Spatial Interpolation of Surface Air Temperatures Using Artificial Neural Networks: Evaluating Their Use for Downscaling GCMs

SETH E. SNELL

Department of Geography, University of New Mexico, Albuquerque, New Mexico

SUCHARITA GOPAL AND ROBERT K. KAUFMANN

Department of Geography, Boston University, Boston, Massachusetts

(Manuscript received 2 February 1998, in final form 28 September 1998)

ABSTRACT

Many climate studies need to generate estimates of a climate variable at a given location based on values from other locations. In this research, a new method for the spatial interpolation of daily maximum surface air temperatures is presented. This new method uses artificial neural networks (ANNs) to generate temperature estimates at 11 locations given information from a lattice of surrounding locations. The out-of-sample performance of the ANNs is evaluated relative to a variety of benchmark methods (spatial average, nearest neighbor, and inverse distance methods). The ANN approach is superior both in terms of predictive accuracy and model encompassing. In 94% of case comparisons, the predictive accuracy of the ANN is superior to the benchmark methods. The ANN approach encompasses the benchmark methods in 77% of case comparisons, while benchmark methods encompass the ANN in only 2%. In light of these results, the potential to use this new method of spatial interpolation to downscale GCM temperature simulations is discussed.

1. Introduction

To assess the effect of changes in the concentration of radiatively active gases on climate, scientists depend heavily on general circulation models (GCMs). The spatial resolution of these models varies from 2.5° × 2.5° up to 8° × 10° lat and long. This resolution is too coarse to assess the regional effects of climate change on ecosystems and economies (Giorgi and Mearns 1991; Clark 1985; Wilby and Wigley 1997). To generate regional information, scientists seek ways to downscale output from GCMs.

Several factors complicate the use of statistical techniques to downscale output from GCMs. First, some climatological fields such as temperature and precipitation vary with elevation and slope. Second, many of these same variables are not first- and/or second-order stationary in space. Finally, many climate variables are influenced by other climatological conditions, such as humidity, pressure systems, and frontal boundaries. As a result, traditional methods for spatial interpolation may perform poorly when used to downscale output from GCMs.

In this paper, we evaluate the potential for artificial neural networks (ANNs) to downscale GCM temperature data. This evaluation is described in six sections. Section 2 describes methods used to interpolate climate data, with special focus on those used to downscale GCM output. Section 3 describes how ANNs may be able to alleviate some of the weaknesses associated with existing methodologies. Section 4 describes a methodology that demonstrates the ability of ANNs to interpolate climatological data [maximum daily surface air temperature (*T*max)] to points within a lattice. The interpolation and its performance relative to traditional methods is described in section 5. Based on the potential indicated by the results in section 5, section 6 describes several methodological issues that confront analysts who seek to use ANNs to downscale output from GCMs.

2. Existing interpolation and downscaling methods

Meyers (1994) reviews a variety of methods for spatial interpolation. These techniques range from inverse distance weighted averages to splines, kriging, and, more recently, radial basis functions. Though these methods vary in complexity, all use information present in known sample locations to interpolate values for other locations. The accuracy with which a method can gen-

Corresponding author address: Seth E. Snell, Department of Geography, University of New Mexico, Bandelier West, Room 111, Albuquerque, NM 87131.
E-mail: sethcs@unm.edu

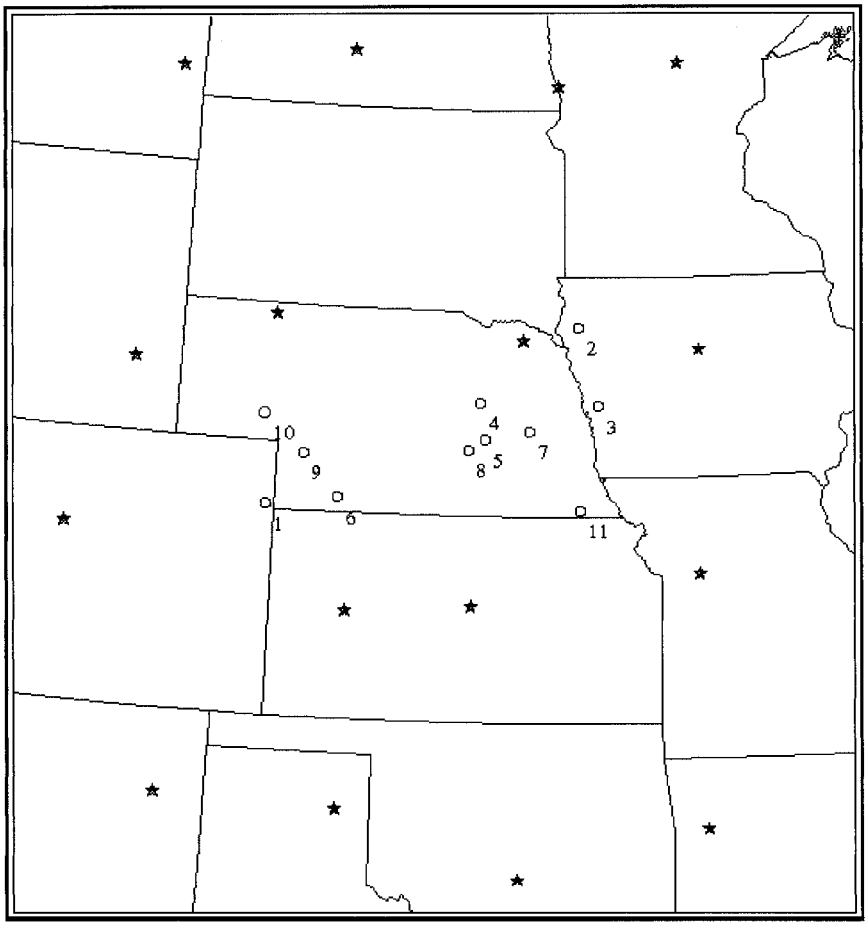

FIG. 1. Map of the study area. Stars depict grid of input NOAA weather stations. Open circles are 11 NOAA weather stations within the grid at which estimates are made.

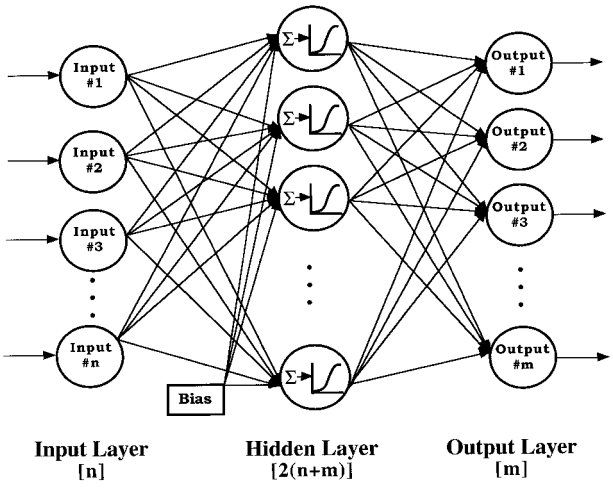

FIG. 2. ANN architecture used to interpolate maximum daily surface air temperature.

erate interior estimates depends on the complexity that underlies the spatial structure of the field. Unfortunately, there are no criteria for analysts to choose among techniques, a priori.

The strengths and weaknesses of existing methods for spatial interpolation are critical to an emerging literature that describes efforts to downscale climate data generated by GCM simulations (Giorgi and Mearns 1991; Hewitson and Crane 1996; Joubert and Hewitson 1997; Wilby and Wigley 1997; Wilby et al. 1998). The techniques used to downscale output from GCMs can be divided between two general categories; process-based and empirical approaches (Hewitson and Crane 1996; Wilby and Wigley 1997). Process-based approaches embed higher-resolution physical models within GCM grids (Dickinson et al. 1989; Giorgi et al. 1994; Russo and Zack 1997). These methods are computationally expensive (Giorgi and Mearns 1991; Hewitson and Crane 1996) and, therefore, may not be practical.

Empirical methods of downscaling generally use transfer functions to relate local conditions to large-scale climate features. The accuracy of these transfer func-

tions is not perfect. Statistical approaches are valid only within the range of the sample data (Katz 1977). The relation between synoptic-scale features and local climatological fields often is highly nonlinear and changes with atmospheric circulation. These difficulties, along with spatial dependencies (e.g., spatial autocorrelation and anisotropy) that are inherent to many climatological fields, create the need for complex mathematical specifications and estimation techniques (Bogardi et al. 1993; Matyasovsky et al. 1994). Finally, it remains uncertain whether these empirical approaches can be applied to the nonstationarity of climate change (Palutikof et al. 1997).

ANNs may address many of the difficulties described above and therefore may be a useful tool to downscale GCM output. Hornik et al. (1989) describe ANNs as universal approximators. ANNs can approximate nonlinear relations and their derivatives without knowing the true nonlinear function; therefore ANNs can make accurate predictions for highly nonlinear systems (Werbos 1974; Rumelhart et al. 1986; Fischer and Gopal 1994; Gopal and Scuderi 1995).

Consistent with these capabilities, investigators have begun to use ANNs to analyze climate data (Gardner and Dorling 1998). The effect of atmospheric circulation on local precipitation is investigated using ANNs (Hewitson and Crane 1994; Cavazos 1997; Crane and Hewitson 1998). McGinnis (1997) uses ANNs to develop transfer functions that predict snowfall from gridscale information generated by GCMs. Zhang and Scofield (1994) use an ANN to estimate convective rainfall and recognize cloud merger from satellite data.

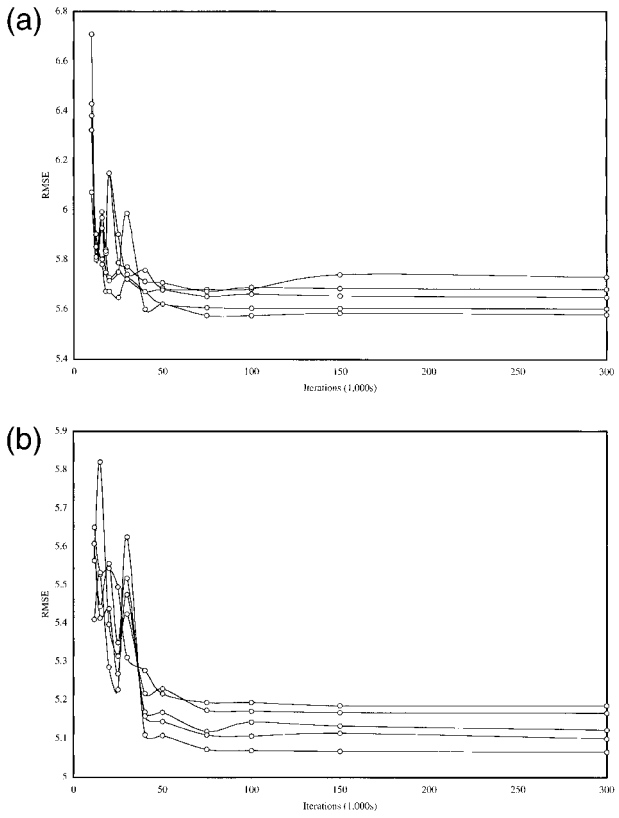

FIG. 3. (a) Rmse on test data at various training intervals for the 4-point ANN for five random divisions of the testing and training datasets. (b) Rmse on test data at various training intervals for the 16-point ANN for five random divisions of the testing and training datasets.

3. Methodology

To investigate the potential for ANNs to downscale GCM output, we train ANNs to predict maximum daily temperature (*T*max) for 11 locations within a four-point grid (4-point ANN) and within a 16-point grid (16-point ANN). The ANNs use *T*max information at the grid locations to predict *T*max at 11 interior locations. The ANNs used in this analysis are multilayer feedforward backpropagation networks.

a. Data sources and preparation

Maximum daily temperature for a 63-yr period from 1931 through 1993 are used as input and output vectors for the ANNs. These data are obtained from National Oceanic and Atmospheric Association (NOAA) ground weather stations (NCDC 1994). The output vector consists of 11 stations in the midcontinental portion of the United States (Fig. 1). The output vector is estimated using two different input vectors: a set of four weather stations that roughly encompass the 11 stations and a set of 16 weather stations (including the first four) that fully encompass the 11 stations (Fig. 1). We consider two input vectors because the optimal spatial extent delineated by the input vector is not obvious.

All *T*max data are checked thoroughly to identify errors. It is well known that errors exist in the TD-3200: Summary of the Day Cooperative Network (Reek et al. 1992). We use the ValHiDD deterministic approach presented by Reek et al. (1992) to identify and correct erroneous values. Missing values for a single day are replaced with the average of *T*max for the day before and the day after the missing day from the same station. This reduces the loss of training data due to missing values. The ANNs used in this analysis can distinguish proficiently between signal and noise (Hecht-Nielson 1990). Therefore, the benefit of including the information that is in the other stations for the day with the missing value outweighs the noise introduced by replacing the missing value with the average of the previous and following day.

b. Artificial neural network architecture

We construct a multilayer feedforward backpropagation ANN [see Hecht-Nielson (1990) and Hertz et al. (1991) for background material on ANNs] to interpolate

TABLE 1. Overall performance measures. RMSE and R^2 for test dataset using the 4-point and 16-point ANNs by station and on average across stations.

	4-point ANN		16-point ANN	
	rmse	R^2	rmse	R^2
Station 1	6.123	0.916	5.889	0.922
Station 2	5.230	0.954	4.131	0.971
Station 3	5.409	0.947	4.125	0.969
Station 4	6.741	0.917	6.036	0.934
Station 5	4.905	0.956	4.664	0.960
Station 6	6.929	0.905	6.697	0.911
Station 7	6.158	0.932	5.340	0.949
Station 8	5.313	0.949	5.110	0.953
Station 9	4.994	0.951	4.761	0.955
Station 10	5.566	0.932	5.064	0.944
Station 11	5.268	0.946	4.488	0.961
Average	5.694	0.937	5.119	0.948
Minimum	4.905	0.905	4.125	0.911
Maximum	6.929	0.956	6.697	0.971

the *T*max field from the input vector to the output vector. All processing elements (PEs) in the ANN have sigmoid activation functions. The neural networks used in this analysis have three layers: an input layer, an output layer, and one hidden layer (Fig. 2). The input layer consists of PEs that represent *T*max at known locations. The interior locations, where estimates are needed, compose the output layer. There are 11 PEs in the output layer. The fully connected hidden layer contains twice as many PEs as the sum of PEs in the other layers. This layer maps the input vector to the output vector. The 4-point ANN has 30 hidden units and the 16-point ANN has 54 hidden units.

The temporal representation embodied in this method is defined by the way data are presented to the ANN. The data can be presented several ways: sequential ordering, random ordering, temporal aggregation of daily values, or with autoregressive components. For this analysis, randomly sorted, daily values are presented to the ANN. Initial testing indicates that random ordering is superior to both sequential ordering and aggregated values. The random ordering of daily values assumes that there is no significant temporal autocorrelation in the pattern of the *T*max field. It is possible that an autoregressive structure could improve performance. Future research efforts may use space–time estimation to include more sophisticated temporal representations (e.g., an autoregressive structure).

c. Artificial neural network training

We train two backpropagation neural networks (a 4-point ANN and a 16-point ANN) to learn the relation between the input and output vectors. During the training phase, connection weights are adjusted to minimize the root-mean-square error between the desired output and the estimate from the ANN.

The ability to represent highly nonlinear relationships highlights an important caution for the training of neural networks: they can be overtrained. Therefore, it is critical to identify the appropriate training interval. If the training interval is too short, the ANN cannot learn all patterns and prediction accuracy is low. If the training interval is too long, neural networks begin to fit the noise present in the training data. When an overtrained ANN is presented with data that it has not yet seen (i.e., data with the same relationship as the training data, but with new and different noise), its performance is lower than if it had only learned the actual relationship during training.

We identify the proper training interval by comparing the performance of the ANN on data that it has not yet seen (i.e., test data) after different training intervals. Network performance should increase as the training interval increases until the ANN becomes overtrained, at which point performance will decrease. Because the overtrained ANN learns the noise of a particular training dataset, we assess the performance using different training and testing datasets. We create the training and testing datasets by randomly separating daily observations. For this research, 80% of the daily observations are used for training (14 624 days) and the remaining 20% are used for testing (2924 days).

For five random separations, the 4-point and 16-point ANN are trained at intervals from 10 000 to 300 000 iterations. At the end of each training level, the performance of the ANN is assessed using the unseen test data. We evaluate the performance using the root-mean-square error (rmse) of the estimate (Figs. 3a and 3b). In the early stages of training, the ANNs perform poorly, as indicated by rmses that are relatively high and unstable. These results imply that the ANNs have not yet learned the relation between input and output vectors.

The rmses reach a minimum at 50 000 iterations for the 4-point ANN and 75 000 iterations for the 16-point ANN. Beyond this point, increasing the number of iterations does not change the rmse significantly. This stability indicates that longer training intervals do not increase performance significantly. Additionally, there is no sign of overtraining for training intervals up to 300 000 iterations. Based on these results, we choose 50 000 iterations for the 4-point ANN and 75 000 iterations for the 16-point ANN as the appropriate training interval.

d. Evaluating ANN performance and comparison with alternative methodologies

The interpolations generated by ANNs are evaluated by their root-mean-square error and the fraction of total variation they can account for (R^2). We also check for bias in the interpolation because the backpropagation algorithm tends to underpredict extreme values (Gopal and Scuderi 1995; McGinnis 1997). We test for this bias by estimating Eq. (1),

TABLE 2. Test results for systematic bias. Regression coefficients from Eq. (1) for each station for the predicted output from the 4-point and 16-point ANNs. Bracketed values are p values for t statistics testing significant differences for α and β from 0 and 1, respectively. Values significant at $p < 0.05$ are in boldface.

	4-point ANN		16-point ANN	
	α	β	α	β
Station 1	0.464 [0.243]	0.992 [0.156]	0.693 [0.068]	0.989 **[0.043]**
Station 2	−1.380 **[0.000]**	1.007 [0.100]	−0.163 [0.434]	1.004 [0.261]
Station 3	−0.514 [0.076]	0.993 [0.123]	0.273 [0.213]	0.999 [0.819]
Station 4	−0.504 [0.166]	0.988 **[0.029]**	0.211 [0.518]	0.997 [0.564]
Station 5	−1.735 **[0.000]**	1.003 [0.459]	−0.148 [0.566]	1.001 [0.811]
Station 6	0.091 [0.828]	0.991 [0.151]	−0.063 [0.876]	0.998 [0.750]
Station 7	−0.343 [0.296]	0.986 **[0.004]**	0.554 [0.052]	0.993 [0.110]
Station 8	−0.805 **[0.005]**	0.999 [0.746]	0.351 [0.205]	0.996 [0.307]
Station 9	−0.299 [0.313]	1.001 [0.726]	0.242 [0.388]	0.996 [0.329]
Station 10	−0.080 [0.814]	0.994 [0.193]	0.477 [0.119]	0.992 [0.087]
Station 11	−1.069 **[0.001]**	1.002 [0.694]	−0.137 [0.602]	1.005 [0.144]

$$T\max_{it} = \alpha + \beta\hat{T}\max_{it} + \mu_{ij}, \quad (1)$$

in which $T\max_{it}$ is the historical value for maximum temperature at station i at time t, $\hat{T}\max_{it}$ is the value for maximum temperature generated by the ANN for station i at time t, μ_{it} is a normally distributed random error, and α and β are regression coefficients. If the values for Tmax generated by the neural networks are unbiased, the intercept (α) will equal zero and the slope coefficient (β) will equal 1.0. We test these null hypotheses using a t-test. Rejecting either of the null hypotheses indicates that the interpolation generated by the ANN is biased in a way that causes it to systematically underpredict or overpredict the actual values.

We also evaluate the performance of the ANN by comparing its interpolation to those generated by traditional techniques. To do so, we interpolate Tmax for the same 11 stations using four techniques: spatial average (AV), nearest neighbor (NN), inverse distance weighted average (IDWA), and inverse distance squared weighted average (ID^{2}WA). Four additional interpolations are generated by including topographical effects. To do so, an environmental lapse rate of -6.5°C km^{-1} is applied to the values for Tmax before the AV, NN, IDWA, and ID^{2}WA are calculated. For each method, the lattice consists of the four stations that surround the predicted station. Only four of the 16 surrounding stations are used because the performance of the traditional techniques declines as the lattice is expanded from the nearest four stations to the nearest 16. The distances used to calculate the nearest neighbor and inverse distance predictions are calculated using distance functions in a geographic information system (ESRI 1998).

The values generated by these eight techniques are compared to the values generated by the 4-point and 16-point ANN for the test data. These test data consist of 2924 days that are selected randomly from the 63-yr period from 1931 to 1993. These days are not presented to the ANNs during training. As such, the test data represent an ''out-of-sample'' forecast for this method of spatial interpolation.

The values of Tmax for the 11 interior stations generated by the 16-point ANN are compared to the values generated by the 4-point ANN and those generated by each of the eight traditional methodologies with regards to their predictive accuracy and the information content of the forecasts. Predictive accuracy evaluates the accuracy of one interpolation relative to another. One interpolation is said to be more accurate if its forecast

TABLE 3. Predictive accuracy tests results. The t statistics for the sign test (S_{2a}) and Wilcoxon's signed-rank test (S_{3a}) for the 16-point ANN compared to nine alternative methods ''e'' subscripts denote methods with elevation incorporation). Boldface values are significant at $p < 0.05$. Negative values indicate a superior forecast by the 16-point ANN.

	AV		AV$_e$		NN		NN$_e$	
	S_{2a}	S_{3a}	S_{2a}	S_{3a}	S_{2a}	S_{3a}	S_{2a}	S_{3a}
Station 1	**−25.11**	**−27.47**	**−5.88**	**−7.11**	**−10.32**	**−14.06**	**−10.87**	**−14.42**
Station 2	**−18.75**	**−23.15**	**−19.01**	**−23.55**	**−9.47**	**−11.57**	**−11.39**	**−14.48**
Station 3	**−8.88**	**−11.22**	**−13.91**	**−15.81**	**−15.42**	**−19.82**	**−16.50**	**−21.50**
Station 4	**−9.99**	**−13.66**	**−14.02**	**−18.55**	**−11.21**	**−13.62**	**−15.16**	**−18.24**
Station 5	**−8.51**	**−9.76**	**−18.23**	**−21.28**	**−10.80**	**−14.20**	**−13.39**	**−18.09**
Station 6	**−3.74**	**−5.46**	**−6.77**	**−8.23**	**−8.40**	**−11.20**	**−9.69**	**−13.03**
Station 7	**−2.92**	**−4.90**	−1.92	**−3.99**	**−13.35**	**−16.61**	**−14.09**	**−17.62**
Station 8	**−7.40**	**−8.63**	**−15.98**	**−17.08**	**−10.87**	**−13.39**	**−14.68**	**−17.83**
Station 9	**−22.97**	**−27.45**	**−12.28**	**−15.30**	**−19.20**	**−25.29**	**−15.79**	**−20.83**
Station 10	**−15.90**	**−17.08**	**−12.80**	**−14.89**	**−13.02**	**−17.37**	**−10.98**	**−15.01**
Station 11	**−7.88**	**−10.08**	**−7.25**	**−8.99**	**−13.54**	**−17.82**	**−10.28**	**−13.94**

error is statistically smaller than the forecast error of an alternative methodology.

The predictive accuracy of the forecasts is evaluated using techniques developed by Diebold and Mariano (1995). We use these techniques because the interpolations violate many of the assumptions that underlie previous tests for predictive accuracy. For example, previous tests assume that the forecast errors are contemporaneously uncorrelated. A simple regression indicates that the forecast errors generated by the 16-point ANN and the 4-point ANN for station 1 are correlated in a highly significant manner [$F(1, 2922) = 15\,128$; $p < 0.000$].

We use two tests to evaluate predictive accuracy; a sign test, which is given by Eq. (2) and Wilcoxon's signed-rank test, which is given by Eq. (3):

$$S_{2a} = \frac{\sum_{t=1}^{T} I_+(d_t) - 0.5T}{(0.25T)^{1/2}}, \tag{2}$$

$$S_{3a} = \frac{\sum_{t=1}^{T} I_+(d_t)\ \text{rank}(|d_t|) - \dfrac{T(T+1)}{4}}{\left[\dfrac{T(T+1)(2T+1)}{24}\right]^{1/2}}, \tag{3}$$

where

$$d_t = |T\text{max}_{it} - \hat{T}\text{max}_{it}| - |T\text{max}_{it} - \hat{T}\text{max}_{itj}|$$

$$I_+(d_t) = \begin{cases} 1 & \text{if } d_t > 0 \\ 0 & \text{otherwise.} \end{cases}$$

In Eqs. (2) and (3), $\hat{T}\text{max}_{it}$ is the value for maximum temperature generated by the 16-point ANN for station i at time t, $\hat{T}\text{max}_{itj}$ is the value for maximum temperature for station i at time t generated by technique j (there are nine techniques including the 4-point ANN), and T is the number of observations being compared (2924).

These two tests postulate a null hypothesis that there is no difference in the predictive accuracy of the techniques. This null is evaluated by comparing the test statistic (S_{2a} or S_{3a}) against a t distribution. Values that exceed the critical threshold indicate that one forecast is superior. The superior forecast is indicated by the sign on the test statistic, which is determined by the order in which the error terms are subtracted from each other (d_t). The order of subtraction used to calculate (d_t) in this research implies that the test statistic will be negative if the forecast error of the 16-point ANN is smaller than the forecast error generated by the alternative technique.

We compare the information content of the interpolations using the notion of encompassing (Nelson 1972). Encompassing tests evaluate whether information in the interpolation generated by one technique is contained wholly in the interpolation generated by another technique. The interpolation generated by the 16-point ANN may be more accurate than the interpolation generated by the inverse distance squared weighted average, but the inverse distance squared weighted average may have information about Tmax that is not in the interpolation generated by the 16-point ANN. In this case, the 16-point ANN does not encompass the inverse distance squared weighted average. As such, encompassing tests are different from tests of predictive accuracy.

We test whether the 16-point ANN encompasses the interpolations generated by the other methodologies (and vice versa) by estimating Eqs. (4) and (5),

$$T\text{max}_{it} - \hat{T}\text{max}_{it} = \lambda \hat{T}\text{max}_{itj} + \mu_{ij} \tag{4}$$

$$T\text{max}_{it} - \hat{T}\text{max}_{itj} = \pi \hat{T}\text{max}_{it} + \mu_{ij}, \tag{5}$$

in which λ and π are regression coefficients, and all other symbols are defined as used previously. The null hypothesis postulates that the interpolation generated by the technique on the right-hand side of Eq. (4) or (5) does not have information about the forecast error for the interpolation generated by the technique on the left-hand side of Eq. (4) or (5) (i.e., λ or π are zero). This null hypothesis is evaluated with a t statistic for the regression coefficient λ or π. If $\pi \neq 0$ and $\lambda = 0$, the 16-point ANN is said to encompass the alternative tech-

TABLE 3. (*Continued*)

IDWA		IDWA$_e$		ID^{2}WA		ID^{2}WA$_e$		4PTANN	
S_{2a}	S_{3a}	S_{2a}	S_{3a}	S_{2a}	S_{3a}	S_{2a}	S_{3a}	S_{2a}	S_{3a}
−22.01	**−23.72**	**−3.11**	**−4.41**	**−18.64**	**−19.19**	−1.66	**−2.14**	**−2.63**	**−2.83**
−6.58	**−7.88**	**−7.73**	**−9.99**	**−3.29**	**−3.98**	**−5.62**	**−6.69**	**−9.32**	**−12.96**
−0.92	−1.02	**−6.36**	**−6.74**	1.81	**2.04**	−0.74	−0.94	**−12.43**	**−16.06**
−4.77	**−7.49**	**−4.77**	**−8.34**	**−4.85**	**−6.98**	**−6.77**	**−8.16**	**−2.59**	**−6.11**
−3.66	**−2.57**	**−11.17**	**−11.45**	1.78	**3.42**	−0.07	0.23	**−4.77**	**−4.80**
−1.96	−1.68	**−3.29**	**−3.46**	0.55	−0.34	0.52	−0.48	−1.18	−1.80
−1.18	**−2.55**	**−2.59**	**−2.98**	**−4.36**	**−4.97**	**−5.25**	**−5.93**	**−5.70**	**−8.24**
−4.88	**−5.49**	**−10.58**	**−11.10**	**−2.29**	**−2.57**	**−4.62**	**−4.50**	**−3.18**	**−2.75**
−18.90	**−22.84**	**−8.51**	**−11.26**	**−15.65**	**−18.60**	**−6.47**	**−7.90**	**−2.70**	**−3.97**
−10.73	**−12.40**	**−7.14**	**−8.56**	**−7.69**	**−10.03**	**−3.18**	**−4.44**	**−3.77**	**−5.69**
−4.07	**−5.75**	**−3.88**	**−4.95**	−1.81	**−2.89**	−1.41	**−2.57**	**−8.95**	**−11.57**

TABLE 4. Model encompassing test results. The t statistics for λ and π in Eqs. (4) and (5) testing significant differences from zero for nine alternative methods compared to the 16-point ANN. Boldface values are significant at $p < 0.05$. Encompassing is found when λ or π (but not both) is significantly different from 0. The 16-point ANN encompasses the alternative method when $\pi \neq 0$ and $\lambda = 0$. The alternative method encompasses the 16-point ANN when $\pi = 0$ and $\lambda \neq 0$.

	AV		AV_e		NN		NN_e	
	$\lambda \neq 0$	$\pi \neq 0$	$\lambda \neq 0$	$\pi \neq 0$	$\lambda \neq 0$	$\pi \neq 0$	$\lambda \neq 0$	$\pi \neq 0$
Station 1	−0.65	**52.85**	−0.63	**7.98**	−0.54	**−7.63**	−0.55	**9.81**
Station 2	1.03	**−25.38**	1.03	**−26.16**	1.74	**−14.07**	1.73	**−19.82**
Station 3	**2.82**	**−16.75**	**2.83**	**−25.47**	**2.75**	0.34	**2.76**	**−10.89**
Station 4	0.14	**−18.92**	0.15	**−29.57**	0.30	0.57	0.30	**13.42**
Station 5	−0.71	**−13.27**	−0.72	**−34.99**	−0.23	**12.21**	−0.21	**21.45**
Station 6	−1.47	**6.10**	−1.47	**12.33**	−1.29	**−15.46**	−1.29	**−21.06**
Station 7	0.82	**−10.64**	0.81	1.82	0.70	**2.72**	0.69	**8.77**
Station 8	0.66	**−17.11**	0.67	**−32.73**	0.93	**7.16**	0.93	**19.54**
Station 9	−0.24	**50.29**	−0.24	**−12.62**	−0.26	**33.09**	−0.26	**20.60**
Station 10	−0.50	**29.36**	−0.48	**−26.65**	−0.19	**18.46**	−0.19	**10.10**
Station 11	**3.33**	**26.18**	**3.33**	**24.66**	**3.83**	**−21.47**	**3.85**	**−10.76**

nique j. If $\pi = 0$ and $\lambda \neq 0$, the alternative technique j is said to encompass the 16-point ANN. Other combinations indicate that neither method fully encompasses the other.

4. Results

The ANNs account for much of the temporal and spatial variation in Tmax at the 11 interior points. Across all stations and days, the 16-point ANN accounts for 94.8% of the variation in Tmax (Table 1). It performs most poorly for station 6, where it accounts for 91.1% of the variation over time, and performs best for station 2, where it accounts for 97.1% of the variation over time. Values for the rmses are consistent with the performance indicated by the values for R^2. Across all stations and days, the 4-point ANN accounts for 93.7% of the variation in Tmax (Table 1). As with the 16-point ANN, the 4-point ANN performs most poorly for station 6 ($R^2 = 0.905$). The 4-point ANN accounts for the most variation at station 5 ($R^2 = 0.956$). Again, as with the 16-point ANN, the rmses generally are consistent with the values for R^2.

The results from Eq. (1) indicate the interpolations generated by the 4- and 16-point ANN differ in the degree of systematic underprediction/overprediction (Table 2). There is little evidence for this sort of bias in the interpolation generated by the 16-point ANN. There is only one station (station 1) where the null hypothesis is rejected. Furthermore, the significance level is just under the 0.05 threshold ($p < 0.043$). On the other hand, the 4-point ANN shows clear signs of bias. The intercept (α) is significantly less than zero for four stations (stations 2, 5, 8, and 11). The slope coefficient (β) is significantly less than one at two stations (stations 4 and 7). Together with the values for rmse and R^2, these results imply that the 16-point ANN is superior to the 4-point ANN.

This conclusion is confirmed clearly by the tests of predictive accuracy [Eqs. (2) and (3)]. The predictive accuracy of the 16-point ANN is statistically superior to the predictive accuracy of the 4-point ANN for all but one station (Table 3). The signs on the test statistics for station 6 are negative, which indicates that the forecast errors for the interpolation generated by the 16-point ANN generally are smaller than the forecast errors for the interpolation generated by the 4-point ANN, but these differences are not statistically significant.

Similarly, the predictive accuracy of the 16-point ANN generally is superior to that of the eight traditional methods (Table 3). The statistical significance of the test statistics generally is quite large ($p < 0.0001$), which implies that the 16-point ANN is clearly superior. The sign test (S_{2a}) indicates that the forecast generated by the 16-point ANN is significantly better than all traditional approaches for 5 of the 11 stations. For the remaining stations, the sign test produces mixed results but does not indicate that any of the forecasts by traditional approaches are significantly better than the 16-point ANN. Wilcoxon's signed-rank test (S_{3a}) indicates that the forecast generated by the 16-point ANN is significantly better than all traditional methods for 8 of the 11 stations. The interpolation generated by the inverse distance squared weighted average generally performs better than the other traditional methods. Indeed, the interpolation generated by this method is superior to the 16-point ANN for two stations, stations 3 and 5. But even this result is mixed compared to the results of the sign test. Only Wilcoxon's signed-rank test rejects the null hypothesis. Interestingly, the inverse distance squared weighted average that incorporates the temperature change with elevation does not fare as well against the 16-point ANN as the same method without the elevation effect.

Two-way tests of encompassing evaluate the performance of the 16-point ANN relative to the eight traditional methods and the 4-point ANN for all 11 stations. Of these 99 evaluations, a majority of the comparisons (75) indicate that the 16-point ANN encompasses the other methods of spatial interpolation (Table

TABLE 4. (*Continued*)

IDWA		IDWA$_e$		ID^{2}WA		ID^{2}WA$_e$		4PTANN	
$\lambda \neq 0$	$\pi \neq 0$	$\lambda \neq 0$	$\pi \neq 0$	$\lambda \neq 0$	$\pi \neq 0$	$\lambda \neq 0$	$\pi \neq 0$	$\lambda \neq 0$	$\pi \neq 0$
−0.64	**46.95**	−0.63	**9.11**	−0.63	**39.37**	−0.62	**10.11**	−0.74	−0.50
1.42	**−18.85**	1.42	**−21.84**	1.64	**−13.90**	1.63	**−18.73**	1.23	**−7.23**
2.89	**−6.62**	**2.90**	**−16.87**	**2.94**	1.81	**2.94**	**−9.22**	**2.60**	**−7.72**
0.20	**−13.72**	0.20	**−16.42**	0.25	**−6.75**	0.25	−0.71	0.06	**−9.46**
−0.59	**−11.79**	−0.60	**−26.17**	−0.47	**−7.25**	−0.48	**−12.79**	−0.61	**−16.30**
−1.43	0.50	−1.43	**4.29**	−1.39	**−4.89**	−1.39	**−3.56**	−1.50	**−3.66**
0.78	**−7.82**	0.77	**3.36**	0.75	**−4.19**	0.74	**5.26**	0.61	**−10.35**
0.72	**−17.50**	0.73	**−26.84**	0.78	**−16.26**	0.78	**−19.26**	0.62	**−8.43**
−0.23	**42.99**	−0.23	**−8.19**	−0.23	**36.77**	−0.23	**−3.42**	−0.27	−1.39
−0.38	**26.07**	−0.38	**−20.41**	−0.29	**22.94**	−0.29	**−12.52**	−0.66	**−4.23**
3.37	**21.30**	**3.37**	**20.19**	**3.42**	**16.43**	**3.42**	**15.88**	**3.11**	**−8.22**

4). The notable exceptions are stations 3 and 11, the two easternmost stations. For these two stations, the predictions generated by all of the traditional methods (and the 4-point ANN) contain information about the forecast error for the 16-point ANN. However, only in two instances (NN, station 3; ID^{2}WA, station 3) does the traditional method encompass the 16-point ANN. In all other instances where traditional methods have information about the 16-point ANN forecast error, the reverse also is true: the 16-point ANN has information about the traditional techniques' forecast error. These results indicate that neither approach fully encompasses the other. The test generates the same result in six cases where neither the 16-point ANN nor the traditional method contain information about the forecast error of the other (Table 4).

5. Discussion

Of the techniques analyzed here, the interpolation generated by the 16-point ANN appears superior. The interpolation generated by the 16-point ANN is unbiased, its predictive accuracy is generally superior to most other techniques, and the interpolation generally encompasses the interpolations generated by the other techniques. This performance is consistent with the ability of ANNs to represent highly nonlinear systems of arbitrary complexity and to learn inherent spatial dependencies such as autocorrelation and anisotropy from training data.

This performance implies that ANNs may be a useful tool to downscale surface air temperature from GCM simulations. Despite this promise, there are several methodological issues that complicate their implementation. The most important issue concerns the construction of the input vector from the historical data in a way that is consistent with the data that will be generated by the GCM. Skelly and Henderson-Sellers (1996) point out that ''the spatial nature of GCM data is unclear.'' From the standpoint of spatial continuity, they argue that GCM simulations of many climate surfaces can be considered as either point or areal data. On the one hand, finite grid GCMs numerically integrate the partial differential equations governing the atmosphere (and ocean) on a discrete grid across the globe (Peixoto and Oort 1992). As such, the temperature data generated by the GCM can be viewed as temperature at a point location in space. Consistent with this perspective, the input vector used to train the ANN can be assembled from historical data that are obtained from a weather station near the GCM grid point.

On the other hand, land surface conditions (e.g., elevation, vegetation) represent an average of these characteristics for the cell surrounding the grid point. In addition, GCMs use subgrid-scale parameterizations to represent atmospheric dynamics that cannot be resolved by the model grid (e.g., convective storms, turbulence). These subgrid-scale parameterizations are compiled as an average over the cell that surrounds the grid point. Consistent with these specifications, the input vector used to train the ANN can be assembled from an average of historical data obtained from two or more weather stations within the cell that surrounds the grid point.

The importance of using an average of historical data obtained from two or more weather stations depends on the spatial heterogeneity of the land surface characteristics in the cell. If the characteristics of the land surface in the cell are highly heterogeneous, an input vector made up of an average from two or more stations from within the cell is more appropriate than a vector that consists of values from a single station near the grid point because there is no reason to believe a priori that the characteristics of the land surface in the center of the cell are consistent with the average for that cell. Conversely, the land surface in the center of the cell probably is consistent with the average for that cell if the land surface characteristics for that cell are relatively homogeneous.

Even in cases where an average for two or more stations is appropriate, methodological difficulties may

force analysts to train ANNs from input vectors that use historical data from a single weather station. The period for which historical data are available varies significantly among stations. This implies that the station with the shortest record will determine the amount of data that can be used to compute an average for a cell. Allowing stations to enter and leave the sample used to calculate the average could introduce a bias that may contaminate the input vector. Ironically, this contamination would be greatest in grid cells where the land surface is highly heterogeneous because the weather data would be highly heterogeneous. Similarly, the density of weather stations included in the Summary of the Day Cooperative Network varies considerably among regions. As a result, the average for some grid cells may contain data from more stations than others. Ironically, grid cells with the greatest spatial heterogeneity (e.g., mountainous areas) may have the fewest weather stations. Together, these limits on data imply that the theoretical shortcomings associated with using a single weather station near the center of the grid cell to assemble the input vector may be offset to some unknown degree by errors and loss of information introduced by assembling an average.

The methodology used to train the ANNs can be improved in several ways. Most importantly, the ANN needs to represent the temporal structure of the data in a more realistic fashion. Such efforts can proceed in two directions that are not mutually exclusive. An autoregressive temporal structure can be integrated with the spatial structure. Additional temporal information may increase predictive accuracy. An autoregressive structure increases the number of input nodes because daily values for *T*max at the output stations depend on current and previous values for *T*max. The appropriate number of lagged values can be identified based on output performance and is guided by the literature on forecasting (Weigend and Gershenfeld 1994). Weigend and Gershenfeld (1994) note that ANNs are superior on predicting time series if the underlying data are characterized by nonlinearities and complicated functions.

Including daily atmospheric circulation patterns (CPs) also may improve the interpolation. Categorical CP information can delineate periods of spatial stability in the underlying covariance structure of a climatological field (Bogardi et al. 1993; Matyasovsky et al. 1994). Including this information could allow ANN connection weights more flexibility in representing daily patterns.

Additionally, we can envision hybrid ANNs that would maintain the basic spatial and temporal structure discussed so far, but would use input from several climatological fields to predict a single output field. For example, information on temperature and precipitation could be used to interpolate precipitation data. The mathematical flexibility of ANNs offers the opportunity to make use of this information. Similarly, the heights of typical pressure surfaces (500 mb, 750 mb, and sea level pressure) have important information about the spatial and temporal structure of temperature and precipitation fields. Crane and Hewitson (1998) use geopotential heights and specific humidity as inputs to an ANN to build transfer functions that downscale climate data generated by the GENESIS GCM for the Susquehanna Basin.

Finally, several newer ANNs could be used in this application with relatively small changes to the spatial and temporal specifications. These newer ANNs could perform significantly better in terms of both processing time and prediction accuracy. In addition, the number of hidden layers could increase relative to the existing ANNs, which may improve performance. ANNs with multiple hidden layers have a greater likelihood for overtraining. There is no evidence for overtraining in this analysis; therefore these ANNs offer the potential for improved performance without loss of generality.

6. Conclusions

The impacts of climate change depend in large part on local effects. To get this information, users have developed several techniques to downscale climate data generated by GCMs. The results of this analysis indicate that ANNs can be used to interpolate temperature data from a grid structure to interior points with a high degree of accuracy. Before this methodology can be used to downscale climate data from GCMs, analysts must weigh the costs and benefits of two competing methods for assembling the training data. In addition, the interpolation may be improved through more complex temporal representations and/or through the addition of other important input variables. These modifications can be implemented with relatively little change in the overall ANN architecture.

Acknowledgments. This research was supported by grants from the National Science Foundation (SBR-9523600 and SBR-9513889) and National Aeronautics and Space Administration (LCLUC-NAG5-6214). We thank two anonymous reviewers for their comments on the manuscript. The authors are solely responsible for any errors that remain.

REFERENCES

Bogardi, I., I. Matyasovsky, A. Bardossy, and L. Duckstein, 1993: Application of a space–time stochastic model for daily precipitation using atmospheric circulation patterns. *J. Geophys. Res.*, **98** (D9), 16 653–16 667.

Cavazos, T., 1997: Downscaling large-scale circulation to local winter rainfall in north-eastern Mexico. *Int. J. Climatol.,* **17,** 1069–1082.

Clark, W., 1985: Scales of climate change. *Climatic Change*, **7,** 5–27.

Crane, R. G., and B. Hewitson, 1998: Doubled CO_2 precipitation changes for the susquehanna basin: Down-scaling from the genesis general circulation model. *Int. J. Climatol.,* **18,** 65–76.

Dickinson, R., R. Errico, F. Giorgi, and G. Bates, 1989: A regional climate model for the western United States. *Climatic Change*, **15,** 383–422.

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.,* **13,** 253–263.
ESRI, 1998: ArcInfo. Version 7.3.2. ESRI, Inc.
Fischer, M. M., and S. Gopal, 1994: Artificial neural networks: A new approach to modeling interregional telecommunication flows. *J. Reg. Sci.,* **34,** 503–527.
Gardner, M. W., and S. R. Dorling, 1998: Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmos. Environ.,* **32** (14/15), 2627–2636.
Giorgi, F., and L. O. Mearns, 1991: Approaches to the simulation of regional climate change: A review. *Rev. Geophys.,* **29,** 191–216.
——, C. S. Brodeur, and G. Bates, 1994: Regional climate change scenarios over the United States produced with a nested regional climate model. *J. Climate,* **7,** 375–399.
Gopal, S., and L. Scuderi, 1995: Application of artificial neural networks in climatology: A case study of sunspot prediction and solar climate trends. *Geogr. Anal.,* **27** (1), 42–59.
Hecht-Nielsen, R., 1990: *Neurocomputing.* Addison-Wesley Publishing Company, 433 pp.
Hertz, J., A. Krogh, and R. Palmer, 1991: *Introduction to the Theory of Neurocomputing.* Addison-Wesley Publishing Company, 327 pp.
Hewitson, B., and R. G. Crane, 1994: Precipitation controls in southern Mexico. *Neural Nets: Applications in Geography,* B. Hewitson and R. Crane, Eds., Kluwer Academic Publishers, 121–143.
——, and ——, 1996: Climate downscaling: Techniques and application. *Climate Res.,* **7,** 85–95.
Hornik, K., M. Stinchcombe, and H. White, 1989: Multilayer feedforward networks are universal approximators. *Neural Networks,* **2,** 359–366.
Joubert, A., and B. Hewitson, 1997: Simulating present and future climates of southern Africa using general circulation models. *Prog. Phys. Geogr.,* **21,** 51–76.
Katz, R., 1977: Assessing the impact of climatic change on food production. *Climatic Change,* **1,** 85–96.
Matyasovsky, I., I. Bogardi, A. Bardossy, and L. Duckstein, 1994: Local temperature estimation under climatic change. *Theor. Appl. Climatol.,* **50** (1–2), 1–13.
McGinnis, D. L., 1997: Estimating climate-change impacts on Colorado plateau snowpack using downscaling methods. *Prof. Geogr.,* **49** (1), 117–125.
Meyers, D. E., 1994: Spatial interpolation: An overview. *Geoderma,* **62,** 17–28.
NCDC, 1994: Summary of the Day Cooperative Network Weather Data. TD-3200, CD-ROM, Earthinfo, Inc.
Nelson, C. R., 1972: The prediction performance of the F.R.B.-M.I.T.-Penn model of the US economy. *Amer. Econ. Rev.,* **62,** 902–917.
Peixoto, J. P., and A. H. Oort, 1992: *Physics of Climate.* American Institute of Physics, 520 pp.
Reek, T., S. Doty, and T. Owen, 1992: A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network. *Bull. Amer. Meteor. Soc.,* **73,** 753–762.
Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1986: Learning representations by back-propagating errors. *Nature,* **323,** 533–536.
Russo, J., and J. Zack, 1997: Downscaling GCM output with a mesoscale model. *J. Environ. Manage.,* **49,** 19–29.
Skelly, W., and A. Henderson-Sellers, 1996: Grid box or grid point: What type of data do GCMs deliver to climate impacts researchers? *Int. J. Climatol.,* **16,** 1079–1086.
Weigend, A. S., and N. A. Gershenfeld, 1994: *Time Series Prediction: Forecasting the Future and Understanding the Past, Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis.* Addison-Wesley Publishing Company, 421 pp.
Werbos, P., 1974: Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. dissertation, Dept. of Applied Mathematics, Harvard University, 453 pp. [Available from Harvard University Archives, Pusey Library, Harvard University, Cambridge, MA 01238.]
Wilby, R. L., and T. Wigley, 1997: Downscaling general circulation model output: A review of methods and limitations. *Prog. Phys. Geogr.,* **21,** 530–548.
Wilby, R. L., T. Wigley, D. Conway, P. Jones, B. Hewitson, J. Main, and D. Wilks, 1998: Statistical downscaling of general circulation model output: A comparison of methods. *Water Resour. Res.,* **34** (11), 2995–3008.
Zhang, M., and R. A. Scofield, 1994: Artificial neural network techniques for estimating heavy convective rainfall and recognizing cloud mergers from satellite data. *Int. J. Remote Sens.,* **15** (6), 3241–3261.