INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **28**: 1357–1368 (2008)
Published online 11 October 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1637

Predicting realizations of daily weather data for climate forecasts using the non-parametric nearest-neighbour re-sampling technique

Mohammad Bannayan* and Gerrit Hoogenboom
Department of Biological and Agricultural Engineering, University of Georgia, Griffin, GA 30223, USA

ABSTRACT: Weather is one of the primary driving variables that prominently impacts agricultural production and associated disciplines, such as resource management. Lack of daily weather data for many locations along with many prognosis requirements for weather for various applications has resulted in continuous efforts to determine the best possible approach for weather sequence prediction. The goal of this study was to verify the k-nearest neighbours (k-NN) approach for the prediction of daily weather sequences. This method can be employed on the assumption that the weather during the target year is analogous to the weather recorded in the past. We used the nearest-neighbour re-sampling method for the simultaneous prediction of daily radiation, maximum and minimum temperature, and precipitation for multiple locations. A vector of weather variables, including precipitation, radiation, maximum and minimum temperature, on day $(t + 1)$ is re-sampled from historical data by conditioning on the vector of the same variables for the preceding day (t). Observed historical weather data for ten different sites located in Georgia were used for evaluation. The selected sites represent different climatic conditions and the number of daily records varied from 46 to 97 years. The predicted daily and monthly data were compared with both the observed daily and monthly average historical weather data and the target year of 2005 for all ten study sites. The statistical analysis included summary statistics, mean square difference (MSD) and its components, and the Kolmogorov-Smirnov (KS) test. The results showed that the k-NN approach was able to reproduce a similar pattern of the target year 2005 from the observed historical weather data. For all weather variables, both the lower and upper quartiles (Q1 and Q3) showed a very good agreement with the data of the observed target year. The cumulative distribution functions (CDFs) for the observed and predicted data were not significantly ($P > 0.05$) different across all sites for precipitation, except for the minimum temperature of seven study sites, radiation for five study sites, and maximum temperature for one study site. Our investigation to determine the minimum number of historical observed weather data required for obtaining reliable prediction revealed that 25 years of data were sufficient to find similar patterns compared to when all available weather data were used across all sites. It can be concluded from this study that the k-NN approach on the basis of pattern recognition can be considered as a reliable method to predict daily weather sequences based on historical weather data. Copyright © 2007 Royal Meteorological Society

KEY WORDS weather simulation; analogue weather; re-sampling methods; agrometeorology; climate; agriculture; hydrology; resource management

Received 9 April 2007; Revised 10 September 2007; Accepted 10 September 2007

1. Introduction

Process-based agricultural, ecological, and hydrological simulation models require weather information as input data for their operation. High quality weather data are quite crucial for the accurate simulation of the underlying crop, soil, and atmospheric processes. The critical agro-meteorological variables associated with crop production are precipitation, air temperature, and solar radiation (Hoogenboom, 2000a). For many agricultural locations, weather data are either not recorded or only a few parameters are available (Hoogenboom, 2000b). Significant advances in data collection and storage has enabled easier extraction of information from large databases. The recent availability of long-term historical data in digital format has provided an opportunity to develop mathematical models that can estimate weather data for locations that have only a small fraction of records. However, even when observed weather data are available, for prognostic applications one needs a set of estimated or generated weather data for the future (Bannayan *et al.*, 2003).

In the absence of observed data, researchers use data from nearby meteorological stations; derive data from other observed weather variables, e.g. solar radiation from air temperature or other variables (Garcia y Garcia and Hoogenboom, 2005); or obtain data with stochastic weather generators (Richardson, 1981; Hutchinson, 1987; Thornton *et al.*, 1997; Stockle *et al.*, 2001; Jarvis *et al.*, 2002). However, access to high quality weather databases required for precise prediction of agricultural production is still limited (Hansen, 2005). Therefore, further research

* Correspondence to: Mohammad Bannayan, Department of Biological and Agricultural Engineering, University of Georgia, Griffin, USA.
E-mail: bannayan@uga.edu

is needed to identify scientific methodologies that can be used to increase the quality of weather data prediction. This may include developing new models, improving weather data generators, using historical weather data, and exploring weather analogue models. The development and evaluation of weather generators have been studied extensively (Meinke *et al*., 1995; Semenov and Barrow, 1997; Soltani *et al*., 2000; Hartkamp *et al*., 2003; Kuchar, 2003). Many studies have also been conducted in which generated daily weather data from weather generators were applied to run the eco-physiological crop models for locations where long-term historical data were not available (Azam-Ali *et al*., 2001; Bannayan *et al*., 2003). Weather generators that are based on parametric statistical techniques typically use precipitation as the driving variable (Nicks and Harp, 1980; Richardson, 1981; Geng *et al*., 1986). In these models, the occurrence and amount of precipitation are generated independently. The other variables are then generated based on the stochastically generated precipitation. A major drawback of this type of models is that persistent events, such as drought or prolonged rainfall, are not well simulated (Hartkamp *et al*., 2003; Sharif and Burn, 2005).

Non-parametric re-sampling procedures are an alternative to generating daily weather data. The k-nearest neighbours (k-NN) is an analogous approach (Lall and Sharma, 1996; Rajagopalan and Lall, 1999). This method has its origin as a non-parametric statistical pattern recognition procedure to distinguish between different patterns according to a selection criterion. Yakowitz (1987) and Karlsson and Yakowitz (1987) constructed a robust theoretical base for the k-NN method. It has been employed in various studies, especially in hydrology (Galeati, 1990; Kember and Flower, 1993; Todini, 2000). The k-NN approach has also been successfully applied in other disciplines, including remote sensing (Chi and Bruzzone, 2005), traffic forecasting (Davis and Nihan, 1991), molecular biology (Wu *et al*., 2005), soil science (Jagtap *et al*., 2004; Nemes *et al*., 2006), and forestry (LeMay and Hailemariam, 2005). The scientific theory has been explained in detail by Brandsma and Buishand (1998), Rajagopalan and Lall (1999), and Gangopadhyay *et al*. (2005).

In our implementation of the k-NN approach, the similarity between the test instances, namely, the target year and the training instances, e.g. the historical weather data, determines k top-ranking nearest instances and the algorithm finds the most similar category where k refers to the number of nearest neighbours on which the selection is based. The k-NN method is based on the assumption that the most similar instance should belong to the same class (k neighbours) as the best match in the training instances. To determine the difference between two instances, Salzberg *et al*. (1991) proposed several distance metrics of which the Euclidean distance metric is the most common. The algorithm is based on the idea that the smaller distance between two instances indicates a higher similarity between them.

Non-parametric methods based on the k-NN bootstrap methods (Young, 1994; Rajagopalan and Lall, 1999; Buishand and Brandsma, 2001) can improve upon the parametric models. Rajagopalan and Lall (1999), using the k-NN approach, found that six simulated daily weather variables showed a higher accuracy in comparison with the parametric method (Richardson, 1981). The non-parametric approach has also been used to predict weather data for flood forecasting (Toth *et al*., 2000). An interesting feature of this approach is that no assumption has to be made about the underlying distributions of each of the variables and of the dependencies between those variables (Brandsma and Konnen, 2006). It can also be easily modified to be conditioned upon El Niño-Southern Oscillation (ENSO) or other ocean–atmosphere phenomena. However, it is rather important to verify such approaches by predicting daily weather data across multi-sites with different climatic conditions.

The primary objective of this study was to evaluate the accuracy of the k-NN approach by using daily weather data of one specific year as the target year, e.g. the test instance and historical daily weather data of multiple sites in Georgia as space instances. The second objective was to determine how many years of historical weather data (space instances) are required as the minimum number to obtain a similar accuracy compared to using the entire available historical data set of a given site.

2. Methodology

2.1. k-NN method

We implemented the k-NN procedure using the following steps:

(1) A feature vector consisting of observed weather data for the target year including solar radiation, precipitation, and maximum and minimum temperature was constructed.

(2) All days within a moving window of width w centred on day t, i.e. the day which is based on observed data and for which we want to predict the next day, were selected as potential candidates for day $t + 1$. For example, a 15-day temporal window for 1 March requires the data of 22 February to 8 March, excluding the data for day t from the historical years. Similar to the studies of Yates *et al*. (2003) and Gangopadhyay *et al*. (2005), a 15-day window, i.e. 7 days' lag and 7 days' lead ($k = 7$), was used in this study. The use of the moving-window approach is to represent smooth variations across seasonal boundaries (Sharma and Lall, 1999). Therefore, for n days as moving window and t days of a year with j weather variables, a data matrix was developed that had $n \times t$ rows and j columns. It is important to note that j can vary, and depends on how many weather variables one wants to use at the same time. In our implementation j was set to 4 and included solar

DOI: 10.1002/joc

radiation, precipitation, and maximum and minimum temperature. The structure of the data matrix is:

$$[A]^f{}_{nt\times 4} = \begin{bmatrix} a_{1,1}\dots a_{1,4} \\ a_{2,1}\dots a_{2,4} \\ \dots \\ a_{nt}\dots a_{nt,1} \end{bmatrix} \tag{1}$$

where $a_{i,j}$ is the value of the weather variable for the time index i $(i = 1, \dots, nt)$ and for weather variable j $(j = 1, \dots, 4)$.

(3) For the variables to be dimensionless and to reduce the seasonal variation, the data were converted to standardized variables (Brandsma and Buishand, 1998). This is also because variables with higher magnitudes disproportionately influence the neighbour selection (Yates *et al.*, 2003). In this study, for each year of the historical weather years, the daily data for each variable was subtracted from its annual mean (m_{d}) and divided by the annual standard deviation (s_{d}):

$$\tilde{x}_t = (x_t - m_{\text{d}})/s_{\text{d}} \tag{2}$$

where x_t and $\tilde{x}_t$ are the original and standardized variables, respectively, for day t.

(4) The Euclidean distance, d_j was computed between the feature vector of the current day's weather and the vector of observed data for each of the 15 days of the individual historical years.

$$d_j = \sqrt{\left[\sum_{j=1}^{d} W_j(V_{ij} - V_{mj})^2\right]} \tag{3}$$

where d_j, refers to the Euclidean distances; V_{ij} and V_{mj} are the jth components of each of vectors (feature and historical); d is the number of weather variables; and W_j are scaling weights ($1/S_{\text{j}}$) where S_{j} is the standard deviation of observed data. The k-NN approach selects from Euclidean distances and assigns probability weights to a subset of k distances with the smallest to the largest Euclidean distance to the feature vector. The Euclidean distances, d_j, were sorted in ascending order and the initial k-NN were retained. The ultimate objective was to select k years that were the most similar to the target year for all predicted days.

(5) The weights of the k neighbours were based on their rank distance to the value of the target weight function that is calculated as:

$$P_j = \frac{1/j}{\sum_{j=1}^{k} 1/j},\ j = 1, \dots, k \tag{4}$$

where j is the rank of the hindcast years in ascending order. The weight function assigns weights to each of the k-NN. The neighbour with the shortest distance was assigned the highest weight, whereas the neighbour with longest distance was assigned the smallest weight. Then a uniform random number U (0, 1) was generated and if $u \geq P_1$ then the day corresponding to distance d_1 was selected. If $u \leq P_k$ then the day corresponding to distance d_k was selected. For $P_1 < u < P_k$, the day t corresponding to d_j was selected for which u was closer to P_j. However, in this study the best match was always the year with the shortest distance.

(6) After processing, the data were re-transformed from the standardized value to their original scale using the mean and standard deviation of the observed data of the same selected year.

To determine the minimum number of required historical observed data, which was the final objective of this study, the model was run using only 5, 10, 15, 20, 25, 30, 35, and 40 years of recent observed data (before 2005) as the only available historical weather data for each study site. The model was enabled to output the rank of historical years from the best (match) to the worst (match) pattern compared to the target year. Therefore, for each set of historical years, it was possible to compare the rank of the predicted year with the most similar year (best match) obtained previously when all available historical years were used. Using the above set of historical years when the predicted year was the same as the best match obtained when all available historical years were available, we were able to determine the minimum number of years required to obtain a reasonable accuracy.

2.2. Study sites

For evaluation of the k-NN approach, we used daily weather data from ten representative sites located in Georgia, USA. These sites included Attapulgus and Camilla in southwest Georgia; Blairsville in the mountainous area; Rome in the northwest; Savannah in the coastal area; and Alma, Griffin, Midville, Plains, and Watkinsville in the central region of the state (Table I).

The difference between the highest amount of annual total rainfall, i.e. Blairsville, and the lowest amount, i.e. Midville, was approximately 200 mm. Rome, Griffin, and Savannah had a somewhat similar annual total rainfall at 1270 mm although these locations ranged from the northwest to the southeast of Georgia. Attapulgus, in the deep south, and Blairsville, the most northern station, also had somewhat similar total annual rainfall at 1422 mm. However, the monthly rainfall patterns were different from one site to another. The highest amount of total monthly rainfall was recorded for Savannah in August at 172 mm, whereas the lowest was recorded in April for Alma at 79 mm.

The average monthly minimum temperature was above 0 °C for all sites, except for Blairsville in January, February, and December, and for Rome in January where it dropped below 0 °C. The average monthly minimum temperature was highest in July, around 20 °C for all sites, except for Blairsville, where it was about 15 °C. The average monthly maximum temperature for January ranged from around 10 °C for Blairsville, Rome, and Watkinsville to around 17 °C for Alma and Attapulgus.

DOI: 10.1002/joc

Table I. Latitude (Lat), longitude (Long), elevation (Elev), and annual average of weather variables for the ten selected sites in Georgia, USA.

Station	Lat (N)	Long (W)	Elev (m)	Average minimum temperature (°C)	Average maximum temperature (°C)	Total precipitation (mm)	Annual number of wet days	Database period	Predicted best match year for 2005
Alma	31°32′	82°31′	63	12.8	25.3	1200	108	1948–2003	1993
Attapulgus	30°45′	84°30′	72	13.3	25.7	1447	111	1909–2003	1981
Blairsville	34°50′	83°56′	587	5.8	19.9	1405	122	1931–2003	1936
Camilla	31°28′	84°29′	50	12.9	26.0	1322	91	1938–2003	1981
Griffin	33°15′	84°17′	287	10.4	22.3	1282	117	1906–2003	1924
Midville	32°52′	82°14′	79	11.3	24.3	1150	92	1957–2003	1985
Plains	32°02′	84°23′	161	11.3	24.2	1246	107	1956–2003	1993
Rome	34°20′	85°08′	187	9.3	22.5	1392	119	1907–2003	1931
Savannah	32°00′	81°17′	8	13.0	24.8	1255	111	1950–2003	1994
Watkinsville	33°52′	83°28′	245	10.5	22.3	1246	112	1944-2003	1981

The average monthly maximum temperature for July ranged from 30 to 34 °C for all sites. The average monthly solar radiation also ranged from about 8 MJ m^{-2} d^{-1} in December for Blairsville to 23 MJ m^{-2} d^{-1} in May for Alma and Attapulgus.

2.3. Comparison of predicted and observed data

We compared the predictions with the observed data using the statistical evaluation methodology proposed by Kobayashi and Salam (2000). In this approach, n sets of predicted (x) and observed (y) values are compared on the basis of the mean squared deviation (MSD) as the measure of the difference between the two:

$$\mathrm{MSD} = \sum_{i=1}^{n}(x_i - y_i)^2/n \tag{5}$$

MSD has three additive components:squared bias (SB), squared difference between standard deviations (SDSD), and lack of correlation weighted by the standard deviations (LCS):

$$\mathrm{MSD} = \mathrm{SB} + \mathrm{SDSD} + \mathrm{LCS} \tag{6}$$

And each component is defined as:

$$SB = (\overline{x} - \overline{y})^2 \tag{7}$$

$$\mathrm{SDSD} = (\mathrm{SD_s} - \mathrm{SD_m})^2 \tag{8}$$

$$\mathrm{LCS} = 2\mathrm{SD_s}\mathrm{SD_m}(1 - r) \tag{9}$$

where $\overline{x}$ and $\overline{y}$ are the means of predicted (x) and observed (y) values, respectively, $\mathrm{SD_s}$ and $\mathrm{SD_m}$ are the standard deviations of x and y, respectively, and r is the correlation coefficient between x and y. The MSD term indicates the overall deviation of the model prediction from observation: a high MSD indicates a large gap between the prediction and observation. The components of MSD represent different aspects of the overall deviation with SB representing the bias of the simulation, SDSD the difference in the variation of predicted and observed values, and LCS providing information on how the temporal pattern of variation in observations was predicted. The square root of MSD is referred to as root mean square difference (RMSD), which has the same dimension as the original variables: x and y. RMSD was divided by the mean of the observations, i.e. $\overline{y}$, to give relative RMSD, which is denoted as $\mathrm{RMSD_r}$ and used for the comparison of RMSD among different weather variables. Similar to Rajagopalan and Lall (1999), the mean, standard deviation, skew, RMSD, and coefficient of variation (CV) were also used to evaluate the performance of predicted data in comparison to the target year. In addition, the daily observed and predicted data sets for each location were compared using cumulative distribution functions (CDFs). The Kolmogorov-Smirnov (KS) two-sample, two-sided test (known as Smirnov test) was used to compare the CDFs of the observed data for 2005 and predicted daily data. The KS test detects the largest difference that exists between two distribution functions based on a statistic, called the D statistic. This statistic is a measure of the discrepancy between the empirical distribution and the hypothesized distribution:

$$\mathrm{D} = Max_y|F_n(y) - F(Y)| \tag{10}$$

where $F_\mathrm{n}(\mathrm{y})$ is the empirical CDF and $F(\mathrm{Y})$ is the hypothesized CDF. However, in this study both cumulative distributions are not known and are considered as empirical functions. The D statistic is the maximum difference between the two distribution functions. If D is sufficiently large, then the null hypothesis (identical distribution) can be rejected. The smaller the D statistic, the smaller the difference between the two distributions at a given probability level (P value). The statistical analyses were performed using SAS Analyst (SAS Institute, 2001).

3. Results and discussion

3.1. Multi-site verification of *k*-NN approach

Precipitation is a crucial component for model evaluation because of its high spatial and temporal variability. The predicted data were able to follow the pattern of the observed total precipitation for all sites (Figure 1). The differences in variation of predicted and observed data (SB) were the main cause for the higher MSD for precipitation compared to other variables, although this variation was within a reasonable range (Table II).

Both the predicted and the observed data showed the highest amount of precipitation, averaged over the period of study, for Rome, and the lowest amount of precipitation for Attapulgus. The *k*-NN approach reproduced the number of wet days for both the monthly and the entire study period reasonably well (Figure 3). The annual total number of predicted and observed wet days differed by 9 days across all sites. The model slightly underestimated the number of wet days, but the *t*-test indicated that they were not significantly different ($P > 0.05$) across the study sites. The highest number of an observed total of wet days in 1 month was found for Blairsville and similar results were obtained from the *k*-NN approach predictions (Table III). The lowest number of a monthly total of wet days was obtained for Midville for both the predicted and observed data. The lowest number of a monthly total of wet days occurred in September, which was also reproduced by the model. This indicated the capability of the model to monitor the observed pattern for the number of wet days of the target year and recognize a similar pattern from the historical data.

These results are very encouraging. However, it should be mentioned that the impact of the error in the number of wet day predictions varies as a function of the application level. For example, the impact of an error of 5 wet days is different for agricultural production compared to the required resolution for aircraft, for instance, and thus should be judged based on the final application of forecasts. One of the real advantages of the *k*-NN approach is that any bias in prediction of the occurrence of the number of wet days does not affect the prediction of the other weather variables. For example, a very slight overestimation of solar radiation of four out of ten study sites has no association with the number of wet days prediction. In the non-parametric methodologies, such as the *k*-NN approach, the proper simulation of wet days is not as crucial as for the parametric approach (Geng *et al*., 1986). In the non-parametric approach, the maximum and minimum temperature and solar radiation are not conditioned on the occurrence of wet and dry days. The solar radiation predicted by the *k*-NN model was very similar to the observed solar radiation as indicated by the various statistics (Table II). The values for SD, skew, CV, and lower and upper quartile of predicted data were close to the target year when compared to the daily average of the historical weather data. There was no significant difference ($P > 0.05$) for both accumulated radiation per month and total accumulated radiation across all sites (Figure 2). The model was able to successfully recognize the pattern of observed radiation of the target year among the historical data. All calculated RMSD values for solar radiation were less than 10% of the mean observation.

Similar to precipitation, the model was able to follow the pattern of both observed minimum and maximum temperature across all sites. Averaged across the study period, the observed data showed the highest maximum temperature for Camilla and the lowest minimum temperature for Blairsville. The *k*-NN approach was able to reproduce the same rankings of the sites (Figure 2). The minimum temperature had a slightly higher MSD compared to the maximum temperature. However, the MSD of both variables was less than 10% of the mean observed values of the study period. Of particular importance for agricultural applications are temperature extremes, such as freeze events and the number of days with a maximum temperature greater than a certain threshold value (Schoof *et al*., 2005). The *k*-NN approach for the prediction of the number of days when minimum temperature is at or below freezing point (0 °C) and maximum temperature was greater than 35 °C across all sites was similar to the extremes of the target year (Figure 2). The *t*-test did not show any significant differences ($P > 0.01$) between the predicted and observed number of freeze events per month or for the entire study period. There was a substantial variation in the number of total freeze events for each site, ranging from more than 19 events at Blairsville to no events at Attapulgus for 2005 for the entire study period. The predicted data showed the highest number of freezing events at 17 for Blairsville and 1 event for Alma and Midville for the study period. The difference between model predictions and observations of the mean number of freeze events during the study period across all sites was less than 1 day. Similar results were obtained for the number of days when the maximum temperature was higher than 35 °C. The difference between model predictions and observations of the mean number of maximum temperature events higher than 35 °C during the study period across all sites was also less than 1 day.

Using observed weather data of 2005 as the target year, the *k*-NN approach was able to successfully find the best match for the pattern of the target year from historical years for each site. The linear correlation between observed and predicted mean total radiation, maximum and minimum temperature, and precipitation for all sites are shown in Figure 3 by a 1 : 1 line. The MSD for all weather variables ranged from 0.30 °C for maximum temperature to 1.51 mm for precipitation across all study sites. A comparison of the statistics for both predicted and observed data for 2005 for each weather variable is shown in Table II. The target year data characterized the highest (17.1 mm) MSD for precipitation and lowest (4.9 °C) MSD for maximum temperature. This indicates that the *k*-NN approach is able to find a similar variation for the target year data based on the historical data. Both the observed and predicted data showed a positive skewness for precipitation and negative skewness for the other variables (Table II). A similar direction and magnitude

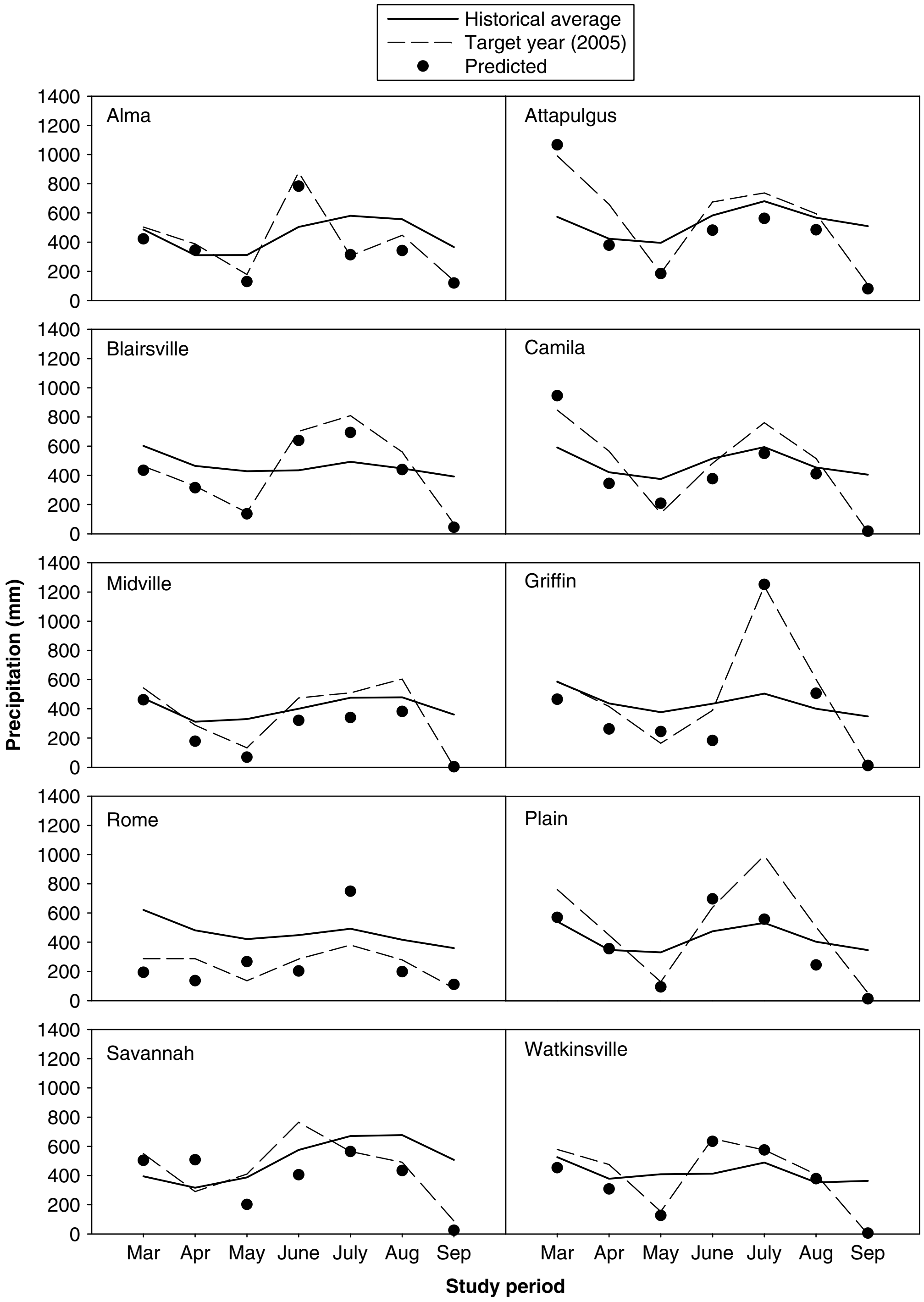

Figure 1. A comparison of the monthly total precipitation for the average of the historical years, predicted data, and the target year (2005) across all sites.

of skewness between observed and predicted data indicated the reliability of the *k*-NN approach on tracking the asymmetry characteristics of the observed weather variables. Similar CV values for both the predicted and observed data further supported the acceptable prediction of variability of weather data based on the *k*-NN approach (Table II). The *k*-NN model also performed well in reproducing the lower (Q1) and upper (Q3) quartiles for the weather variables for all sites with respect to the observed values of the target year (Table II).

The CDFs for daily precipitation, solar radiation, and maximum and minimum temperature were analysed for both the predicted and observed data (Figure 4). The predicted data generally matched the observed data well using the KS test. The D value ranged from 0.07 to 0.15 for radiation, 0.05 to 0.27 for maximum temperature, 0.06 to 0.21 for minimum temperature, and 0.03 to 0.10 for precipitation. The statistical analysis showed that for all sites, the CDFs of predicted and observed precipitation were not significantly different ($P > 0.05$). However, the

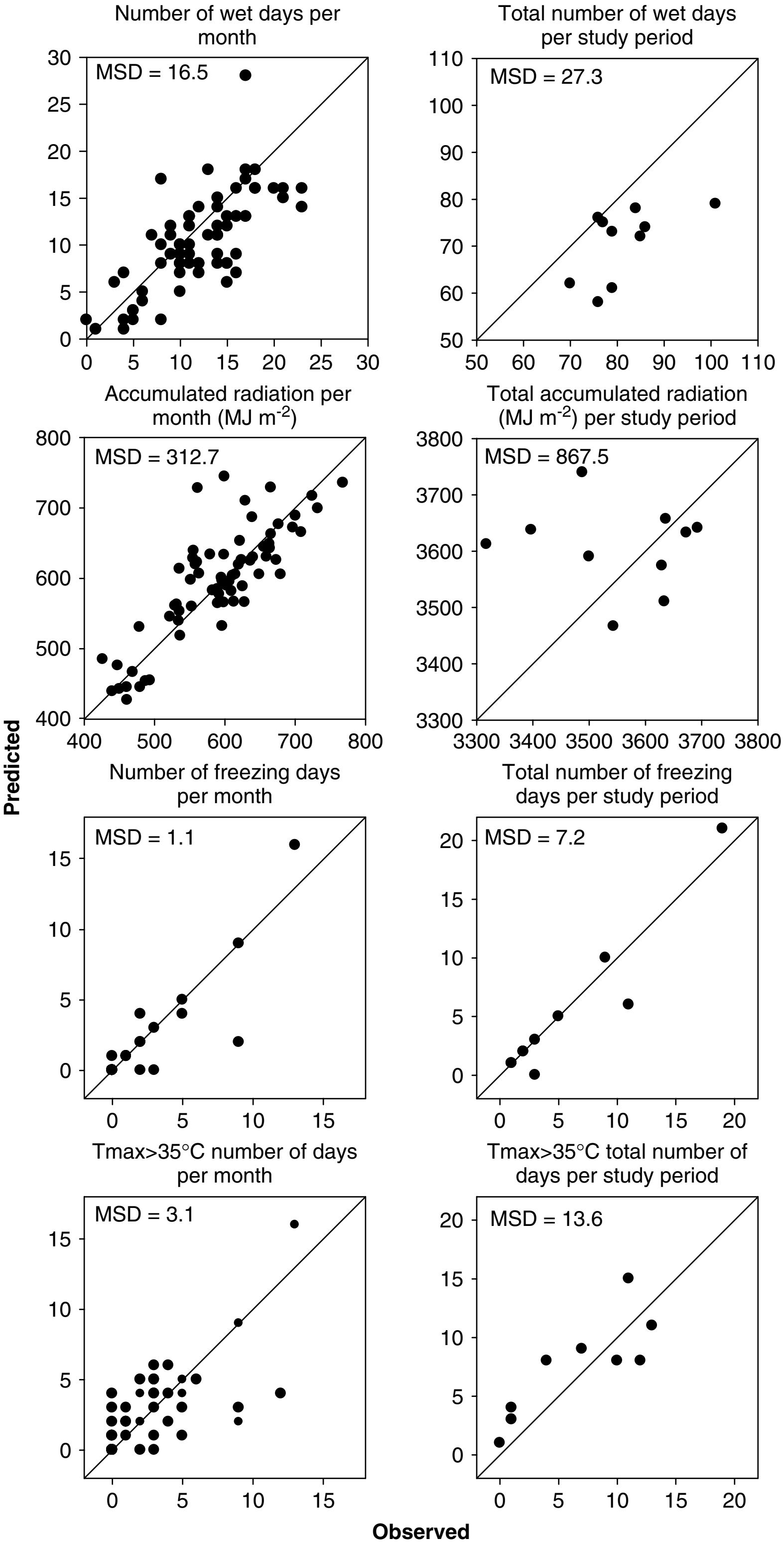

Figure 2. A comparison of monthly and annual predicted and observed (2005) number of wet days, total accumulated solar radiation, number of freezing events, and the number of days when the maximum temperature was above 35 °C across all sites.

difference was significant for a minimum temperature for seven out of ten sites, for maximum temperature for one out of ten sites, and for radiation for five out of ten sites.

Providing a complete set of weather variables is important for crop model applications, which require a fairly accurate combination of weather data input in order to provide accurate simulations of crop growth, development, and yield. An advantage of the *k*-NN approach to some other techniques is that extra weather variables can easily be included in the feature vector. While the *k*-NN approach was designed to reproduce daily or short-term statistics, the statistics of higher time

Table II. Calculated statistics for predicted (Prd) and target year (Obs) solar radiation, maximum and minimum temperature, and precipitation over the entire study period, across all study sites.

Site	Alma		Attapulgus		Blairsville		Camilla		Griffin		Midville		Plains		Rome		Savannah		Watkinsville	
Statistic	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs
Solar radiation																				
SD	5.3	6.4	5.2	6.5	6.0	6.8	5.1	6.2	4.8	6.3	4.9	5.9	5.1	6.8	4.8	6.3	5.6	6.1	5.3	6.7
Skew	−0.9	−0.9	−0.9	−0.9	−0.6	−0.6	−1.1	−1.1	−1.1	−0.7	−0.9	−0.9	−0.8	−0.8	−0.9	−0.5	−0.6	−0.8	−0.8	−0.8
CV	0.3	0.3	0.3	0.3	0.3	0.4	0.3	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Q1	17.0	16.7	17.1	17.1	15.3	15.0	17.4	17.1	17.3	15.0	17.0	14.7	16.4	15.5	16.6	14.1	15.9	15.1	15.8	15.5
Q3	23.4	24.2	23.6	24.8	23.0	24.1	23.2	24.3	23.4	23.5	22.8	22.1	22.8	24.7	22.3	22.8	24.0	23.6	22.9	24.5
Maximum temperature																				
SD	5.3	5.8	4.8	5.3	6.2	6.3	4.9	5.5	4.9	5.9	5.5	5.9	5.7	5.7	4.9	6.5	5.7	6.2	6.2	6.2
Skew	−1.2	−1.2	−1.2	−1.3	−1.2	−1.2	−1.1	−1.3	−0.9	−1.1	−1.0	−1.2	−1.2	−1.2	−0.7	−1.1	−1.2	−1.2	−1.2	−1.1
CV	0.2	0.2	0.2	0.2	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Q1	25.9	25.4	27.0	26.2	22.0	21.8	27.0	26.1	24.2	23.3	25.3	24.8	25.3	25.0	24.9	23.4	25.9	25.3	23.2	22.9
Q3	33.0	32.6	32.9	29.2	29.2	29.0	33.0	32.8	30.1	30.7	33.0	32.5	32.5	32.1	31.4	32.0	32.5	33.2	31.4	30.6
Minimum temperature																				
SD	5.9	6.5	5.8	5.9	6.8	7.1	6.0	6.4	5.0	6.6	6.2	6.5	6.4	6.6	6.1	7.1	6.2	7.7	6.5	6.8
Skew	−1.0	−1.0	−1.1	−0.9	−0.8	−0.6	−1.0	−1.0	−1.0	−0.8	−0.9	−0.9	−1.0	−0.9	−0.9	−0.7	−1.0	−0.7	−0.9	−0.9
CV	0.3	0.4	0.3	0.4	0.6	0.6	0.3	0.4	0.3	0.4	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.5	0.4	0.5
Q1	13.2	12.7	14.3	12.2	5.0	6.1	13.2	12.9	13.2	10.4	11.2	11.6	11.6	11.9	11.0	8.4	13.2	10.6	9.9	10.0
Q3	21.5	22.4	21.6	20.4	16.5	18.0	21.5	22.2	19.6	21.0	20.4	22.0	20.9	21.6	19.7	20.8	22.0	23.5	20.4	20.6
Precipitation																				
SD	9.0	10.1	11.6	15.1	9.7	11.1	10.8	12.3	8.5	13.1	8.2	10.9	12.7	17.1	5.7	5.2	10.6	10.0	8.9	10.8
Skew	3.6	3.7	3.2	3.7	3.8	4.2	3.5	4.1	5.7	4.2	5.9	5.2	6.5	6.7	3.2	2.7	4.7	3.7	3.4	3.6
CV	2.6	2.5	2.5	2.7	2.5	2.5	2.6	2.6	3.0	2.7	3.3	3.0	3.5	3.4	2.2	2.1	2.8	2.6	2.5	2.4
Q1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Q3	0.8	1.5	2.2	2.2	2.2	3.2	1.3	2.7	0.7	2.3	0.5	0.7	1.2	2.0	2.3	2.0	1.5	1.3	1.3	2.5

SD, standard deviation; CV, coefficient of variation; Q1, lower quartile; Q3, upper quartile.

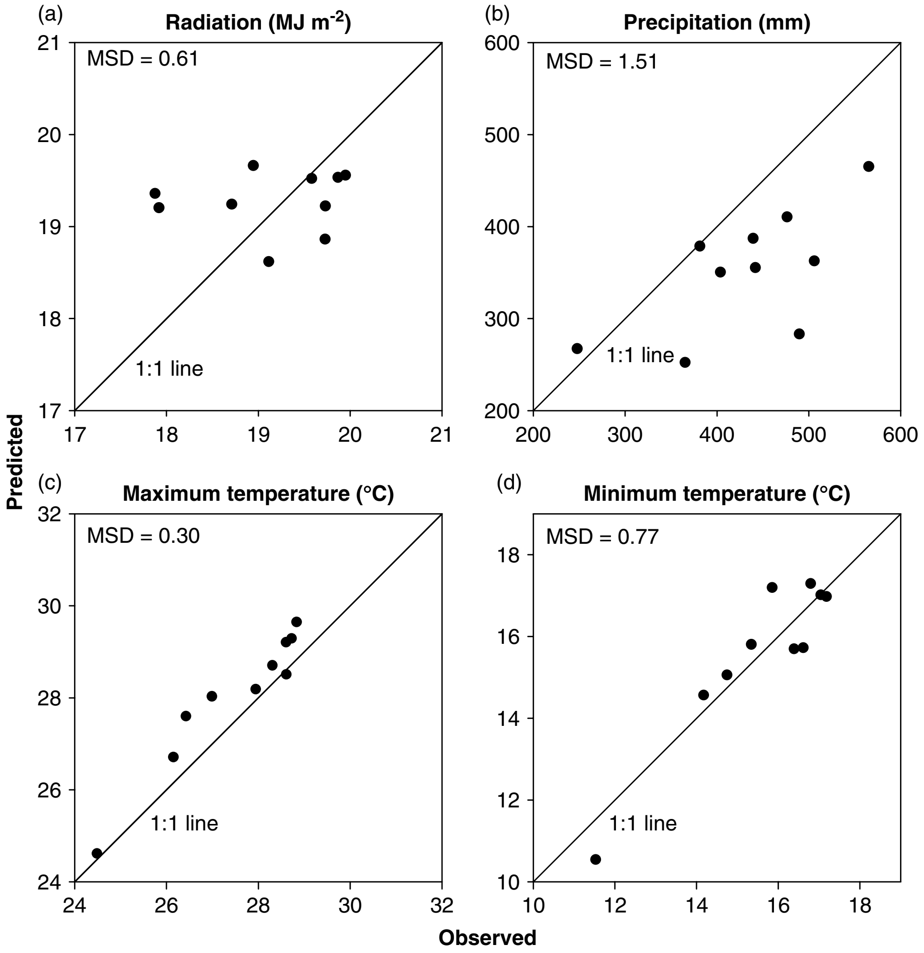

Figure 3. A comparison of predicted and observed (2005) (a) daily solar radiation, (b) total precipitation, (c) daily maximum, and (d) daily minimum temperature averaged over the entire study period.

Table III. Comparison of the predicted (Prd) and observed (Obs) number (#) of wet days per month for all study sites.

	Wet Days (#)													
Site	March		April		May		June		July		August		September	
	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs	Prd	Obs
Alma	12	15	10	10	8	10	18	18	12	9	12	14	3	5
Attapulgus	13	11	8	8	9	10	14	14	18	13	13	17	3	5
Blairsville	15	14	13	16	10	10	15	21	16	23	9	14	2	5
Camilla	13	11	5	10	7	10	13	16	11	13	13	17	1	4
Griffin	8	12	9	10	11	7	8	14	7	16	9	16	6	3
Midville	8	11	7	12	7	4	16	18	14	12	11	14	1	1
Plains	10	11	11	9	5	6	18	17	16	18	14	23	2	4
Rome	8	15	8	15	9	9	6	15	28	17	9	10	4	6
Savannah	12	11	8	12	9	11	16	20	17	8	16	16	2	8
Watkinsville	11	14	8	10	10	8	16	21	17	17	13	15	2	0

scale, mainly at the monthly time scale, appeared to be reproduced effectively as well.

3.2. Minimum required number of historical weather years

A major limitation of the k-NN approach is that it cannot reproduce values that are not part of the historical data base. The accuracy of the k-NN approach, therefore, partially relies on the availability of long-term historical weather data. Our study indicated that the k-NN approach reproduced the observed weather data reasonably well across all study sites. In our approach, the number of historical weather years ranged from 46 years data (minimum) for Midville to 97 years data (maximum)

DOI: 10.1002/joc

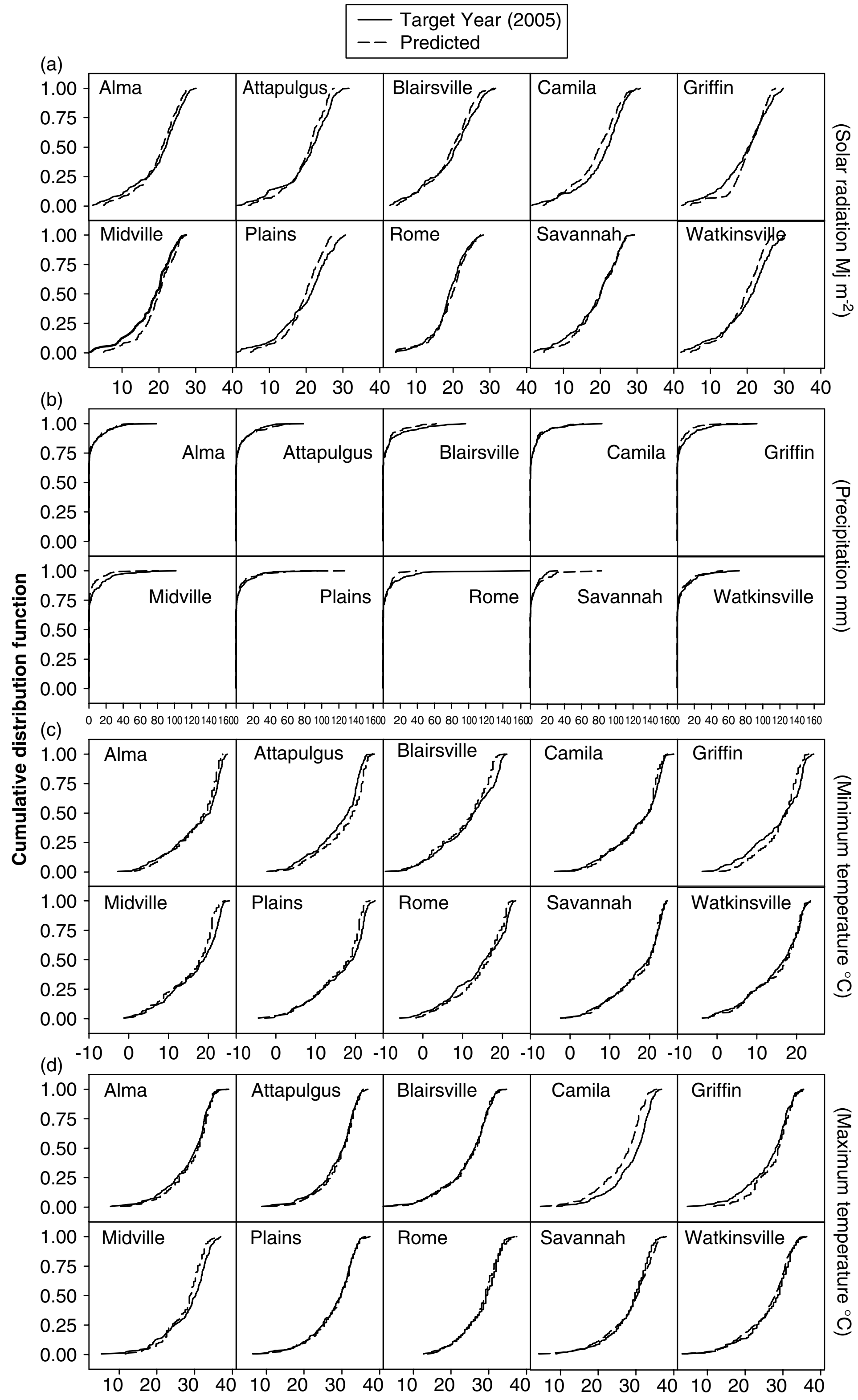

Figure 4. A comparison of predicted and observed (2005) cumulative distribution of (a) solar radiation, (b) precipitation, (c) minimum temperature, and (d) maximum temperature across all study sites.

for Griffin. Our study to determine the smallest number of observed historical data showed that by using only 5 years of historical data, e.g. 1998–2003, the most similar year was the year which ranked lower from the best match obtained when all available historical years were used. Using either 10 or 20 years of data (1993–2003, 1983–2003), the k-NN approach found that the most similar year ranked from first to third for the best match compared to when all available data were used. However, using 25 years of historical data, e.g.

1978–2003, the *k*-NN approach found the same best match (except for three out of ten sites) as when the model used all available historical weather data. Using ten different sites in this study, our results showed that *k*-NN approach was able to find the best matching year with smaller number of observed historical weather data (25 years) compared to when all available data (Table I) were used. However, we believe that a higher accuracy might be possible when the databases contain as many years of data as possible, but with a minimum of 25 years.

4. Summary and conclusion

In this study the prediction of four different weather variables using the nearest neighbour re-sampling approach was compared with observed weather data for 2005 as the target year for ten different sites in Georgia, USA. The *k*-NN method predicted the weather sequences for these multiple sites successfully and it was able to reproduce the temporal and spatial statistics of the target year. The lowest MSD values were obtained for maximum temperature and the highest MSD values were obtained for precipitation. A good agreement between the lower and upper quartile of the predicted and observed data of target year was obtained for precipitation. This indicated the potential capability of the *k*-NN approach to capture the daily variability of the weather data sequences. The sensitivity analysis to determine the minimum number of observed historical years showed that 25 years of data were sufficient to obtain a similar best match compared to when we used all available years of historical weather data. Further work for improvement of these realizations includes employing longer records than daily data, adding more climate indicator variables, and/or finding similarities using a two-step approach, such as precipitation first and the other variables at a second step. Refinement of the model also requires verification for different climatic zones across the world. It would also be an advantage to evaluate the model outputs by linking the predicted weather data to ecological, hydrological, agricultural, and economic models.

Acknowledgements

This work was conducted under the auspices of the Southeast Climate Consortium (SECC; secc.coaps.fsu.edu) and supported by a partnership with the United States Department of Agriculture-Risk Management Agency (USDA-RMA), by grants from the US National Oceanic and Atmospheric Administration-Office of Global Programs (NOAA-OGP) and USDA-Cooperative State Research, Education and Extension Services (USDA-CSREES) and by State and Federal funds allocated to Georgia Agricultural Experiment Station Hatch project GEO01654.

References

Azam-Ali SN, Aguilar-Manjarrez J, Bannayan M. 2001. *A Global Mapping System for Bambara Groundnut (Vigna subterranea L. Verdc) Production, FAO Agricultural Information Management Series*. FAO Rome, Italy.

Bannayan M, Crout NMJ, Hoogenboom G. 2003. Application of the CERES-wheat model for within-season prediction of winter wheat yield in the United Kingdom. *Agronomy Journal* **95**: 114–125.

Brandsma T, Buishand TA. 1998. Simulation of extreme precipitation in the Rhine basin by nearest neighbor resampling. *Hydrology and Earth System Sciences* **2**: 195–209.

Brandsma T, Konnen GP. 2006. Application of nearest-neighbor resampling for homogenizing temperature records on a daily to sub-daily level. *International Journal of Climatology* **26**: 75–89.

Buishand TA, Brandsma T. 2001. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Water Resources Research* **37**(11): 2761–2776.

Chi M, Bruzzone L. 2005. An ensemble-driven k-NN approach to ill-posed classification problems. *Pattern Recognition Letters* **27**: 301–307.

Davis GA, Nihan NL. 1991. Nonparametric regression and short-term freeway traffic forecasting. *Journal of Transportation and Engineering-ASCE* **117**(2): 178–188.

Galeati G. 1990. A comparison of parametric and non-parametric methods for runoff forecasting. *Hydrological Sciences Journal* **35**(1): 79–94.

Gangopadhyay S, Clark M, Rajagopalan B. 2005. Statistical downscaling using k-nearest neighbors. *Water Resources Research* **41**: 1–23.

Garcia y Garcia A, Hoogenboom G. 2005. Evaluation of and improve daily solar radiation generator for the southeastern USA. *Climate Research* **29**: 91–102.

Geng S, Penning de Vries FWT, Supit I. 1986. A simple method for generating daily rainfall data. *Agricultural and Forest Meteorology* **36**: 363–376.

Hansen JW. 2005. Integrating seasonal climate prediction and agricultural models for insights into agricultural practice. *Philosophical Transactions of the Royal Society B: Biological Science* **360**: 2037–2047.

Hartkamp AD, White JW, Hoogenboom G. 2003. Comparison of three weather generators for crop modeling: a case study for subtropical environments. *Agricultural Systems* **76**(2): 539–560.

Hoogenboom G. 2000a. The Georgia automated environmental monitoring network. 2000. *Preprints 24th Conference on Agricultural and Forest Meteorology*. American Meteorological Society: Boston, Massachusetts; 24–25.

Hoogenboom G. 2000b. Contribution of agrometeorology to the simulation of crop production and its applications. *Agricultural and Forest Meteorology* **103**(1–2): 137–157.

Hutchinson MF. 1987. Methods of generation of weather sequences. In *Agricultural Environments: Characterization, Classification and Mapping. Proceedings of a Workshop on Agro-Ecological Characterization, Classification and Mapping. Rome, Italy, 14–18 April 1986*, Bunting AH (ed). CAB International: Wallingford; 149–158.

Jagtap SS, Lall U, Jones JW, Gismn AJ, Ritchie JT. 2004. Dynamic nearest neighbor method for estimating soil water parameters. *Journal of the American Society of Agricultural and Biological Engineers* **47**(5): 1437–1444.

Jarvis CH, Stuart N, Hims MJ. 2002. Towards a British framework for enhancing the availability and value of agro-meteorological data. *Applied Geography* **22**: 157–174.

Karlsson M, Yakowitz S. 1987. Nearest-neighbor methods for nonparametric rainfall–runoff forecasting. *Water Resources Research* **23**(7): 1300–1308.

Kember G, Flower AC. 1993. Forecasting river flow using nonlinear dynamics. *Stochastic Hydrology and Hydraulics* **7**: 205–212.

Kobayashi K, Salam MU. 2000. Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal* **92**: 345–352.

Kuchar L. 2003. Using WGENK to generate synthetic daily weather data for modelling of agricultural processes. *Mathematics and Computers in Simulation* **65**: 69–75.

Lall U, Sharma A. 1996. A nearest neighbor bootstrap for time series resampling. *Water Resources Research* **32**(3): 679–693.

LeMay V, Hailemariam T. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. *Forest Science* **51**(2): 109–119.

Meinke H, Carberry PS, McCaskill MR, Hills MA, McLeod I. 1995. Evaluation of radiation and temperature data generators in the Australian tropics and sub-tropics using crop simulation models. *Agricultural and Forest Meteorology* **72**: 295–316.

Nemes A, Rawls WJ, Pachepsky YA. 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. *Soil Science Society of America Journal* **70**: 327–336.

Nicks AD, Harp JF. 1980. Stochastic generation of temperature and solar radiation data. *Journal of Hydrology* **48**: 1–7.

Rajagopalan B, Lall U. 1999. A k-nearest-neighbor simulator for daily precipitation and other variables. *Water Resources Research* **35**(10): 3089–3101.

Richardson CW. 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research* **17**(1): 182–190.

Salzberg S, Delcher A, Heath D, Kasif S. 1991. Best case for nearest neighbor learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**: 599–610.

SAS Institute. 2001. *SAS System*, 8th edn. SAS Institute: Cary.

Schoof JT, Arguez A, Brolley J, O'Brien JJ. 2005. A new weather generator based on spectral properties of surface air temperatures. *Agricultural and Forest Meteorology* **135**: 241–251.

Semenov MA, Barrow EM. 1997. Use of a stochastic weather generator in the development of climate change scenarios. *Climatic Change* **35**: 397–414.

Sharif M, Burn DH. 2005. Simulating climate change scenarios using an improved K-nearest neighbor model. *Journal of Hydrology* **325**: 179–196.

Sharma A, Lall U. 1999. A nonparametric approach to daily rainfall simulation. *Mathematics and Computers in Simulation* **48**: 367–371.

Soltani A, Latifi M, Nasiri M. 2000. Evaluation of WGEN for generating long–term weather data for crop simulations. *Agricultural and Forest Meteorology* **100**: 1–12.

Stockle CO, Nelson R, Donatelli M, Castellvi F. 2001. ClimGen: a flexible weather generation program. *Proceedings 2nd International Symposium Modelling Cropping Systems*. European Society of Agronomy: 16–18 July: Florence; 229–230.

Thornton PE, Running SW, White MA. 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology* **190**: 214–251.

Todini E. 2000. Real-time flood forecasting: operational experience and recent advances. In *Flood Issues in Contemporary Water Management*, Marsalek J, Watt WE, Zeman E, Sieker F (eds). Kluwer Academic Publisher: The Netherlands; 261–270.

Toth E, Brath A, Montanari A. 2000. Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology* **239**: 132–147.

Yakowitz S. 1987. Nearest neighbor method for time series analysis. *Journal of Time Series Analysis* **8**(2): 235–247.

Yates D, Gangopadhyay S, Rajagopalan B, Strzepek K. 2003. A technique for generating regional climate scenarios using a nearest neighbor algorithm. *Water Resources Research* **39**(7): 1199.

Young KC. 1994. A multivariate chain model for simulating climatic parameters from daily data. *Journal of Applied Meteorology* **33**: 661–671.

Wu W, Xing EP, Myers C, Mian IS, Bissell MJ. 2005. Evaluation of normalization methods for cDNA microarray data by k-NN classification. *Bioinformatics* **6**: 191.