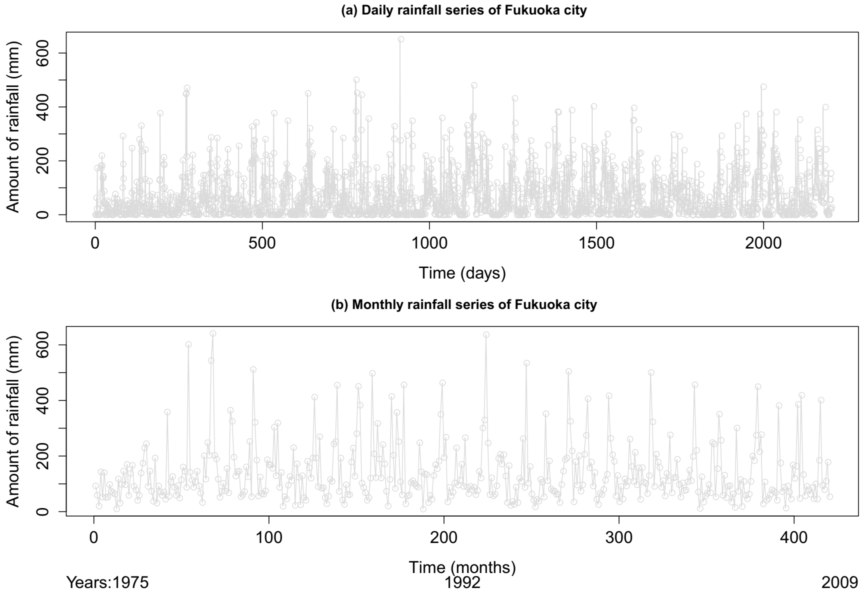


Fig. 1. Daily rainfall series in the rainy season (June and July) (a) and monthly rainfall series of the Fukuoka city (b) from 1975 to 2009.

(Xiong *et al.*, 2001; Abrahart and See, 2002; Kim *et al.*, 2006; Baruque *et al.*, 2011; Siwek *et al.*, 2009; Zaman and Hirose, 2011). For example, Xiong *et al.* (2001) used a Takagi–Sugeno–Kang fuzzy technique to couple several conceptual rainfall-runoff models. Coulibaly *et al.* (2005) employed an improved weighted-average method to coalesce forecasted daily reservoir inflows from the k-Nearest Neighbor (k-NN), the conceptual model, and the Artificial Neural Network (ANN). Kim *et al.* (2006) investigated five ensemble methods for improving stream flow prediction.

The idea of ensemble learning is popular in other time series applications as well. Wichard and co-workers applied an ensemble of multi-models to construct hybrid models for NN5 time series competition (Wichard and Ogorzalek, 2007; Wichard, 2011). Deng *et al.* (2005) applied a parallel ensemble of support vector regression in two simulated time series datasets, the Sunspot and Mickey Glass datasets. A novel neural network ensemble approach called the generalized regression neural network ensemble for time series forecasting (GEFTSGRNN) which is a concatenation of existing machine learning algorithms has been applied in benchmark time series forecasting datasets by Gheyas and Smith (2011). Everingham *et al.* (2009) constructed an ensemble method comprising statistical data mining models, to forecast crop productions in north eastern Australia.

In this article, we make a comparison of several machine learning methods of forecasting an average daily and monthly rainfall of the Fukuoka city in Japan. All the methods are coupled with two data-preprocessing techniques. Prior to applying the methods, two input selection techniques are used. For the modelling of the rainfall, a novel hybrid multi-model method is proposed. The constituent models of the hybrid method are the ANN, Multivariate Adaptive Regression Splines (MARS), the k-nearest neighbour, and radial basis Support Vector Regression (SVR). The hybrid method generates sub-models first from each of the above methods with different parameter settings. Second, all the sub-models are ranked with a variable selection technique called least angle regression (LARS). Third, the higher ranked models are selected based on their Leave-One-Out Cross-Validation (LOOCV) error. The forecasting using the out of samples is done by a weighted combination (Timmermann, 2006) of the finally selected models. For evaluation of this hybrid method, we have constructed all these methods with their respective optimal parameters and applied to out of sample forecasting.

The rest of the paper is organised as follows. In Section 2, we discuss briefly the study area and the rainfall series used in this paper. In Section 3, we describe the hybrid forecast model including the input selection technique and the variable selection method, and how the weights are extracted. This is followed by discussions about the experimental setup (Section 4) and results (Section 5). Conclusive discussions of the paper appear in Section 6.

2. Study area

In this paper, we have taken a daily rainfall series of rainy season and a monthly rainfall series of the Fukuoka city. The rainfall data are taken from nearby weather stations, which each weather station being within the range of 48 km from the Fukuoka city. For the distance, the rainfall data are taken from six forecast stations (as the forecast point) in the Fukuoka and Saga prefectures in Japan. Both the daily and monthly rainfall series are plotted in Fig. 1. Each series contains rainfall updates from 1975 to 2009. Our objective is to forecast a 1-step ahead rainfall for the rainy season and a monthly rainfall in the Fukuoka city.

3. Methodology

3.1. Data-preprocessing techniques.

3.1.1. Moving Average (MA). The MA method is based on the idea that any large irregular component at any point in time will exert a smaller effect if we average the point with its immediate neighbours (Newbold *et al.*, 2007). The MA smooths data by replacing each data point with the average of the k neighbouring data points, where k may be termed the length of a memory window. The equally weighted MA is most commonly used, in which each value of the data carries the same weight in the smoothing process. There are three types of moving modes, including centering, backward and forward. In a forecasting scenario, only the backward mode is used since the other two modes may necessitate future observed values. For a time series $\{x_1, x_2, \dots, x_N\}$, when the backward moving mode is adopted (Lee *et al.*, 2000), the k-term unweighed moving average y_t^* is written as

$$y_t^* = \frac{1}{k}\sum_{i=0}^{k-1} y_{t-i}, \qquad (1)$$

where $t = k, \dots, N$. The choice of the window length k is made with a trial and error procedure with a minimization of the prediction error.

3.1.2. Principle Component Analysis (PCA). The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. The PCA approach uses all of the original variables to obtain a smaller set of Principal Components (PCs) which can be used to approximate the original variables. PCs are uncorrelated and are ordered so that the first few retain most of the variation present in the original set.

Consider a data matrix X which has n rows (observations) and p columns (variables). Let the covariance matrix of X be Σ, where $\Sigma = \mathrm{cov}(X) = E(X^TX)$. The linear transformed orthogonal matrix Z is represented as

$$Z = XA, \qquad (2)$$

where Z is the PCs with elements (i, j) of the i-th observation and the j-th principal component while A is a $(p\times p)$ matrix with eigenvector elements of the covariance of X and having $A^TA = AA^T = 1$.

Since the matrix X^TX is real and symmetric, it can be expressed as $X^TX = A\Lambda A^T$, where Λ is a diagonal matrix whose non-negative entries are the eigenvalues $(\lambda_i, i = 1, \dots, p)$ of X^TX. The total variance of the data matrix X is represented as

$$\mathrm{trace}(\Sigma) = \mathrm{trace}(A\Lambda A^T) = \mathrm{trace}(\Lambda) = \sum_{i=1}^{p}\lambda_i. \qquad (3)$$

The covariance matrix of principal components Z is expressed as

$$\mathrm{cov}(Z) = E(Z^TZ) = E(A^TX^TXA) = \Lambda, \qquad (4)$$

$$\mathrm{trace}(Z) = \mathrm{trace}(\Lambda) = \sum_{i=1}^{p}\lambda_i. \qquad (5)$$

Therefore, the total variance of the data matrix X is identical with the total variance after PCA transformation Z.

The solution of PCA, using Singular Value Decomposition (SVD) or determinants of the covariance matrix of X, can provide the eigenvectors A with their eigenvalues, $\lambda_i, i = 1, \dots, p$, representing the variance of each component after PCA transformation. If the eigenvalues are ordered by $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0$, the first few PCs can capture most of the variance of the original data while the remaining PCs mainly represent the noise in the data. The percentage of total variance explained by the first m-th PCs is

$$V = \frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \times 100. \qquad (6)$$

The higher value of total data variance V implies that more properties of the data matrix are preserved. For the sake of the dimensionality reduction, a small number of PCs is selected, though most of the data variance in selected components still remain. The original data matrix A can be reconstructed by a reverse operation of Eqn. (2) as

$$X = ZA^T. \qquad (7)$$

By choosing a suitable $m\,(\leq\ p)$ PCs from Z and accompanying m eigenvectors from A, the original data can be filtered.

3.2. Construction of input/output pairs.

Let $\{x_1, x_2, \dots, x_N\}$ stand for a rainfall time series. It can be reconstructed into a series of delay vectors as

$$X_t = x_t, x_{t+\tau}, x_{t+2\tau}, \dots, x_{t+(m-1)\tau},$$

where $X_t \in \mathbb{R}^m$, τ is the delay time as a multiple of the sampling period, and m is the embedded dimension. Suppose that the rainfall $x_{t+T+(m-1)\tau}$ at T-step lead is related to the vector X_t. Then the available historical data may be summarized into a set of pairs as $\{X_t, x_{t+T+(m-1)\tau} : t = 1, \ldots, n\}$, where n stands for the number of pairs, and $n = N - (m-1)\tau$.

The functional relationship between the input vector X_t at time t and the predicted output $x^F_{t+T+(m-1)\tau}$ at time $t+T$ can be written as follows:

$$x^F_{t+T+(m-1)\tau} = f(X_t) + e_t, \tag{8}$$

where e_t is a typical noise term, $x^F_{t+T+(m-1)\tau}$ is the prediction of $x_{t+T+(m-1)\tau}$, and $f(\cdot)$ is the mapping function. The difference among various data-driven forecasting models used in the current study consists in the way of approximating $f(\cdot)$ once model inputs are attained with the appropriate selection of (τ, m).

3.3. Individual machine learning methods.

Four machine learning models are selected to construct the hybrid multi-model forecasting method. The models are the ANN, the k-NN, MARS, and SVR. These are usually called data-driven models because of the ability to capture the mapping between input (e.g., antecedent rainfall) and output variables (forecasted rainfall) without directly considering the physical laws that underlie the mechanism of rainfall. These models are purely based on the information retrieved from the collected rainfall data.

3.3.1. Artificial neural network.

The multilayer perceptron network is by far the most popular ANN paradigm, which usually uses the technique of error back propagation to train the network configuration. The architecture of the ANN consists of a number of hidden layers and a number of neurons in the input layer, hidden layers and output layer. ANNs with one hidden layer are commonly used in hydrologic modelling (Dawson and Wilby, 2001; De Vos and Rientjes, 2005) since these networks are considered to provide enough complexity to accurately simulate the nonlinear properties of the hydrologic process. The ANN forecasting model is formulated as

$$\begin{aligned} & x^F_{t+T+(m-1)\tau} \\ & \quad = f(X_t, w, \theta, m, h) \\ & \quad = \theta_0 + \sum_{j=1}^{h} w_j^{\text{out}} \phi\Big(\sum_{i=1}^{m} w_{ji} x_{t+(i-1)\tau} + \theta_j \Big), \end{aligned} \tag{9}$$

where ϕ denotes transfer functions; w_{ji} are the weights defining the link between the i-th node of the input layer and the j-th node of the hidden layer; θ_j are biases associated with the j-th node of the hidden layer; w_j^{out} are the weights associated to the connection between the j-th node of the hidden layer and the node of the output layer; and θ_0 is the bias at the output node. To apply Eqn. (8) to rainfall predictions, an appropriate training algorithm is required to optimize w and θ.

3.3.2. *k*-nearest neighbor.

The k-NN is a nonparametric method that bases its prediction on the target outputs of the k-nearest neighbors of the given query point (Hastie *et al.*, 2009). Specifically, given a data point, we compute the Euclidean distance between that point and all points in the training set. We then pick the closest k training data points and set the prediction as the average of the target output values for these k points. The prediction of $x_{t+T+(m-1)\tau}$ by the k-NN method is formulated as

$$x^F_{t+T+(m-1)\tau} = \frac{1}{k} \sum_{t \in S(X,n)} x_{t+T+(m-1)\tau}, \tag{10}$$

where $S(X, n)$ denotes the set of indices t of the k-nearest neighbors to the feature vector $X(n)$. Therefore, if i belongs to $S(X, n)$ and j is not in $S(X, n)$, then, according to the Euclidean distance, $\| X_n - X_i \| \leq \| X_n - X_j \|$. Intuitively speaking, the forecast $x^F_{t+T+(m-1)\tau}$ in Eqn. (10) is the sample average of the output rainfall of the k-nearest neighbors to $X(n)$.

3.3.3. Multivariate adaptive regression splines.

MARS was first proposed by Friedman (1991) as a tree-based local modeling technique, dividing the data space in several, possibly overlapping regions and fitting truncated spline functions in each region. It is very useful for high dimensional problems and constitutes a great promise for fitting non-linear multivariate functions. A special advantage of MARS lies in its ability to estimate the contributions of the basis functions so that both the additive and the interactive effects of the predictors are allowed to determine the response variable.

For each of the descriptive variables in a data set, MARS selects the pair of spline functions and the knot location that best describes the response variable. Specifically, all the spline functions are combined in a complex non-linear model, describing the response as a function of the descriptive variables. The corresponding model has the form

$$\hat{y} = \alpha_0 + \sum_{i=1}^{M} \alpha_i \beta_i(x), \tag{11}$$

where $\hat{y}$ is the predicted value for the response variable, α_0 is the coefficient of the constant term, M is the number of spline functions, and β_i and α_i are the i-th spline function and its coefficient, respectively (Friedman, 1991).

3.3.4. Support vector regression. Support vector regression (Schölkopf and Smola, 2002; 2004), is a successful method penalizing the ensuing complexity using a penalty term added to the error function. Considering a linear model for illustration, the prediction is given by

$$f(x) = w^T x + b, \tag{12}$$

where w is the weight vector, b is the bias and x is the input vector. Let x_m and y_m denote respectively the m-th training input vector and target output, $m = 1, \ldots, M$. The error function is given by

$$J = \frac{1}{2}\|w\|^2 + C\sum_{m=1}^{M} |y_m - f(x_m)|_\epsilon. \tag{13}$$

The first term in the error function is a term that penalizes model complexity. The second term is the ϵ-insensitive loss function, defined as $|y_m - f(x_m)|_\epsilon = max\{0, |y_m - f(x_m)| - \epsilon\}$. It does not penalize errors below ϵ, allowing some wiggle room for the parameters to move to reduce model complexity. It can be explained that the solution that minimizes the error function is given by

$$f(x) = \sum_{m=1}^{M} (\alpha_m^* - \alpha_m) x_m^T x + b, \tag{14}$$

where α_m and α_m^* are Lagrange multipliers. The training vectors giving non-zero Lagrange multipliers are called *support vectors*, and this is a key concept in SVR theory. Non-support vectors do not contribute directly to the solution, and the number of support vectors is some measure of model complexity (Cherkassky and Ma, 2004; Chalimourda *et al.*, 2004). This model is extended to the non-linear case through the concept of *kernel* κ, yielding

$$f(x) = \sum_{m=1}^{M} (\alpha_m^* - \alpha_m) \kappa(x_m^T x) + b. \tag{15}$$

In this paper, we have used the Gaussian kernel, which is a common kernel. Its width, σ_K, is the standard deviation of the Gaussian function.

3.3.5. Parameter optimization of individual models. All the methods discussed above require a parameter tuning process to extract the optimal performance. In this paper we have employed a different form of cross-validation, the popular model validation technique to tune the parameters of the above methods. For model validation, special care should be taken when the data are serially correlated (i.e, for time series data). More specifically, data points adjacent to or near the omitted observation(s) usually tend to be more similar to them than randomly selected ones, so the omitted observation(s) will be more easily predicted than the uncorrelated future observations they are meant to simulate. We have employed the hv-block cross-validation technique (Racine, 2000) to tune the parameters of all the methods. In this technique, a model is trained on a set of observations of size N_h and validated on a set of size N_v while the h-blocking assert near-independence of the training and validation data.

For a given method, let a tuning parameter α for a set of models $f(x, \alpha)$ be indexed. Then the hv-cross-validation function can be defined as

$$CV(\alpha) = \frac{1}{(N_h - 2v)N_v} \sum_{i=v}^{N_h - v} \left|\left| Y_{i:v} - \hat{f}^{-(i:h,v)}(x_{i:v}, \alpha) \right|\right|^2,$$

where $\hat{f}^{-(i:h,v)}(x_{i:v}, \alpha)$ is the α-th model fit with $2h + 2v + 1$ observations removed. The parameter h controls the dependence of the validation and training sets and is set to insure near-independence of these sets. The parameter v controls the relationship between the training set, validation set, and sample size. The parameter $\hat{\alpha}$ which minimizes the function $CV(\alpha)$ is chosen to construct the model $f(x, \hat{\alpha})$.

In this paper, we have used $h = 6$ and $v = 3$. That is, the size of each block $N_h = N/6$ or 15% of the original training set size and the size of the validation set $N_c = N/3$ or 33% of the original training set. For consistency of the results, we have repeated the hv-cross-validation process 20 times for each parameter of each method.

3.4. Multi-model hybrid forecasting method. In constructing the hybrid method, less complex *sub-models* of the aforementioned candidate models are preferred. The sub-models are constructed from the parameter grids of the candidate models. After construction of sufficient sub-models, a model ranking technique is employed to extract top ranked sub-models based on the predicted rainfall values. Then a model selection method is applied to select the most accurate top ranked sub-models to construct the hybrid multi-model forecasting method. The pseudo-code of the hybrid multi-model forecasting method is given in Algorithm 1.

3.4.1. Sub-model ranking using least angle regression. Efron *et al.* (2004) proposed least angle regression which provides the ranking of the variables according to their predictive performance. A convenient feature of LARS is that the resulting sequence of the covariates can be derived from the correlation matrix of the data (without the observations themselves).

Let $Y, X_1, \ldots, X_d$ be the standardized variables. Let r_j denote the correlation between X_j and Y, and R_X be the correlation matrix of the covariates $X_1, \ldots, X_d$.

Algorithm 1. Hybrid multi-model rainfall forecasting method.

```
Given a training set $\mathscr{L}$, in which $X \in \mathscr{R}^d$ is the matrix of lagged rainfall values and $y \in \mathscr{R}$ is the response rainfall, with $M$ models.
for $i \leftarrow 1$ to $M$ do
    Construct a parameter grid $G_i$ for the $i$-th model $P_i$. Construct $g$ submodels by picking different parameter settings randomly from the grid $G_i$.
    Train each submodel on $\mathscr{L}$.
end for
Use LARS to rank the total $g \times M$ submodels on the basis of the training responses.
Extract $\mathcal{M}^*$ top ranked models.
for $j \leftarrow 1$ to $\mathcal{M}^*$ do
    while $t < t_p$ ($t_p$ = time length of the training set) do
        Compute $loo_{jt} = \left( \frac{y_t - \hat{y_{jt}}}{1 - x_t^T (X^T X)^{-1} x_t} \right)^2$
    end while
    Compute the LOOCV error of the $j$-th model $P_j$, $\texttt{LOOCV}(P_j)$, as $LOO_j = \frac{1}{t_p} \sum_{t=1}^{t_p} (loo_{jt})$
    Store $LOO_j$ in matrix $L$
end for
Compute the threshold value, $\theta = \texttt{min}(L) + 3\texttt{Std}(L)$
// here min = minimum and Std = standard deviation
Select the submodel $P_k$, if $\texttt{LOOCV}(P_k) \leq \theta; \forall k = 1, \ldots, \mathcal{M}^*$
Construct the multi-model forecasting method with the selected submodels $\mathcal{M}^\theta$ ;
```

Suppose that X_m has the maximum absolute correlation r with Y and denote $s_m = \mathrm{sign}(r_m)$. Then X_m becomes the first *active variable* and the current prediction $\hat{\mu} \longleftarrow 0$ should be modified by moving along the direction of $s_m X_m$ up to a certain distance γ that can be expressed in terms of correlations between the variables. By determining γ, LARS simultaneously identifies the new covariate that will enter the model, that is, the second active variable.

In this paper, we have employed LARS to rank the sub-models based on the forecasting performance on the training responses. In addition to this, to reduce the computational complexity, we have used blocked LARS (Fraley and Hesterberg, 2009). This method performs the original LARS in blocks of the original variables, which speeds up the process. These weights are later utilized for weighted average combination (Timmermann, 2006) of the forecasts of the sub-models in the testing phase.

3.4.2. Sub-model selection using the leave-one-out error. Using LARS, only the top ranked sub-models are extracted, but to select the actual *best* sub-models we have employed the leave-one-out cross-validation method. The main disadvantage of the LOOCV method is that it is computationally burdensome if the dataset is large. Fortunately, the PREdiction Sum of Squares (PRESS) (Myers, 1990) statistic provides a direct and exact formula for the calculation of the LOO error for linear models (see the work of Syed (2011) for implementation).

In this paper, a threshold value of the LOO error is used for final selection of the top ranked sub-models. The threshold value θ (see Algorithm 1) is set in such a way that the sub-models with the LOO error within the +3SD (standard deviation) limit of the lowest LOO error are selected for creating the hybrid multi-model forecasting method.

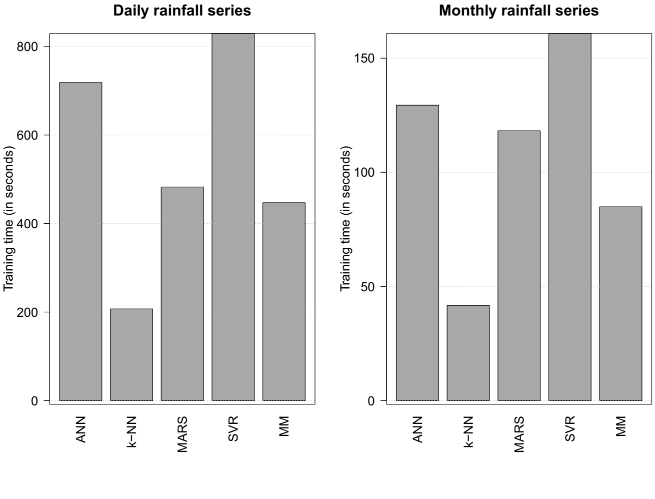


Fig. 2. Comparison of training times of individual forecasting methods with the multi-model: daily rainfall series (training set size = 1830) (left), monthly rainfall series (training set size = 360) (right).

3.4.3. Advantage of the proposed hybrid multi-model forecasting method. The proposed hybrid multi-model forecasting method consists in ranking the sub-models and finally selecting the sub-models with the LOOCV error within a threshold value. The final sub-model selection by the LOOCV error incurs an extra amount of variance in the prediction of each individual sub-models (Hastie *et al.*, 2009). The higher amount of variance causes individual sub-models to predict (forecast) different parts of the forecasting problem. As a result, when all these forecast results are combined (multi-model forecasting), this variance is reduced significantly, which is the underlying success of ensemble learners with high variance sub-learners (Hastie *et al.*, 2009). In this way, an approximately accurate forecasting can be performed by combining forecasting results from multi-models.

The proposed forecasting method is designed to construct sub-models from a narrow span of parameter values. The parameter values which exert lesser complexity are employed to construct each sub-model. This training procedure is carried out to facilitate the multi-model method making lesser computational burden, as parameter optimization of each sub-model would make the training process of the multi-model intractable. In this way, the proposed method becomes computationally more advantageous than single models. In Fig. 2, it can be seen that for both the rainfall series the multi-model (MM) forecasting method is computationally less expensive than the ANN, SVR and MARS.

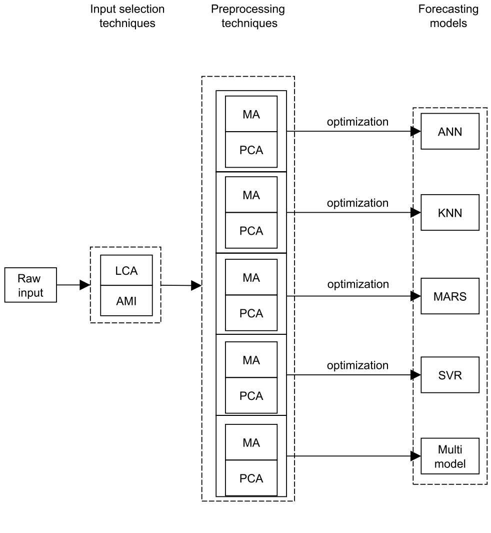


Fig. 3. Framework of training machine learning methods.

3.5. Statistical test for comparing forecast accuracy. We have employed a distribution-free test known as the Diebold–Marino test (Diebold and Mariano, 1995) to compare the difference between the forecasts of two competing methods. The main reason behind this test is that, it is a model-free test of forecast accuracy and can be directly applied for non-quadratic loss functions, multi-period forecasts, and forecast errors that are non-Gaussian, non-zero-mean, serially correlated, and contemporaneously correlated. In this paper, the test is carried out using the `R` software package named `forecast` (Hyndman *et al.*, 2012). In the experiments, we have used a 95% significance level for testing the forecast accuracy of the methods.

4. Experiment

In the experiments, we have split the data into two parts: (a) the training set, which is from 1975 to 2004, and (b) the test set, which is from 2005 to 2009. In the training phase, each of the individual models is trained with extensive parameter optimization. This means that every model is constructed with optimal values of the respective parameters. It should be noted that the optimization procedures for the ANN, MARS and SVR are computationally expensive. To decrease the computational complexity, a multi-model forecasting method is applied and a parameter grid is constructed with parameter values which enable the individual methods to train faster. For example, in the ANN, with small values of the decay parameter, the ANN will train slower, whereas with smaller values of kernel width (σ), the SVR method will train faster. This is to reduce the computational complexity of the hybrid multi-model forecasting method.

4.1. Implementation framework of training the models. Figure 3 illustrates the implementation framework of rainfall forecasting methods, where four individual machine learning methods and a hybrid multi-model method is conducted with two data preprocessing methods (dashed box). These acronyms in the column of “methods for model inputs” represent two methods to determine model inputs: LCA (Linear Correlation Analysis) (Sudheer *et al.*, 2002) and AMI (Average Mutual Information) (Fraser and Swinney, 1986).

4.2. Evaluation of model performances. Pearson’s correlation coefficient (r) or the coefficient of determination ($R^2 = r^2$), has been identified as inappropriate measures in hydrologic model evaluation by Legates and McCabe (1999). The Coefficient of Efficiency (CE) (Nash and Sutcliffe, 1970) is a good alternative to r or R^2 as a “goodness-of-fit” or a relative error measure in that it is sensitive to differences in the observed and forecasted means and variances. Legates and McCabe (1999) also suggested that a complete assessment of model performance should include at least one absolute error measure (e.g., Root Mean Square Error (RMSE)) as a necessary supplement to a relative error measure. Besides, the Persistence Index (PI) (Kitanidis and Bras, 1980) is adopted for the purpose of checking

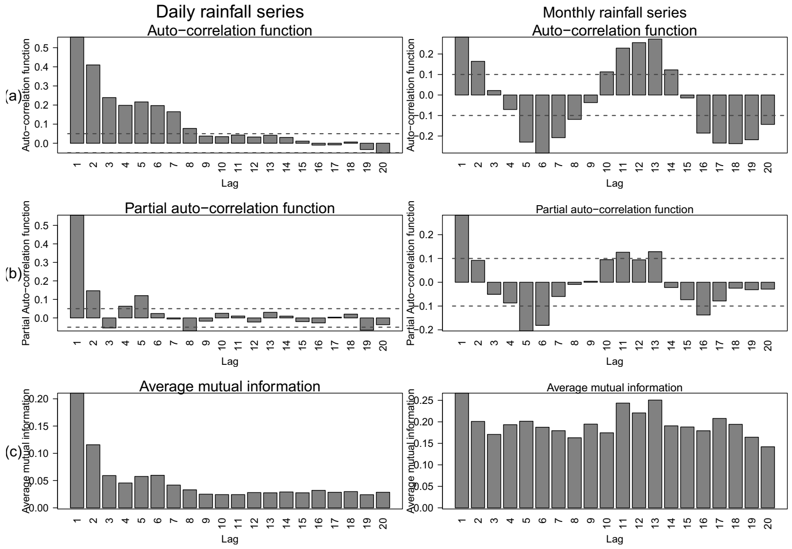


Fig. 4. Daily rainfall series (left) and monthly rainfall series (right): auto-correlation function plot (a), partial ACF plot (b), average mutual information plot (c).

the prediction lag effect. Three measures are therefore used in this study. They are listed below:

$$\mathrm{CE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{16}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{17}$$

$$\mathrm{PI} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y_{i-l})^2}. \tag{18}$$

In these equations, n is the number of observations, $\hat{y}_i$ stands for the forecasted flow, y_i represents the observed flow, $\bar{y}$ denotes the average observed flow, and y_{i-l} is the flow estimate from the so-call persistence model (or naive model) that basically takes the last flow observation (at time i minus the lead time l) as a prediction. CE and PI values of 1 stand for perfect fits. A small value of the PI may imply occurrence of lagged prediction.

5. Results and discussion

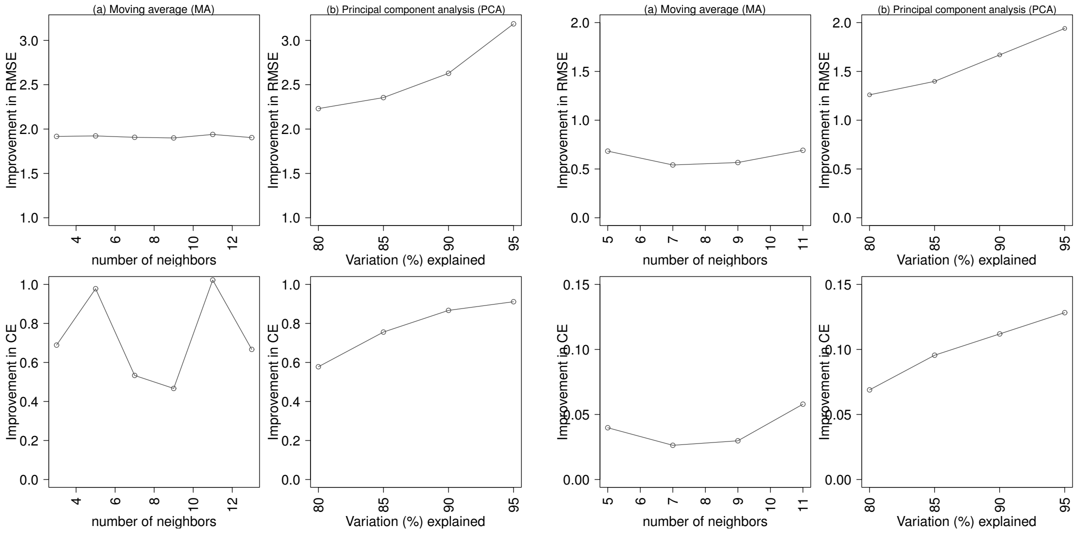
5.1. Determination of model inputs. The proposed hybrid multi-model method is used as the benchmark model to examine the input methods LCA and AMI in terms of the RMSE. The results are presented in Table 1 and are based on one-step ahead forecasting. We can see that for both daily and monthly rainfall forecasting, LCA has a slightly lower RMSE than the AMI method and in monthly forecasting LCA is significantly better than AMI. Considering this higher accuracy in forecasting and the convenience of operation, the LCA method is preferred in this study.

Table 1. Comparison of methods to determine model inputs using the hybrid multi-model forecasting method.

Data	Method	Inputs	RMSE
Daily rainfall	AMI	Last 2	31.34
	LCA	Last 3	**29.87**
Monthly rainfall	AMI	Last 13	36.75
	LCA	Last 12	**33.42●**

● Corresponding input selection method is significantly better than the competing input selection method at significance level = 95%

Figure 4 estimates the Autocorrelation Functions (ACFs), Partial Autocorrelation Functions (PACFs) and average mutual information, from lag 1 to lag 20 for the two rainfall series. AMI measures the general dependence of two variables whereas the ACF and the PACF show the dependence from the perspective of linearity. The first order autocorrelation and AMI of each data is large. The rapid decaying pattern of the PACF confirms the dominance of the autoregressive process, relative to the moving-averaging process revealed by the ACF. From Fig. 4(a), the ACF exhibits the peak at lag 3 and lag 13 for the daily and monthly rainfall series, respectively. In addition, Fig. 4(b) shows a significant correlation of the PACF at a 95% confidence level interval up to a 3 days and 12 months lag for the daily and monthly rainfall series. Therefore, 3 days and 12 months prior rainfall values have the most information to predict future rainfall and



(1) Daily rainfall series

(2) Monthly rainfall series

Fig. 5. Relative improvement in the RMSE and CE of the multi-model forecasting method after using MA (a) and PCA (b) as the pre-processing technique for the daily rainfall series (1) and monthly rainfall series (2). The improvement is computed from forecasting without any pre-processing technique.

are considered as input for daily and monthly rainfall time series modelling. From Fig. 4(c), the AMI value reaches the peak at lag 2 and lag 13 for the daily and monthly rainfall series, respectively.

5.2. Preprocessing techniques. In the forecasting experiment, the MA based preprocessing entails the window length k in Eqn. (1) to smooth the raw rainfall data. In the search of for an appropriate k, we have employed different values of k from 3 to 13. The smoothed data are then used to feed into the multi-model forecasting method. The targeted value of k corresponds to the optimal model performance in terms of the RMSE. PCA based on preprocessing is carried out for noise reduction by choosing the leading components (contributing most of the variance of the original rainfall data) to reconstruct rainfall series (depending on Eqn. (7)). The percentage V of total variance (according to Eqn. (6)) is set at four horizons, 80%, 85%, 90%, and 95%, for principal component selection.

For comparison between the two preprocessing techniques, the smoothed rainfall by the MA (with different k) and the reconstructed rainfall series by PCA (at different horizons) are fed into the multi-model forecasting method and the corresponding CE and RMSE are computed. The multi-model forecast method without any preprocessing is used as the benchmark model for this comparison. The relative difference (improvement) in CE and RMSE values after applying the preprocessing techniques is computed and reported in Fig. 5. From Fig. 5, it can be seen that regarding the RMSE values multi-model method based on PCA preprocessing has a higher amount of relative improvement than the MA based multi-model. The optimal value of k is 9 for the daily rainfall series and 7 for the monthly rainfall series. As expected, with an increasing value of V(%), the relative improvement also increased for PCA. Regarding the relative improvement in the CE, MA produced slightly higher improvement for a daily series, but PCA based multi-model produced much higher improvement for the monthly rainfall data. Considering the efficiency of PCA, the reconstructed rainfall series by PCA is fed to the other machine learning models.

5.3. Modelling rainfall. Table 2 presents the values of the metrics RMSE, CE and PI of the four individual models (ANN, k-NN, MARS, SVR) and the hybrid multi-model. All the metrics are calculated based on single step ahead forecasting. PCA based preprocessing is employed before the models are applied to extract the forecast values. Four horizons of the percentage of variation (V(%)) by PCA based preprocessing are used. The most suitable value of each metric (e.g., for the RMSE the lowest, for the CE and PI the highest) is marked bold for each data. We have employed the Diebold–Mariano test to check the significance of the difference between the forecast accuracy of the most accurate method (producing the lowest RMSE) and the other competing methods. In the table the methods which are significantly ‘worse’ than the most accurate method are marked with “○”. The comparison of forecast accuracy of all the methods is conducted at V(%) = 95. It should be noted that, at

Table 2. Forecasting results of the ANN, the k-NN, MARS, SVR and the multi-model forecasting method for daily and monthly rainfall series.

Datasets	V(%)	Metrics	ANN	k-NN	MARS	SVR	MM
Daily rainfall	80	RMSE	24.132	25.467	23.879	29.452	23.705
Series	85		19.382	18.316	17.745	27.581	15.630
	90		17.007	14.740	14.677	26.645	11.593
	95		14.633 ○	11.165 ○	11.610 ○	25.710 ○	**7.555**
	80	CE	0.9663	0.9647	0.9661	0.9477	0.9676
	85		0.9771	0.9800	0.9820	0.9557	0.9824
	90		0.9826	0.9876	0.9900	0.9598	0.9899
	95		0.9880	0.9952	**0.9980**	0.9638	0.9973
	80	PI	0.9159	0.8961	0.9258	0.8591	0.9369
	85		0.9407	0.9269	0.9486	0.8966	0.9560
	90		0.9530	0.9423	0.9601	0.9153	0.9656
	95		0.9654	0.9577	0.9715	0.9340	**0.9752**
Monthly rainfall	80	RMSE	40.927	69.899	47.925	38.413	39.672
series	85		39.477	69.693	38.776	33.645	35.180
	90		38.752	69.589	34.201	31.261	32.934
	95		38.027 ○	69.486 ○	29.857	**28.815**	29.688
	80	CE	0.8518	0.5677	0.7968	0.8695	0.8608
	85		0.8649	0.5733	0.8635	0.9059	0.8897
	90		0.8695	0.5751	0.8968	0.9221	0.9062
	95		0.8721	0.5728	0.9181	**0.9364**	0.9297
	80	PI	0.7903	0.7962	0.7921	0.7912	0.7923
	85		0.8292	0.8424	0.8447	0.8401	0.8502
	90		0.8466	0.8644	0.8721	0.8646	0.8782
	95		0.8620	0.8825	0.8934	0.8850	**0.9022**

○ Corresponding method is significantly worse than the best method.

other horizons of V(%), the values of the metrics of the competing methods are considerably higher than that of V(%) = 95.

It can be seen from Table 2 that, in the case of the daily rainfall series, the hybrid multi-model produced the most accurate forecast out of all the individual models, considering that it produced the lowest RMSE. The accuracy of the forecasts of other methods are insignificant compared the multi-model forecast. The forecast values of the multi-model method are also least affected by the time lag compared the forecast of other models as the PI of the multi-model method is highest among all the methods. The MARS method produced the best CE value, indicating better fit from among the methods; the multi-model produced a slightly lower CE than MARS.

For the monthly rainfall series, SVR produced the most accurate forecast (lowest RMSE) and the best rainfall mapping (highest CE). Forecast values produced by the ANN and the k-NN are significantly worse than those of SVR. The forecast accuracy of MARS and the hybrid multi-model has no significant difference than the SVR. The multi-model method again is least affected by the time lag. The forecast values of the ANN and the

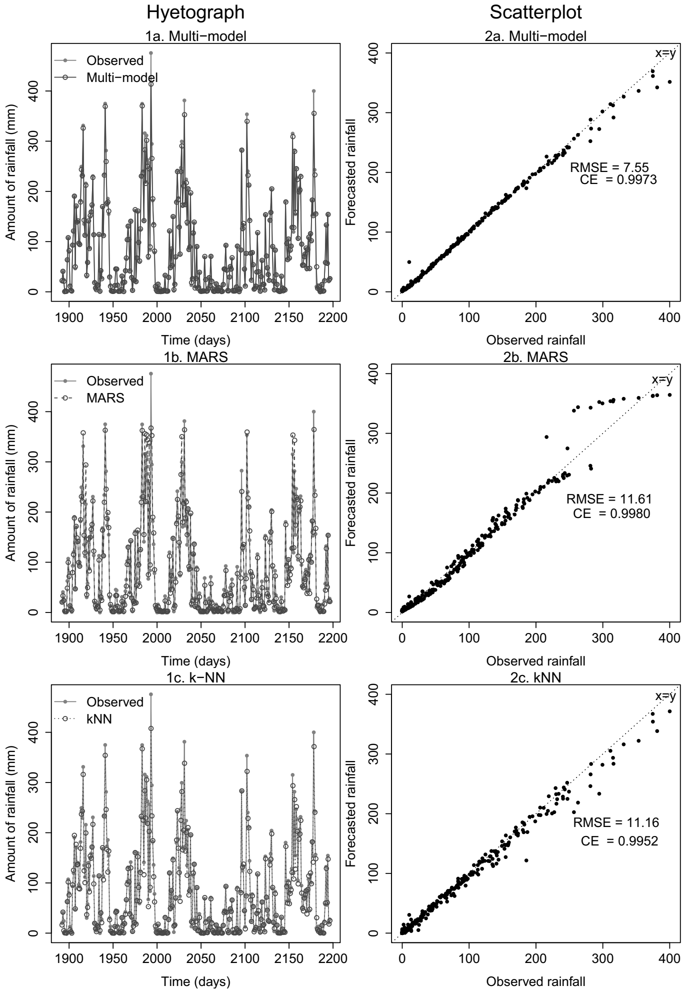


Fig. 6. Hyetograph (a) and scatter plot (b) of the observed and forecasted rainfall of the daily rainfall series of top performed methods.

k-NN are negligible for the monthly rainfall problem.

In Figs. 6 and 7, the scatter plots and hyetographs of the forecasted and observed rainfall values of the daily and monthly rainfall series are given for a visual inspection of the forecasting performance of the methods. In each figure, the top three methods are included according to the metric values of Table 2. The hyetographs are plotted in a selected range for better visual comparison.

For the daily rainfall, other than the multi-model method, MARS and the k-NN are included. It can be seen that the multi-model method estimates the higher-intensity daily rainfalls better than other methods. The lower value of the RMSE demonstrated this fact, too. In the case of the monthly rainfall, SVR and MARS are included with the proposed method. From the scatter plots, the medium-intensity monthly rainfalls are estimated better by the multi-model method, although high-intensity rainfalls (or peak values) are still underestimated compared to SVR.

6. Conclusion

This paper investigates the use of several machine learning methods and particularly suggests to employ a hybrid multi-model method coupled with model ranking and selection for improving two rainfall forecasting problems in the Fukuoka city. The rainfall series include the daily and monthly rainfall of the Fukuoka city. For reasonable evaluation of the performance of the hybrid method, its constituent models (ANN, k-NN, MARS and SVR) are separately constructed and used for the purpose of comparison. In the process of model construction, model inputs and data-preprocessing techniques are carefully analysed and discussed. The following conclusions are obtained based on this study:

(a) LCA can be assessed as a more effective and efficient method among the two input techniques due to simplicity in computation and superior capability of forecasting.

(b) Regarding the data preprocessing techniques,

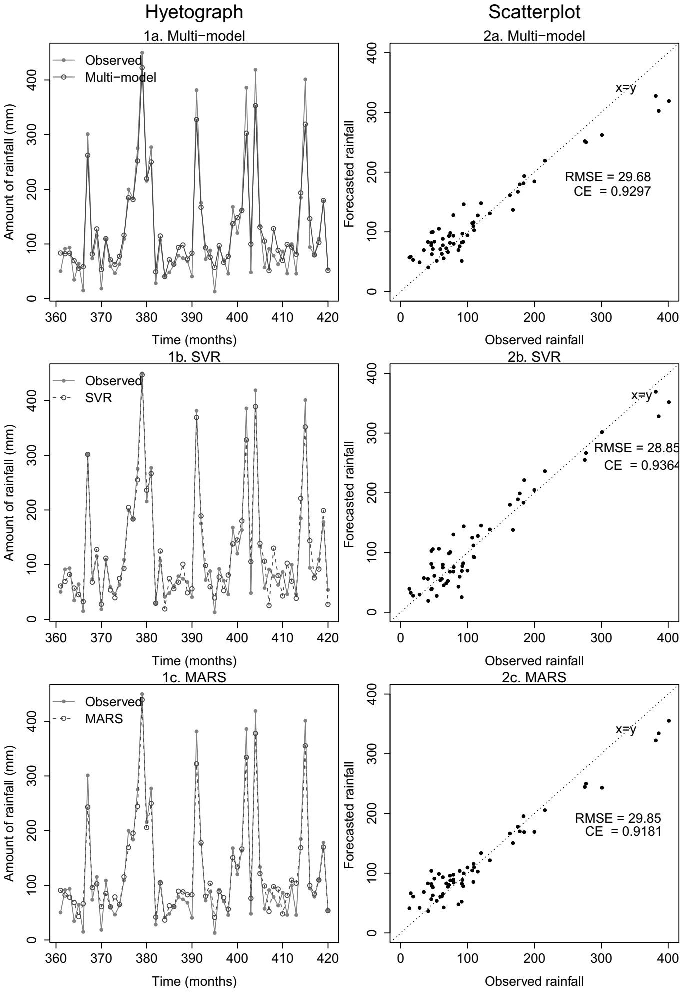


Fig. 7. Hyetograph (a) and scatter plot (b) of the observed and forecasted rainfall of the monthly rainfall series of the top performed methods.

the effect of MA is negligible (compared with the no-preprocessing mode) in improving the performance of the hybrid forecasting method.

(c) PCA is more efficient as a data preprocessing technique. Specifically, this is the case for PCA for the purpose of noise reduction. Results show that PCA improves the hybrid method performance.

(d) The hybrid method produces more accurate forecast than the single models for the daily rainfall series. Among the single models, SVR performs better and produced a better forecast than the hybrid method for the monthly rainfall series.

Acknowledgment

The authors are thankful to anonymous reviewers for their helpful comments and suggestions.

References

Abrahart, R.J. and See, L. (2002). Multi-model data fusion for river flow forecasting: An evaluation of six alternative methods based on two contrasting catchments, *Hydrology and Earth System Sciences* **6**(4): 655–670.

Baruque, B., Porras, S. and Corchado, E. (2011). Hybrid classification ensemble using topology-preserving clustering, *New Generation Computing* **29**(3): 329–344.

Chalimourda, A., Schölkopf, B. and Smola, A.J. (2004). Experimentally optimal ν in support vector regression for different noise models and parameter settings, *Neural Networks: The Official Journal of the International Neural Network Society* **17**(1): 127–41.

Cherkassky, V. and Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks: The Official Journal of the International Neural Network Society* **17**(1): 113–26.

Coulibaly, P., Haché, M., Fortin, V. and Bobée, B. (2005). Improving daily reservoir inflow forecasts with model combination, *Journal of Hydrologic Engineering* **10**(2): 91.

Dawson, C.W. and Wilby, R.L. (2001). Hydrological modelling using artificial neural networks, *Progress in Physical Geography* **25**(1): 80–108.

De Vos, N.J. and Rientjes, T.H.M. (2005). Constraints of artificial neural networks for rainfall-runoff modelling: Trade-offs in hydrological state representation and model evaluation, *Hydrology and Earth System Sciences* **9**(1–2): 111–126.

Deng, Y.-F., Jin, X. and Zhong, Y.-X. (2005). Ensemble SVR for prediction of time series, *Proceedings of the International Conference on Machine Learning and Cybernetics, Guangzhou, China*, Vol. 2, pp. 734–748.

Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy, *Journal of Business & Economic Statistics* **13**(3): 253–263.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics* **32**(2): 407–499.

Everingham, Y.L., Smyth, C.W. and Inman-Bamber, N.G. (2009). Ensemble data mining approaches to forecast regional sugarcane crop production, *Agricultural and Forest Meteorology* **149**(3–4): 689–696.

Fraley, C. and Hesterberg, T. (2009). Least angle regression and LASSO for large datasets, *Statistical Analysis and Data Mining* **1**(4): 251–259.

Fraser, A.M. and Swinney, H.L. (1986). Independent coordinates for strange attractors from mutual information, *Physical Review A* **33**(2): 1134–1140.

Friedman, J.H. (1991). Multivariate adaptive regression splines, *Annals of Statistics* **19**(1): 1–67.

Gheyas, I.A. and Smith, L.S. (2011). A novel neural network ensemble architecture for time series forecasting, *Neurocomputing* **74**(18): 3855–3864.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd Edn., Springer, New York, NY.

Hong, W. (2008). Rainfall forecasting by technological machine learning models, *Applied Mathematics and Computation* **200**(1): 41–57.

Hyndman, R.J., Slava R. and Schmidt, D. (2012). *forecast: Forecasting functions for time series and linear models*, R package version 3.19, http://CRAN.R-project.org/package=forecast.

Kim, T., Heo, J.-H. and Jeong, C.-S. (2006). Multireservoir system optimization in the Han River basin using multi-objective genetic algorithms, *Hydrological Processes* **20**(9): 2057–2075.

Kitanidis, P.K. and Bras, R.L. (1980). Real-time forecasting with a conceptual hydrologic model, 2: Application and results, *Water Resources Research* **16**(6): 1034–1044.

Lee, C.F., Lee, J.C. and Lee, A.C. (2000). *Statistics for Business and Financial Economics,* 2nd Edn., World Scientific, Singapore.

Legates, D.R. and McCabe, G.J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, *Water Resources Research* **35**(1): 233–241.

Li, P.W. and Lai, E.S.T. (2004). Short-range quantitative precipitation forecasting in Hong Kong, *Development* **288**(1–2): 189–209.

Myers, R.H. (1990). *Classical and Modern Regression with Applications*, Duxbury, Boston, MA.

Nash, J. and Sutcliffe, J. (1970). River flow forecasting through conceptual models, I: A discussion of principles, *Journal of Hydrology* **10**(3): 282–290.

Newbold, P., Carlson, W. and Thorne, B. (2007). *Statistics for Business and Economics*, 6th Edn., Prentice Hall, Upper Saddle River, NJ.

Pucheta, J., Patino, D. and Kuchen, B. (2009). A statistically dependent approach for the monthly rainfall forecast from one point observations, *in* D. Li and Z. Chunjiang (Eds.), *Computer and Computing Technologies in Agriculture II, Volume 2*, IFIP Advances in Information and Communication Technology, Vol. 294, Springer, Boston, MA, pp. 787–798.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation, *Journal of Econometrics* **99**(1): 39–61.

Siwek, K., Osowski, S., Szupiluk, R. (2009). Ensemble neural network approach for accurate load forecasting in a power system, *International Journal of Applied Mathematics and Computer Science* **19**(2): 303–315, DOI: 10.2478/v10006-009-0026-2.

Schölkopf, B. and Smola, A.J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Adaptive Computation and Machine Learning, Vol. 98, MIT Press, Cambridge, MA.

Schölkopf, B. and Smola, A.J. (2004). A tutorial on support vector regression, *Statistics and Computing* **14**(3): 199–122.

Shrestha, D.L. and Solomatine, D.P. (2006). Machine learning approaches for estimation of prediction interval for the model output, *Neural Networks* **19**(2): 225–235.

Solomatine, D.P. and Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches, *Journal of Hydroinformatics* **10**(1): 3.

Sudheer, K.P., Gosain, A.K. and Ramasastri, K.S. (2002). A data-driven algorithm for constructing artificial neural network rainfall-runoff models, *Hydrological Processes* **16**(6): 1325–1330.

Syed, A.R. (2011). A review of cross validation and adaptive model selection, *Statistics*, Mathematics Theses, Georgia State University, Arlanta, GA, Paper 99.

Timmermann, A. (2006). Forecast combinations, *in* G. Elliott, C. Granger and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Elsevier, Amsterdam, Chapter 4, pp. 135-196.

Wichard, J. (2011). Forecasting the NN5 time series with hybrid models, *International Journal of Forecasting* **27**(3): 700–707.

Wichard, J. and Ogorzalek, M. (2007). Time series prediction with ensemble models applied to the CATS benchmark, *Neurocomputing* **70**(13–15): 2371–2378.

Wu, C., Chau, K. and Li, Y. (2008). River stage prediction based on a distributed support vector regression, *Journal of Hydrology* **358**(1–2): 96–111.

Xiong, L., Shamseldin, A. Y. and Oconnor, K. (2001). A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi–Sugeno fuzzy system, *Journal of Hydrology* **245**(1-4): 196–217.

Yang, Y., Lin, H., Guo, Z. and Jiang, J. (2007). A data mining approach for heavy rainfall forecasting based on satellite image sequence analysis, *Computers Geosciences* **33**(1): 20–30.

Zaman, M. and Hirose, H. (2011). Classification performance of bagging and boosting type ensemble methods with small training sets, *New Generation Computing* **29**(3): 277–292.

S. Monira Sumi was born in 1981. She received the B.Sc. and M.Sc. degrees from the Department of Statistics, Rajshahi University, in 2005 and 2006, respectively. She joined Shafi Consultancy in 2006 as a statistical programmer, where she analyzed medical trial data using SAS. She started her Ph.D. study at the Kyushu Institute of Technology in 2010.

M. Faisal Zaman was born in 1981. He received the B.Sc. and M.Sc. degrees from the Department of Statistics, Rajshahi University, in 2005 and 2006, respectively. He joined Shafi Consultancy in 2006 as a statistical programmer, where he analyzed medical trial data using SAS. He obtained his Ph.D degree in 2011. Currently he is a post-doc researcher at the Kyushu Institute of Technology. He is member of the IEEE Computer Science Society. His research interests lie in constructing statistical data mining models for biological and climate data.

Hideo Hirose was born in 1951 in Japan. He obtained the Baccalaureate degree in mathematics from Kyushu University and the Dr.Eng. degree from Nagoya University in 1977 and 1988, respectively. He worked for Takaoka Electric Manufacturing Co., Ltd. from 1977 to 1995, and served as the vice research director from 1988 to 1995. He was a professor at Hiroshima City University from 1995 to 1998. Since 1998, he has been a professor at Kyushu Institute of Technology. His interests include a variety of numerical computations such as finite element analysis, computational fluid dynamics, transient analysis of electrical networks, reliability engineering, and statistical data analysis. He is a member of the Institute of Electrical Engineers of Japan, the Information Processing Society of Japan, the American Statistical Association, the Institute of Mathematical Statistics, the American Mathematical Society, the Society for Industrial and Applied Mathematics, the Mathematical Programming Society, and the Association for Computing Machinery..

Received: 10 November 2011
Revised: 11 May 2012