Journal of Hydrology (NZ) 39(2): 83-106, 2000

Estimating missing daily maximum and minimum temperatures for Mount Cook, South Island, New Zealand, using a statistical model and 'aiNet' neural network models

Nazrul Islam, Crile Doscher and Tim Davies
Department of Natural Resources Engineering, EMD Division, Lincoln University

Abstract

'aiNet' is a modern machine-learning modeling tool based on the self-organizing systems called neural network-like systems. It is similar to the methods of nearest neighbor, Learning Vector Quantization Network and probabilistic neural networks. This paper presents aiNet as a useful weather modelling technique and reports on its use in developing a serially complete data set of daily maximum and minimum temperatures for the Mount Cook meteorological station in the South Island, New Zealand. Multiple linear regression was used to determine the best subset of available data from meteorological stations to use in the aiNet models. The aiNet models were developed using daily maximum and minimum temperatures from Tekapo, Queenstown, and Hokitika stations as input vectors and Mount Cook station's daily maximum and minimum temperatures as output vectors. Evaluated over the period of 1972-1998 these models gave regression coefficients (R^2 adj) of 89.3% and 93.4% for maximum and minimum temperature respectively, compared to the best possible regression coefficients of 82.3% and 90.8% for maximum and minimum temperatures obtained by regression analysis. The major attributes that distinguish aiNet from more conventional neural networks are its analysis speed and improved prediction and forecasting accuracy.

Introduction

Planning and management studies of water resources require the analysis of various hydrological variables such as temperature, precipitation, evaporation, inflow and outflow. For example, flow studies of a glacierized river system may require a serially complete historical data set of maximum

and minimum temperature, precipitation and catchment outflow to evaluate likely future responses. Evapotranspiration, snowmelt, soil decomposition, and plant productivity studies also require serially complete temperature data series.

Hydrological data sets often have missing records. Missing daily input data is a serious hindrance to using many dependent models. Often, surrogate data are generated solely to eliminate the problems caused by the missing data, with little regard given to the accurate estimation of the missing observations. The arbitrary replacement of missing data values can lead to unrealistic and discontinuous results, confounding the comparison of solutions based on different modelling approaches.

In the last two decades the treatment of hydrometeorological missing data has improved, and researchers have developed and used a variety of procedures to estimate missing daily temperature data. An assortment of serially complete data sets, each fitting a particular need and location, has been produced. Salinger and Rhodes (1993) investigated the homogeneity of temperature and rainfall records for site changes in a New Zealand catchment. In their proposed homogenisation process, they checked the monthly data for obvious errors and estimated any missing data in the series, where possible using data from a neighbouring station. In Europe, Huth and Nemesova (1995) proposed a method for estimating missing daily temperatures based on weather classification, using principal component and cluster analysis. At the time of each observation (0700, 1400, and 2100 local time) the weather is characterised by temperature, relative humidity, wind speed, and cloudiness. The coefficients of regression equations for these factors enabled the missing temperatures to be determined from the known temperatures at nearby stations, computed for each weather class. Although novel, this method is deficient due to its reliance on a number of weather variables for which data may not be readily available. It also disregards the fact that observation times may differ for different variables.

This paper is part of a study examining the relationship between the flow of the glacier-fed Hooker River and catchment inputs such as maximum and minimum temperature, and precipitation. For this purpose the records for all the variables must be of equal length and serially complete. As flow records are available since 1960, it was necessary to generate a serially complete data set of maximum and minimum temperature, and precipitation for the same period (1960-98). The meteorological station Mount Cook, station number H30711, adjacent to the Hooker River catchment, has daily temperature records since 1972 and precipitation records since 1901, with some missing records. This paper describes a procedure for estimating those missing records. The detailed procedure for generating missing values in the maximum and minimum temperature records of Mount Cook station

within the recorded period of 1972-98, and extrapolating those daily temperature records of Mount Cook station back to 1960 are presented here as example of application of the 'aiNet' model in hydrological studies.

Existing methods for estimating missing data

Very few meteorological stations worldwide have complete records. This difficulty has led researchers to develop a variety of procedures to estimate missing records, each fulfilling a particular need for a given location. Kemp *et al.* (1983) compared seven methods for estimating missing maximum and minimum temperature records. They were broadly classified into three categories: 1) within-station, 2) regression-based, and 3) between-station techniques. In within-station techniques, missing data values are calculated using records from the previous and subsequent time steps. An example of this approach would be the calculation of missing maximum temperature on 11 May using the average of the maximum temperatures recorded on 10 and 12 May. Similarly, an average can be calculated using temperatures for more than one day on either side of the missing day. Although these methods are satisfactory when calculating means for monthly or longer periods, shorter periods such as daily and hourly estimates can show significant errors.

Regression-based methods for estimating data tend to be more accurate than within-station methods (Kemp *et al.*, 1983). In the regression-based approach, linear regression equations are developed using the available data from one or more nearby stations; they are subsequently used to estimate missing records. Using this approach Kemp *et al.* (1983) reported a 50% reduction of mean absolute error associated with daily minimum temperature. While such methods are useful over limited areas, Degaetano *et al.* (1994) reported the limitations of such methods when data must be estimated for a large number of stations over a long period of time. This is especially true when it is necessary to update the regression equations with the change of observation time step and station relocation. This method also does not work when both the dependent and predictor variables have missing records at the same time step. One type of between-station method, as described by Kemp *et al.* (1983), is based on the assumption that the differences between daily temperatures at adjacent stations are equal to the differences between the monthly average temperatures of the sites. They reported that the mean absolute errors produced by the between-station method were lower than those associated with the within-station methods, but were slightly higher than those of regression-based methods. Regression-based methods thus turn out to be the best-suited methods so far reported.

Many of the available techniques for hydrological time series analysis assume linear relationships among variables. In the real world, data do not

exhibit simple regularities over time and are difficult to analyse and predict accurately. The linear models and their combinations for describing hydrological time series are often inadequate. To describe complex time series researchers have successfully adopted the Artificial Neural Network (ANN) approach, which is inspired by the neurons and synapses of the human brain. This method emulates the parallel-distributed processing of the brain and has proven to be very successful in dealing with complicated problems, such as function approximation, pattern recognition and time-series prediction. A number of studies have used ANN to solve problems in precipitation and rainfall/runoff processes. For example, French *et al.* (1992) used ANN to forecast rainfall intensity fields in space and time, while Hsu *et al.* (1995) applied an ANN to model the rainfall-runoff process. Although there are no reports of using ANN for forecasting and generating missing temperature records, the present study considered ANN, and in particular the neural-network model 'aiNet', as a useful tool for generating serially complete temperature time-series, based on its capability to perform non-linear input-output mapping.

Artificial neural network models

In this section a brief overview of Artificial Neural Network (ANN) is provided. More thorough discussions are available in French *et al.* (1992) and Hsu *et al.* (1995) among others.

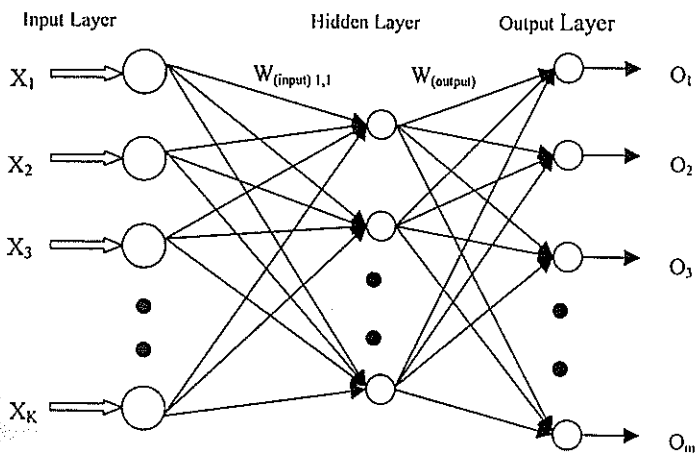


Figure 1 – Structure of a three-layer neural network model.

In essence, an ANN is a computing system made up of a number of interconnected nodes (called neurons) arranged into three basic layers—input, hidden, and output (there are variations on this topology, but they are beyond the scope of this paper). Each layer is made up of several nodes; layers are interconnected by sets of correlation weights. In the feed-forward network no computations are performed in the input nodes, but they are used to distribute inputs into the network. The neurons in the hidden or output layers receive input either from the initial inputs or from the interconnections. Figure 1 provides an overview of such ANN topology.

The number of input nodes, k, and the number of output nodes, m, in an ANN are dependent on the problem to which the network is being applied. There are no fixed rules as to how many nodes should be included in the hidden layer. For a given N data input sets, the input values X_i are multiplied by the first interconnection weights, W_{ij}, j=1...h at the hidden nodes, and the products are summed over the index, i, and become the inputs to the hidden layers. Each hidden node value is then transformed through a sigmoid function to produce a hidden node output Ho_j. The output, Ho_j serves as the input to the succeeding layer and this procedure is continued till the out layer is reached. The output of the hidden layer is multiplied by the output weights, W_{jn}, n=1...m at the output nodes, and the products are summed over the index, j, to become the input of the final or output layer. These input values are processed through the sigmoid function to give the neural network output values, O_n.

The output, O_n at the output layer will not be the same as the target value, T_n. For each input set, the sum of squares of error, e_p for the nth input set is:

$$e_p = \frac{1}{2}\sum_{n=1}^{m}(T_n - O_n)^2$$

Finally for all the input set the mean square error, MSE is calculated as follows:

$$MSE = \frac{1}{2N}\sum_{p=1}^{N}\sum_{n=1}^{m}(T_{pn} - O_{pn})^2$$

In the neural network modeling this mean square error is minimised iteratively by providing inputs to the model, computing the output, and adjusting the interconnection weights until the desired output is reached. This procedure is called the training phase of the neural network. In the training phase some historical observations are reserved for use in the testing phase. When the network is satisfactorily trained, these data are used to test the network and monitor its performance on test samples in terms of mean-square error criteria.

aiNet model

aiNet is a modern machine-learning modeling tool that is derived from the fields of neural networks and statistical models. As reported by its developers (Krajnc and Perus, 1995) it is a kind of Artificial Neural Network (ANN) based on a self-organising system presented by Grabec (1990). It is similar to the methods of the nearest neighbour, Learning Vector Quantisation Networks (Kohonen, 1988), and also to the probabilistic neural network proposed by Specht (1988). All these methods have the same foundation and similar rules for describing various phenomena. In this section we briefly describe aiNet in a non-mathematical way; more detailed information is available in the users' manual of aiNet (Krajnc and Perus, 1995).

Modelling of a phenomenon means the selection of the right attributes of a phenomenon and the suitable coding of the right number of measurements and/or facts in the mathematical form. aiNet is a tool that allows modeling in a defined sense, and they are presented in brief below from the users' manual.

Basic principles and derivation

To develop a model within aiNet, it is assumed that a phenomenon can be described by a number of partial observations, L, each containing m_i variables. These partial observations are composed of both inputs to and outputs from the phenomenon of interest. Each observation is called a model vector and can be written in the form:

$$\mathbf{mv} = (mv_1, mv_2, \ldots, mv_L)$$

A proper description of the phenomenon requires many observations. With N total observations, a database (or knowledge base) is developed which can be represented by a finite set of model vectors:

$$\text{model} = \{\mathbf{mv}_1, \mathbf{mv}_2, \ldots, \mathbf{mv}_N\}$$

This can then be written in matrix form as:

$\mathbf{mv}_1$	=	m_{11}	m_{12}	...	m_{1L}
$\mathbf{mv}_2$	=	m_{21}	m_{22}	...	m_{2L}
...		...			...
...		...			...
$\mathbf{mv}_N$	=	m_{N1}	m_{N2}	...	m_{NL}

For use of this database in the neural net we can assume that each model vector consists of two partial vectors, the first representing M input variables and the second the L–M outputs. Labelling the first as **P** and the second as **Q**,

$$\mathbf{P} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M)$$

$$\mathbf{Q} = (\mathbf{m}_{M+1}, \mathbf{m}_{M+2}, \ldots, \mathbf{m}_L)$$

Concatenation of both vectors gives the original model vector. This can be written as:

$$\mathbf{mv} = \mathbf{P} \oplus \mathbf{Q} = \left(m_1, m_2, \ldots, m_M, m_{M+1}, \ldots, m_L\right)$$

or in matrix form:

$\mathbf{mv}_1$	=	m_{11}	m_{12}	...	m_{1M}	$M_{1,M+1}$	...	m_{1L}
$\mathbf{mv}_2$	=	m_{21}	m_{22}	...	m_{2M}	$M_{2,M+1}$	...	m_{2L}
...		...						...
...		...						...
$\mathbf{mv}_N$	=	m_{N1}	m_{N2}	...	m_{NM}	$M_{N,M+1}$	...	m_{NL}

Shadowed part belongs to the partial vector **Q**.

P and Q constitute a database for the phenomenon being observed and can be used to 'train' the neural network. One way in which aiNet differs from conventional neural nets is that rather than a number of hidden layers with weights calculated for each neuron, a single factor controls the training of the net. This factor, the penalty coefficient, is indirectly related to the learning error and determines the shape of the solution curve in two-dimensional problems, and the shape of the hyper-plane in three- or more-dimensional problems. As aiNet is a non-parametric method, the assumptions of an underlying shape function that best suits the phenomenon is not required. aiNet adapts the model automatically when the data is changed, or new data are added to the database with the newly defined penalty coefficient. aiNet allows this penalty coefficient be static, dynamic or nearest neighbour. In the latter two cases its value is additionally modified for each point in the hyperspace by different built-in methods. This makes the computation of solutions using aiNet comparably faster and more attractive than using conventional neural nets. In aiNet, conventional training is replaced by the introduction of model vectors into the model by loading a database of model

vectors. Predicted output values are then estimated for known input vectors to the model.

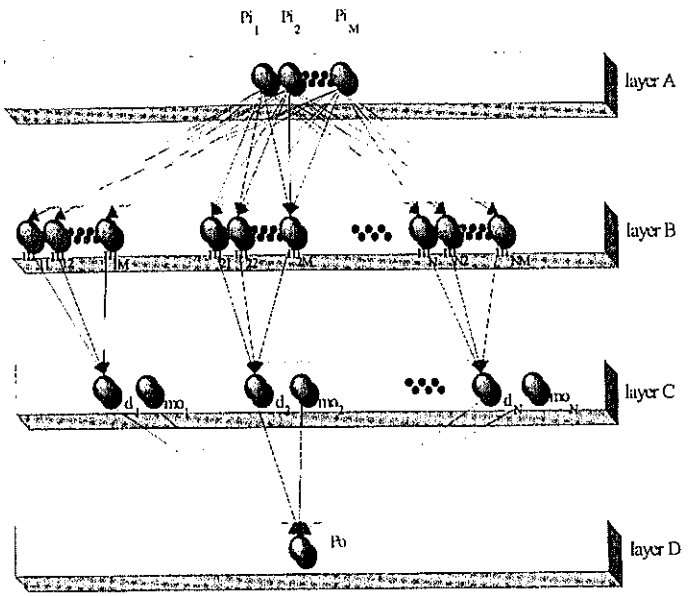
During the prediction phase, the penalty coefficient is determined on a trial-and-error basis, with the minimum value being adopted for prediction. This optimum value corresponds to the minimum value of the root mean square (RMS) error in prediction in the filtration and verification processes. During filtration, an output variable is predicted for each model vector using all available model vectors from the training set in the database. This process gives a straightforward estimation of the noise in the data. During verification, the model vector under consideration is excluded from the predictions, giving a better global error estimation. Given a series of input vectors, output variables can then be calculated. A model developed in this way can be considered as a kind of primitive intelligence, which corresponds to the training and learning process in the development of ANNs. In one of such treatments a particular model vector $\mathbf{m}_n$ is characterised as a neuron. When driven by a particular input vector **P**, the neuron is excited as described by the amplitude c_n and contributes to the complete output of the network, **Q**.

Graphical presentation of the aiNet

Figure 2 provides a brief overview of the aiNet topology in a manner similar to the other ANNs. An aiNet network is made up of a number of interconnected nodes (called neurons) arranged into three basic layers — input (layer A), hidden (layer B and C) and output (layer D). In this network, information passes one way through the network from the input layer, to the hidden layers and finally to the output layer.

The number of input nodes, M, and the number of output nodes, O, in an aiNet topology depend on the problem to which the network is being applied. Unlike other ANNs the aiNet has two hidden layers. The number of nodes in these hidden layers depends on the model vectors and the number of variables in the model vector. The number of neurons in layer B is (N*M) and equals to the product of the number of all the model vectors, N, and the number of input variables, M. The number of neurons in layer C is twice the number of model vectors (N). They are denoted by d_1 to d_N and mo_1 to mo_N in Figure 2.

Figure 3 provides a closer look at an individual neuron in the hidden or output layer. Each neuron, J, receives number of inputs (Pi_1, ... ,Pi_N) according to the connection topology. The neurons in the hidden layer B receive inputs from previous input layer A. Similarly, hidden layer C receives input from previous hidden layer B and output layer D receives input from the hidden layer C. These inputs are accompanied by a weight, w_{ij}, which represents a factor by which any values passing into the neuron are



Notations in Figure 2 have the following meaning:

- Pi_k, k=1,....M is a neuron belonging to the input vector where M is the number of input variables of the phenomenon.
- m_{jk}, j=1,....,N j=1,.....,M is a neuron in the first hidden layer B where N is the total number of model vector and M is the total number of input variables.
- d_j and mo_j, j=1,.....,N are the neurons in the second hidden layer C where N is the total number of model vector.
- Po is a neuron belonging to the output layer D.

Figure 2 – Graphical presentation of the aiNet model.

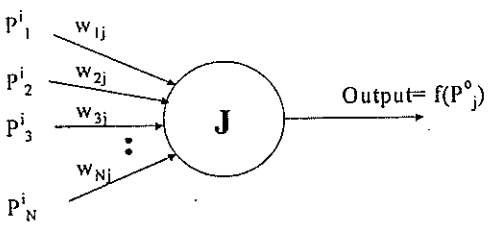


Figure 3 – An Artificial Neuron in Hidden or Output Layer.

multiplied. In aiNet the value of weights on connections is either equal to one or equal to zero. A neuron, J, sums all the input values according to the following equation:

$$P_j^o = \sum_{k=1}^{N} P_k^i w_{kj}$$

These input values are then transformed by a linear function to produce the output of the neuron, as it is transformed by a sigmoid function in other ANNs. In layer C the output of the neurons is a linear function of the input value and of a penalty coefficient. For a particular penalty coefficient, the model calculates the RMS error for each set of predictions. Consequently, different RMS errors could be calculated for different penalty coefficients. The optimum penalty coefficient, or, in other words, the best model, is then selected corresponding to the minimum RMS error.

Figure 2 shows the graphical presentation of aiNet; it is obviously very similar to the other ANNs (see Figure 1). To compare it exactly with other supervised ANNs, we can distinguish between its training and prediction phases. The training phase in aiNet is very quick and corresponds to the presentation of the model vectors (loading the data base or aiNet data file) to the network – aiNet. The calculation of the unknown output values of the prediction vector (in the case of prediction) or output values of model vectors (in the case of filtration of verification for optimising the penalty coefficient value) correspond to the prediction phase of the ANN model.

Statistical modelling of temperature data

The meteorological station located within the Hooker River catchment area is H30711, 'Mount Cook at the Hermitage', and is operated by the Department of Conservation. The station is located at latitude 43° 44' S and longitude 170° 06' E at an altitude of 765 m (Fig. 4, Table 1). Two other stations (H30712 and H30713) in the area operated for a short period. Nearby meteorological stations with temperature records are Franz Josef, Haast, and Hokitika, located on the west side of the Main Divide, and Queenstown, Lake Tekapo, and Twizel located on the east side of Main Divide. The location, length of record and distance from Mount Cook station are given in Table 1.

The Mount Cook meteorological station adjacent to the Hooker River catchment has daily maximum and minimum temperature records since 1972 and daily precipitation records for the entire study period of 1960-98. The Hooker River flow measurement gauge (site no. 71125) at the Ball Hut Road Bridge site has flow records since 1960. To make full use of these

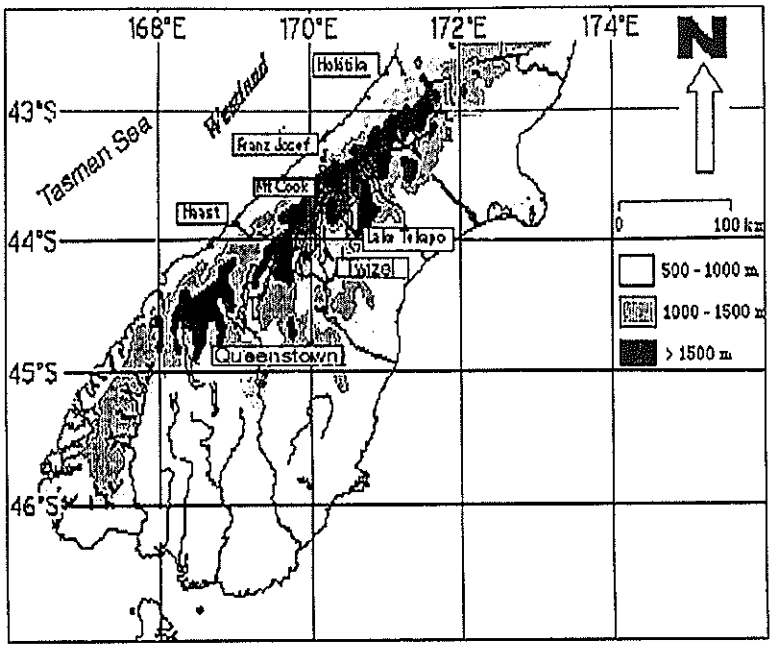


Figure 4 – Southern part of the South Island, New Zealand, showing the location of Mount Cook and other neighbouring meteorological stations with temperature records.

Table 1 – Nearby meteorological stations and their distance from Mount Cook Station.

Station Number	Station	Location *WRT Main Divide	Temperature Record Length	Latitude	Longitude	Distance from Mount Cook in km
F30312	Franz Josef	West	1972-98	43.383	170.167	28.8
F39801	Haast	West	1972-76	43.850	169.000	89.6
F20793	Hokitika	West	1964-98	42.712	170.983	109.3
I58061	Queenstown	East	1866-98	45.033	168.667	156.7
H40041	Lake Tekapo	East	1970-98	44.017	170.467	37.5
H40213	Twizel	East	1972-91	44.250	170.100	41.8
H30711	Mt. Cook	East	1972-98	43.733	170.100	0.0

*WRT – With respect to

Source: NIWA archive

flow and precipitation records the temperature time series must be extended back to 1960, and the missing records within the period 1972-98 must be generated with adequate dependability.

The meteorological stations at Franz Josef, Haast, Hokitika, Queenstown, Twizel and Lake Tekapo have daily temperature records since 1972 or earlier (Table 1). Of those, the stations at Twizel and Haast have temperature records up to 1991 and 1976 respectively. The inclusion of Twizel and Haast data does not improve the model accuracy, so after preliminary investigation these stations were omitted from the study. The daily maximum and minimum temperature records of the remaining meteorological stations, Franz Josef, Hokitika, Queenstown and Lake Tekapo, were used in estimating the missing temperature records of Mount Cook station within the 1972-98 period. The records of Queenstown and Hokitika were also used to extrapolate the Mount Cook station maximum and minimum temperatures back to 1960.

Using the Minitab statistical software, the descriptive statistics and correlation coefficients for Franz Josef, Lake Tekapo, Hokitika, Queenstown, and Mount Cook stations maximum and minimum temperatures for the period 1972-98 were calculated, and are shown in Tables 2 and 3. The notations used in all the tables in Statistical Modeling and aiNet Modeling sections have the following meanings:

- Fzmn — Minimum temperature of Franz Josef station.
- Fzmx — Maximum temperature of Franz Josef station.
- Tekmn — Minimum temperature of Lake Tekapo station.
- Tekmx — Maximum temperature of Lake Tekapo station.
- Hokimn — Minimum temperature of Hokitika station.
- Hokimx — Maximum temperature of Hokitika station.
- Qnmn — Minimum temperature of Queenstown station.
- Qnmx — Maximum temperature of Queenstown station.
- Mtcmn — Minimum temperature of Mount Cook station.
- Mtcmx — Maximum temperature of Mount Cook station.
- N — Total number of available records in the recorded period.
- N* — Total number of missing records in the recorded period.
- Tr Mean — Truncated Mean.
- StDev — Standard deviation of the sample.
- SE Mean — Standard error in mean.
- S — Estimated standard deviation about the regression line.

Table 2 – Descriptive statistics for maximum and minimum temperatures at Franz Josef, Lake Tekapo, Hokitika, Queenstown and Mount Cook stations for the period 1972-98.

Variable	N	N*	Mean	Median	Tr Mean	StDev	SE Mean
Fzmn	9418	291	6.5459	6.7000	6.5687	3.9085	0.0403
Fzmx	9436	273	15.527	15.100	15.450	3.817	0.039
Tekmn	9673	36	3.3437	3.1000	3.3523	5.2078	0.0530
Tekmx	9673	36	14.018	14.000	14.031	6.697	0.068
Hokimn	9701	8	7.5426	7.7000	7.5533	4.2601	0.0433
Hokimx	9702	7	15.659	15.500	15.613	3.339	0.034
Qnmn	9709	0	5.5746	5.5000	5.5514	4.5839	0.0465
Qnmx	9708	1	15.630	15.500	15.564	6.065	0.062
Mtcmn	9624	85	3.5973	3.5000	3.5692	4.9706	0.0507
Mtcmx	9574	135	13.755	13.600	13.725	6.303	0.064

Table 3 – Pearson correlation coefficients between stations

	Fzmn	Fzmx	Tekmn	Tekmx	Hokimn	Hokimx	Qnmn	Qnmx	Mtcmn
Fzmx	0.662								
Tekmn	0.827	0.605							
Tekmx	0.750	0.632	0.829						
Hokimn	0.923	0.624	0.844	0.761					
Hokimx	0.758	0.904	0.695	0.690	0.724				
Qnmn	0.861	0.684	0.914	0.840	0.866	0.764			
Qnmx	0.710	0.766	0.721	0.775	0.685	0.791	0.794		
Mtcmn	0.848	0.640	0.922	0.861	0.854	0.725	0.928	0.770	
Mtcmx	0.627	0.746	0.641	0.752	0.596	0.736	0.711	0.896	0.711

Among all the records of maximum and minimum temperatures for the stations, only Queenstown station minimum temperature does not has any missing record for the period 1972-98; it has a maximum correlation coefficient of 0.928 with Mount Cook station minimum temperatures. The first step then was to develop a serially complete data set for Mount Cook station minimum temperatures so that it can be used to generate the missing values in the Mount Cook station maximum temperatures also.

To select the best variables in the regression equation to generate the missing values in the Mount Cook station minimum temperature record, the Best Subset regression was done between Mount Cook station minimum temperatures and the maximum and minimum temperatures of Franz Josef, Lake Tekapo, Hokitika, and Queenstown stations. The best regression coefficient (R^2 adj) that could be obtained was 0.909, using the maximum and minimum temperatures of Lake Tekapo and Queenstown stations and minimum temperature of Hokitika station as independent variables. This

equation was also used for extrapolating Mount Cook station minimum temperatures to 1970. For extrapolating the Mount Cook station minimum temperatures between 1960-69 the best regression coefficient (R^2 adj) that could be obtained was 0.86, using the minimum temperature of Queenstown station as the independent variable. The results of Best Subset Regression are summarized in Table 4.

Table 4 – Summary of Best Subsets Regression for generating Mount Cook station minimum temperatures using the maximum and minimum temperatures at Franz Josef, Hokitika, Lake Tekapo and Queenstown stations.

Vars	R-Sq	R-Sq (adj)	C-p	S	Tekmn	Tekmx	Hokimn	Hokimx	Qnmn	Qnmx	
5	90.8	90.8	140.9	1.5005	X	X	X		X	X	. (1)
1	86.1	86.1	1489.4	1.8516					X		. (2)

The C-p value is a measure of goodness of fit when comparing the model with different combinations of predictors. In general we look for a model with a lower C-p value and the highest R^2 (adj) value, which is equivalent to choosing the model with the smallest mean square error (MSE). The statistics of the generated minimum temperatures of Mount Cook station (RegMtcmn) for the period 1960-98 are provided in the aiNet Model section to compare with output of the aiNet models.

Using the same procedure as that for the generation of missing values in Mount Cook station minimum temperature, the Best Subset Regression was done between the Mount Cook station maximum temperature and the maximum and minimum temperatures of Franz Josef, Lake Tekapo, Hokitika, Queenstown, and the Mount Cook station minimum temperature. The best regression coefficient (R^2 adj) that could be obtained was 0.823, using the maximum and minimum temperatures of Lake Tekapo, Hokitika, Queenstown and the minimum temperatures of Mount Cook station as predictor variables. The regression equation based on the best regression coefficient (0.823) was used to generate the missing values of Mount Cook maximum temperatures for the period 1972-98 and extrapolating those maximum temperatures back to 1970. For extrapolating Mount Cook station maximum temperatures between 1964-69 the best regression coefficient (R^2 adj) that could be obtained was 0.805, using the maximum temperature of Hokitika and Queenstown stations as independent variables. For the

remaining period (1960-63) Queenstown station maximum temperatures were used to predict Mount Cook station maximum temperatures (regression coefficient of 0.803). Two regression equations were used to extrapolate Mount Cook station maximum temperatures from 1964 to 1969 and 1960 to 1963, with the maximum temperatures of Hokitika and Queenstown stations as predictor variables for the first period and the maximum temperatures of Queenstown station as the predictor variable for the second period. Though their regression coefficients do not differ much, the C-p value is 1164 in the first case when two predictor variables are used, whereas the C-p value increased to 1263.5 when only Queenstown station maximum temperatures are used as the predictor variable. The results of Best Subset Regression for generating Mount Cook station maximum temperatures are summarized in Table 5.

Table 5 – Summary of Best Subsets Regression for generating Mount Cook station maximum temperatures using the maximum and minimum temperatures at Franz Josef, Hokitika, Lake Tekapo and Queenstown stations and the minimum temperatures at Mount Cook Station.

Vars	R-Sq	R-Sq (adj)	C-p	S	TekmnTekmnxHokminHokminxQnminQnmxMtcmn	
7	82.3	82.3	8.0	2.6218	X X X X X X X	(3)
2	80.5	80.5	1164.9	2.7781	X X	(4)
1	80.3	80.3	1263.5	2.7910	X	(5)

The maximum temperatures of Mount Cook station (RegMtcmx) were generated for the period 1960-98 using equations 3, 4, and 5 and are now compared with that of aiNet model output.

Application of aiNet

Maximum and minimum temperatures of Mount Cook station show good correlation with the maximum and minimum temperatures of Tekapo, Hokitika, and Queenstown (Table 3). Because of their good correlations, two model vectors for the Mount Cook minimum temperature were developed for generating the missing values and extrapolating values back to 1960. The first model vector was developed using the maximum and minimum temperatures of Tekapo, Hokitika, and Queenstown stations for the period 1972-97 as input vectors and the minimum temperature of

Mount Cook station for the same period as the output vector. This model was used in generating the Mount Cook station minimum temperature for the period 1970-98, and the output of the model is designated as nn72_97 hereafter. The second model vector was developed for extrapolating the Mount Cook minimum temperature back to 1960 using maximum and minimum temperatures of Hokitika and Queenstown stations for the period 1972-85 as input vectors and the minimum temperature of Mount Cook station for the same period as the output vector. This model is designated as nn72_85 hereafter. The penalty coefficients were set to 0.09 and 0.15 in dynamic mode corresponding to the minimum Root Mean Square Error (RMSE) of the filtration and verification process.

To compare the aiNet models outputs (nn72_97 and nn72_85) with those of the statistically generated minimum temperatures (RegMtcmn) and the recorded minimum temperatures (Mtcmn) at Mount Cook station, their descriptive statistics, correlation and regression analysis were done. They are shown in Tables 6, 7, and 8. The prime objective of this study is to generate serially complete data series, but the statistically generated record, RegMtcmn, has 45 missing values, as there are missing records in the maximum and minimum temperatures of Tekapo, Hokitika and Queenstown stations. The aiNet model outputs (nn72_85 and nn72_97) do not have any missing values, despite the missing values in the input vector. The aiNet model output nn72_97 shows the best correlation with the measured record of Mount Cook minimum temperatures (Mtcmn). Evaluated for the period 1972-98 it shows a correlation coefficient of 0.966 and a regression coefficient (R^2 adj.) of 0.934, compared to the correlation coefficient of 0.953 and a regression coefficient (R^2 adj.) of 0.908 of the statistically generated minimum temperatures RegMtcmn. The aiNet model (nn72_85) built using the shorter length of record (1972-85) and four input vectors has a correlation coefficient of 0.952 and a regression coefficient of 0.907 with the measured minimum temperatures (Mtcmn), but it has better C-p statistics compared to that of the statistically generated minimum temperatures (RegMtcmn).

Table 6 – Descriptive Statistics of generated and measured minimum temperatures at Mount Cook station for the period 1972-98.

Variable	N	N*	Mean	Median	Tr Mean	StDev	SE Mean
Mtcmn	9624	85	3.5973	3.5000	3.5692	4.9706	0.0507
nn72_85	9709	0	3.5798	3.5000	3.5455	4.6239	0.0469
nn72_97	9709	0	3.5803	3.5000	3.5864	4.5599	0.0463
RegMtcmn	9664	45	3.5940	3.5480	3.5993	4.7354	0.0482

Table 7 – Pearson correlation coefficients between generated and measured minimum temperatures at Mount Cook station for the period 1972-98.

	Mtcmn	RegMtcmn	nn72_85
RegMtcmn	0.953		
nn72_85	0.952	0.941	
nn72_97	0.966	0.9930	.951

Table 8 – Best Subsets Regression between measured Mount Cook station minimum temperatures and generated minimum temperatures for the period 1972-98.

Response is Mtcmn
9580 cases used; 129 cases contain missing values.

Vars	R-Sq	R-Sq (adj)	C-p	S	FIT1	nn85	nn72_97
1	93.4	93.4	2475.2	1.2798			X
1	90.8	90.8	7077.6	1.5044	X		
1	90.7	90.7	6698.6	1.5166		X	

Similarly, two model vectors for the maximum temperatures of the Mount Cook station were developed for generating missing values and extrapolating them back to 1960. In the first model, the maximum and minimum temperatures from 1972-90 of the Tekapo, Hokitika, Queenstown stations and the minimum temperatures of Mount Cook station were used as input vectors, and maximum temperatures of Mount Cook station for the same period was used as the output vector. The second model was developed using the maximum and minimum temperatures of Queenstown station and minimum temperatures of Mount Cook station for the period 1972-86 as input vectors and the maximum temperatures of Mount Cook station for the same period as the output vector. The first model was used for generating the maximum temperatures at Mount Cook station for the period 1970-98 and the second model was used to extrapolate the maximum temperatures at the Mount Cook station back to 1960. The model outputs are denoted as ndMtcmx1 and ndMtcmx2 respectively.

Tables 9, 10, and 11 below show the descriptive statistics, correlation matrix, and best subsets regression-output between the measured maximum temperatures at Mount Cook station (Mtcmx), two aiNet model outputs (ndMtcmx1 & ndMtcmx2) and a statistical model output, RegMtcmx. Evaluated for the period 1972-98, the measured maximum temperature and statistical model outputs have 135 and 45 missing records respectively due to missing values in the predictor variables. The aiNet model outputs, ndMtcmx1 and ndMtcmx2, do not have any missing records, as the model automatically can cope with the missing values in input vectors. The aiNet model output, ndMtcmx1, has a correlation coefficient of 0.945 and a regression coefficient (R^2 adj.) of 0.893 with the measured record, Mtcmx. The statistically generated maximum temperature gives a correlation coefficient of 0.907 and a regression coefficient (R^2 adj.) of 0.823 with the measured record. The regression coefficient of the second aiNet model output (ndMtcmx2) is 0.819 with the measured records, which is little less than that of the statistical model output, but the C-p value of ndMtcmx2 is much lower than that of the statistical model output.

Table 9 – Descriptive Statistics of the measured and generated maximum temperatures at Mount Cook station for the period 1972-98.

Variable	N	N*	Mean	Median	Tr Mean	StDev	SE Mean
Mtcmx	9574	135	13.755	13.600	13.725	6.303	0.064
ndMtcmx1	9709	0	13.707	13.600	13.670	5.864	0.060
ndMtcmx2	9709	0	13.811	13.600	13.759	5.561	0.056
RegMtcmx	9664	45	13.754	13.600	13.693	5.719	0.058

Table 10 – Pearson correlation coefficients between generated and measured maximum temperatures at Mount Cook station for the period 1972-98.

	Mtcmx	ndMtcmx1	ndMtcmx2
ndMtcmx1	0.945		
ndmtcmx2	0.905	0.919	
RegMtcmx	0.907	0.883	0.866

Table 11 – Best Subsets Regression between measured Mount Cook station maximum temperatures and generated maximum temperatures for the period 1972-98.

Response is Mtcmx
9530 cases used; 179 cases contain missing values.

Vars	R-Sq	R-Sq (adj)	C-p	S	ndMtcmx1	ndMtcmx2	RegMtcmx
1	89.3	89.3	2969.6	2.0563	X		
1	81.9	81.9	8060.0	2.6837		X	
1	82.3	82.3	1E+04	2.6458			X

Summary

Since many physically and statistically-based models of weather–sensitive processes require serially complete temperature data, a new method to estimate missing records and also to extend the data records in relation to the records of nearby meteorological stations has been developed for the Mount Cook station. This method is used to generate missing values in the 1972-98 Mount Cook station maximum and minimum temperature records and to extrapolate those records back to 1960. To evaluate the efficiency of the aiNet model in estimating maximum and minimum temperatures at the Hooker River catchment, its performance is compared with that of commonly used multiple linear regression methods using the values of neighbouring stations. Two aiNet models were developed for generating maximum and minimum temperatures at Mount Cook station. The first aiNet model uses maximum and minimum temperatures at Tekapo, Hokitika and Queenstown stations for the period 1972-97 as input vectors and Mount Cook minimum temperatures for the same period as the output vector to generate the minimum temperatures for the period 1970-98. The second aiNet model for extrapolating the minimum temperature back to 1960 was developed using fewer input vectors and a shorter length of record. It used the maximum and minimum temperatures at Hokitika and Queenstown stations for the period 1972-85 as input vectors and the Mount Cook minimum temperatures for the same period as the output vector. Similarly, the first aiNet model of Mount Cook maximum temperatures was developed using the same input vectors as that of minimum temperature, but for the period 1972-90; it

included the Mount Cook minimum temperatures for the same period in the input vectors. The second aiNet model for Mount Cook maximum temperature was developed with only maximum and minimum temperatures at Queenstown station and minimum temperatures at Mount Cook station for the period 1972-86 as an input vectors. All the aiNet models were then refined by adding a few data points of extreme values for the remaining period (1987-98) of available data.

The performance of these four aiNet models and two statistical models are compared with those measured values. Figure 5 shows the scatter plots for the results of minimum temperature with the best-fit line for reference. Figure 6 shows the same for the maximum temperatures at Mount Cook station. These figures show that aiNet model outputs fall close to the best-fit line when compared with statistical model outputs. Additional information about the relative performance of these models can be obtained from the statistics in Table 12.

Table 12 – Evaluation statistics for the aiNet model and statistical model outputs for the period 1972-98.

Technique	Model name	RMSE (°C)	Regression Coff/ R^2(adj)	Correlation Coefficient	Number of Missing Values
		Minimum Temperature			
Regression		1.499	0.908	0.953	45
AiNet	nn72_97	1.287	0.934	0.966	nil
AiNet	nn72_85	1.476	0.907	0.952	nil
		Maximum Temperature			
Regression		2.646	0.823	0.907	45
AiNet	ndMtcmx1	2.058	0.893	0.945	nil
AiNet	ndMtcmx2	2.687	0.819	0.905	nil

Examination of the above table provides a number of useful insights. First of all, the first aiNet model for both maximum and minimum temperatures outperformed the statistical model in estimating the maximum and minimum temperatures of Mount Cook station. Furthermore, the second aiNet model's statistics are quite comparable to that of the statistical model, even though they were developed using fewer variables and shorter lengths of record. Finally, and perhaps most importantly, all the aiNet models were able to generalise relationships for the input data sets whilst remaining relatively robust despite noisy and missing inputs.

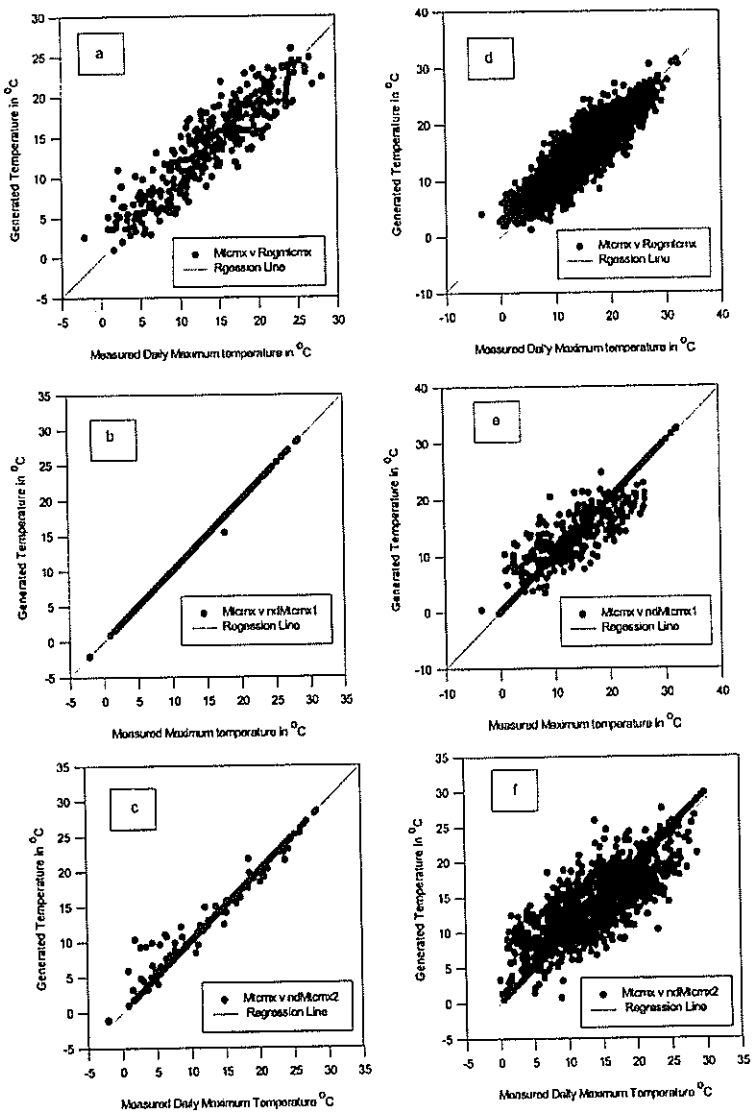


Figure 5 – Scatter plots for generated vs. measured maximum temperatures at Mount Cook station, where a, b, and c represents the year 1972 and d, e, and f represents the years 1985-91.

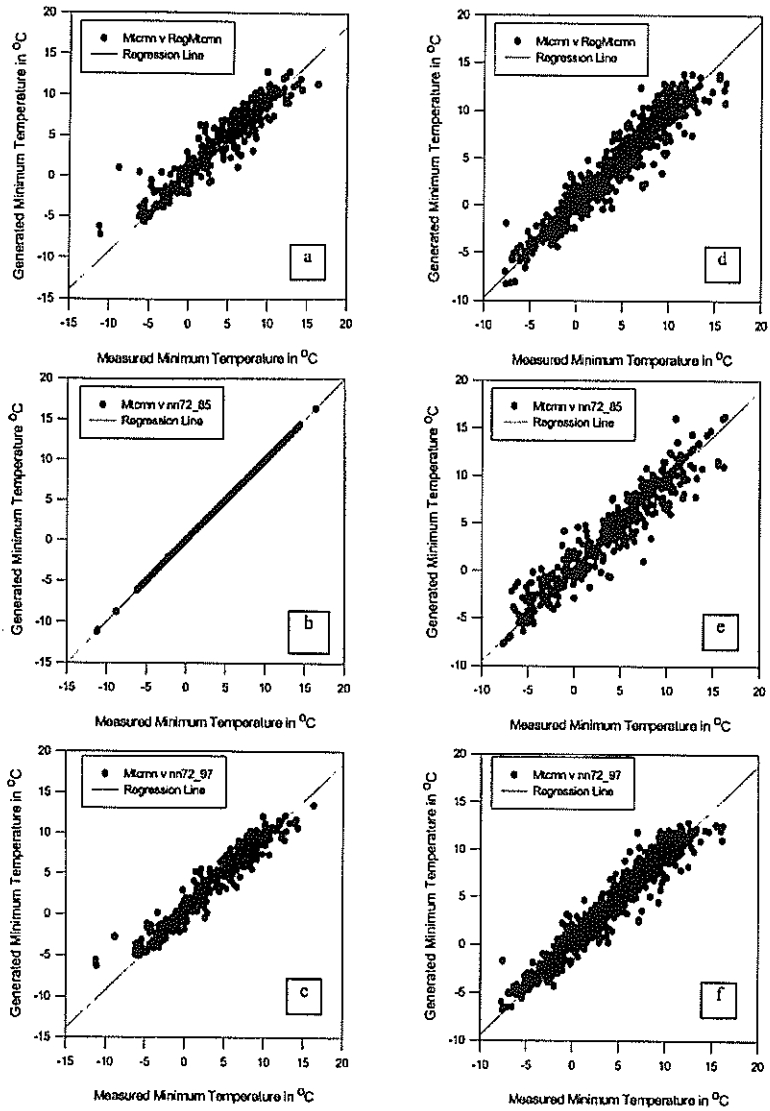


Figure 6 – Scatter plots for generated vs. measured minimum temperatures at Mount Cook station, where a, b, and c represents the year 1972 and d, e, and f represents the years 1985-86.

Conclusion

The aiNet model applied to the generation of daily maximum and minimum temperatures for Mount Cook station has given encouraging results. The same technique could be applied to the development of serially complete daily precipitation time series for the same station and for other hydro-meteorological variables in other catchments. The aiNet model can automatically cope with missing records in input vectors and it is faster in computation than statistical and other ANN models. The efficiency of the aiNet model depends on the appropriate selection of input vectors on the one side and a sufficient number of model vectors on the other side.

Acknowledgments

The research reported in this article is supported by the Commonwealth Postgraduate Scholarship to New Zealand. This support is gratefully acknowledged. The authors are indebted to Kevin McGill of NIWA for providing a useful database for the study.

The authors also gratefully acknowledge the helpful comments from Dr. David Painter and Dr. Stuart Larsen who read the manuscript and made several valuable suggestions. The authors also wish to thank the anonymous reviewers for their comments and helpful suggestions.

References

Auer, A.A. 1992: Wave Cloud Formations in the Lee of Southern New Zealand, *Weather 47:* 103-105.

Barry, R.G. 1994: *Mountain Weather and Climate*, Methuen, London, New York.

Cherry, N.J. 1972: Winds and Lee Waves Over Canterbury, New Zealand, During 1970, *New Zealand Journal of Science 15*: 587-600.

Degaetano, A.T.; Knapp, W. W.; Egglestonet, K. L. 1995: A Method to Estimate Missing Daily Maximum and Minimum temperature Observations, *Journal of Applied Meteorology 34*: 371-380.

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.

Grabec, I. 1990: Modeling of Natural Phenomena by a Self-Organizing System, Proceedings ECPD Neurocomputing 1(1) (mentioned in the aiNet user's manual Krajnc, A. and Iztok Perus, 1995).

Hsu, K.; Gupta, H.V.; Sorooshin, S. 1995: Artificial Neural Network Modeling of the Rainfall-Runoff Process, *Water Resources Research 31(10)*: 2517-2530.

Huth, R.; Nemesova, I.1995: Estimation of Missing Daily Temperatures-Can a Weather Categorization Improve its Accuracy, *Journal of Climate 8(7)*: 1901-1916.

Kemp, W.P.; Burnell, D. G.; Everson, D. O.; Thomson, A. J. 1983: Estimating Missing Daily Maximum and Minimum Temperatures, *Journal of Appled Meteorology 22*: 1587-1593.

Kohonen, T. 1988: *Self-Organization and Associated Memory*, Second Edition, Springer-Verlag, Berlin.

Krajnc, A; Iztok Perus 1995: User's Manual: aiNet.

Ryan, A.P. 1987: The Climate and Weather of Canterbury (Including Aorangi), Misc. Pub. Vol 115 (17), New Zealand Meteorological Service: 66.

Salinger, M.J.; Rhodes, D. A. 1993: Adjustment of temperature and Rainfall Records for Site Changes, *International Journal of Climatology 13:* 899-913.

Specht, D.F. 1988: Probabilistic Neural Networks for Classification, Mapping or Associative Memory, International Conference on Neural Network, Conference Proceedings.

Manuscript received: 11 October 1999; accepted for publication: 19 May 2000.