Neural Networks for Rainfall Forecasting by Atmospheric Downscaling

J. Olsson[1]; C. B. Uvo[2]; K. Jinno[3]; A. Kawamura[4]; K. Nishiyama[5]; N. Koreeda[6]; T. Nakashima[7]; and O. Morita[8]

Abstract: Several studies have used artificial neural networks (NNs) to estimate local or regional precipitation/rainfall on the basis of relationships with coarse-resolution atmospheric variables. None of these experiments satisfactorily reproduced temporal intermittency and variability in rainfall. We attempt to improve performance by using two approaches: (1) couple two NNs in series, the first to determine rainfall occurrence, and the second to determine rainfall intensity during rainy periods; and (2) categorize rainfall into intensity categories and train the NN to reproduce these rather than the actual intensities. The experiments focused on estimating 12-h mean rainfall in the Chikugo River basin, Kyushu Island, southern Japan, from large-scale values of wind speeds at 850 hPa and precipitable water. The results indicated that (1) two NNs in series may greatly improve the reproduction of intermittency; (2) longer data series are required to reproduce variability; (3) intensity categorization may be useful for probabilistic forecasting; and (4) overall performance in this region is better during winter and spring than during summer and autumn.

DOI: 10.1061/(ASCE)1084-0699(2004)9:1(1)

CE Database subject headings: Neural networks; Rainfall; Forecasting; Japan.

Introduction

While numerical atmospheric models can accurately reproduce free-atmosphere circulation, they are far less successful in predicting precipitation [for example, Giorgi and Mearns 1991] mainly because of the simplified parametrization schemes used and the large grid size, which causes relevant geographical features to be crudely represented in the models. Therefore, statistical weather forecasting has been used by meteorologists for many years to complement and enhance the output from atmospheric models [for example, Klein (1982); Glahn (1985); Wilks (1995)]. The concept here is to derive statistical relationships between observed precipitation at a certain point or in a region (the predictand) and relevant free-atmosphere variables (the predictors). The methodology has recently gained interest as a tool for relating large-scale, coarse-resolution atmospheric data to local or regional hydrometeorological variables—so-called (atmospheric) downscaling [for example, Hewitson and Crane (1996); Wilby et al. (1998)]. In particular, it has been applied to the output from general circulation models in order to assess local hydrometeorological impacts under different scenarios for future climate change [for example, Giorgi and Mearns (1991); Zorita and von Storch (1999)].

The concepts of statistical forecasting and downscaling also open up possibilities for hydrologists and engineers to develop rainfall prediction models with limited effort, but this prospect does not seem to have been given much attention. A particularly attractive feature is that the forecasting models are tailor-made for a certain catchment or region, so that the derived relationships implicitly take the specific geographical forcing into account. For model calibration, access to a sufficient amount of historical data on both the potential predictors and the desired predictand is required. Predictor data are generally obtained from standard (coarse-resolution) atmospheric data sets covering the region of interest. For model application, access to real-time or forecast predictor data is required. Note that coarse-resolution global atmospheric forecasts, potentially applicable for the present purpose, are expected to become routinely distributed in the not-too-remote future [for example, Kidson and Thompson (1998)].

A wide range of statistical techniques have been used to identify predictor-predictand relationships, such as ordinary and inflated regression and canonical correlation [for example, Karl et al. (1990); Wigley et al. (1990)]. Recently, artificial neural networks (NNs) also have been used, which are particularly attractive owing to their natural ability to accommodate the nonlinear interactions taking place between atmospheric processes at different scales [for example, Giorgi and Mearns (1991); Zorita and von Storch (1999)]. In an early application of NNs for short-term downscaling, Hewitson and Crane (1992) used a back-propagating NN to estimate daily mean precipitation in a region in Mexico on the basis of sea-level pressure and 500 mb geopotential height, obtaining a correlation with observed precipitation

[1]Senior Researcher, Swedish Meteorological and Hydrological Institute, SE-601 76 Norrköping, Sweden. E-mail: jonas.olsson@smhi.se

[2]Associate Professor, Dept. of Water Resources Engineering, Lund Univ., Box 118, SE-221 00 Lund, Sweden.

[3]Professor, Institute of Environmental Systems, Kyushu Univ., 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.

[4]Associate Professor, Institute of Environmental Systems, Kyushu Univ., 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.

[5]Research Associate, Institute of Environmental Systems, Kyushu Univ., 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.

[6]Manager, CTI Engineering Co., Ltd., 2-4-12 Daimyo, Chuo-ku, Fukuoka 810-0041, Japan.

[7]Senior Engineer, CTI Engineering Co., Ltd., 2-4-12 Daimyo, Chuo-ku, Fukuoka 810-0041, Japan.

[8]Professor, Dept. of Earth and Planetary Sciences, Kyushu Univ., 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.

Note. Discussion open until June 1, 2004. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on June 5, 2001; approved on March 4, 2003. This paper is part of the ***Journal of Hydrologic Engineering***, Vol. 9, No. 1, January 1, 2004. ©ASCE, ISSN 1084-0699/2004/1-1–12/$18.00.

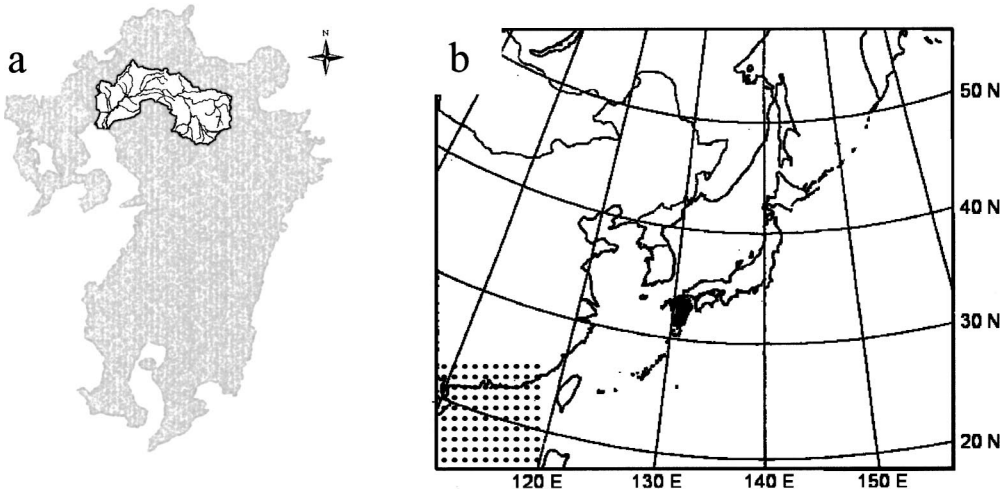

Fig. 1. (a) Kyushu Island, southern Japan, and location of Chikugo River basin; (b) area in which grid point value meteorological data were available (Kyushu Island is marked in black). Dots in bottom left corner of (b) represent density of grid point value grid points (100×100 km resolution)

of ~0.8. Similar results for the same region were obtained by Cavazos (1997). Crane and Hewitson (1998) used NNs for daily downscaling in the Susquehanna basin in the eastern United States, reaching similar correlation strengths.

A few studies have compared the performance of NNs with daily downscaling based on other methods, such as weather generators, historical analogs, and vorticity index techniques (Wilby et al. 1998; Zorita and von Storch 1999). In these comparative studies the performance of NNs was generally somewhat poorer than the other methods, mainly owing to an inability of the NNs to reproduce two key features of high-resolution precipitation time series: intermittency and variability. Intermittency refers to the incidence of both ''wet'' and ''dry'' time intervals, that is, with and without observed precipitation; NNs tend to generate small trace precipitation in actual dry intervals, thereby underestimating the observed zero-depth probability [for example, Wilby et al. (1998)]. Variability, on the other hand, is particularly manifested in sudden and short-lived extreme intensities of magnitudes many times the mean intensity; NNs typically underestimate the observed extreme intensities [for example, Zorita and von Storch (1999)].

The overall aim of the present study is to further investigate the potential of NNs for rainfall prediction by atmospheric downscaling, and in particular to explore ways to improve the performance with respect to intermittency and variability. The experiments are carried out for the Chikugo River basin on Kyushu Island, southern Japan, and the temporal resolution is 12 h. The predictand is the mean catchment rainfall, and the predictors are wind speeds at 850 hPa and precipitable water, selected in a previous investigation (Uvo et al. 2001). As the predictors' explanatory power varies in both space and time [for example, Wilby and Wigley (2000)], the areas used to specify the predictors' values are determined by correlation analysis on a seasonal basis. We interpret the results with respect to rainfall-generating mechanisms in the region.

Study Region and Databases

The experiments aimed at predicting the mean rainfall in the Chikugo River basin (~3,000 km^2), located in the northern part of Kyushu Island, Japan [Fig. 1(a)]. The river flows from a volcanic caldera in central Kyushu and plays an important role as a freshwater resource. The river is particularly characterized by a wide range of flows between its low and high extremes, which has been manifested, on the one hand, in severe droughts for the surrounding communities, notably in 1978 and 1994, and on the other hand, in devastating floods and debris flows, exacerbated by the catchment's steep mountain slopes and river beds [for example, Merabtene et al. (1998)].

The main climatological rainfall-generating mechanisms of the basin, and of Kyushu Island in general, are as follows:

- Spring: transient midlatitude synoptic cyclones moving along well-defined storm tracks;
- Summer: so-called Bai-u front, a polar front gradually passing Kyushu Island from south to north, associated with mesoscale rainbands and cyclones bringing heavy precipitation;
- Fall: similar to spring with the addition of frequent typhoons; and
- Winter: winter monsoon, occasionally associated with snowfall, and synoptic disturbances originating from the so-called Taiwan low.

The present study specified the large-scale atmospheric properties by grid point value (GPV) meteorological data in a region spanning approximately 105–160°E and 20–55°N, provided by the Japan Meteorological Agency [Fig. 1(b)]. The data consist of sounding measurements at 21 vertical levels, obtained at 00*Z* and 12*Z*. Originally measured variables include sea-level pressure or geopotential height, zonal and meridional wind speeds, temperature, and dew point depression. These were interpolated from the irregularly spaced measurement stations into a 20×20 km resolution grid by means of optimum interpolation [for example, Daley (1991)]. From the original data, additional variables including vorticity, precipitable water (that is, vertically integrated humidity), and total-totals index (a measure of atmospheric stability) were derived. Note that the importance of including some measure of humidity in statistical atmospheric downscaling has been recently emphasized [for example, Crane and Hewitson (1998)]. Finally, all data were reduced to a 100×100 km resolution by arithmetic averaging [Fig. 1(b)].

Mean rainfall in the Chikugo River basin (CRb rainfall) was

determined as the arithmetic average of 11 precipitation gauges from the Japanese national meteorological network, which are evenly distributed within and just outside the catchment area [Fig. 1(a)]. Measurements were made on an hourly basis, but 12-h totals were used to correspond with the GPV data. Each 12-h accumulation period started at the time of the GPV soundings; that is, the GPV values were used to estimate the rainfall during the next 12-h period.

Uvo et al. (2001) used singular value decomposition (SVD) to investigate the potential of the (original and derived) GPV variables to function as predictors for the CRb rainfall. It was found that zonal and meridional wind speeds at 850 hPa (u_{850},v_{850}) and precipitable water (PW) are highest correlated to the CRb rainfall and thus the most efficient predictors, especially when combined. However, SVD is based on linear combinations of variables, and thus the possibly nonlinear response in CRb rainfall to large-scale atmospheric changes could not be accurately represented, notably resulting in a systematic underestimation of high rainfall intensities. In the present study we therefore replace SVD with NNs as the method used for the prediction of CRb rainfall, as NNs allow for nonlinear relationships to be interpreted.

Data were available between April 1996 and August 1999, leading to a total of 2,417 12-h periods, with 614 in spring (March–May), 717 in summer (June–August), 546 in fall (September–November), and 540 in winter (December–January). For further details of the study region and databases, see Uvo et al. (2001).

Correlation Analysis

While using the predictors specified at every grid node in the GPV area as input to the NN is theoretically possible, this would make the model less efficient as not all GPV nodes' predictor values are strongly related to the CRb rainfall. Moreover, a large number of input nodes require large amounts of calibration data, which further makes NN training slow. Therefore it is desirable to isolate the predictor areas exerting the highest influence on CRb rainfall and use these as the basis for determining the NN input, as also argued and demonstrated by Wilby and Wigley (2000). Note that in previous rainfall downscaling experiments this issue has generally not been taken into consideration, but an arbitrary synoptic area covering the region of the predictand has been used when specifying the predictor values.

The present study used the standard (linear) correlation coefficient (CC) between the time series of CRb rainfall (the predictand) and the potential predictors' values at each GPV grid point to evaluate the spatial correlation pattern and in turn determine the influencing predictor areas. It may appear contradictory to use a linear technique to define the inputs to a nonlinear model, but it should be emphasized that the predictor-predictand relationships (as visualized, for example, in scatter plots) are not clearly and strongly nonlinear. A nonlinear regression is likely to be somewhat more accurate than a linear one, but the difference is generally small; therefore linear correlation was found sufficient for comparing correlation strengths, both between predictors and within the GPV area for the same predictor. Note that linear correlation is only used for the identification of predictor regions; the NN model uses the actual predictor-predictand relationships rather than any simplified linear approximations. From the resulting correlation fields in the entire GPV area, subareas in which the correlation was significant at a 95% level, corresponding to a maximum absolute correlation coefficient ($|CC|$) of at least 0.25, were used for calculating the final NN inputs. For each subarea and time step (that is, 12-h period), the final input was a weighted areal mean; in the averaging the predictor value at each grid point was weighted according to the grid point $|CC|$.

In agreement with the SVD analysis of Uvo et al. (2001), u_{850}, v_{850}, and PW proved to be the most meaningful predictors, both owing to their correlation strengths and to the fact that their spatial correlation patterns have a clear physical interpretation. Figs. 2(a–c) shows the results from the correlation analysis for summer. Concerning PW, a significantly correlated area elongated in an east-west direction was found covering Kyushu Island [Fig. 2(a)]. This location corresponds to the Bai-u front (see the section on the study region and databases). Fig. 2(b) displays the correlation field for u_{850}, and the overall picture is similar to PW, although the significant area extends to southern China in the west and far beyond eastern Japan in the east, emphasizing the strong control of the Bai-u front over summer rainfall. The field of v_{850}, however, displays quite different features, with a positively correlated area centered over Kyushu Island and elongated in a southwest-northeast direction and a negatively correlated area located over the Yellow Sea [Fig. 2(c)]. This picture suggests a cyclonic flow of synoptic disturbances migrating approximately along the Bai-u front, a well-known climatological feature of the area [for example, Uvo et al. 2001]. For the positively correlated areas in Figs. 2(a–c), the maximum CC reached ~0.4, and the weighted means of these areas were thus used as NN input. For the negatively correlated area in Fig. 2(c), maximum $|CC|$ was <0.25, and thus this area was not selected as input for the NN.

It must be emphasized that the correlation fields differ markedly with the season in both the location of significant areas and their correlation strength (CC). For example, compare Fig. 2(b) (u_{850}, summer) with Fig. 2(d), which shows the correlation field for u_{850} in winter. Rather than the elongated, east-west oriented pattern found in summer, three distinct areas appear in winter; a positively correlated area northeast of Taiwan and two negative areas north and east of it. This pattern suggests a cyclonic flow west of Kyushu Island, bringing in cyclones from the Taiwan low (see the section on the study area and databases). A convergence zone with northerly winds is located south of Kyushu Island, and further east of Japan the influence of the Pacific high becomes apparent. In Fig. 2(d), the maximum $|CC|$ was >0.25 in the negatively correlated area northwest of Kyushu Island but <0.25 in the negative area southeast of Kyushu Island.

For different seasons, the number of NN inputs varied between three (summer, autumn), four (spring), and five (winter), depending on the number of significantly correlated areas corresponding to each predictor. This underlines the importance of seasonal separation. The maximum value of $|CC|$ reached was 0.62 (PW, winter).

Neural Network Experiments: General Design and Initial Application

Prior to NN application, the original input and target time series were preprocessed. Since the rainfall target was given as a spatial average, the definition of wet and dry periods required some limit below which the average rainfall is considered insignificant. This limit must be related to the application. For example, rainfall fields of a very limited spatial extension, such as those of convective origin, can most probably not be traced on the basis of larger-scale atmospheric features. By studying the data and considering the rainfall station network density, it was decided to require a nonzero rainfall intensity in at least three stations for the spatial

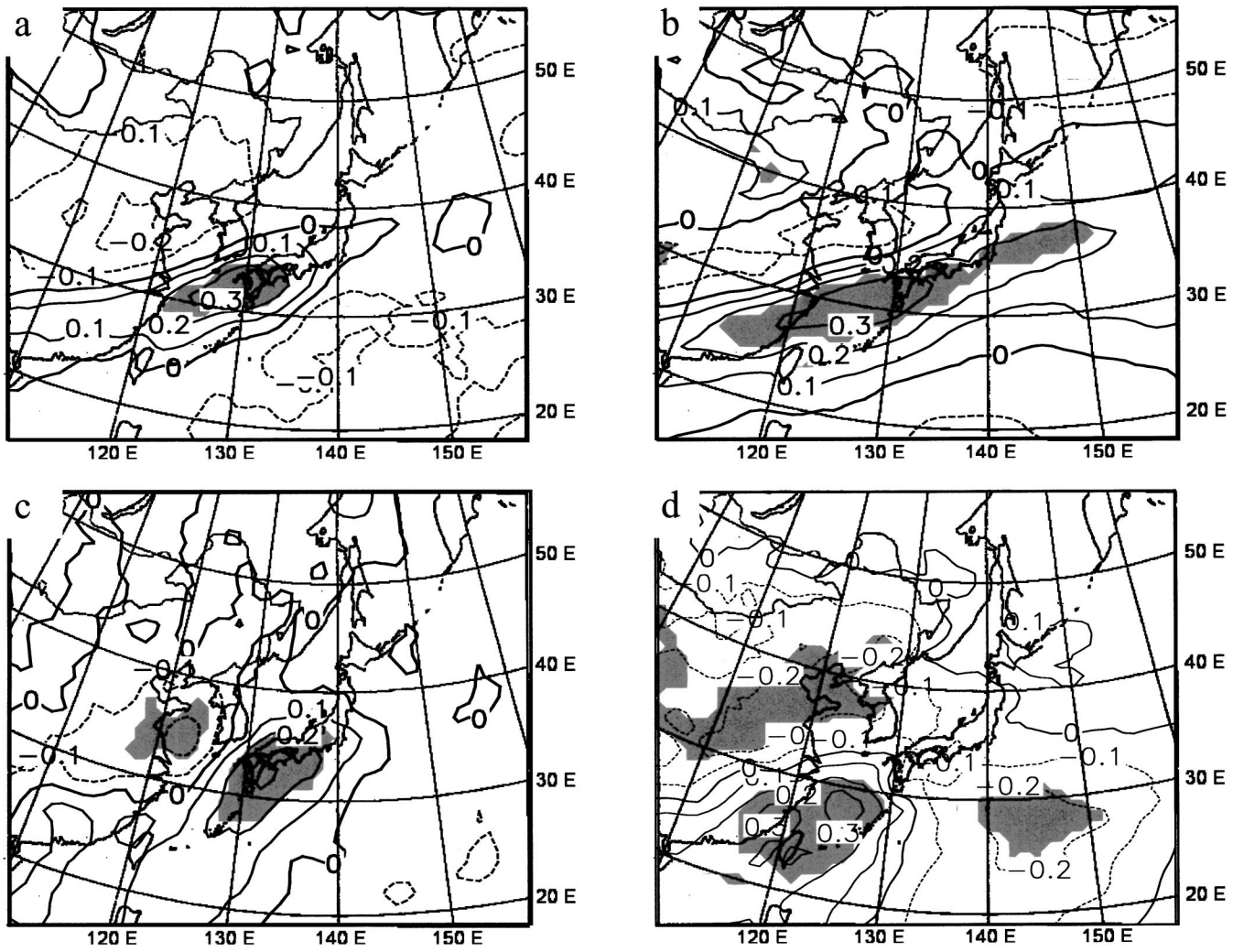

Fig. 2. Correlation fields representing covariation between mean rainfall in Chikugo River basin and grid point value variables: (a) precipitable water; (b) zonal wind speed; (c) meridional wind speed (850 hPa) in summer; (d) zonal wind speed (850 hPa) in winter. Isolines correspond to correlation coefficient and shaded areas denote 95% statistical significance

rainfall to be significant, that is, to define the 12-h period as a wet period. If rain occurred in less than three stations, the period was considered dry. This limit roughly corresponded to a spatial average rainfall of 1 mm.

To ensure that every input receives equal attention during the training, the input series were standardized by subtracting the series' mean and dividing by the standard deviation [for example, Maier and Dandy (2000)]. The target values were rescaled, generally in the range 0.1–0.9, in order to be directly attainable as output from the log-sigmoid transfer function, avoiding the asymptotic limits 0 and 1 [for example, Smith (1993)]. One-fourth of the series were reserved for independent verification

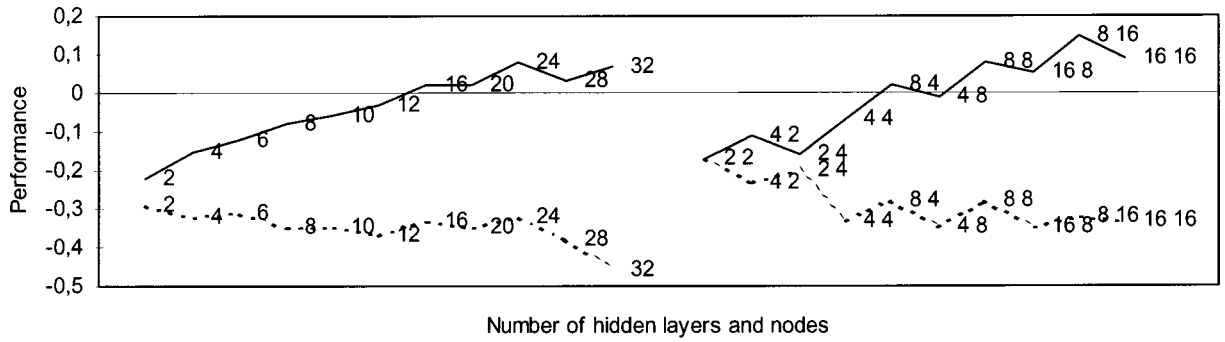

Fig. 3. Neural network performance expressed as correlation coefficient-root mean square error for training (solid line) and validation (dashed line) sets as function of number of hidden layers (left curves: one; right curves: two) and nodes in each layer (numbers)

(verification period) and were not used at all in the NN calibration process. As the data span almost 4 years and the application was done on a seasonal basis, the validation data generally corresponded to one season. The remaining 75%—the calibration period—were further divided into training and validation periods, with 80% in the former (including the maximum input values during the entire calibration period) and 20% in the latter.

General properties of NNs, as well as their applications in hydrology and water resources, have been thoroughly covered in a number of publications [for example, Smith (1993); Bishop (1995); Maier and Dandy (2000)]. For background information the reader is referred to this literature; only specific properties of the NNs employed are given here.

We used feed-forward NNs trained by a back-propagation algorithm using the Levenberg-Marquardt optimization (Hagan and Menhaj 1994). In plain language, back-propagation can be explained as the adjustment of NN parameters (weights and biases, in NN terminology) by back-propagating the differences between the NN output and the actual target (that is, CRb rainfall). Typically the data available for NN calibration are split into two periods, one for training and one for validation. This division is made in order to use cross-validation or early stopping, a common and practical method to avoid overfitting but ensure proper generalization [for example, Bishop (1995)]. Only the training set is used for the back-propagation, and parameter adjustment based on one sequence through all values is termed one epoch. Cross-validation means that during the training, the performance for the validation period is checked after each epoch. When a consistent decrease in the performance for the validation period is observed, training is stopped and the NN is considered calibrated.

A log-sigmoid transfer function was allocated to every neuron and defined as

$$ov = \frac{1}{1+e^{-iv}} \tag{1}$$

where iv = input value to the neuron and ov = output value from the neuron.

Determination of the number of neurons and how to divide them into separate layers are delicate issues. Large (or complex) NNs require a lot of data to generalize well and are computationally intensive; small (or simple) NNs may not be able to reproduce intricate input-output (I/O) relationships [for example, Bishop (1995)]. Some attempts have been made to relate the optimal number of neurons to the number of training samples [for example, Rogers and Dowla (1994)]. Some systematic approaches also have been made that include so-called constructive algorithms, which essentially start from a minimal NN and add neurons until performance ceases to increase [for example, Kwok and Yeung (1997)]. Conversely, so-called pruning algorithms start from a large NN and remove neurons until performance starts to decrease [for example, Reed (1993)]. NN size and geometry are, however, highly problem-dependent, and therefore trial-and-error still appears to be the most widely used method [for example, Maier and Dandy (2000)]. In the present study we use the principle of constructive algorithms in combination with cross-validation to find the most suitable NN size.

It is worth noting that cross-validation generally can not compensate for the tendency to overfitting arising from an excessively large and powerful NN, which runs the risk of also learning to reproduce noise in the calibration set [for example, Smith (1993)]. For the present data, this is illustrated in Fig. 3. The figure shows a typical example of how the performance for the training and validation sets, respectively, varies with the number of neurons and hidden layers. Performance of the NNs was assessed in terms of both the correlation coefficient (CC) and the root mean square error (RMSE) of the rescaled data. We started from an NN with two neurons in one hidden layer and then increased the NN size up to two hidden layers with 16 neurons in each, keeping training parameters and methods constant. From Fig. 3 it is obvious that despite using cross-validation, generalization ability decreases with increasing NN size. The picture may change somewhat with location of the validation set (for example, the beginning or end of the calibration period), season, and NN configuration (see the section on dealing with intermittency and variability), but overall it is clear that a small NN is preferable. As two hidden layers generally proved slightly superior to one, the final NN geometry employed consisted of two hidden layers with four and two nodes, respectively.

The performance of calibrated NNs for the validation period exhibited some variation, even when a small NN was used. For example, for ill-conditioned configurations of initial weights and biases determined by the method of Nguyen and Widrow (1990), which essentially ensures that the initial neurons are active within and roughly cover the input space, proper convergence was not attained during training, making the NN output arbitrary. To tackle the NN output variation required developing a procedure for extracting a final output from a number of calibrated NNs. First, from n_{tot}-calibrated NNs, the n_{sel} NNs that exhibited the highest performance for the entire calibration period (training +validation) were selected. Second, the outputs of these n_{sel} NNs for the verification period were averaged, and to evaluate performance this averaged series was compared with the observed verification data. Averaging was required since there was still some variation in performance among the n_{sel} NNs. Generally, the averaged series performed similarly to the output of the better n_{sel} NNs. To reach a stable and optimally accurate averaged output, using $n_{tot}=25$ and $n_{sel}=5$ generally proved sufficient. In the following, NN output refers to the averaged output for the verification period from the 5 out of 25 trained NNs of the above-given size and geometry that performed best for the calibration period.

Fig. 4 shows a typical example of NN output using the original summer time series as target (CC=0.70 and RMSE=10.8). Generally, the NN well captures the timing of rainfall events, but it is clear that peak intensities are severely underestimated, and the fraction of zero rainfall in the NN output, $f_o(0)$, is also severely underestimated. The actual zero-rainfall periods are in the NN output represented by a small positive intensity making $f_o(0)=0.00$, to be compared with $f_t(0)=0.54$ for the target series. It should be remarked that the output averaging is likely to have some negative impact on performance, particularly in terms of $f_o(0)$. Nevertheless, the overall reasonable agreement shown in Fig. 4 suggests that modifying the NN application strategy may be possible to obtain a more accurate and useful output, and this possibility is explored in the following section.

Dealing with Intermittency and Variability

In this section we explore two ways to deal with intermittency and variability: (1) using two serially coupled NNs, one for rainfall occurrence and one for rainfall intensity; and (2) classifying the target (that is, rainfall) into intensity categories. For this purpose, results for summer are presented as this is the dominant season in the present rainfall regime.

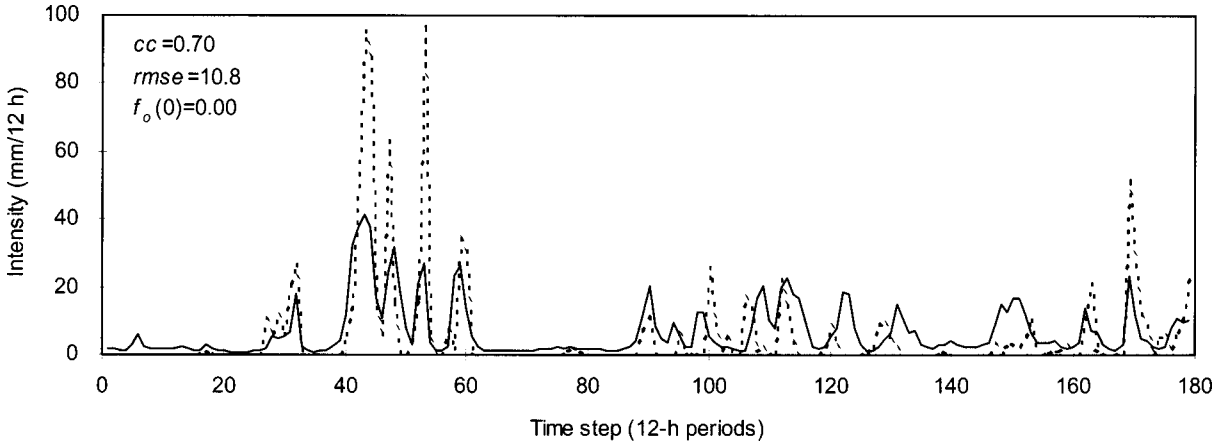

Fig. 4. Neural network output (solid line) for summer using original observed time series (dashed line) as target

Two Neural Networks in Series

An idea previously put forward is to use two NNs in series, the first to determine rainfall occurrence, and the second to determine the intensity during rainy periods [for example, Wilby et al. (1998)]. The first NN, trained on a binary wet-or-dry series, would fix the zero-depth probability, and the second, trained on a series of only rain events, would have a better chance to reproduce high intensities than would using the entire time series for training.

NN1: Rainfall Occurrence

To evaluate the NN's ability to separate periods with zero rainfall (dry periods) from periods during which rainfall occurred (wet periods), the original Chikugo River basin mean rainfall time series were initially converted into a binary series with 0 representing a dry period and 1 representing a wet period.

A type of NNs tailor-made to produce binary output are the so-called perceptron networks [for example, Rosenblatt (1961)]. For the perceptron neuron, $ov=1$ for $iv \geq 0$ and $ov=0$ for $iv<0$. However, successful application of a perceptron model requires the binary classes in the target to be linearly separable. Preliminary tests indicated that this was not the case for the present data, which would lead to a poor performance of perceptron NNs. Instead, feed-forward NNs with nonlinear (log-sigmoid) transfer functions were employed. Since for the log-sigmoid function $0<ov<1$, to be attainable the dry and wet periods were represented by 0.1 and 0.9, respectively [for example, Smith (1993)]. In the final output from a trained NN the exact values 0.1 and 0.9 are generally not attained, but the values vary between 0 and 1. A threshold of 0.5 was used to determine the most probable output; that is, periods with lower values were considered dry and periods with higher values wet.

To evaluate the performance of the NNs, denote by $n_t(C)$ and $n_o(C)$ the number of values of category C (W for wet periods, D for dry) in the target and output series, respectively. Further, denote by $n_{o=t}(C)$ the number of hits, that is, the number of correctly reproduced values, of category C. The fraction of correctly reproduced values or hit rate hr in the verification period, is defined as

$$hr=\frac{n_{o=t}(W)+n_{o=t}(D)}{n_t(W)+n_t(D)} \tag{2}$$

which varies between 0 and 1, with high values indicating a high accuracy, and the bias of wet periods, $b(W)$, defined as

$$b(W)=\frac{n_o(W)}{n_t(W)} \tag{3}$$

for which values close to 1 indicate an accurate number of simulated wet periods [for example, Wilks (1995)].

Fig. 5 shows a typical NN output using the binary wet-or-dry summer time series as target. To make the figure clear, only half of the validation period is shown. For extended wet and dry spells the agreement is generally good, but the NN in some cases fails to identify single or a few wet periods embedded within longer dry spells, and vice versa. The overall satisfactory agreement is supported by the values $hr=0.78$ and $b(W)=0.99$. This technique

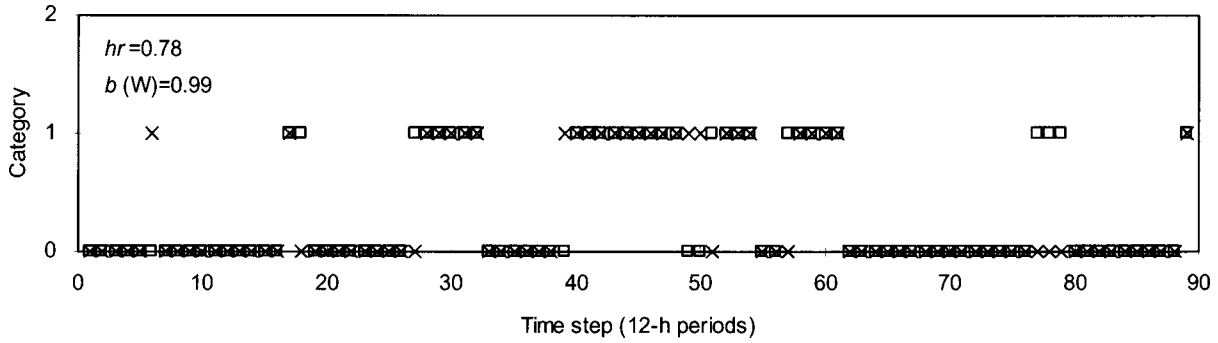

Fig. 5. Example of neural network output (crosses) for summer using observed time series converted into a binary wet-or-dry series (squares) as target (NN1)

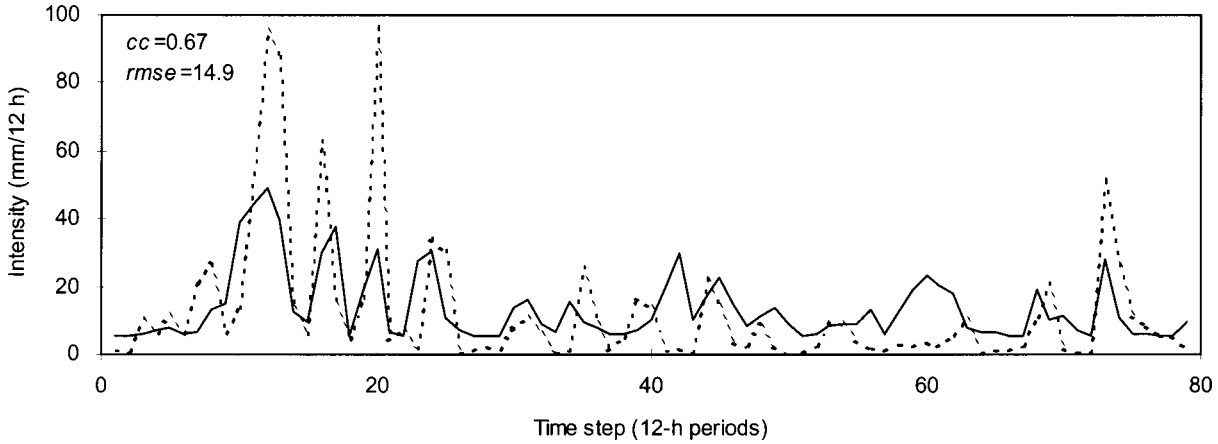

Fig. 6. Neural network output (solid line) for summer using observed time series with zero values removed (dashed line) as target (NN2)

thus appears to be a viable way to improve the reproduction of zero-depth probability; what remains to be estimated is the rainfall intensity during wet periods.

NN2: Rainfall Intensity

To study NN performance for wet periods, all 12-h periods with zero values in the target, as previously defined as rainfall at less than three stations, were omitted from the original series. The resulting series were then divided into training, validation, and verification periods of the relative proportions given in the section on the general design and initial application of NNs.

Fig. 6 shows a typical NN output for the present case. Overall the agreement is very similar to the agreement during wet periods when using the original series as a target, shown in Fig. 4. The values CC=0.67 and RMSE=14.9 for NN2 are essentially identical to the CC and RMSE obtained for the original series after removing periods with zero target from the results. Thus, for the present data, excluding the zero targets does not substantially improve the estimation of wet-period intensities.

Combining NN1 and NN2 (2NN/int)

NN1 and NN2, described in the sections on rainfall occurrence and rainfall intensity, must be run in series to produce a complete output (we term this experiment 2NN/int: two NNs in series with actual intensity as output). This may be done by training them separately, as described in the respective sections, running NN1 to determine whether a period is wet or dry, and running NN2 to estimate the rainfall intensity during periods classified as wet in NN1. A result from such a two-stage run is shown in Fig. 7. Visually the agreement between observed and simulated series is nearly identical to Fig. 4, and so are the values of CC (0.67) and RMSE (11.1). However, for the two-stage NN $f_o(0)=0.56$, that is, very close to the observed $f_t(0)=0.54$.

We conclude that the present approach of using two serially coupled NNs appears to be an effective way to improve performance with respect to the simulated fraction of zero rainfall, $f_o(0)$. Even if NN1 introduces some inaccuracy concerning the division into wet and dry periods, when combined with NN2 for wet period intensities, the result is similar to the result from using the original series as a target, except for a greatly improved estimation of $f_t(0)$.

Target Classification

As is obvious from the previous section, even if training the NN using nonzero values only, its ability to reproduce the exact rainfall intensity is limited for the present data. An alternative is to categorize the rainfall using hydrologically meaningful intensity

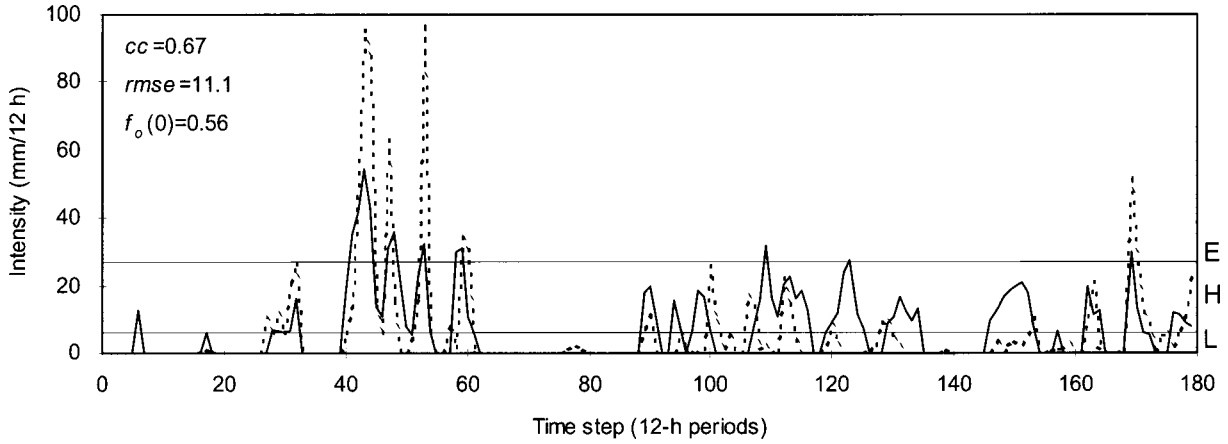

Fig. 7. Neural network output from experiment 2NN/int (solid line) for summer and observed time series (dashed line). Horizontal lines mark intensity limits used to separate classes *L*, *H*, and *E*

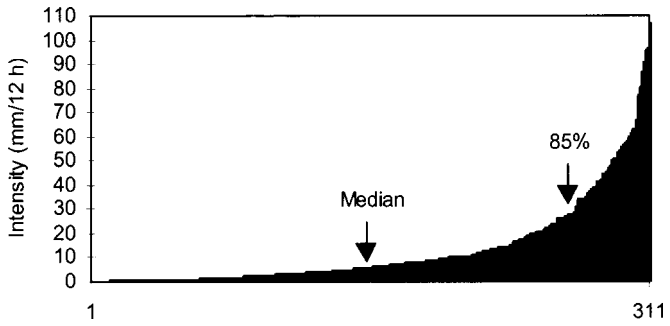

Fig. 8. Ordered plot of nonzero target intensities during summer with median and 85th percentile marked

thresholds and to train the NN with these categories. If a separate class for dry periods is included, classification can simultaneously address the problem of intermittency.

Categories and Evaluation

The experiments covered in the section on two NNs in series clearly demonstrate the difficulties involved in accurately predicting short-term rainfall intensities on the basis of large-scale atmospheric features. However, in many hydrological applications an exact value of the rainfall intensity may not be required, but a more qualitative assessment is sufficient. One important example is flood warning systems, in which flood forecasting is often triggered by a forecast (spatial average) rainfall intensity above some critical threshold [for example, Krzysztofowicz (1993)]. Thus, consideration may be given to dividing the rainfall intensity into meaningful categories and training the NN to reproduce these rather than the exact intensities. In light of the reasonably accurate reproduction of relative intensity magnitudes in the above experiments (for example, Fig. 4), we assume that classification has a good chance of improving NN output accuracy.

We chose to employ a classification of intensities into four categories: zero (Z), low (L), high (H), and extreme (E) intensity. Category Z was defined as rainfall in less than three stations. To separate L from H, the median of observed intensities is an obvious candidate. The separation of H from E is more difficult due to the conflict between the desired isolation of very high intensities and the need to have a sufficient number of occurrences of each category for a meaningful NN training and evaluation. The 85th percentile was found a reasonable choice for the present data. Fig. 8 shows an ordered plot of observed (target) intensities I_t for the wet periods, with the median and the 85th percentile marked. Whereas the median is located in a region of slowly increasing gradient, the 85th percentile roughly corresponds to a location where the gradient increases rather abruptly, indicating a suitable separation point of extreme values.

In the following application to summer data, categories L and H are separated by the median, corresponding to an intensity of 6 mm per 12-h period, and H and E by the 85th percentile, 27 mm per 12-h period. Thus, target intensities $I_t \leqslant 6$ were converted to category L, $6 < I_t \leqslant 27$ to H, and $27 < I_t$ to E (Fig. 7).

From the definition, the relative proportions of categories L, H, and E is 50/35/15. Thus, a particular training set is likely to contain more than three times as many L as E, which essentially means that the NN will learn the I/O relationship of category L three times better than that of E. Naturally the situation is similar when using actual values as output; the NN will be better trained on (the frequent) low than on (the rare) extreme intensities. When using categories, however, this effect may be compensated for rather straightforwardly in different ways [for example, Smith (1993)].

One possibility is to employ what is termed a stratified sample for the NN training. From the entire training set a nonrepresentative subsample designed to contain an identical or similar number of occurrences of each category is extracted and used in the NN training. We tested this approach for the present data, but did not reach any noticeable improvement. The potential gain in accuracy for high-intensity categories was generally offset by an overall reduction resulting from the shortening of the training set. We conclude that for the present data the entire training set is required for accurate NN calibration, and we assume that the definition of the categories produces a sufficient number of each.

For the NN training, the four categories Z, L, H, and E were represented by the values 0.2, 0.4, 0.6, and 0.8. As the (logsigmoid) NN output node can produce any value between 0 and 1 (in practice generally between the outer category values, that is, 0.2 and 0.8), the output value is generally between two categories (say, C_1 and C_2) and closer to one of them (C_1) than the other (C_2). For example, if the output value is 0.35, $C_1 = L$ and $C_2 = Z$. In light of this, it is natural to define two NN output series, $o1$ and $o2$. Series $o1$ consists of C_1, that is, the rainfall category closest to the actual NN output, for each 12-h period in the verification set. The entries in this series thus represent the most probable intensity category. Output series $o2$ contains C_2, essentially specifying whether the corresponding C_1 is inclined toward lower or higher intensities.

To evaluate NN performance for the case of multiple categories C_i, the hit rate expressed in Eq. (2) may be generalized as

$$hr_k = \frac{\Sigma_{i=1}^{N_C} n_{o=t}(C_i)}{\Sigma_{i=1}^{N_C} n_t(C_i)} \tag{4}$$

where k denotes the output series (1 or 2) and N_C is the number of categories. While being a proper measure of the overall output accuracy, hr_k is insensitive to variations in performance between categories. In particular, the performance for rare categories, which often are of most interest (for example, extreme rainfall), has very little impact on hr_k. An NN producing only the most frequently occurring category as output may get a high hr_k despite being practically useless. To better take each category into account, a weighted hit rate whr_k in which the hits were weighted with respect to the category's occurrence frequency was also used. This is defined as

$$whr_k = \sum_{i=1}^{N_C} \frac{n_{o=t}(C_i)}{N_C n_t(C_i)} \tag{5}$$

which, while varying similarly to hr_k between 0 and 1, gives equal weight to the performance of each category. It should be remarked that whr_k may suffer from being overly sensitive to rare categories for which a few hits or misses may have an excessive impact on the total value of whr_k. Thus hr_k and whr_k need to be used complementarily for the overall assessment. Finally, in the evaluation we use the bias $b_k(C)$ of each output series and category, defined as equivalent to Eq. (3).

Neural Network Experiment (1NN/cat)

In the NN experiment based on categorized data, 1NN/cat (single NN with output in terms of categories), the actual target values in the entire original series were replaced by their corresponding category value ($Z = 0.2$, $L = 0.4$, $H = 0.6$, or $E = 0.8$), after which

Table 1. Average Performance Measures and Fraction of Zero Rainfall for Experiments 1NN/cat and 2NN/cat, Summer Season. Standard Deviation of Biases=~0.1 and of Other Parameters =~0.02

Experiment	hr_1	hr_2	whr_1	$b_1(Z)$	$b_1(L)$	$b_1(H)$	$b_1(E)$	$f_o(0)$
1NN/cat	0.56	0.31	0.44	0.8	1.6	1.1	0.2	0.43
2NN/cat	0.63	N.A.	0.61	1.0	0.2	1.9	1.3	0.54

data separation, NN calibration, and NN simulation were performed as outlined in the section on the general design and initial application of NN experiments.

Typical values of the performance indicators are given in Table 1. Looking first at the most probable output series $o1$, $hr_1 = 0.56$, that is, just over half of all target categories were hit by the NN. The lower value of whr_1, 0.45, indicates an unequal performance between categories. From the biases, we find that the NN produces far too many items in category L [$b_1(L) = 1.6$] and far too few in category E [$b_1(E) = 0.2$]. Generally, the bias values reflect an inability of the NN to fully reach the outer categories Z and E; Z in the target is sometimes estimated by the NN to be L. As category Z is more frequent than L, this affects $b_1(L)$ more than $b_1(Z)$. Similarly for categories H and E, even if most E in the target are actually estimated to be H, this has only a small influence on $b_1(H)$.

The value $hr_2 = 0.31$ shows that if considering both $o1$ and $o2$, only 13%—that is $[1-(hr_1+hr_2)]\cdot 100$—of the target values are entirely missed by the NN. In other words, out of the 44% of the targets missed by $o1$, more than 70% are captured by $o2$. Thus the present approach does allow a rough quantitative estimate of the rainfall intensity with reasonable confidence.

Comparative Evaluation

To compare experiments 2NN/int and 1NN/cat, one possibility is to convert the output of 2NN/int into the corresponding intensity category (Z, L, H, or E). The median (6 mm) and 85th percentile (27 mm) of observed nonzero intensities were used for this conversion (Fig. 7). Thus, as in the target classification discussed in the section on catagories and elevation, NN output intensities $I_o = 0$ were converted to category Z; $0 < I_o \leq 6$ to L; $6 < I_o \leq 27$ to H, and $27 < I_o$ to E. We denote the converted output 2NN/cat.

The output 2NN/cat was compared with the target, similarly categorized, to assess performance using the hit rates and biases previously described. The results are summarized in Table 1 (for 2NN/cat, subscript 1 of hr, whr, and b corresponds to the one and only output series). In fact both hr_1 and whr_1 indicate a higher performance of 2NN/cat than of 1NN/cat. Whereas the standard hit rate hr_1 is rather close to the value obtained in the categorical experiments (~0.6), the value of the weighted hit rate whr_1 is substantially higher, 0.61, as compared with 0.44 in 1NN/cat. This indicates an improved performance of one or more categories. From visual inspection, and as indicated in Fig. 7, it is found that 2NN/cat performs significantly better than 1NN/cat for extreme values (E). This is reflected in the value of $b_1(E)$, being close to 1 for 2NN/cat. From the biases it may further be deduced that the contributions to whr_1 from categories Z and H are larger for 2NN/cat than for 1NN/cat, as more Z and H are produced in 2NN/cat. The number of L is, however, severely underestimated in 2NN/cat. Overall, 1NN/cat is noticeably biased toward low intensities and 2NN/cat toward high intensities.

Judging from the performance measures in Table 1, the approach of separating into two NNs is preferable to intensity categorization. However, the two approaches should be viewed as contrasting rather than complementary. The most appropriate approach is much related to any succeeding application of the estimated rainfall. If only one value of the intensity is accepted, 2NN/val is able to produce a reasonably accurate estimate that, however, may differ considerably from the actual intensity. The output from 1NN/cat is a probability-based range of intensities that contains the actual intensity with a high degree of confidence. Thus 1NN/cat may be more suitable for applications based on probabilistic forecasts.

Seasonal Application

Table 2 summarizes the results from seasonal applications of approaches 2NN/int and 1NN/cat. Looking first at the results from 2NN/int for summer, we find, as previously shown (Fig. 7), an overall reasonable agreement between observed and simulated series. However, high peak values between time steps 40 and 60 (June 22–July 2, 1999) are severely underestimated. During this entire period the Bai-u front was stationed just over Kyushu Island, generating favorable conditions for heavy rainfall. Further, intensities during the extended wet spell starting at time step 90 (July 17) are often overestimated. This period was characterized by very high values of PW and v_{850} (southerly winds), which made the NN forecast rather high intensities. The values of u_{850}, however, indicated only weak westerly or even easterly winds, and in reality the highest rainfall intensities were likely found east of Kyushu Island.

The value of the CC is markedly lower for autumn than for the other seasons. This is likely related to the presence of typhoons, which are associated with very extreme and often erratic conditions in terms of both the NN input and the target. For a limited data set such as the present, these values may heavily influence the training and lead to unanticipated NN output. This is clearly illustrated in Fig. 9(a), showing a typical output from 2NN/int for

Table 2. Average Performance Measures and Fraction of Zero Rainfall for Experiments 2NN/int and 1NN/cat and Fraction of Zero Rainfall in Target (Observed), All Seasons

	Experiment							
	2NN/int			1NN/cat				
Season	Correlation coefficient	Root mean square error	$f_o(0)$	hr_1	hr_2	whr_1	$f_o(0)$	Observed $f_t(0)$
Summer	0.64	11.4	0.54	0.56	0.31	0.44	0.43	0.54
Autumn	0.50	9.3	0.72	0.78	0.12	0.50	0.72	0.74
Winter	0.68	1.6	0.89	0.84	0.11	0.40	0.85	0.85
Spring	0.77	5.5	0.74	0.75	0.18	0.59	0.70	0.71

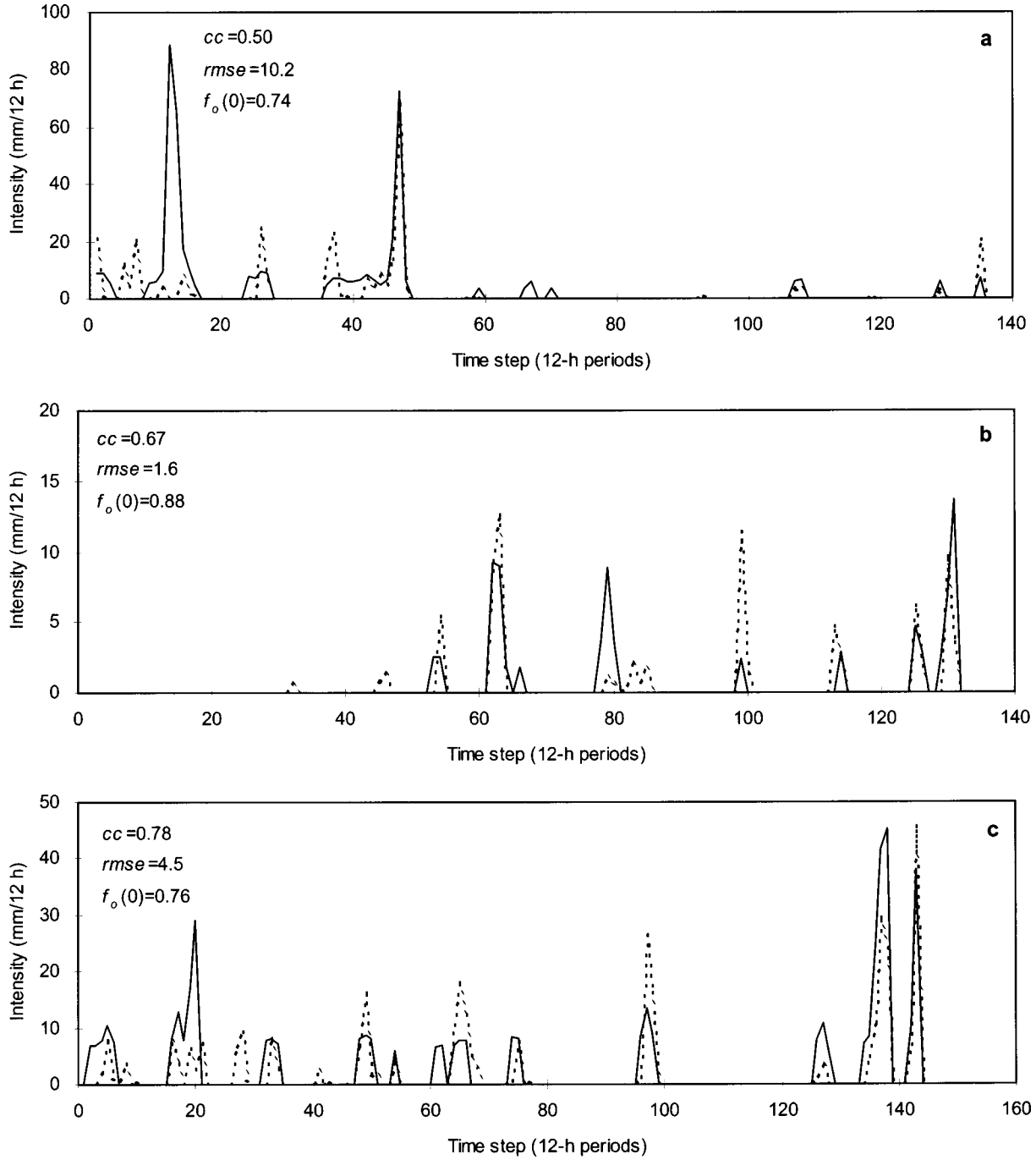

Fig. 9. Neural network output from experiment 2NN/int (solid line) and observed time series (dashed line) for (a) autumn; (b) winter; and (c) spring

autumn. The two sharp peaks in the NN output actually correspond to the passage of two typhoons: Yanni (~September 30, 1998) and Zeb (~October 17). Both were associated with an exceptionally strong southwesterly wind and a large amount of precipitable water, conditions typically associated with heavy rainfall over the Chikugo River basin. For Yanni, however, the rainfall potential never materialized, possibly because the typhoon was at the time rapidly approaching its dissipation stage.

Despite the lower CC for autumn, the RMSE is lower than for summer. It should however be noted that the RMSE is closely related to the frequency of rainfall occurrence. For two seasonal NNs that predict rainfall intensity with the same accuracy, the RMSE will be lower for the season with the higher $f_t(0)$ as zero values are usually accurately predicted by NN1. This is further illustrated for winter, with 85% dry periods and an RMSE of only 1.6 mm. Despite the small number of wet periods on which to train the NNs for this season, both occurrence and intensity are generally fairly well predicted, as shown in Fig. 9(b).

For spring the value of the CC reaches nearly 0.8, which reflects the dominance of midlatitude cyclones in this season and the consequent strong relationship between synoptic-scale features and basin-scale rainfall. An output example is shown in Fig. 9(c). The overall pattern is well reproduced and the peaks reached with reasonable accuracy.

Turning to experiment 1NN/cat, the strong relationship between hr_1 and $f_t(0)$ may be noted first of all. Equivalent to the case of the RMSE in 2NN/int, the more zero values in the data, the higher the hit rate of output series $o1$. The seasonal variation of hr_1 is, however, accompanied by a reverse variation of hr_2, making their sum $\sim$0.9 for all seasons. On average $<$10% of the intensity classes are entirely missed by the NN. The variation of whr_1 shows no obvious seasonal pattern, but the CC in 2NN/int is highest during spring, indicating optimal performance during this season. It should be emphasized that whr_1 is often strongly influenced by a few values in the validation period and is therefore generally more suited for comparing performance of different NNs for the same season than for the same NN during different seasons. Finally, the pronounced underestimation of $f_t(0)$ by 1NN/cat during summer turned out to be an exception. For the other seasons, $f_t(0)$ is in fact better estimated by 1NN/cat than by 2NN/int.

Summary and Conclusions

Neural networks (NNs) were used for prediction of 12-h mean rainfall in the Chikugo River basin (CRb), Kyushu Island, southern Japan, from values of wind speeds at 850 hPa (u_{850},v_{850}) and precipitable water (PW) specified in a 100$\times$100 km grid surrounding the island. Prior to NN application, a seasonal correlation analysis was performed to identify the location of areas in which each predictand significantly influences CRb rain fall. NN inputs were weighted averages of predictor values in each of these areas.

An initial, straightforward NN application highlighted two distinct problems related to intermittency and variability in the target data, namely pronounced underestimations of the fraction of zero rainfall $f_t(0)$ and extreme intensities. To improve performance, two approaches were tested. The first was to use two NNs in series, the first to determine rainfall occurrence, and the second to determine the intensity during rainy periods. This strategy greatly improved the reproduction of $f_t(0)$, but extreme values remained underestimated. The second approach was to categorize rainfall into intensity categories and train an NN to reproduce these rather than the actual intensities. This did not lead to any apparent improvement, but may constitute a viable alternative in cases where a probability-based range of intensities can be accepted as output. A seasonal application showed a somewhat higher performance during spring, in line with the strong relationship between synoptic-scale features and basin-scale rainfall in this season.

Overall we conclude that NNs are a potentially powerful tool for relating spatial rainfall to relevant large-scale free-atmosphere variables. Particularly encouraging results were reached using a binary NN to distinguish between wet and dry 12-h periods, that is, periods with zero or nonzero rainfall intensity, discussed in the section on rainfall occurrence when using NN1. In some forecasting applications, knowledge of whether the following period will be wet or dry is useful information in itself. Accurately forecasting the intensity during wet periods is, however, clearly far more problematic. The developed NN models also performed reasonably well in this respect during winter and spring, but severely underestimated extreme intensities during summer and autumn.

There are a number of reasons for this seasonal variation in performance. One is the temporally and spatially localized nature of the heavy rainfalls in summer (triggered by convection) and autumn (associated with typhoons), which are often only weakly related to the large-scale atmospheric state, as represented by the GPV data used in the present study. The only possible way to overcome this limitation is to use GPV data of higher resolution in space and time. Another reason is the smoothing nature of the NNs' transfer functions. Similar to other deterministic models, this leads to underestimations of variance and extreme values [for example, Zorita and von Storch (1999)]. Since the summer and autumn extremes differ more from the rest of the distribution than do the extremes in other seasons, this effect is most pronounced in summer and autumn. Developing an NN design and application to specifically target downscaling of extreme events is a recommended area for future research.

A more general limitation of NNs is related to the large amounts of data required for complete NN training, that is, training that produces fully converged weights and biases. It is clear that the performance of the NN models used in the present study was constrained by the limited length of the time series available for the present experiments. Rainfall generation is a complex process, and rainfall occurrence can be associated with a wide range of atmospheric states. A limited data set first of all means that these different states may be incompletely represented in the set used for NN training. The limited data further brought about some ''emergency'' solutions, such as using NNs with very few layers and nodes and averaging the output from a number of calibrated NNs. These features, in turn, restricted the models' flexibility and prevented them from learning all rain-producing states and all aspects of their relationship with the basin rainfall. Thus, if more data would have been available, the confidence in (and most likely performance of) the models would have been higher. We, however, wish to emphasize that such data limitations are not uncommon in hydrology and hydrometeorology, and it is therefore of interest to also evaluate the potential of NNs in these situations and share the experience gained.

Finally, it is clear that the performance of the present approach can only be fully assessed by comparing the results with alternative methods for rainfall estimation on the same time and space scales. Of most interest perhaps is comparison with a physically based, numerical prediction model, and currently the present results are being compared with the CRb rainfall predicted by the Japan Meteorological Agency's regional model. Preliminary results indicate a similar performance, and these results will be fully reported elsewhere.

Acknowledgments

The study was funded by the European Commission Science and Technology Fellowship Program in Japan, the Swedish Natural Science Research Council (NFR), the Scandinavia-Japan Sasakawa Foundation, and Ångpanneföreningens Forskningsstiftelse, Sweden. We thank the reviewers for constructive and helpful criticism.

References

Bishop, C. M. (1995). *Neural network for pattern recognition*, Oxford University Press, New York.

Cavazos, T. (1997). ''Downscaling large-scale circulation to local winter rainfall in north-eastern Mexico.'' *Int. J. Climatology*, 17(10), 1069–1082.

Crane, R. G., and Hewitson, B. C. (1998). ''Doubled CO_2 precipitation changes for the Susquehanna basin: Down-scaling from the GENESIS general circulation model.'' *Int. J. Climatology*, 18(1), 65–76.

Daley, R. (1991). *Atmospheric data analysis*, Cambridge University Press, New York.

Giorgi, F., and Mearns, L. O. (1991). ''Approaches to the simulation of regional climate change: A review.'' *Rev. Geophys.*, 29(2), 191–216.
Glahn, H. R. (1985). ''Statistical weather forecasting.'' *Probability, statistics, and decision making in the atmospheric sciences*, A. H. Murphy and R. W. Katz, eds., Westview, Boulder, Colo., 289–335.
Hagan, M. T., and Menhaj, M. (1994). ''Training feedforward networks with the Marquardt algorithm.'' *IEEE Trans. Neural Netw.*, 5(6), 989–993.
Hewitson, B. C., and Crane, R. G. (1992). ''Large-scale atmospheric control on local precipitation in tropical Mexico.'' *Geophys. Res. Lett.*, 19(18), 1835–1838.
Hewitson, B. C., and Crane, R. G. (1996). ''Climate downscaling: Techniques and application.'' *Clim. Res.*, 7(2), 85–95.
Karl, T. R., Wang, W.-C., Schlesinger, M. E., Knight, R. W., and Portman, D. (1990). ''A method of relating general circulation model simulated climate to the observed local climate. I: Seasonal statistics.'' *J. Clim.*, 3(10), 1053–1079.
Kidson, J. W., and Thompson, C. S. (1998). ''A comparison of statistical and model-based downscaling techniques for estimating local climate variations.'' *J. Clim.*, 11(4), 735–753.
Klein, W. H. (1982). ''Statistical weather forecasting on different time scales.'' *Bull. Am. Meteorol. Soc.*, 63(2), 170–177.
Krzysztofowicz, R. (1993). ''A theory of flood warning systems.'' *Water Resour. Res.*, 29(12), 3981–3994.
Kwok, T.-Y., and Yeung, D.-Y. (1997). ''Constructive algorithms for structure learning in feedforward neural networks for regression problems.'' *IEEE Trans. Neural Netw.*, 8(3), 630–645.
Maier, H. R., and Dandy, G. C. (2000). ''Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications.'' *Env. Modelling and Software*, 15(1), 101–123.
Merabtene, T., Jinno, K., Kawamura, A., and Olsson, J. (1998). ''Drought management of water supply systems: A decision support system approach.'' *Memoirs Fac. Eng. Kyushu Univ.*, 58(14), 183–197.
Nguyen, D., and Widrow, B. (1990). ''Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights.'' *IJCNN Int. Joint Conf. on Neural Networks*, 3, 21–26.
Reed, R. (1993). ''Pruning algorithms—A review.'' *IEEE Trans. Neural Netw.*, 4(5), 740–747.
Rogers, L. L., and Dowla, F. U. (1994). ''Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling.'' *Water Resour. Res.*, 30(2), 457–481.
Rosenblatt, F. (1961). *Principles of neurodynamics*, Spartan Press, Washington, D.C.
Smith, M. (1993). *Neural networks for statistical modeling*, Van Nostrand Reinhold, New York.
Uvo, C. B., et al. (2001). ''Statistical atmospheric downscaling for rainfall estimation in Kyushu Island, Japan.'' *Hydro. Earth System Sci.* 5(2), 259–271.
Wigley, T. M. L., Jones, P. D., Briffa, K. R., and Smith, G. (1990). ''Obtaining sub-grid-scale information from coarse-resolution general circulation model output.'' *J. Geophys. Res.*, 95(D2), 1943–1953.
Wilby, R. L., and Wigley, T. M. L. (2000). ''Precipitation predictors for downscaling: Observed and general circulation model relationships.'' *Int. J. Climatology*, 20(6), 641–661.
Wilby, R. L., et al. (1998). ''Statistical downscaling of general circulation model output: A comparison of methods.'' *Water Resour. Res.*, 34(11), 2995–3008.
Wilks, D. S. (1995). *Statistical methods in the atmospheric sciences*, Academic, San Diego.
Zorita, E., and von Storch, H. (1999). ''The analog method as a simple statistical downscaling technique: Comparison with more complicated methods.'' *J. Clim.*, 12(8), 2474–2489.