ELSEVIER

Journal of Hydrology 241 (2001) 153–176

Journal of Hydrology

www.elsevier.com/locate/jhydrol

# Groups and neural networks based streamflow data infilling procedures

M. Khalil[a,1], U.S. Panu[a,b,*], W.C. Lennox[a,1]

[a]*Civil Engineering Department, University of Waterloo, Waterloo, Ont., Canada N2L 3G1*
[b]*Civil Engineering Department, Lakehead University, Thunder Bay, Ont., Canada P7B 3G1*

Received 5 August 1999; revised 17 March 2000; accepted 17 July 2000

**Abstract**

Hydrologic data sets are often of short duration and also suffer from missing data values. For estimation and/or extrapolation, the presence of missing data not only affects the choice of a particular method of analysis but also the resulting decision making process. Existing methods are based on the single-valued data approach and thus do not involve the effect of seasonal grouping (or segmentation) in hydrologic data prediction. Based on concepts and properties of groups and artificial neural networks, this paper develops a segment estimation model for infilling of missing hydrologic records. Efficacy of the proposed model is demonstrated through applications to a number of natural watersheds. The group-based neural network models are shown to retain relevant properties of the historical streamflows both at the auto- and cross-variate series levels. Further, the group-based neural network models are found to closely infill the missing peak flows and also the moderate flows. The results suggest that infilling of data gaps of streamflows based on the concept of neural networks and group-valued data approach is a reasonable alternative, and warrants further investigations. 

*Keywords*: Data infilling; Data groups; Nonlinear modeling; Seasonal segmentation; Neural networks; Multivariate time series

## 1. Introduction

Effective planning, management, and control of water resource systems require considerable data on numerous hydrological variables such as rainfall, streamflow, and temperature. Invariably, the data sets are recorded in time and are referred to as time series. These series are analyzed using statistical methods to evaluate the parameter of interest so as to arrive at a suitable decision support system for management and control purposes. However, a majority of the time series often exhibits some form of deficiency due to the presence of gaps, discontinuities, and inadequate length. Such deficiencies in hydrological time series are attributable, among other, to the malfunctioning of monitoring equipment (electric or mechanical), effects of natural phenomena (e.g. earthquakes, hurricanes, landslides, etc.), data transmission, storage, and retrieval processes. Time series methods, among other, do not tolerate missing observations and therefore, numerous data infilling procedures have evolved in various scientific disciplines to deal with incomplete data sets.

There are two basic problems when dealing with inadequate hydrological time series. In the first case, the time series are of adequate time length but suffer from the presence of data gaps. In this case, data

* Corresponding author. Fax: +1-807-343-8928.
*E-mail addresses:* mmkhalil@engmail.uwaterloo.ca (M. Khalil), uspanu@lakeheadu.ca (U.S. Panu), wclennox@uwaterloo.ca (W.C. Lennox).
[1] Fax: +1-519-888-6197.

0022-1694/01/$ - see front matter 
PII: S0022-1694(00)00332-2

infilling has been referred to as data augmentation. In the second case, the historical time span of the data series is inadequate and thus efforts are made to extend the historical time span to a desired one. This latter case of data infilling has commonly been referred to as data extension.

In order to design water resources systems that provide satisfactory performance in the future, all relevant characteristics of the time series should be taken into consideration while modeling the hydrological data. For example, monthly streamflow time series exhibit the presence of heterogeneous relationships among data points (Panu and Unny, 1980) and also exhibit nonlinear relationship among data paints. This paper develops a segment estimation model for infilling of missing hydrologic records that takes into account the nonlinear characteristics of hydrologic data. By incorporating seasonal segments as integral components of the model structure, the proposed model overcomes the problems that invariably arise from ignoring the role of heterogeneous relationships among data points. Efficacy of the proposed model(s) is examined in relation to their infilling ability of missing streamflows.

## 2. Groups and groupings in hydrologic data

Groups are considered to possess certain characteristics that distinguish them uniquely as separate entities. That is, the variables (or elements) that form a group represent a certain uniqueness of the group. In time series, time dependent observations, which form groups collectively, reflect the unique characteristics of such groups. In this paper, for example, hydrologic observations corresponding to seasonal periods of dry and wet are considered to form groups (Khalil et al., 1998a,b; Unny et al., 1981; Panu et al., 1978; and others). On the other hand, the term grouping refers to the formation of cluster in-groups.

## 3. Assessment of existing streamflow data infilling techniques

A concise and comprehensive review of existing techniques of data infilling is provided by Khalil et al. (1998a). Majority of data infilling procedures and techniques are based on single-valued approach with the exception of the group-valued approach proposed by Panu and associates (Panu, 1991; Panu and Afza, 1993; Goodier and Panu, 1994; Khalil et al., 1998a,b). A brief but relevant discussion of various pertinent data infilling procedures can be found elsewhere (Panu et al., 2000).

### *3.1. Single-valued data approach*

In this approach, all models have one thing in common that each data value within a historical record is used as an informative unit by itself and thus effectively ignores the information that may propagate if similar data values are treated as groups. This means the properties of the data series are not changed with time for both the mean and variance (second order stationarity). Additional discussion of relevant procedures, such as those of the bivariate, multivariate, mixed-variate regression methods, and autovariate time series and multivariate time series methods has been provided by Panu et al. (2000).

### *3.2. Group-valued data approach*

In this approach, the data with similar attributes are collected together to form groups. Each group is considered to satisfy the requirement of weak stationarity. The formation of a group by the use of data values of same properties assures group homogeneity (Panu et al., 1978). This homogeneity helps in the extraction of satisfactory information for better estimation accuracy of data (Unny et al., 1981). It is in this regard that there are some advantages of forming seasonal groups towards the development of data infilling models. Such advantages include among others: (a) the stability of confidence limits over the duration of the group; (b) the estimation of correlation coefficient and other similar parameters of interest over the homogeneous sub-set rather than across the entire set of data; and (c) the formation of data groups does not require the assumption of stationarity as presently used in statistical formulations. In essence, it would require stationarity of the other kind, for example, separate stationarity for each type of groups within the data set (Khalil et al., 1998a,b; Goodier and Panu, 1994; Panu and Afza, 1993). The pattern recognition (PR) based methods are capable of dealing with groups and their properties for enhanced data infilling

of missing values. For brevity, such methods are briefly presented below.

*3.2.1. PR based methods*

The variability of duration in a data gap is inconsequential for the single-valued data approach because a single value is estimated at each time step. The resulting error of estimation in single-valued approach increases after the very first infilled value. However, in the group-valued data approach, the variability in duration of a data gap plays a significant role (Panu, 1991). It is noted that the estimation error over the duration of a seasonal segment is time invariant (Panu, 1991; Panu and Afza, 1993; Goodier and Panu, 1994). Additional considerations pertaining to the definition of a gap (Panu, 1991) can significantly reduce the error of estimation in infilling of missing values. It is in this vein that Afza and Panu (1992) proposed two types of models, namely the auto series and cross series models of monthly streamflows to investigate the role of seasonal groups and their characteristics in the estimation of missing data values. In these models, characteristics of groups of data of the subject river (i.e. the river with missing values) and the base river (i.e. the nearby river(s) without any missing values of concurrent records) are utilized. In the auto series models, the missing data values are estimated based on the conditional projection of the most probable flow group (i.e. a flow pattern). The relationships among flow patterns are considered to follow the Markovian dependence (Panu, 1991; Afza and Panu, 1992). In the cross series models, the flow pattern is projected into the data gap conditionally based on the occurrence of a flow pattern during the data gap in a participating base river. The configuration of the projected flow pattern in the data gap is obtained based on the assumption of joint multivariate normality (i.e. mean vector and covariance matrix). The elemental values of the projected flow pattern are obtained based on the multivariate probability distribution (Johnson and Wichern, 1988).

Based on such considerations, the missing values can be infilled as a group (i.e. a vector) rather than as individual values (Afza and Panu, 1992). Conventionally, the regression and maintenance of variance extension (MOVE) models (Hirsch, 1982) compute parameters from the sample data consisting of all available data irrespective of the existence of different seasons (or data groups). Three sampling schemes were utilized (Afza and Panu, 1992) in the evaluation of the effectiveness of data groups for infilling purposes. The calibration of regression and MOVE models in the first, second, and third sampling schemes were, respectively, based on sample data consisting of all available data and thus these models are called AREG and AMOVE; sample data from seasonal groups and hence models are referred to as SREG and SMOVE; and sample data from specified sub-groups within a seasonal group and therefore models are denoted by SSREG and SSMOVE. These models were evaluated for the autovariate series and cross-variate series cases. In all such sampling schemes, the cross-variate series models performed better than the autovariate series models (Panu, 1991; Afza and Panu, 1992). Goodier and Panu (1994) have reported additional modifications of these models for the mixed multivariate scenario with satisfactory results.

These models deal only with completely missing seasonal segments, and thus do not evaluate the case of partially missing segments. Also grouping of the monthly flow into a six-month period may not strictly be similar, and a multivariate model of lesser dimension (three or four months within the group) can be more reliable. Another improvement may be achieved by combining the two algorithms (i.e. the stochastic structures dealing with the autovariate-structure and the cross-variate-structure) together to form a model that deals with both type of relationships between two nearby rivers having the same seasonal properties.

*3.2.2. Artificial neural networks based methods*

The literature related to the use of artificial neural networks (ANN) with missing or incomplete data is sparse. Only a limited number of reports and research publications are available. Karunanithi et al. (1994) used ANN to predict river flows, and compared the results with those obtained from parametric models. It was further observed that the results obtained from ANN-based models were more favorable. Besides applications in water resources, the use of ANN-based models by Gupta and Lam (1996) has been reported for estimating missing values in a multivariate data set. The results of ANN-based models were found to be more accurate than those obtained by iterative regression analysis. Further, the focus of

some contributions has been on improving the learning capabilities of ANN-based methods from incomplete training data sets, and also for handling inputs with missing attributes (Ishibuchi et al., 1993; Pedreira and Parente, 1995). For streamflow predictions, Elshorbagy et al. (2000) found that ANN-based models performed better than those based on linear and nonlinear regressions.

### *3.3. Concluding remarks on data infilling procedures*

With the exception of the group-valued data approach (i.e. PR based methods), the previously proposed procedures implicitly or explicitly invoke the assumption of linear relationships between variables. Tong (1983) described the drawbacks of linear models and pointed out their inadequacy in the prediction of the occurrence of sudden bursts of streamflows with large amplitudes at random time intervals. Tiao and Tsay (1989) also described some of the difficulties that may occur with linear relationships in multivariate models. Due to these difficulties, many nonlinear statistical models have been developed (Granger and Newbold, 1986; Tong, 1990). Despite the significant progress achieved during the last decade, it is still difficult to formulate reasonable nonlinear models (Tong, 1983), because of simplification made in the modeling stage.

Recent advances in ANN are encouraging (Tong, 1983; Granger and Newbold, 1986; Tiao and Tsay, 1989; Tong, 1990). The success of ANN in modeling dynamical systems in other fields of sciences suggests that the ANN approach may prove effective and efficient in the development of data infilling procedures in hydrological data series.

## 4. Significance of neural networks in monthly streamflow data

Monthly streamflows are often short and exhibit a nonlinear multi-variable nature and this can make it difficult for linear models (e.g. multiple regression, seasonal ARIMA, and PR based techniques) to accurately infill the data gaps. Alternative methods that can deal with this complexity are the focus of considerable research. ANN, coupled with seasonal grouping, appear to be one such alternative to the linear statistical methods in identifying and studying the intricate nature of such time series.

The use of an ANNs to manage the stochastic variations of the hydrologic data has been proposed by many researchers (Hsu et al., 1995; Raman and Sunilkumar, 1995; Kang et al., 1993). Some studies in which ANN models have been applied to solve problems concerning missing data and data forecasting in hydrology have been reported (Tanaka, 1996). As an example, Hsu et al. (1995) proposed a very comprehensive study for developing and applying an ANN algorithm to a rainfall–runoff time-series. Using a combination of linear least squares and multi-start simplex optimization, Hsu et al., used a new algorithm named "linear least squares simplex" (LLSSIM) for identifying the structure and parameters of a three layer feed-forward ANN model. Comparing this algorithm with a linear auto-regressive moving average with exogenous input (ARMAX), the ANN model was found to provide a better representation of the system than the ARMAX time series approach. However, it is envisioned that the inclusion of seasonal segmentation technique in ANN based rainfall–runoff modeling could significantly improve the prediction accuracy.

Raman and Sunilkumar (1995) presented another comparison between the AR and ANN models in simulating the monthly flows for the case of two reservoirs. Again, the ANN models provided more promising results than the AR model. French et al. (1992) were capable of forecasting the complex temporal and spatial distribution of rainfall using ANN. Kang et al. (1993) used ANN for the daily and hourly streamflow forecasting. Wong et al. (1994) successfully predicted the missing pH values in the Great Lakes data using the cluster techniques in identifying the most correlated parameters. Tanaka (1996) used stochastic neural networks and the EM algorithm (Streit and Luginbuhl, 1994) in the estimation of missing data in a plant model. None of the previous studies have investigated the stopping criteria and overfitting error. Such an error may contribute to an erroneous estimation of missing data values. Lachtermacher and Fuller (1994) were aware of this problem in their neural based time series forecasting model.

Based on the argument that the number of

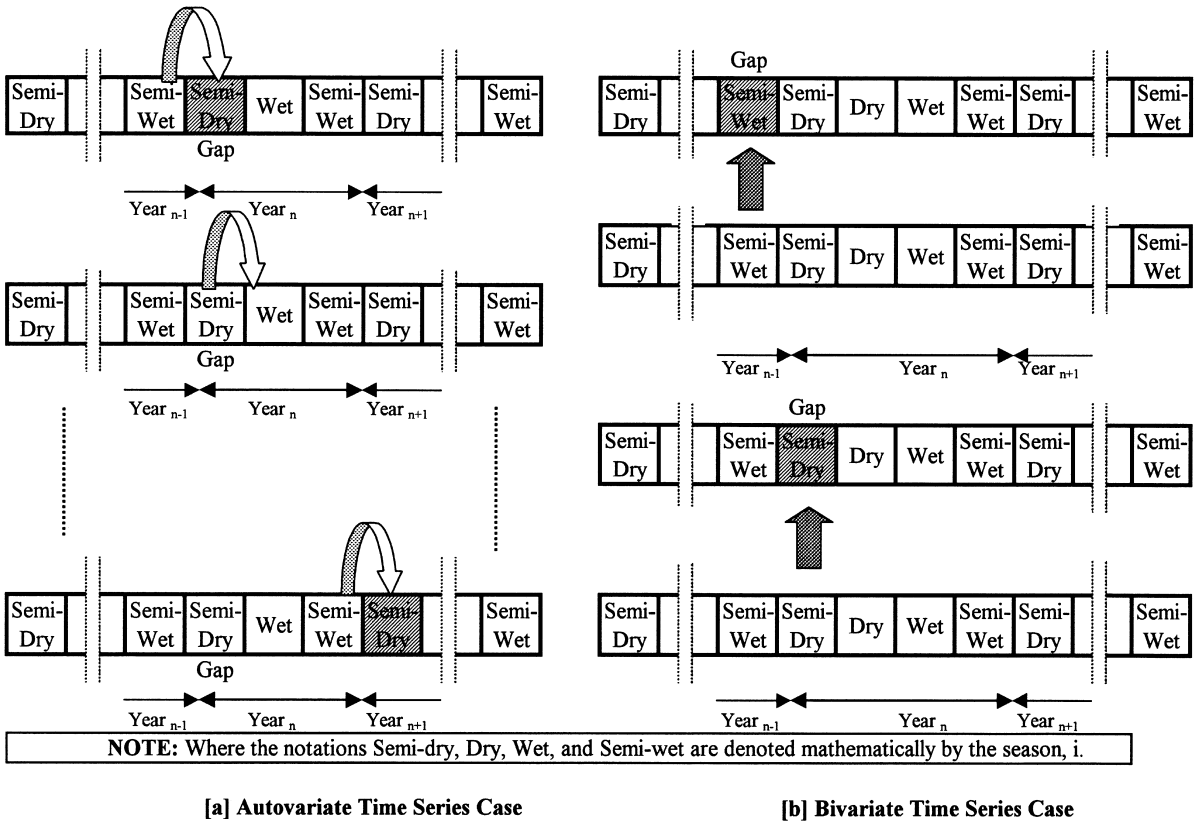



Fig. 1. Stochastic Infilling of Data Gaps: (a) Autovariate Time Series Case and (b) Bivariate Time Series Case.

parameters in a regression equation should, at the most, be equal to one fourth of the number of observations (McCuen, 1993), Tokar (1996) used a simple rule (as described later in Eq. (8)) in rainfall/runoff modeling, and also applied the rule to ANN techniques to compute the maximum number of nodes in a network. In addition, the neural networks were found to provide better prediction than regression techniques for classifying the daily data into wet, dry, and average days. It is with this view that the internal structure (i.e. the intra-structure) of groups is examined with a view to describing this structure through the use of ANN, PR, and AR procedures. The stochastic interrelationship (i.e. the inter-structure) among various groups is explored in an effort to develop a stochastic methodology for infilling one or more of either full or partial missing segments.

## 5. Development of data infilling models based on neural and group concepts

Based on the consideration of structural composition of neural networks and the group-valued data approach, data infilling methods (or models) are proposed for the cases: (i) a single data series with gaps; and (ii) a data series with gaps along with the availability of one (or more) concurrent but complete data series from neighboring stations. Two types of methods, namely, the autovariate time series (i.e. the series with gaps) and the bivariate time series (i.e. the series with gaps and another series with complete but concurrent records) are proposed.

In the autovariate-series case, the relationships among groups within the data series with missing values are utilized. In the bivariate-series case, the relationships among groups of two (or more)

concurrently occurring data series (one series with missing data values and another series with complete data values) are utilized. Further, in the bivariate-series case, one or more data series with complete records may be available at an upstream/downstream location, or at a tributary of the same river, or at a tributary of an adjacent river. Concurrent as well as complete data series need not be of streamflows but could be of rainfall, temperature, evaporation, etc. In this paper, the concurrent but complete data series considered are only of streamflows.

The structural composition of groups and relationships within groups are characterized through consideration of neural concepts. The structural complexity of groups is specifically addressed through the use of a specific set of neurons in the input and output layers. For example, the number of elements in the input-seasonal-segments and the number of elements in the output-seasonal-segments, respectively, specify the number of neurons in the input-layer and output-layer. Additional complexity in seasonal segments within a hydrological year and over the entire data series is considered through the specification of a number of hidden layers and the number of neurons in each hidden layer. The optimal number of hidden layers and the number of neurons in the hidden layers are normally obtained through experimentation with the data set.

In this paper, the notation followed to describe a multi-layer neural network is NN ($n$, $n_{hi}$, $t$), where $n$ is the number of neurons in the input-layer, $n_{hi}$, is the number of neurons in the $i$th hidden layer, and $t$ is the number of neurons in the output layer. The number of neurons in the input-layer and output-layer depends upon the number elements (i.e. months) defining a seasonal segment (i.e. a group). The parameters of the neural network (e.g. the number of hidden layer(s), the number of nodes in the hidden layer(s), the number of epochs used to stop the network) were experimentally determined (Khalil et al., 1998b). The parameters, which minimize the mean squared error (MSE) between the actual and the infilled values, were retained in the models.

### 5.1. Multi-layer-feed-forward autovariate series model

The formulation of an autovariate-series model (ASM) to infill the missing seasonal gap(s) in a single seasonal streamflow data set is conceptualized based on the group-valued data approach, the autovariate lag-one series model of the type suggested by Panu (1991) and developed by Afza and Panu (1992). In this paper, a Markovian relationship between the seasonal segment prior to the missing seasonal segment and the missing seasonal segment is assumed (Fig. 1a). The governing equation for this type of model can be expressed as:

$$Q_i = f(Q_{i-1}) \tag{1}$$

where, $Q_i$ is a vector of monthly streamflows at season $i$, while $Q_{i-1}$ is a vector of monthly streamflows at season $i-1$. The multi-layer-feed-forward autovariate lag-one series model (MASM) is formulated as a fully connected back propagation network, involving input, hidden, and output layers. The number of nodes in the input and in the hidden layers depends on the number of months that define a season in the autovariate data set of the subject river. The network is based on a NN (3,7,3) configuration (Khalil et al., 1998b). A nonlinear logistic (sigmoid) function is used for activation to map the nonlinearity of streamflow data.

### 5.2. Multi-layer feed-forward bivariate series model

The bivariate series model (BSM) deals with infilling of the missing values in the subject river (i.e. the river with missing data values) and uses the prior information obtained from one or more similar data series from nearby rivers also called the base rivers (Fig. 1b). The streamflows from base rivers exhibit synchronous seasonality and are cross-correlated to each other. A lag-zero BSM of the type suggested by Panu (1991) and developed by Afza and Panu (1992) is conceptualized to infill the missing seasonal gap(s) in a seasonal streamflow data set. In this model, the data series at base sites can be any other nearby streamflow data, any precipitation data or any information that strongly affects the streamflows at the subject river. This model can be expressed as follows:

$$Q_{si} = f(Q_{1bi}, Q_{2bi}, Q_{3bi}, \ldots, Q_{nbi}, P_i) \tag{2}$$

where, $Q_{si}$, is a vector of monthly streamflows of the subject river in season $i$, $Q_{jbi}$, a vector of the monthly streamflows of the base rivers ($j = 1, 2, \ldots, n$), and $P_i$

is a vector of the precipitation data in season $i$ collected on the watershed of the river with data gaps.

The multi-layer-feed-forward bivariate series model (MBSM) is formulated as fully connected back propogation network consisting of an input layer, hidden layer(s), and an output layer. Similar to the MASM model, the number of nodes in the input and the hidden layers depends on the number of months that define a season in both the subject river and the base rivers. It is noted that nodes in the input layer and the output layer, respectively, are corresponding to the group-characteristics of the base river and the subject river. The nodes in the hidden layer reflect the degree of synchroneous in group-characteristics of the base river and the subject river. One can develop neural networks based on NN (3,7,3) configuration. As noted earlier, a nonlinear logistic (sigmoid) function is used as an activation function.

## 6. Model performance indicators

To evaluate the adequacy of the proposed models, the performance of the models should be analytically measured, and related to stable statistics. Such testing procedures for model reliability are presented below.

The relative mean error (RME) is obtained as follows:

$$\mathrm{RME} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{y_i} \tag{3}$$

where, $y_i$ is the observed value, $\hat{y}_i$ the estimated value, and $N$ is the number of observations. A value of RME near zero implies that the model is providing a good estimate of the missing values. This equation can be rewritten to test the model goodness-of-fit at seasonal levels (i.e. dry, wet, and average seasons) as follows:

$$\mathrm{RME} = \frac{1}{M}\sum_{i=1}^{M}\frac{|\hat{y}_i^k - y_i^k|}{y_i^k} \tag{4}$$

where, $M < N$ is the number of observations in the specified $k$th season.

The correlation coefficient has commonly been used as a goodness-of-fit statistic. However, the correlation coefficient is only a measure of the linear association between the variables. The standard error of estimate can be applied to both linear and nonlinear models. For a nonlinear information processing system, Tokar (1996) suggested the use of the standard error of estimate, *Se*, for performance evaluation of the neural networks. A measure (Noise to Signal Ratio) involving the term *Se* is defined as follows:

$$\mathrm{Noise} = Se/Sy \tag{5}$$

The term, *Sy* is the standard deviation of the observed values of the dependent variable and is obtained as follows:

$$Sy = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}} \tag{6}$$

In Eq. (5), *Se* is the biased standard error of estimate and is obtained as follows:

$$Se = \sqrt{\frac{1}{v}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{7}$$

The degree of freedom, $v$, in Eq. (7) is obtained as follows:

$$v = N - [nN_{\mathrm{out}} + nN_{\mathrm{in}} + n + N_{\mathrm{out}}] \tag{8}$$

where, $n$ is the number of nodes in the hidden layer, $N_{\mathrm{in}}$ the number of the input nodes, $N_{\mathrm{out}}$ the number of output nodes, $N$ the number of observations, $y_i$ the observed data, $\hat{y}_i$ the estimated data, and $\bar{y}_i$ is the mean of the observed data. When *Se* is significantly smaller than *Sy*, the model is considered to provide good estimates of the missing values. If *Se* is nearly equal to or larger than *Sy*, the model is considered unsatisfactory. Eqs. (7) and (8) can be rewritten to test the goodness-of-fit of models at the seasonal level (i.e. dry, wet, and average seasons).

## 7. Evaluation basis of the proposed data infilling models

Based on the above statistical considerations, the proposed group-based neural network models (hereafter referred to as ANN models), namely the MASM and the MBSM, are compared to evaluate the best reliable approach for the selection of a suitable data infilling model. To assess the suitability and the efficacy of the proposed ANN-based models, it is necessary to compare the estimation capabilities of the

proposed models with those of existing models. Testing all the models of interest on the same data set usually makes this comparison. Comparable models such as those of the bivariate multi-dimensional regression (MR) and the PR models are used for comparative analysis. It is noted that PR models by definition are multi-dimensional models.

### *7.1. Performance assessment of ANN- and MR-based models*

The MR-based models are statistical linear procedures for estimating values of one or more dependent variables from a collection of predictor (independent) variables. These models can also be used for assessing the effects of the predictor variables on the response variable (Johnson and Wichern, 1988). Due to simplicity and importance of the application of the MR procedures, especially in streamflow data estimation, many researchers in estimating the missing data (Tang et al., 1996; Beauchamp et al., 1989; Boakye and Schultz, 1994) have used these procedures.

In order to evaluate the proposed ANN-based models, the estimated values from the MASM and MBSM are compared to the estimated values of the MR models using the group-valued approach. One such MR model is presented as follows:

$$\underset{(n\times m)}{Y} = \underset{(n\times(r+1))}{Z}\ \underset{((r+1)\times m)}{\beta} + \underset{(n\times m)}{\varepsilon} \tag{9}$$

where, $m$ is the number of response estimated variables $Y$, $n$ the number of observations, and $(r+1)$ is the rank of *design matrix*, $Z$. The unknown parameter, $\beta$ is estimated as follows:

$$\hat{\beta} = (Z'Z)^{-1}Z'Y \tag{10}$$

$\varepsilon$ is the residual whose sum of squares ($\varepsilon'\varepsilon$) is minimized and has the form

$$\hat{\varepsilon}'\hat{\varepsilon} = Y'Y - \hat{Y}'\hat{Y} = Y'Y - \hat{\beta}'Z'Z\hat{\beta} \tag{11}$$

and $\hat{Y}$ is the estimated values and has the following form:

$$\hat{Y} = Z\hat{\beta} = Z(Z'Z)^{-1}Z'Y \tag{12}$$

It should be noted that the multiple regression equation implicitly assumes that expectation of errors are equal to zero, the errors are statistically independent, the variance of errors is constant, and the errors are normally distributed.

### *7.2. Performance assessment of MR-, PR-, and ANN-based models*

Based on concepts of PR and the information contained in the seasonal segments, Goodier and Panu (1994) developed models for infilling of data gaps in both ASM and BSM. The results of these existing models are available for the case of two six-month seasons in the English River (Goodier and Panu, 1994). For comparative purposes, therefore, the ANN-based and the MR-based models using seasonal segments of six-month were also developed. The MR-based models (ASM and BSM) at the Sioux Lookout station of the English River (as the Subject River) and Umferville station of the English River (as the Base River) were developed. A neural network NN (6,7,6) of the English River was developed to map the same problem using the ANN-based models (MASM and MBSM). The results of the PR-based, MR-based and ANN-based models are then compared and the goodness of fit for each season is computed.

The graphical assessment for the ANN-based models is evaluated by comparing the estimated value of the missing data with that of the observed values during the dry, semi-dry, semi-wet, wet, and combined all seasons. Also the estimated value from the existing MR-based and PR-based models are plotted in the same figures for comparative purpose. In addition, the statistics for dry, semi-dry, semi-wet, wet, and combined all seasons are tabulated to assess the best infilling model.

## 8. Application of ANN-based models to streamflow data sets

In applying ANN-based models to watersheds, the issues addressed were: (1) the effect of seasonal cycles on the estimation accuracy; and (2) the effects of ANN-based models on the quality of infilled data using seasonal groups in case of (a) the multi-layer autovariate time series, and (b) the multi-layer bivariate time series.

The models are applied on several rivers to evaluate their data infilling efficacy in terms of statistical and graphical assessment. Based on the consideration of

the group-valued data approach and the relevant structural composition of ANN, two types of models are evaluated in this paper namely the MASM and the MBSM. The MASM and MBSM models are, respectively, for the cases involving only one data series with data gaps, and for the data series with data gaps in which vicinity one or more concurrent but complete data series are available. The complete data series may be available at an upstream or downstream location, or in a tributary of the same river, or in a tributary of an adjacent river. The concurrent complete data series does not need to be streamflows but could be data on rainfall, temperature, evaporation etc. In this paper, only streamflows are considered as concurrent and complete data series.

### 8.1. Selection of streamflow information

In order to apply the proposed models, there are certain conditions required for the appropriate selection of streamflow data. Such aspects are: (1) availability of data with suitable lengths; (2) unregulated and uninterrupted streamflow data to minimize the effect of external influence; and (3) nearby streamflow data that is subject to similar physiographical properties as the subject river site. An extensive search was carried out to find suitable streamflow data across Canada. Table 1 shows the names, locations, and properties of the five chosen sites (watersheds).

The reason for choosing the upstream/downstream related sites are that such sites are expected to exhibit a linear relationship between the subject and the base sites. It is noted that although the occurrence of seasonal groups in a bivariate data set is linearly related but their underlying processes would have different degree of nonlinear compositions defining the internal structures of seasonal groups in each time series. The degree of nonlinearity in two time series of a bivariate data corresponding to two watersheds would be higher when such watersheds belong to two different river systems compared to two sub-watersheds of a river system. It is in this vein that the selection of the nearby sites for two different watersheds can provide an indication of the manner in which nonlinearly related data can affect the data infilling process. Selection of a tributary (i.e. a sub-river) as the subject site and the main river as the base site help to enhance the symbiotic and highly nonlinear relation between such data time series and how such a relationship would influence the data infilling process.

### 8.2. Preparation of streamflow data for infilling purposes

All streamflow data sets used in this paper are complete and thus exhibit no data gaps. However, for testing of the various models, a variety of data gaps were randomly created in desired streamflow data. To evaluate the overall performance of any proposed model, missing data of one seasonal length was assumed to occur in any year. This assumption is continued (i.e. repeated) for all the years until all seasons are assumed to be successively missing and then succeedingly infilled with the models. However, it is noted that the first season in the data set for the MASM model is not estimated because the model is formulated to infill only for the forward shift (e.g. season $i-1$ to season $i$).

For the training and testing phases of the ANN-based models (or the phases of calibration and verification of parametric models), the data set for each site or a site-pair (i.e. a subject site and the corresponding base-site(s)) were divided into two parts. The first part comprising 80% of the data set was used in the training phase, while the other part comprising 20% of the data set was used in the testing phase of the ANN-based models employed in this paper. Thus, 173, 134, 134, 173, and 403 monthly data points, respectively, for site or site-pair 1, 2, 3, 4, 5 (Table 1) were used during the training phase of the neural network models. Although, a larger data set is always desirable, however, small data sets comprising of 100 or less data points have been successfully employed for various synthesis and forecasting purposes (Gupta and Lam, 1996; Chow and Cho, 1997; Loke et al., 1997; Elshorbagy et al., 2000). Maier and Dandy (2000) concisely discuss various modeling issues and applications of neural networks in water resources.

### 8.3. Seasonality assessment of data sets

Seasonality was determined for the subject rivers (Table 1) through the application of both correlation and spectral analyses. The resulting correlogram and periodogram are exhibited in Fig. 2. A strong indication of the presence of twelve-month seasonality is

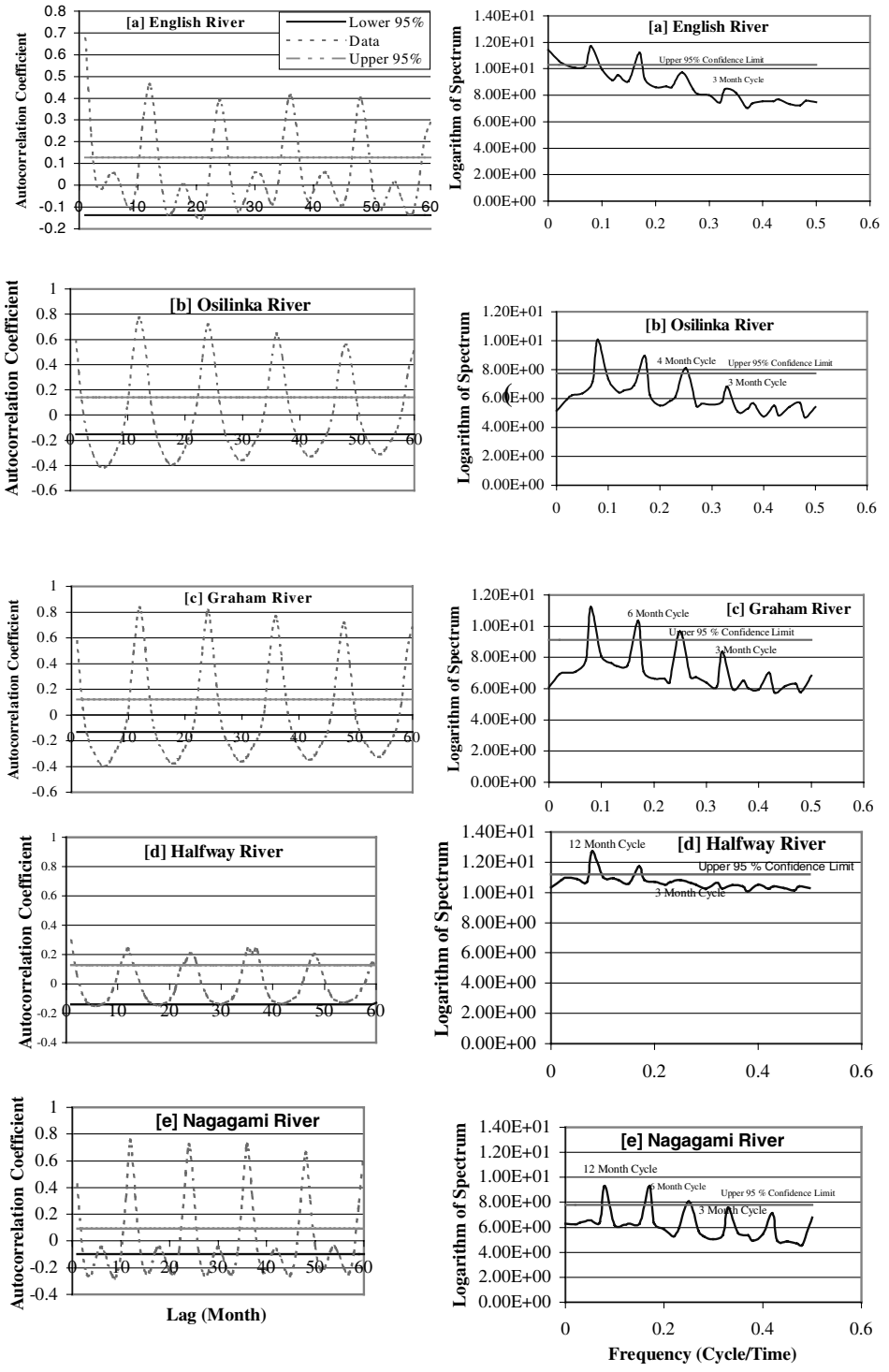



Fig. 2. Correlograms and Periodograms of the rivers: (a) English River, (b) Oslinka River, (c) Graham River, (d) Halfway River, and (e) Nagagami River.

Table 1
Geographical location across Canada of subject and base rivers used in the formation of various site-pairs (CCF means cross correlation coefficient between the subject and the base rivers and ACF means the auto-correlation coefficient for the subject river)

| Streamflow station | Station number | Latitude (N) | Longitude (W) | Remark on location | CCF | Period of records (years) | Mean ($m^3/s$) | Area ($km^{2)}$ | ACF | General remark | Site-pair number |
|---|---|---|---|---|---|---|---|---|---|---|---|
| English River at Sioux Lookout | 05QA002 | 49 52 30 | 91 27 30 | Down-stream | 0.90 | 1963–1981 (18) | 120 | 13 900 | 0.79 | Subject site | 1 |
| English River at Umferville | 05QA001 | 50 04 15 | 91 56 40 | Up-stream | | | 57.1 | 6230 | | Base site | |
| Osilinka River | 07EC004 | 56 07 35 | 124 47 47 | Nearby sites, two different rivers | 0.99 | 1981–1995 (14) | 35.3 | 1960 | 0.71 | Subject site | 2 |
| Mesilinka River | 07EC003 | 56 14 38 | 124 38 36 | | | | 44.5 | 2980 | | Base site | |
| Graham River | 07FA005 | 56 27 31 | 122 21 22 | Sub-river from the main Halfway River | 0.93 | 1981–1995 (14) | 25.4 | 2200 | 0.75 | Subject site | 3 |
| Halfway River | 07FA003 | 56 30 30 | 122 14 28 | | | | 35.6 | 3780 | | Base site | |
| Halfway River | 07FA006 07FA001 | 56 13 40 | 121 28 50 | Down-stream | 0.961 | 1977–1995 (18) | 75.8 | 9400 | 0.77 | Subject site | 4 |
| Halfway River | 07FA003 | 56 30 30 | 122 14 28 | Up-stream | | | 35.6 | 3780 | | Base site | |
| Nagagami River | 04JC002 | 49 46 44 | 84 31 48 | Nearby sites, two different rivers | 0.95 | 1951–1993 (42) | 24.7 | 2410 | 0.62 | Subject site | 5 |
| Kabinakagami River | 04JA002 | 49 44 39 | 84 06 13 | | | | 48.0 | 3780 | | Base site | |

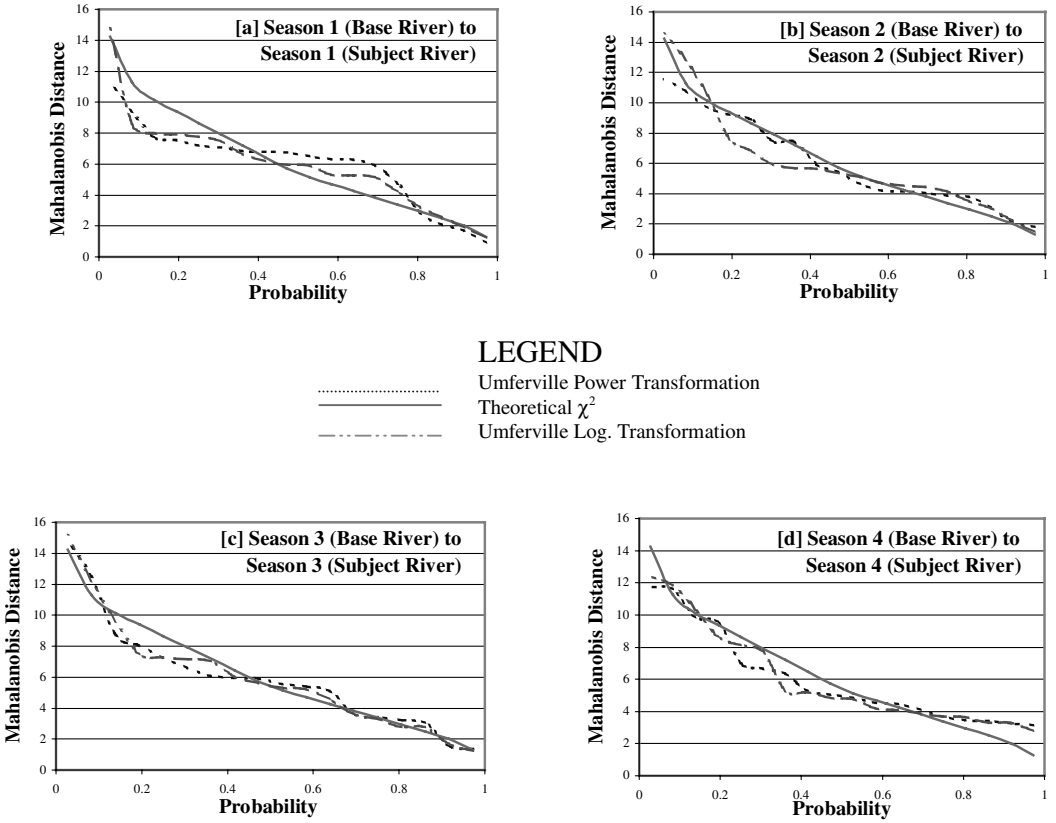
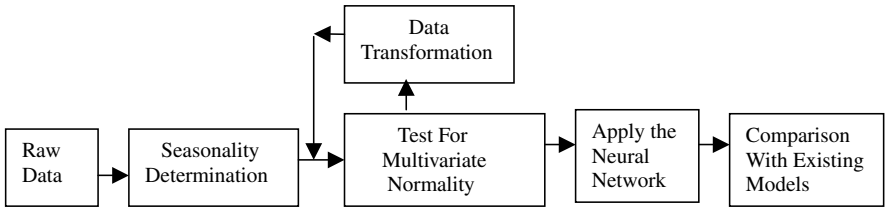



Fig. 3. A plot of Mahalanobis distance and $\chi^2$ distribution for testing multivariate normality of the four seasonal segments of the English River with its base river: (a) Season-1 to Season-1, (b) Season-2 to Season-2, (c) Season-3 to Season-3, and (d) Season-4 to Season-4.

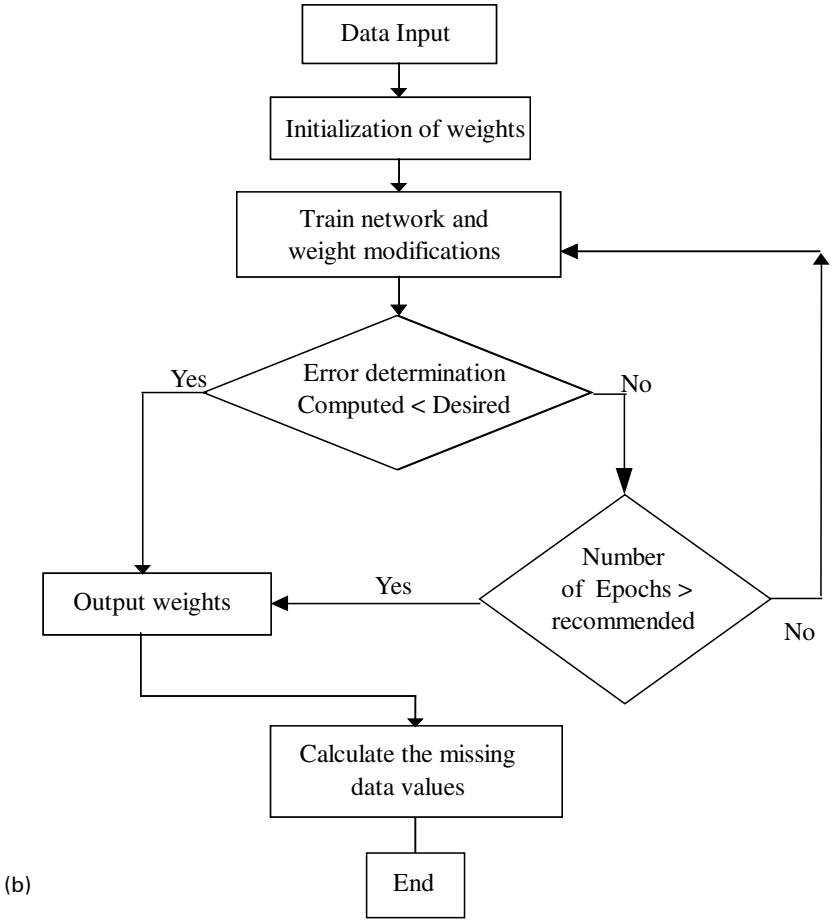
apparent in the five streamflows. Two six-month seasons (i.e. two seasons of six-month duration) are also indicated for all streamflows presented. Based on 95% confidence intervals, the presence of three seasons of four-month duration is evident in the Osilinka, Graham, and Nagagami streamflows. The periodograms also show evidence of four three-month seasons, especially in the three previously mentioned rivers. A weak indication of the presence of four seasons of three-month length is indicated in the English River as shown in Fig. 2a. This is probably due to the storage capacity of such a large watershed. An experiment to infill the missing data using two seasons of six-month length was also conducted for this site. The results of this experiment will then be compared to an analysis using the PR models presented by Goodier and Panu (1994). Using an iterative method (Khalil et al., 1998b), the starting and ending months for the four seasons of three-month length were determined to be November–January, February–April, May–July, and August–October for all the sites except the Nagagami River. The seasons are classified as semi-dry, dry, wet, and semi-wet, respectively. The seasons for Nagagami River were detected to be January–March, April–June, July–September, and October–December. The seasons are classified as dry, wet, semi-wet and semi-dry, respectively.

### *8.4. Multivariate-normality assessment of data sets*

The grouped data should be tested for multivariate normality using the multivariate $p$-dimensional normal density. The $p$-dimensional normal distribution of such a random vector has the form called Mahalanobis distance (MD) (Gnanadesikan, 1997). The MD is




(a)




(b)

Fig. 4. (a) Flow chart for the application of the ANN-based Models, (b) Flow chart for Training the ANN-based Models.

distributed as $\chi_p^2$ (i.e. the chi-square distribution) with $p$ degrees of freedom and $\Sigma$ as the covariance matrix. By comparing the theoretical value of the $\chi_p^2$ statistic to that obtained by calculating the MD of the actual data, the multivariate normality can be assessed. As real data is rarely normal, especially for a set of natural streamflows, a Box and Cox power transformation is used in order to transform the data to normality.

Table 2
Comparative summary statistics of the ANN-based models (MASM and MBSM) during the testing phase

| Scenario | English River at Sioux Lookout | | Oslinka River | | Graham River | | Halfway River | | Nagagami River | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RME | Noise | RME | Noise | RME | Noise | RME | Noise | RME | Noise |
| | *MASM model* | | | | | | | | | |
| Semi-dry | 0.51 | 1.11 | 0.41 | 0.26 | 0.46 | 0.24 | 0.84 | 1.30 | 0.55 | 0.82 |
| Dry | 2.66 | 3.08 | 14.93 | 4.61 | 11.18 | 4.06 | 13.27 | 3.41 | 8.16 | 3.24 |
| Wet | 0.50 | 1.28 | 0.68 | 1.30 | 0.62 | 1.27 | 0.65 | 1.21 | 0.74 | 1.24 |
| Semi-wet | 0.45 | 1.21 | 0.62 | 1.60 | 0.66 | 1.78 | 0.69 | 1.48 | 0.56 | 0.80 |
| **All seasons** | **1.03** | **1.67** | **4.16** | **1.94** | **3.23** | **1.84** | **3.86** | **1.85** | **2.50** | **1.53** |
| | *MBSM model* | | | | | | | | | |
| Semi-dry | 0.24 | 0.64 | 0.11 | 0.07 | 0.24 | 0.14 | 0.23 | 0.09 | 0.29 | 0.59 |
| Dry | 0.46 | 0.65 | 0.10 | 0.04 | 0.37 | 0.39 | 0.32 | 0.19 | 0.18 | 0.08 |
| Wet | 0.29 | 0.58 | 0.08 | 0.17 | 0.19 | 0.33 | 0.14 | 0.26 | 0.16 | 0.24 |
| Semi-wet | 0.46 | 1.19 | 0.11 | 0.31 | 0.25 | 0.51 | 0.37 | 0.58 | 0.24 | 0.55 |
| **All seasons** | **0.36** | **0.73** | **0.10** | **0.18** | **0.26** | **0.34** | **0.27** | **0.31** | **0.22** | **0.28** |

Multivariate normality testing was carried out for each of the four seasonal segments at all sites, due to the underlying requirement of multivariate normal data sets by the MR and PR models. It was determined that streamflow data normality was best achieved through the use of a natural logarithm transform (Panu and Unny, 1980; Unny et al., 1981). The sample $\chi^2$-statistics and the MD were computed using the SPSS computer package (SPSS, 1995) for the English River (Fig. 3). There is an apparent indication of the presence of multivariate normality for all seasonal segments of the English River and similar results with respect to multivariate normality were obtained for the remainder of the rivers (Khalil et al., 1998b) presented in this paper. A general flow chart utilized in various applications of models to a specific site or site-pair is given in Fig. 4a.

### *8.5. Training of neural network based data infilling models*

Training comprises of presentation of grouped data pertaining to input and output to the network and obtaining the inter-connection weights for the back propagation network. Initially, the transfer function parameters are defined and the network is assigned arbitrary values between 0 and 1 to the inter-connection weights. The input-vector and corresponding output-vector were normalized to fall within the interval of 0 and 1. The error is computed as the difference between the actual and the desired output (i.e. observed data).

One epoch represents presentation of complete training data sets once to the network. In this paper, such a data set comprised of 80% of the data set for a particular site or the site-pair. A part of the training data set (approximately 15%) was used for performance evaluation during the training process of both MASM and BASM models. These neural models were presented with all the input-vectors and their corresponding output-vectors till the desired level of performance was achieved on the portion (usually 15%) of data set not used in the training process.

The structural composition of a neural network and the associated consideration of improvement of neural network generalization are briefly described in Appendix A. The neural network algorithm used in this paper for brevity is summarized in Fig. 4b. To improve the generalization property of the neural network models, the cross validation techniques were used. During the training process, as indicated earlier, only a portion of the data set (i.e. 80%) and of which 15% data set was used to evaluate the performance of the model during the training process. The reason for doing this was to validate the model on a portion of the data set which is different from the data set (i.e. the remainder of the data set, approximately 20%) later used for testing the efficacy of the models in estimating missing values.

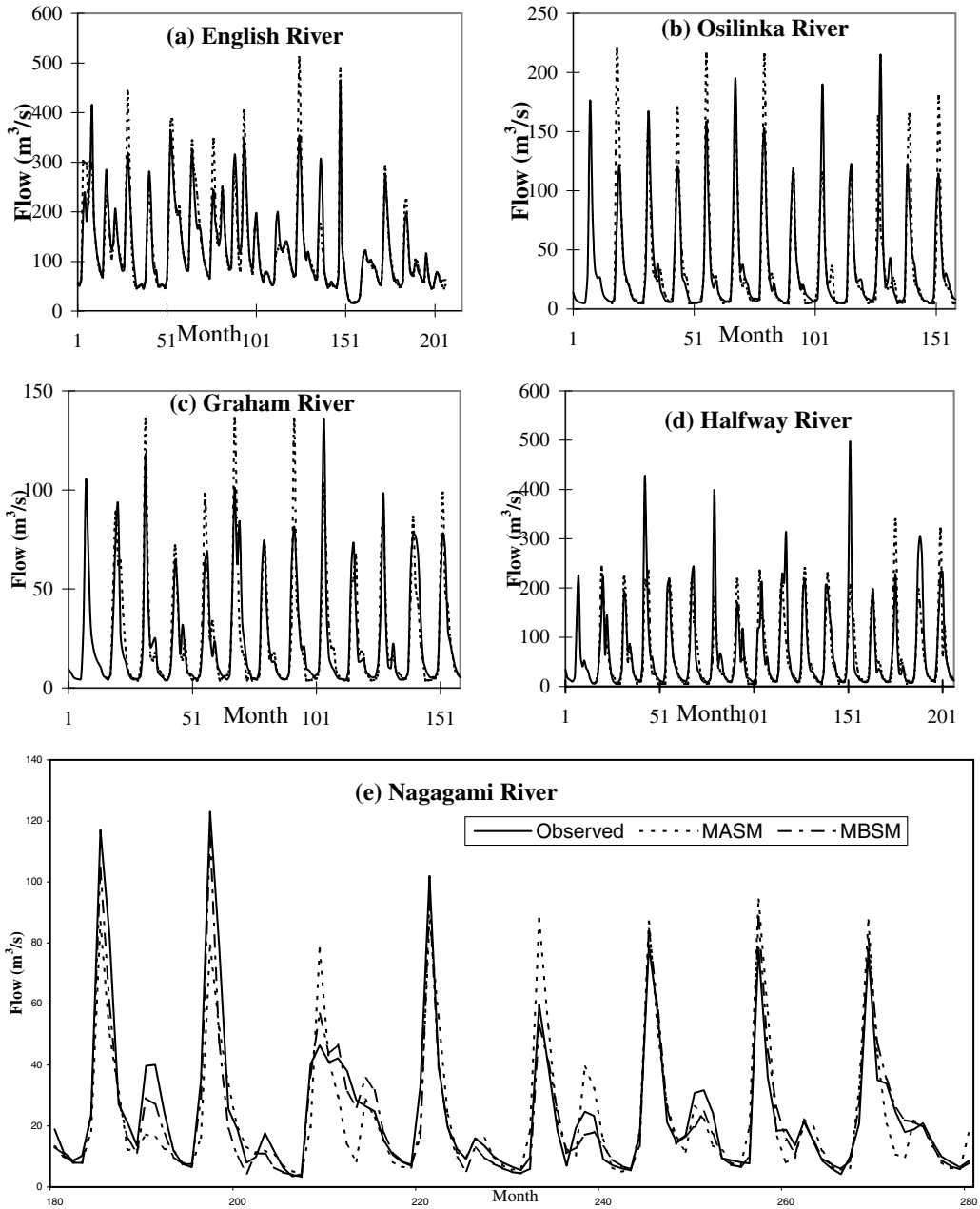



Fig. 5. Comparison between MASM and MBSM models: (a) English River, (b) Oslinka River, (c) Graham River, (d) Halfway River, and (e) Nagagami River.

## 9. Performance assessment of data infilling models

The proposed ANN-based models and the existing MR-based and PR-based models are applied to various cases involving the MASM and MBSM as summarized in Table 1. The MASM and MBSM models were first assessed among themselves and in turn were assessed in comparison with the existing MR-based and PR-based models. For this purpose, the data set (approximately 20%) which was not

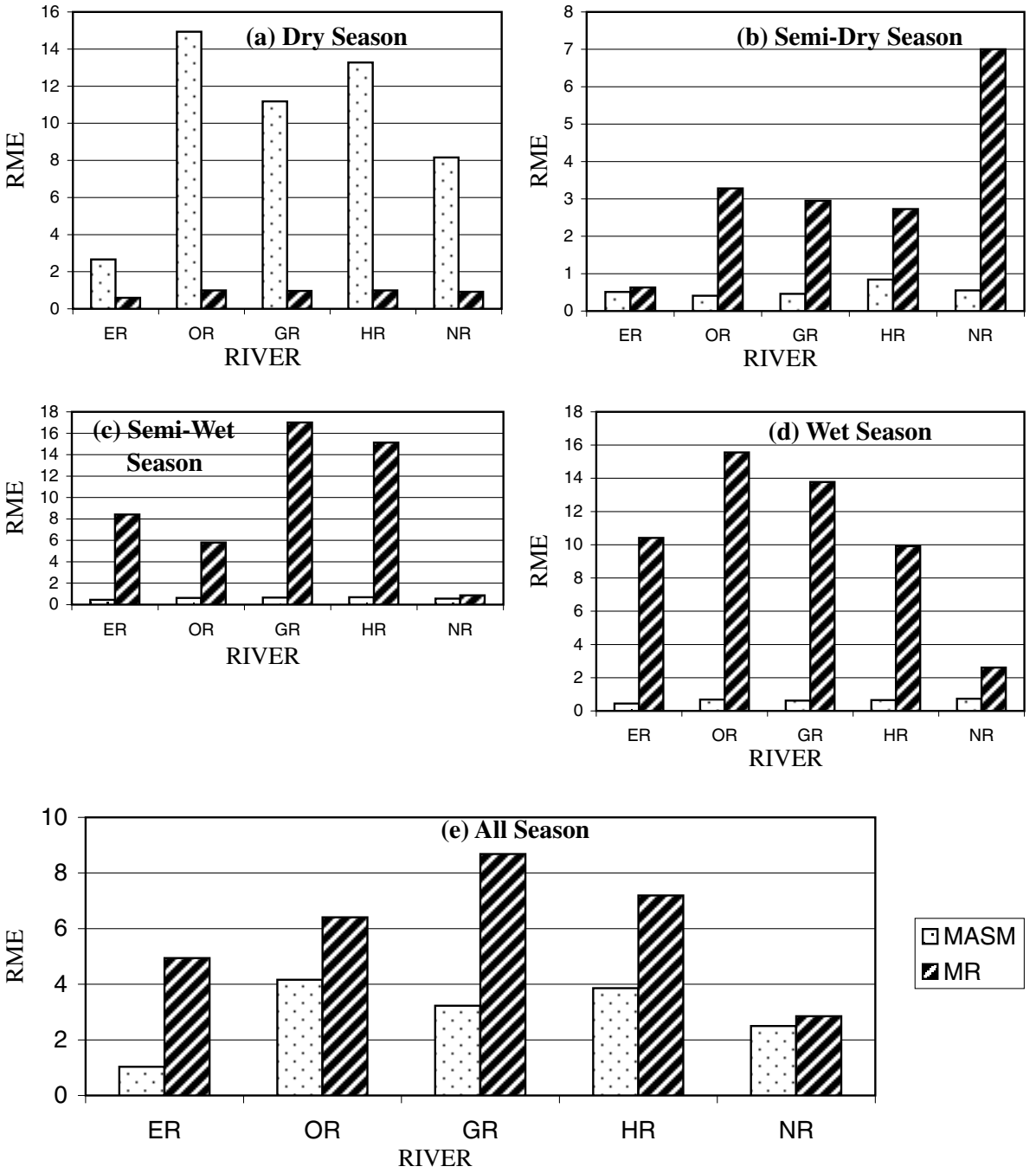



Fig. 6. Four seasons comparative summary statistics between MASM and MR models in the (a) Dry Season, (b) Semi-Dry Season, (c) Semi-Wet Season, (d) Wet Season, and (e) All Seasons.

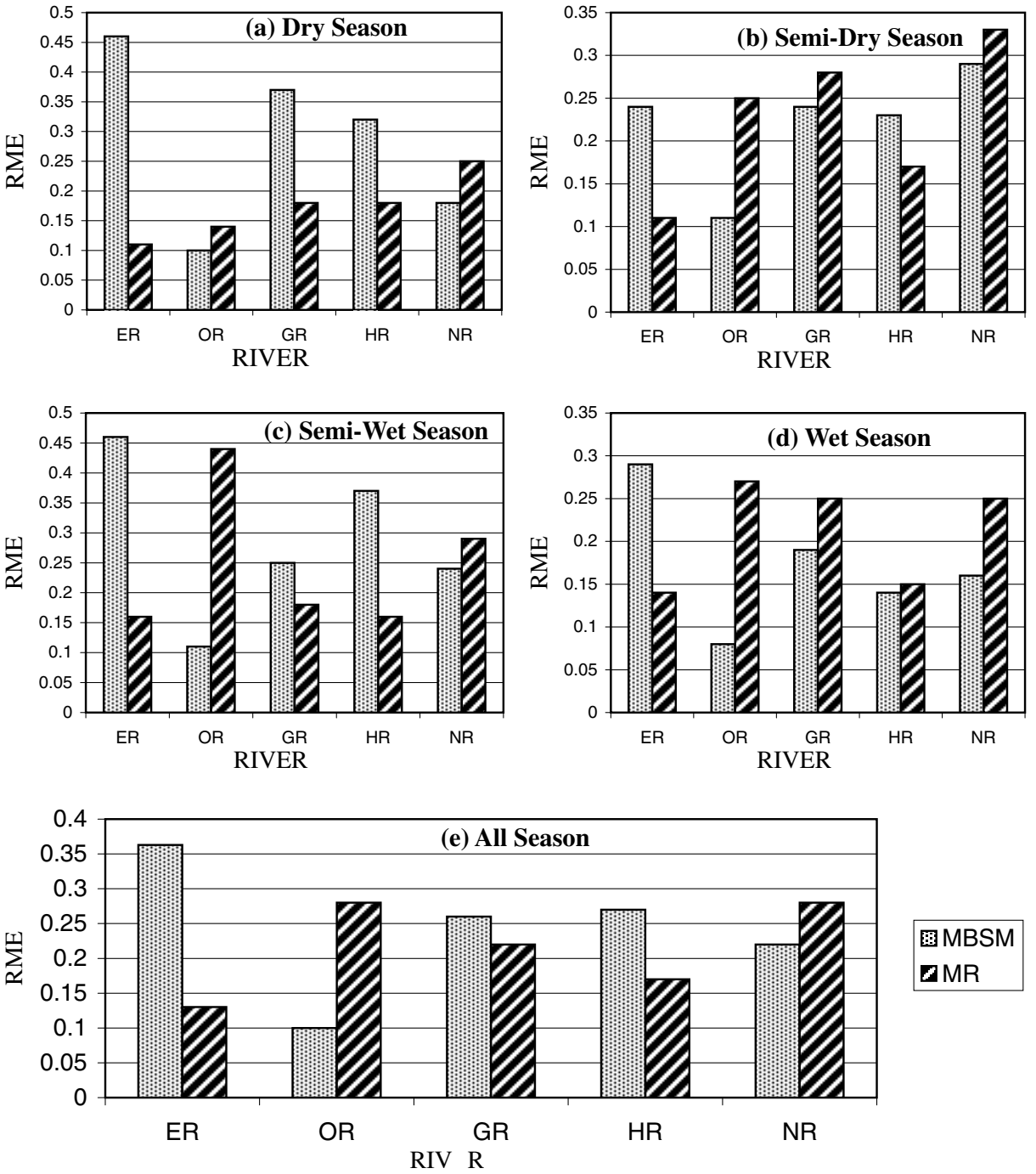



Fig. 7. Four seasons comparative summary statistics between MBSM and MR models in the (a) Dry Season, (b) Semi-Dry Season, (c) Semi-Wet Season, (d) Wet Season, and (e) All Seasons.

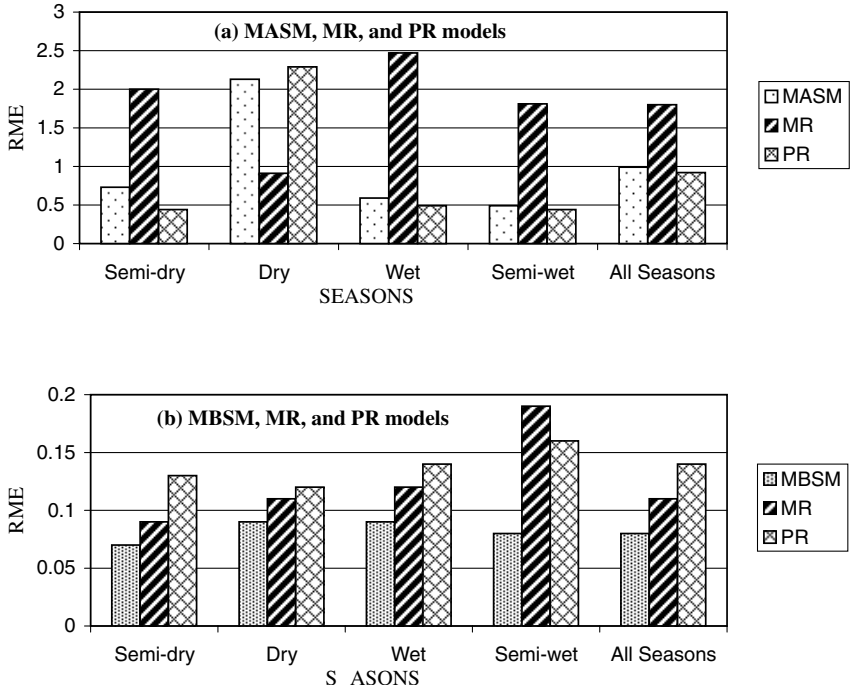



Fig. 8. Two seasons (Wet and Dry) comparative summary statistics in English River of ANN-based, MR-based, and PR-based models (a) MASM, MR, and PR models, and (b) MBSM, MR, and PR models.

used for model calibration and validation was used. Based on the results of these applications, the proposed and existing models were assessed through the graphical and statistical analyses. In all graphical presentations, the abbreviations ER, OR, GR, HR, and NR, respectively, represent the English River, Oslinka River, Graham River, Halfway River, and Nagagami River.

### *9.1. Comparisons between the MASM and MBSM models*

The test results of relative performance of MASM and MBSM models are summarized in Table 2. As expected in all rivers, the MBSM models perform much better than the MASM models. The values of cross-correlation between participating rivers for MBSM models are higher (Table 1) than the values of Lag-one auto-correlation for the MASM models. Graphical comparisons in Fig. 5 between MASM and MBSM models indicate that the bivariate techniques are more suitable for streamflow data infilling purposes.

### *9.2. Comparisons of MASM and MR models*

A graphical comparison in Fig. 6 between MASM and MR models indicates that MASM is more capable of handling the more variable character of monthly streamflows forming the seasonal groups. Due to small variation during the dry season, the MR model was found to be better in estimating the missing data than the MASM. In general, the values of RME (Fig. 7) for MASM are smaller in all seasons except the dry season. The values of the RME for the all seasons of MASM ranges between 1.03 and 4.16, while the values of the RME for the MR model ranges between 2.69 and 8.84. In other words, MASM appears to be a promising estimator of the missing values for the wet, semi-wet, and semi-dry seasons in autovariate series.

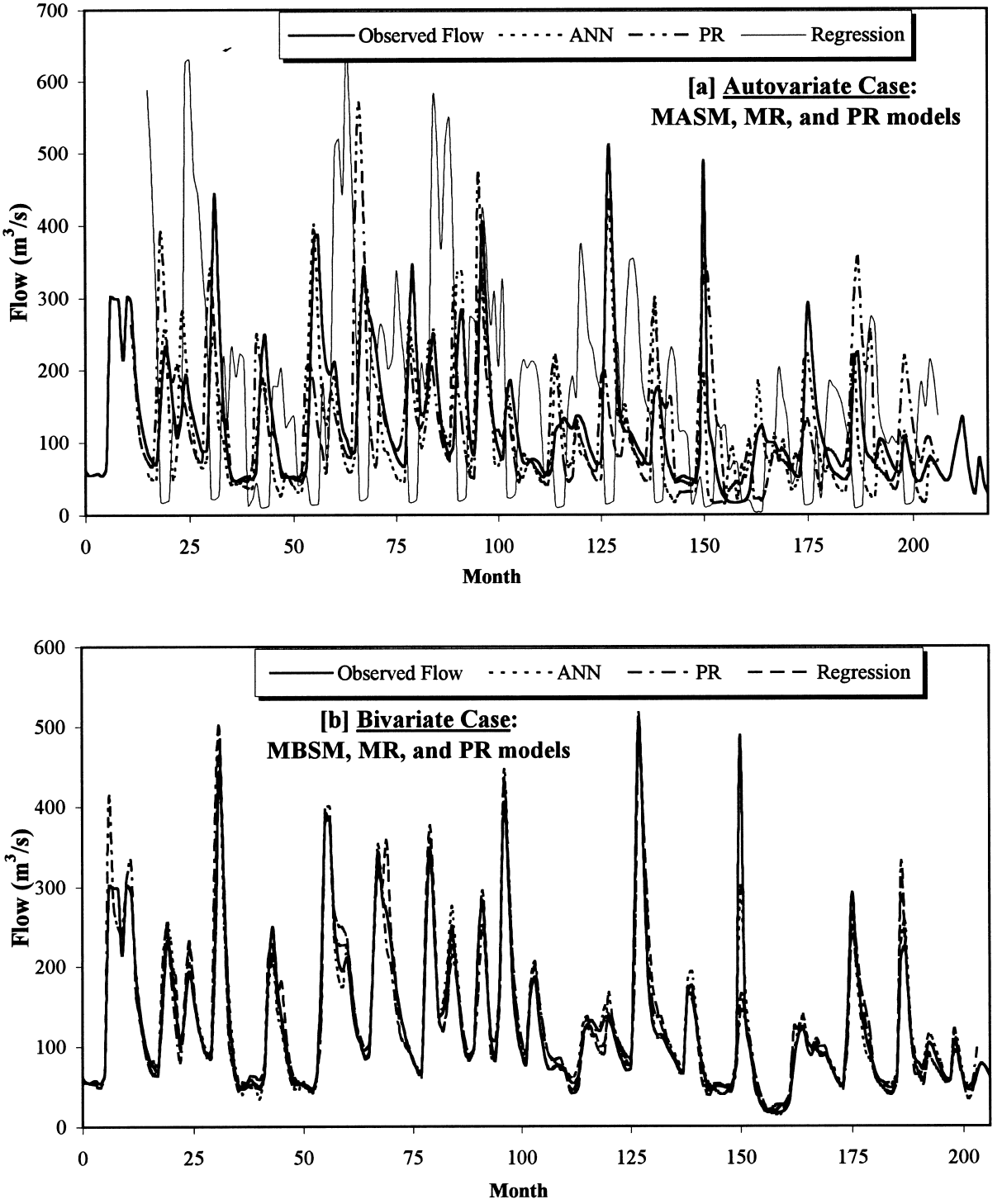



Fig. 9. Comparison of ANN-, MR-, and PR- Models Using Two Seasons (Wet and Dry) in English River (a) MASM, MR, and PR models, and (b) MBSM, MR, and PR models.

### 9.3. Comparisons of MBSM and MR models

A graphical comparison in Fig. 7 between MBSM and MR models indicates that MBSM is slightly more capable of handling the more variable character of monthly streamflows forming the seasonal groups. The MR model was found to be better in estimating the missing data than the MBSM during the dry season. In general, the values of RME (Fig. 7) for MBSM are smaller in all the analyzed seasons except the dry season. In other words, MBSM appears a promising estimator of the missing values for the wet, semi-wet, and semi-dry seasons in bivariate series.

The MR model shows better estimation capabilities for the missing data in rivers that have upstream/downstream relationships (English River and Halfway River). This may be due to the presence of linear relationship between upstream and downstream sites. However, the ANN based models show relatively good estimation in the wet season for all sites except for English River. In other words, ANN based models have a higher capability of mapping during the high flows.

### 9.4. Comparisons of ANN-based, MR-based, and PR-based models

The results of the PR based models for two six-month seasons (Goodier and Panu, 1994) were available for the case of English River at Sioux Lookout. For this river, the correlogram and the periodogram showed no evidence of the existence of seasonality of four seasons of three-month length. Three types of model (ANN-based, MR-based, and PR-based models) were applied in cases, i.e. the autovariate series and bivariate series. ANN-based models of NN (6,7,6) were used for comparison.

In the autovariate series case, the results of the PR-based and ANN-based models are comparable. In general, the PR-based models performed better for all seasons. For example, the values of RME are, respectively, 0.92 and 0.99 (Fig. 8a). However, for the specific case of dry season, the ANN models gave the best results. Again it is noted that the MR model was the best estimator for the dry season.

In the bivariate case, it is apparent from Fig. 8b that the ANN models are comparatively better than the PR models and the MR models with values of RME of 0.08, 0.14 and 0.13, respectively. In general, the ANN models are found to perform much better in all the seasons. The MR models are found to be marginally better than the PR models except for the semi-wet seasons. Such observations are clearly apparent from a graphical comparison in Fig. 9.

It is noted that the values of RME are slightly improved in the MR models when the English River is modeled using two six-month seasons. Such an improvement is also noticeable (Figs. 7–9) for the ANN models.

## 10. Conclusions

The concepts of ANN and seasonal groups and their characteristics have been investigated for the estimation of missing data values in monthly streamflows. Five watersheds with varying degrees of physical characteristics have been used to assess the efficacy of proposed models for estimating data gaps in monthly streamflow time series. The ANN techniques and concepts have been shown to be good candidates for the infilling of the missing seasonal groups in monthly streamflow data series.

The ANN-based models produced relatively more accurate estimates of the data gaps for most sites except the English River and the Halfway River. As measured by the RME, the average improvement during the testing period was 54% for both types of models thus indicating that ANN-based models are more accurate in data infilling for seasonal group than the regression (MR) model.

The MR based model using the autovariate series method had shown relatively poor estimation ability for streamflow data infilling. This poor estimation is quite pronounced for the small watersheds (i.e. Graham, Halfway, and Osilinka). However, the MR based-model gave relatively better estimation for streamflow data infilling for the larger watersheds (i.e. English and Nagagami). This observation can be related to the low auto correlation coefficient within the lag-three months. Further, the larger watersheds are more stable and thus accidental events have less influence on the seasonal streamflows.

Comparing the results of MR-based, PR-based and ANN-based models on the six months groups, the

ANN-based models gave relatively improved results during the testing phase in both MASM and MBSM. The average improvement for ANN-based models over MR-based models is 55%, and over the PR-based models is 34%. The average improvement for PR-based models over MR-based models is 45%. In other words, the ANN-based models, in general, are more accurate in data infilling for seasonal grouping than the other models, and PR-based models are more accurate than MR-based models.

Based on the results and discussions presented in this paper, the proposed methodology in general and the ANN based models in particular, appears promising for streamflow data infilling. Further investigation may be needed for infilling missing data using more than one base site and also using prior information from mixed records such as precipitation data, temperature data, and/or snowmelt data from nearby sites.

## Appendix A. ANN algorithm

### *A.1. Generalization improvement of the ANN algorithm*

To improve the generalization property of ANN-based models, the cross validation techniques are used. First the available data set is randomly partitioned into a training set and a test set. The training set is further partitioned into two subsets:

1. a subset used for training the model;
2. a subset used for evaluation of performance of the model.

About 15% of the training set is used to evaluate the network. The reason for doing this is to validate the model on a data set different from the data which will be used later in estimating missing values.

An experiment was conducted to estimate the number of neurons, which gave the best generalization ability. For the validation set, a plot of RME versus the number of neurons as shown in Fig. A1. The values of the RME are determined by subtracting the estimated and the observed values of the data series, and then dividing their absolute values by the observed data. The plot shows that the minimum error is achieved when using seven neurons in the first hidden layer. Table A1 summarizes the value of RME for a number of neurons in the first hidden layer.

Another experiment was done to evaluate the reliability of using two hidden layers. The same subset is used for training the network and using seven neurons in the first hidden layer and a different number of neurons for the second hidden layer. Fig. A2 shows that best generalization is achieved when using four neurons in the second hidden layer. It is also apparent that the second hidden layer did not significantly decrease the RME. Table A2 summarizes the values of RME for the number of neurons in the second hidden layer.

A value of one to ten for the relation between the weight and the patterns (number of input data) was suggested by Weigned et al. (1991) in the following the heuristic rule to obtain good generalization properties in a network:

$$\frac{1.1P}{10} \leq n(I+1) < \frac{3P}{10} \tag{A1}$$

where $P$ is the total number of patterns used for training ($P = 204$), $n$ the number of neurons in the hidden layers ($n = 7$) in this study, and $I$ the number of input neurons used ($I = 3$) in this study. The chosen seven hidden neurons computed before was found to satisfy the heuristic rule suggested by Weigned et al. (1991).

An experiment was conducted to examine the maximum number of epochs which will be used to stop the model before fitting the training data. The RME is plotted against the number of epochs used to train the model. The number of epochs for which the value of RME is a minimum is determined. This is the number of epochs at which the

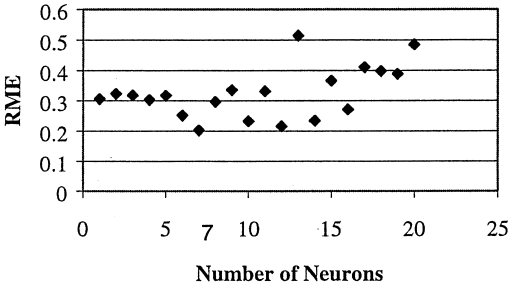



Fig. A1. Number of neurons versus the relative mean error.

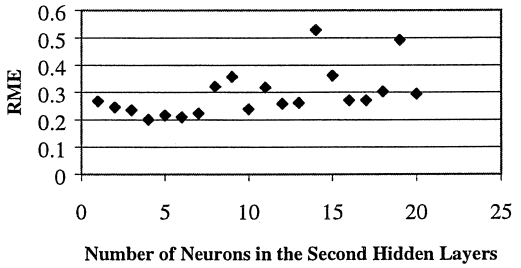



Fig. A2. Number of neurons versus the relative mean error.

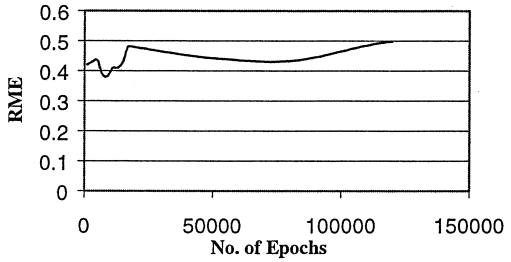



Fig. A3. Number of epochs versus the relative mean error.

model would have the maximum and the best generalization ability. The subset used for evaluation of the model and for the determination of the maximum number of epochs can be re-used again for training the final model. Fig. A3 shows that the best generalization error is encountered at 10,000 epochs corresponding to the RME of 0.3798. Table A3 summarizes the value of RME for various number of epochs used in a hidden layer consisting of seven neurons.

Using the optimal number of epochs as the upper boundary for training the model, the sum square error (SSE) of the model is also monitored while running the model to minimize the over-fitting error. If the SSE starts to increase or to decrease very slowly, it is desirable to stop the process to avoid over-fitting errors.

Table A1
Summary of RME for number of neurons in the first hidden layer

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| No of neurons | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| RME | 0.3059 | 0.323 | 0.3185 | 0.3020 | 0.3179 | 0.2522 | 0.2023 | 0.2963 | 0.3358 | 0.2324 |
| No of neurons | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| RME | 0.3319 | 0.2156 | 0.5141 | 0.2341 | 0.3656 | 0.2710 | 0.4097 | 0.3975 | 0.3883 | 0.4847 |

Table A2
Summary of RME for number of neurons in the second hidden layer

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| No of neurons | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| RME | 0.2670 | 0.2447 | 0.2348 | 0.2005 | 0.2159 | 0.2087 | 0.2227 | 0.3214 | 0.3574 | 0.2384 |
| No of neurons | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| RME | 0.3181 | 0.2579 | 0.2613 | 0.5277 | 0.3614 | 0.2706 | 0.2704 | 0.3024 | 0.4919 | 0.2941 |

Table A3
A summary of RME values for various epochs

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| No of epochs | 1000 | 3000 | 5000 | 7000 | 9000 | 10 000 | 11 000 | 13 000 | 15 000 | 17 000 |
| RME | 0.4203 | 0.4297 | 0.4347 | 0.3880 | 0.3811 | 0.3798 | 0.4088 | 0.4098 | 0.4293 | 0.4805 |
| No of epochs | 20 000 | 50 000 | 60 000 | 70 000 | 80 000 | 90 000 | 10 000 | 11 000 | 12 000 | 13 000 |
| RME | 0.4776 | 0.4421 | 0.4394 | 0.4357 | 0.4319 | 0.4578 | 0.4706 | 0.4846 | 0.4988 | 0.5001 |

By establishing the number of layers, neurons, and epochs that would give the best model performance for a given data, the development of the data in-filling process can be presented for the group-valued data approach.

## References


Afza, N., Panu, U.S., 1992. In-filling Missing Monthly Streamflow Data for Rivers with Seasonal Runoff, Civil Engineering Technical Report, No. CE-92-3, Lakehead University, Ontario, Canada.

Beauchamp, J.J., Dowing, D.J., Railsback, S.F., 1989. Comparison of regression and time-series methods for synthesizing missing streamflow records. Water Resources Bulletin 25 (5), 961–975.

Boakye, P.G., Schultz, G.A., 1994. Filling gaps in runoff time series in west Africa. Hydrological Science Journal 39 (6), 621–636.

Chow, T.W.S., Cho, S.Y., 1997. Development of a recurrent Sigma–pi neural network rainfall forecasting system in Hong Kong. Neural Computing and Applications 5 (2), 66–75.

Elshorbagy, A., Simonovic, S.P., Panu, U.S., 2000. Performance evaluation of artificial neural networks for runoff prediction. ASCE Journal of Hydrologic Engineering 5 (4), 424–427.

French, M.N., Krajewski, W.F., Cuykendall, R.R., 1992. Rainfall forecasting in space and time using a neural network. Journal of Hydrology 137, 1–31.

Goodier, C., Panu, U., 1994. Infilling missing monthly streamflow data using a multivariate approach. Stochastic and Statistical Methods in Hydrology and Environmental Engineering 3, 191–202.

Gnanadesikan, R., 1997. Methods for Statistical Data Analysis of Multivariate Observation. Wiley, New York.

Granger, C.W.J., Newbold, P., 1986. Forecasting Economic Time Series. . 2nd ed.Academic Press, Orlando, FL.

Gupta, A., Lam, M., 1996. Estimating missing values using neural networks. Journal of Operations Research 47 (2), 229–238.

Hirsch, R.M., 1982. A comparison of four streamflow record extension techniques. Water Resources Research 18 (4), 1081–1088.

Hsu, Kuo-lin, Gupta, H., Sorooshian, S., 1995. Artificial neural network modelling of the rainfall–runoff process. Water Resources Research 31 (10), 2517–2530.

Ishibuchi, H., Miyazaki, A., Kwon, K., Tanaka, H., 1993. Learning from incomplete training data with missing values and medical application. Proceedings of the International Joint Conference on Neural Networks, Japan 2, 1871–1874.

Johnson, R.A., Wichern, D.W., 1988. Applied Multivariate Statistical Analysis. Prentice Hall, New York.

Kang, K.W., Park, C.Y., Kim, J.H., 1993. Neural network and its application to rainfall–runoff forecasting. Korean Journal of Hydroscience 4, 1–9.

Karunanithi, N., Grenney, W.J., Whitley, D., Bovee, K., 1994. Neural networks for river flow prediction. Journal of Computing in Civil Engineering, ASCE 8 (2), 201–220.

Khalil, M., Panu, U., Lennox, W., 1998a. Estimation of Missing Streamflows: A Historical prospective. Canadian Society for Civil Engineering, Annual Conference, Halifax, NS, vol. 1, pp. 235–246.

Khalil, M., Panu, U.S., Panu, Lennox, W.C., 1998b. Infilling of Missing Streamflow Values Based on Concepts of Groups and Neural Networks, Civil Engineering Technical Report No CE-98-2, Lakehead University, Thunder Bay, Ont., Canada.

Lachtermacher, G., Fuller, J.D., 1994. Back-propagation in hydrological time series forecasting. Stochastic and Statistical Methods in Hydrology and Environment Engineering 3, 229–242.

Loke, E., Warnaars, E.A., Jacobsen, P., Nelen, F., Almeida, M.D., 1997. Artificail neural networks as a tool in urban storm drainage. Water Science and Technology 36 (8/9), 101–109.

Maier, R.H., Dandy, G.C., 2000. Neural networks for prediction and forecasting of water resources variables: a review of modeling issues and applications. Environmental Modeling and Software 15, 101–124.

McCuen, R.A., 1993. Statistical Hydrology. Prentice-Hall, Englewood Cliffs, NJ.

Panu, U.S., Khalil, M., Elshorbagy, A., 2000. Streamflow data infilling techniques based on concepts of groups and neural networks. In: Govindraju, R.S., Rao, A.R. (Eds.). Artificial Neural Networks in Hydrology. Kluwer Academic Publishers, Dordrecht, 235-258.

Panu, U.S., 1991. Application of some entropic measures in hydrological data infilling procedures. Entropy and Energy Dissipation in Water Resources, 175–192.

Panu, U.S., Afza, N., 1993. Entropic evaluation of streamflow data infilling procedures. Proceedings of Stochastic and Statistical Methods in Hydrology and Environmental Engineering, 410–412.

Panu, U.S., Unny, T.E., 1980. Extension and application of feature prediction model for synthesis of hydrologic records. Water Resources Research 16 (1), 77–96.

Panu, U.S., Unny, T.E., Regade, R., 1978. A feature prediction model in synthetic hydrology based on concepts of pattern recognition. WRR 14 (2), 335–344.

Pedreira, C.E., Parente, E., 1995. Neural networks with missing values attributes. Proceedings of IEEE International Conference on Neural Networks 6, 3021–3023.

Raman, H., Sunilkumar, N., 1995. Multivariate modelling of water resources time series using artificial neural networks. Hydrological Sciences Journal 40 (2), 145–163.

SPSS, 1995. Base Systems User's Guide (Part-II), SPSS Inc., Chicago, IL, USA.

Streit, R.L., Luginbuhl, T.E., 1994. Maximum likelihood training of probabilistic neural networks. IEEE Transactions on Neural Networks 5 (5), 764–783.

Tanaka, M., 1996. Identification of nonlinear systems with missing data using stochastic neural network, decision and control. Proceedings of the 35th IEEE Conference-Journal 1, 933–934.

Tang, W.Y., Kassim, A.H.M., Abubakar, S.H., 1996. Comparative studies of various data treatment methods — Malaysian experience. Atmospheric Research Journal 42, 247–262.

Tiao, G.C., Tsay, R.S., 1989. Model specification in multivariate time series. Journal of the Royal Statistical Society B51, 157–213.

Tokar, A.S., 1996. Rainfall–runoff modeling in an uncertain

environment. PhD thesis. University of Maryland. UMI Dissertation Service. Bell-Howell Company.

Tong, H., 1983. Threshold Models in Nonlinear Time Series Analysis. Lecture Notes in Statistics, 21. Springer, New York.

Tong, H., 1990. Nonlinear Time Series: A Dynamical System Approach. Oxford University Press, Oxford.

Unny, T.E., Panu, U.S., McInnes, C.D., Wong, A.K.C., 1981. Pattern analysis and synthesis of time dependent hydrologic data. , Advances in Hydrosciences, vol. 12. Academic Press, New York (pp. 222–244).

Weigned, A.S., 1991. Connectionist Architectures for Time Series Prediction of Dynamical Systems. A Ph.D. thesis, Department of Physics, Stanford University, University Microfilms International, Ann Arbor, Michigan, USA.

Wong, I., Lam, D., Storey, A., Fong, P., 1994. A Neural Network Approach to Predict Missing Environmental Data, World Congress in Neural Networks, vol. 1, Conference, San Diego, CA.