Available online at www.sciencedirect.com
ScienceDirect
Environmental Modelling & Software 23 (2008) 703–713
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Weather analogue: A tool for real-time prediction of daily weather data realizations based on a modified *k*-nearest neighbor approach

Mohammad Bannayan*, Gerrit Hoogenboom

Department of Biological and Agricultural Engineering, University of Georgia, Griffin, GA 30223, USA

Received 17 April 2007; received in revised form 12 September 2007; accepted 16 September 2007
Available online 19 November 2007

Abstract

Quantifying the response of any given system beyond their current condition to weather alone or together with other factors requires predicting realizations of future weather conditions. The predicted weather data should not only be sufficiently accurate, but also their time scale should be in accordance to the decision support system in which the studied system is being applied. Inclusion of predicted weather data with, for example, a crop process-based simulation model could provide valuable and timely information for evaluation of various management techniques to avoid potential losses or increase crop production and income. Weather analogue as a nonparametric approach is easy and accurate to use to achieve this goal. In this study a weather analogue modeling tool is presented for predicting daily weather data realizations that are based on a modification of the *k*-nearest neighbor approach. Our intent was to develop a tool to predict a realization of real-time daily weather data by introducing two different methodologies for the *k*-nearest neighbor approach. In the first approach (*k*-mean), weather prediction for day $t+1$ was assumed as the average of all days found as the *k best match* days for the target day. In the second approach we assumed that only a fraction of the observed data (target year) was available (e.g. 90, 120, and 150 days) and that the realization for the remainder of the year is of interest. Based on this approach, the model should recognize the most similar pattern to the available data of the target year among the same sequence of historical data. Daily weather data of the selected year as the *best match* would be considered for the remainder of the target year. Both approaches were compared with observed data from 16 locations in the USA, Europe, Africa, and Asia, representing different climatic regions. Employing the first approach (*k*-mean), the *k*-NN model was quite promising and was able to recognize the pattern of the target year among the historical observed weather data for solar radiation, maximum and minimum temperature. However, the *k*-mean approach only reproduced the observed pattern of precipitation successfully when there was not a high variability in the pattern of precipitation occurrences. Using the second approach, as expected, a larger share of observed data in the target year beyond 90 days greatly improved the accuracy of prediction. However, after using 150 days both bias measures, e.g., MSD and MASE, slightly increased due to a change of the *best match* year. The results from this study showed that this weather analogue program could be a valuable tool for realization of any weather dependent function. There is also scope for incorporation of this tool with application of agricultural, ecological, and hydrological process-based simulation models.

Keywords: Climate forecasting; Weather data generator; Weather software; Climate variability; Real-time weather realization

Software availability

Name of product: Weather Analogue
Developed by: Mohammad Bannayan and Gerrit Hoogenboom
Contact address: Department of Biological and Agricultural Engineering, University of Georgia, Griffin, GA 30223, USA. Tel.: +1 770 229 3436; fax: +1 770 228 7218. bannayan@uga.edu
Available since: 2007
Programming language: Delphi 2006
Hardware requirement: Personal Computer with Windows XP or equivalent
Program size: 500 kbytes
Availability: gerrit@uga.edu

* Corresponding author. Tel.: +1 770 229 3436; fax: +1 770 228 7218.
E-mail address: bannayan@uga.edu (M. Bannayan).

1364-8152/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2007.09.011

1. Introduction

Agriculture is one of the most weather dependent human activities (Oram, 1989). Farmers usually use a conservative approach for cultivation due to the unknown weather they might face. Vulnerability of agriculture to weather variability makes weather an important component of the agricultural production system. Agricultural businesses are responsive to weather fluctuations mainly due to the impacts of weather on production and associated management interventions. The availability of weather realizations with an adequate accuracy for the unrecorded future is an essential component of crop production forecasts (Bannayan et al., 2003). Incorporating the predicted weather data with an agricultural, ecological, and hydrological process-based simulation model is considered an added value to the prediction of weather data (Hoogenboom, 2000; Jones et al., 2000). However, there is usually a mismatch between the spatial and temporal scale of the output of dynamic climate models (Goel and Dash, 2007) and the required input for process-based simulation models. A means of predicting future weather data is required, either as standalone or as part of a decision support system, for improved and timely management of farming systems to reduce the risk of production loss and thus increase the gross margins and net returns (Tsuji et al., 1998).

Weather data generators have been developed and employed for generating the required daily weather data for model applications (Geng et al., 1986; Jones and Thornton, 2000; Yates et al., 2003; Kilsby et al., 2007). These generators can be broadly classified into two categories as parametric and nonparametric approaches. Weather generators based on parametric statistical techniques typically use precipitation as the driving variable (Richardson, 1981; Nicks and Harp, 1980). In such models, precipitation occurrence and amount are generated independently and other variables are then generated based on the stochastically generated precipitation. A major drawback of these weather generators is that persistent events, such as drought or prolonged rainfall, are not well simulated (Hartkamp et al., 2003; Sharif and Burn, 2005). This aspect was addressed in models presented by Rackso et al. (1991) and Semenov et al. (1998), but their models are still site specific and require the specification of the model parameters (Soltani and Hoogenboom, 2003; Sharif and Burn, 2005). Nonparametric resampling procedures form an alternative approach to predict daily weather data. An interesting feature of nonparametric approaches 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). Among nonparametric approaches the *k*-nearest neighbor (*k*-NN) approach has shown to be promising and has been applied in various prediction studies, including remote sensing (Chi and Bruzzone, 2005), traffic forecasting (Davis and Nihan, 1991), molecular biology (Wu et al., 2005), soil science (Nemes et al., 2006), forest science (LeMay and Hailemariam, 2005), and hydrology (Todini, 2000). In this approach, *k* refers to the number of nearest neighbors on which the selection is based and NN abbreviates nearest neighbors. The *k*-NN method is based on recognizing a similar pattern of target file within the historical observed weather data which could be used as prediction of the target year. The target year is the initial seed of data which, together with the historical data, are required as input files for running the model. This method relies on the assumption that the actual weather data observed during the target year could be a replication of weather recorded in the past. Nonparametric methods based on *k*-NN bootstrap methods (Young, 1994; Rajagopalan and Lall, 1999; Buishand and Brandsma, 2001) can improve upon the parametric models. Buishand and Brandsma (2001) extended this approach to predict weather data across multi-locations. The *k*-NN algorithm typically selects a specified number of days similar to the pattern of weather variables of the day of interest. One of these selected days is randomly resampled as prediction of the weather for the next day.

Various predictions with dynamic agricultural, ecological, and hydrological models require sufficient lead time forecasts of weather variables to enable informed management decisions. By providing, in advance, information early enough, it might be possible to adjust critical agricultural, ecological, and hydrological decisions which would result in significant improvement in efficiency of agro-environmental management and food security. Previously it has been shown (Bannayan and Hoogenboom, in press) that the *k*-NN approach is able to reproduce a similar pattern of the observed weather data featured as the target year from historical weather data. The study also determined the minimum number of historical years required for obtaining a similar accuracy when the full database of historical data was used. This study introduces a new tool for predicting daily weather data realizations based on the modified *k*-NN approach. The overall goal was to determine if the new approaches are able to predict the future "unrecorded" data even if only a partial number of days of observed data exists for the target year.

2. Materials and methods

The *k*-NN approach is rooted in pattern recognition in which a target object with a defined vector of features could be used to find a similar pattern among the objects space (e.g. historical years of observed weather data). The *k*-NN procedure determines the similarity between different patterns according to one or more selected criteria. Yakowitz (1987) and Karlsson and Yakowitz (1987) constructed a robust theoretical base for this method. The original algorithm has been explained in detail by Brandsma and Buishand (1998), Rajagopalan and Lall (1999), and Gangopadhyay and Rajagopalan (2005). The prediction flowchart for the default approach for a 1-day prediction and the tool interface are shown in Figs. 1 and 2, respectively. For *n* successive number of days to be predicted, the processes shown in the flowchart (Fig. 1) should be run for *n* times.

The feature vector of the target year, as the available latest year of observed weather data for each of our study sites (Table 1), consisted of observed data for solar radiation, precipitation, and maximum and minimum temperature. The pattern of observed data for each day (*t*) of each site is compared to the pattern of the same variables for the same day in each year of the historical weather data. The comparison process computes the Euclidean distance between the target pattern (each day) and each historical pattern (same day as target day). For each set of historical data, the Euclidean distance computation is needed to determine the *k*-nearest neighbor of the target data. Then,

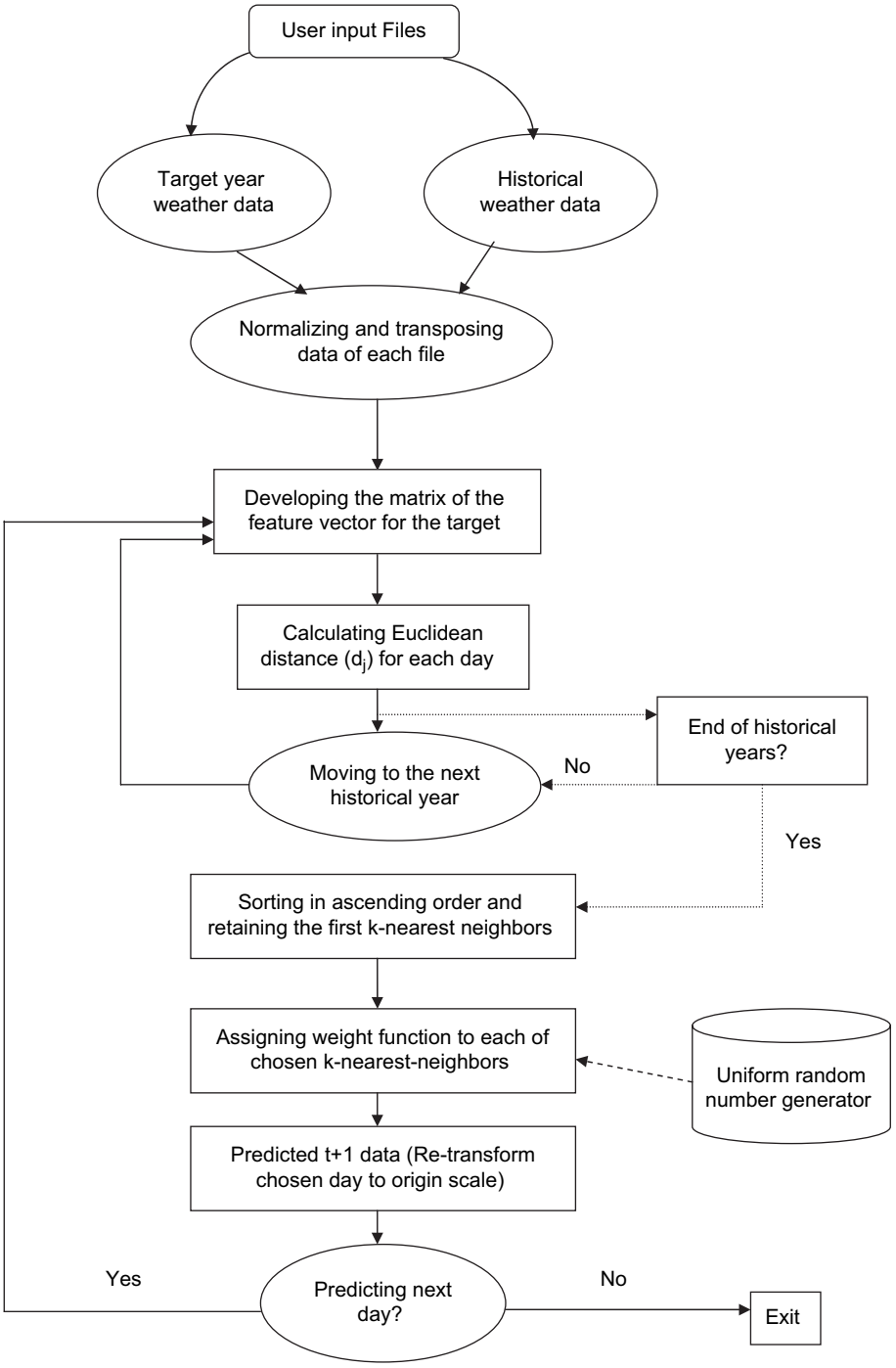

Fig. 1. Sequence of steps for the *k*-NN approach.

a probability weight is assigned to each distance with the smallest weight to the largest Euclidean distance order.

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

where d_j is the Euclidean distance, V_j is the jth component of either of vectors (feature, i, and historical, m), d is the number of weather variables and W_j are weights. The weight function of the k neighbors is calculated as:

$$W_j = \frac{1/j}{\sum_{j=1}^{k} 1/j}, \quad j = 1, \ldots, k \tag{2}$$

The final selection from k patterns (including all four weather variables) as prediction for the next day (*day* $t+1$) is based on a unique random number sampling procedure. Therefore, based on a generated uniform random number U (0, 1) one of the k neighbors is selected as the predicted data for *day* $t+1$.

We introduced two new approaches into the original k-NN algorithm. The first approach considers the average of the k days (k-mean) as representative of weather data of $t+1$ instead of resampling *1 day* out of k days by probability weighting. In other words, if the target day (t) is March 1 and k is defined as

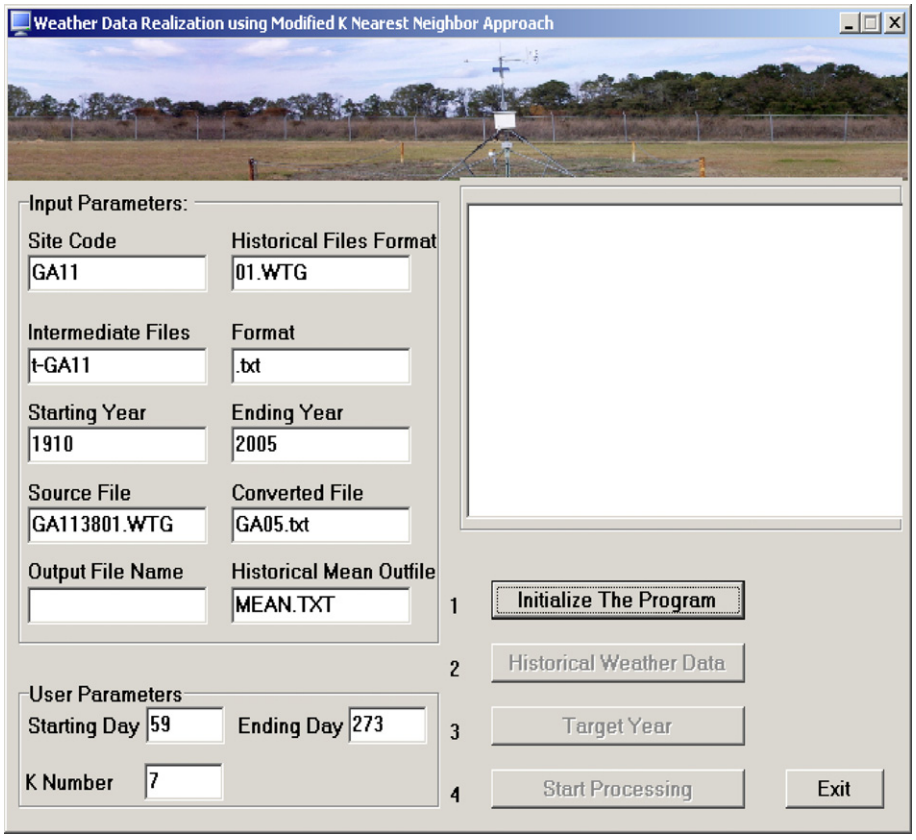

Fig. 2. Interface of the *k*-NN software tool.

seven, k days includes the period from February 21 to March 7. In this approach, we used the average of all these days as prediction of the weather data for day $t+1$. The model was run employing this approach and the weather data of the last observed year (Table 1) from the set of historical years were used as the target year.

The second approach is based on the idea that we have recorded weather data for part of the year and the weather data for the remainder of the year have not been observed and recorded yet. In this approach, the model is able to find the *best match* for the remainder of the year from observed historical data based on a similar pattern with recorded weather data of the target year. For example, when the target year consists of only 90 observed days (starting on January 1) the model calculates d_j values between the target year and each historical year separately for the first 90 days of the year. Afterwards, all calculated values for d_j for all historical years are sorted in ascending order. The historical year with the lowest d_j will be selected as the *best match*. At this stage, the output file contains $365+1$ days of weather data, consisting of the first 90 days from the target year and the remainder from the *best match* year. Obviously, as we proceed through the target year (more than 90 days), the

Table 1
Physiographic features, length of the historical database, target year, historical annual average of radiation, and temperature and total annual precipitation of the study sites (latitude: Lat, longitude: Long, and elevation: Elev)

Country	State/province/town	Station	Lat (N)	Long (W)	Elev (m)	Radiation (MJ m^{-2})	T_{max} (°C)	T_{min} (°C)	Precipitation (mm)	Database period	Target year
USA	Alabama	Baldwin	30°88′	−87°78′	83	16.3	25.3	13.6	1696.7	1913–2004	2004
Burkina Faso	Bale	Boromo	11°73′	−2°92′	264	20.9	34.9	21.2	912.1	1945–1999	1999
USA	Alabama	Covington	31°30′	−86°52′	76	16.1	23.0	10.0	1269.0	1912–2004	2004
USA	California	Davis	38°53′	−121°78′	18	17.6	23.6	7.7	442.1	1909–2002	2002
Burkina Faso	Mouhoun	Dedougou	12°46′	−3°48′	299	22.6	35.2	21.9	810.9	1945–1999	1999
USA	Alabama	Limestone	34°68′	−86°88′	183	15.2	22.3	9.2	1362.5	1950–2002	2002
Iran	Khorasan	Mashhad	36°15′	59°28′	985	*	22.4	8.2	255.5	1962–2004	2004
USA	Michigan	Michigan	42°40′	−85° 38′	277	14.5	15.0	3.7	951.4	1929–2002	2002
USA	Georgia	Midville	32°88′	−82°22′	85	10.1	24.3	11.3	1150.0	1957–2004	2004
USA	Florida	Monroe	25°00′	−80°52′	2	18.9	29.1	18.2	1361.2	1936–2003	2003
USA	Florida	Nassau	30°67′	−81°47′	4	16.7	25.2	15.5	1296.9	1910–2003	2003
USA	Georgia	Plains	32°05′	−84°37′	152	16.3	24.2	11.3	1246.0	1956–2004	2004
USA	Florida	Sumter	28°67′	−82°08′	23	17.2	28.2	15.0	1301.0	1918–2002	2002
USA	Alabama	Tallapoosa	32°82′	−85°65′	207	15.9	14.4	1.3	1227.8	1910–2004	2004
USA	Georgia	Tifton	31°45′	−83°48′	116	16.5	25.1	12.7	1200.1	1911–2004	2004
UK	Harpenden	Rothamsted	52°50′	−5°00′	100	9.7	13.4	5.7	701.6	1959–1999	1999

T_{max}: maximum temperature, T_{min}: minimum temperature.

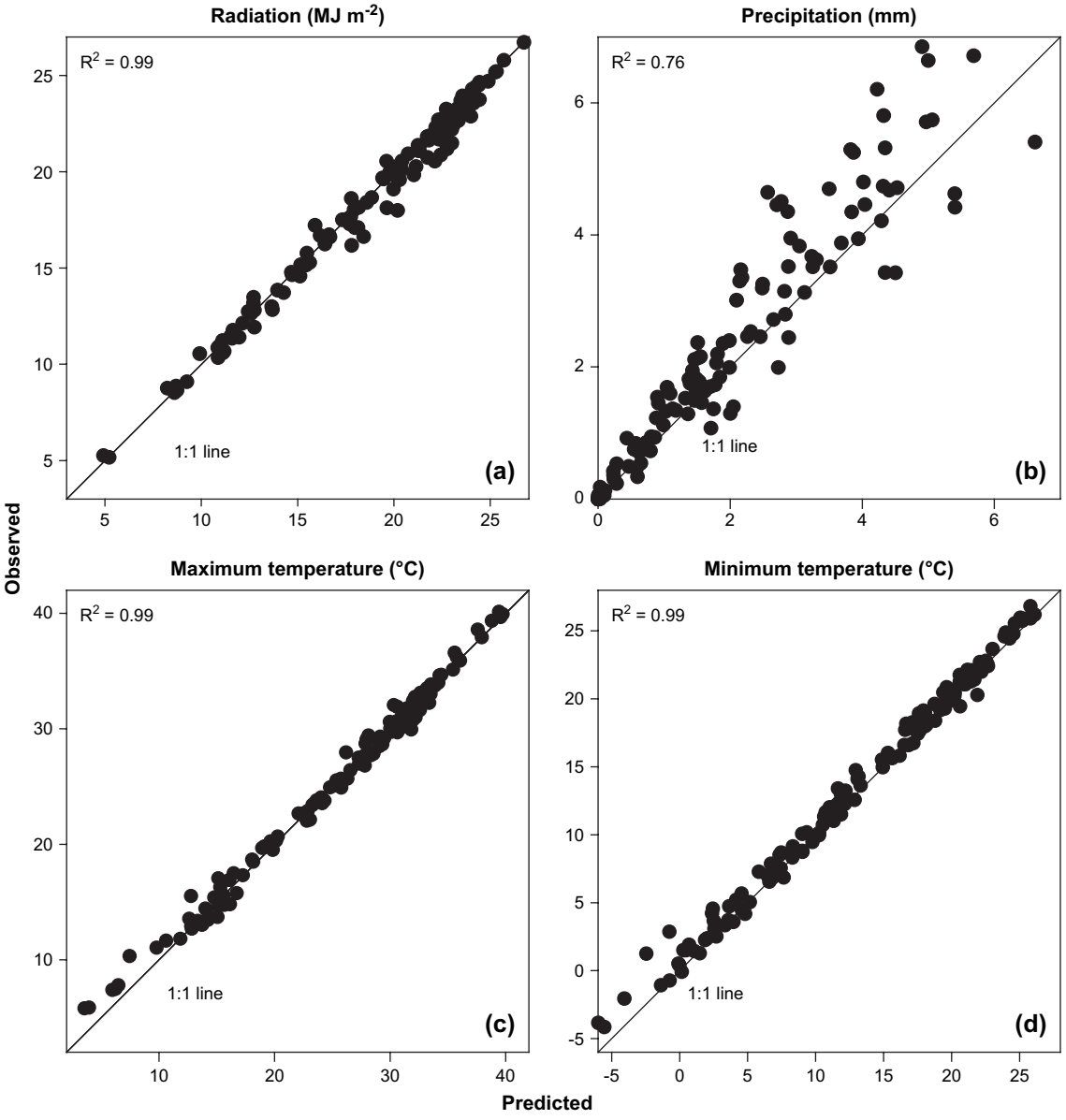

Fig. 3. Mean monthly observed and predicted solar radiation, precipitation, and maximum and minimum temperature data across all study sites, using the *k*-mean approach.

share of observed data in the output file will be higher. However, for each time update of the observed daily data, the model should be run again.

2.1. Comparison of observed vs. predicted data

The Mean Square Difference (MSD) was used to compare the observed and predicted data. The calculation of MSD is based on using n sets of predicted (x) and observed (y) values, which are compared as the measure of the difference between the two. The MSD was calculated as:

$$\mathrm{MSD} = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{n} \tag{3}$$

However, as MSD is sensitive to outliers (Armstrong, 2001), Mean Absolute Scaled Error (MASE) (Hyndman and Koehler, 2006) may eliminate such a problem due to its independency of the scale of the data. It was, therefore, also employed in the accuracy calculation. MASE was calculated as:

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|} \tag{4}$$

and MASE = mean ($|q_t|$).

Where Y_i and Y_{i-1} denote the observation at time i and $i-1$, e_t denotes the difference between the observed and the predicted data at time t, and q_t is the scaled error. MASE is the absolute value of mean of q_t.

2.2. Study site

This study employed daily weather data from 16 sites, from the USA, Europe, Africa, and Asia (Table 1). Latitude, longitude, elevation, target year, and the length of historical observed weather data for each site are presented in Table 1. Monthly rainfall patterns were different from one site to another. The historical average minimum temperature was as low as 3 °C for Michigan and as high as 22 °C for Dedougou, Burkina Faso. The highest historical average maximum temperature (40 °C) was recorded for Boromo, Burkina Faso, while the lowest (13.4 °C) was obtained for Davis, CA, USA.

The historical average daily solar radiation also ranged from about 14.5 MJ m^{-2} d^{-1} in Michigan, USA to 22.6 MJ m^{-2} d^{-1} in Dedougou, Burkina Faso.

3. Results and discussion

3.1. Approach I

Applying the *k*-mean modification of *k*-NN approach showed that the model was able to reproduce similar values and patterns of solar radiation, minimum and maximum temperature, but underestimated the precipitation for most of our study sites (Fig. 3). Observed solar radiation within the target year of each site ranged from 2.5 to 26.7 MJ d^{-1} and was accurately predicted by the *k*-mean approach as 2.4–26.8 MJ d^{-1}. Similar to solar radiation, this approach was able to follow the pattern of both observed minimum and maximum temperature across all sites. The range of values for observed minimum and maximum temperature were −4.2 to 26.2 °C and 5.7–39.9 °C, respectively. The predicted range for minimum temperature was −5.5 to 26.1 °C and for maximum temperature was 3.6–39.8 °C. This indicated the capability of the tool to monitor the observed pattern of weather data of the target year, find the *k* best match days for each day of the target year, using the *k*-mean approach, and to predict the realization of weather data for the current year. Employing the *k*-mean approach, the tool was able to find a similar pattern for the low precipitation values across all sites, but underestimated the high values for precipitation.

The Mean Square Difference (MSD) for each weather variable for all sites is shown in Table 2. Precipitation had the highest values for MSD compared to the other weather variables. The two highest values for precipitation were obtained for Covington and Plains, while the two lowest values were obtained for Davis and Rothamsted (Table 2). A further analysis of the precipitation patterns of these four sites showed that for Covington and Plains, the outliers for the amount of precipitation, e.g., close to 200 mm for Covington and 140 mm for Plains, occurred in both the target year and the historical data. In contrast for Davis and Rothamsted the precipitation values were smoother and lower and there were no outlier data such as a heavy precipitation occurrence for any given day. These results suggest that applying the *k*-mean approach, the model was able to follow the pattern of solar radiation, maximum and minimum temperature quite well. This method also showed promising results when very high values for precipitation, e.g., more than 100 mm, did not occur for a given site. However, if a site had a history of very heavy and intensive precipitation on any given day, then due to the nature of the average function, the *k*-mean approach was not able to follow the exact pattern of observed precipitation.

3.2. Approach II

Our results across all sites (Figs. 4 and 5) showed that the model was able to find the most similar target year pattern among the historical observed data when only a fraction of weather data of a year was available. All weather variables, including extreme values of precipitation occurrence, were reasonably reproduced. Table 3 shows a good performance of the model to capture both the annual mean and variation of observed data when the target year contained 120 days of observed weather data. The highest correlation coefficient (0.97) between the observed and predicted annual mean was obtained for minimum temperature while the lowest correlation coefficient (0.62) was obtained for precipitation.

The highest MSD was obtained for precipitation and the lowest was obtained for solar radiation (Figs. 4 and 5). It was our expectation that a higher number of contributed

Table 2
Mean Square Difference (MSD) of the predicted vs. observed values for solar radiation, precipitation, and maximum and minimum temperature using the *k*-mean approach

Country	State/province/town	Station	MSD			
			Solar radiation (MJ m^{-2})	Maximum temperature (°C)	Minimum temperature (°C)	Precipitation (mm)
USA	Alabama	Baldwin	0.7	1.4	1.03	33.6
Burkina Faso	Bale	Boromo	2.9	0.7	0.86	19.4
USA	Alabama	Covington	0.7	0.8	2.00	65.3
USA	California	Davis	0.3	0.9	1.6	2.1
Burkina Faso	Mouhoun	Dedougou	5.8	0.7	0.6	3.2
USA	Alabama	Limestone	0.9	2.6	1.6	9.6
Iran	Khorasan	Mashhad	*	2.9	3.2	3.2
USA	Michigan	Michigan	1.3	4.4	3.2	1.4
USA	Georgia	Midville	0.7	1.4	1.6	17.7
USA	Florida	Monroe	0.3	0.2	0.7	6.9
USA	Florida	Nassau	0.6	1.1	0.6	6.6
USA	Georgia	Plains	1.0	1.1	1.2	44.2
USA	Florida	Sumter	0.7	1.1	0.5	10.9
USA	Alabama	Tallapoosa	0.5	1.4	1.5	19.2
USA	Georgia	Tifton	0.6	0.9	0.7	3.5
UK	Harpenden	Rothamsted	2.5	1.8	1.6	1.3

*Not available.

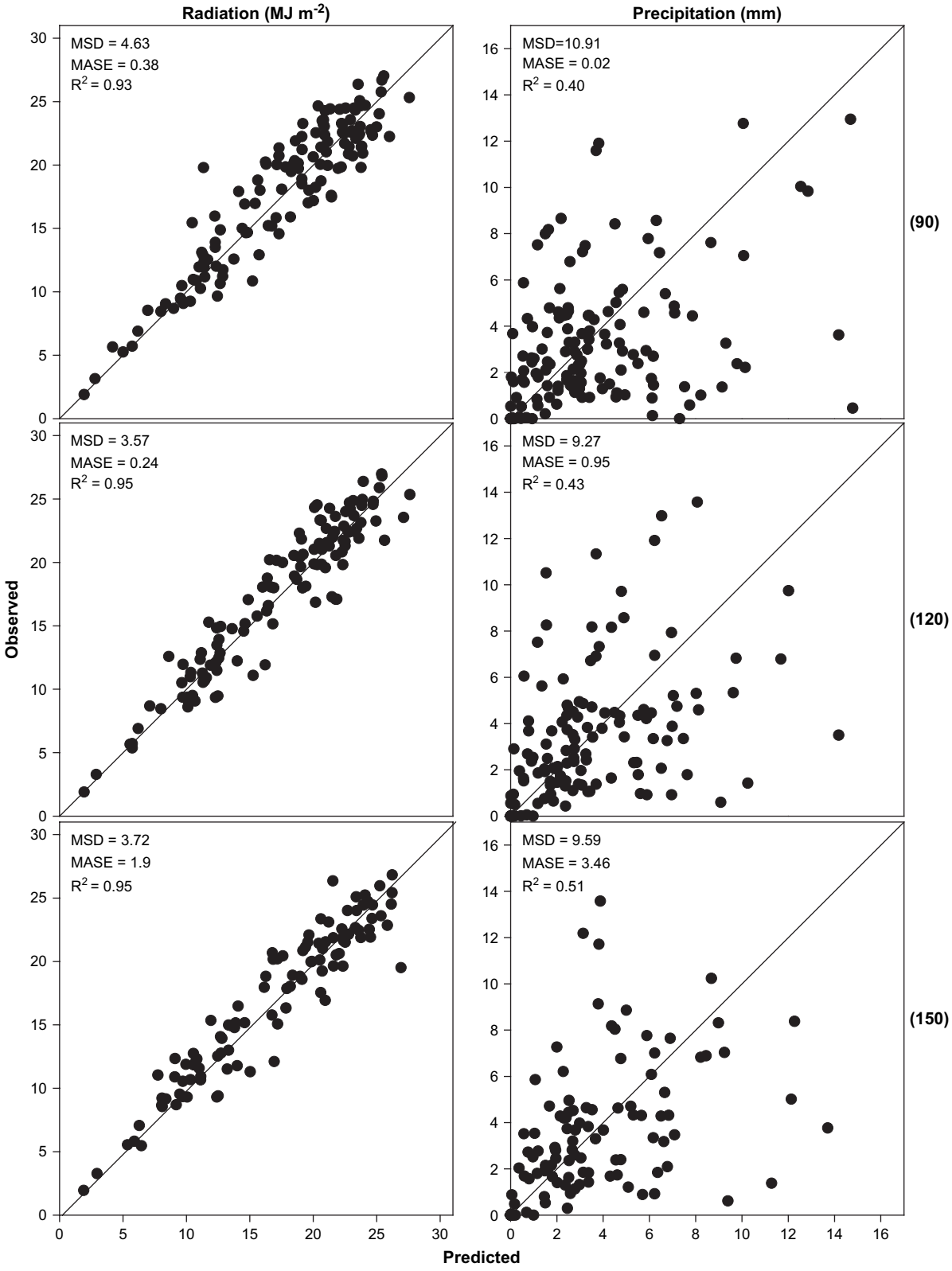

Fig. 4. Mean monthly observed and predicted solar radiation and precipitation across all study sites when the target year contained 90, 120, and 150 observed days.

days as observations in the target year would reduce the statistical mismatch of predicted data vs. the target year. As the number of observed weather days increased from 90 to 120 days, the MSD values of prediction decreased for all weather variables. However, MSD values were slightly higher for minimum temperature, solar radiation, and precipitation using 150 vs. 120 days of observed data. This could be due to finding a different *best matching* year when using a different amount

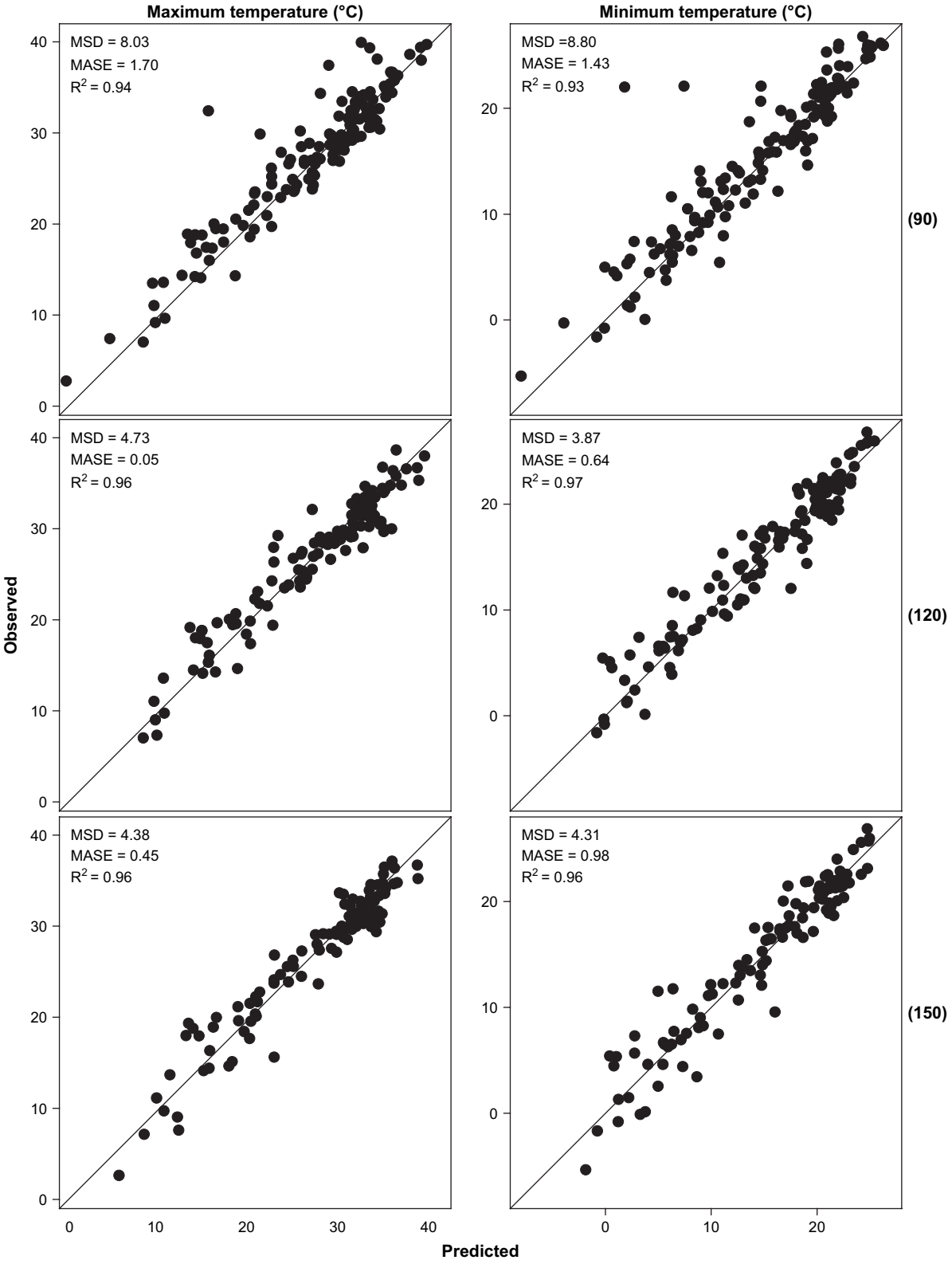

Fig. 5. Mean monthly observed and predicted maximum and minimum temperature across all study sites when the target year contained 90, 120, and 150 observed days.

of observed weather data in the target year. However, the correlation between predicted and observed data increased by increasing the share of observed data in the target year. A comparison of observed and predicted data for the months of July and December across all study sites when the target year contained 120 days of observed data is shown in Fig. 6. These two months were chosen to represent different summer and winter weather conditions in a year for all sites. Although

Table 3
Annual observed (Obs) and predicted (Pre) mean and standard deviation (SD) for various weather variables when the target year contained 120 observed days

Country	State/province/town	Site	Radiation (MJ m^{-2})				Precipitation (mm)				T_{max} (°C)				T_{min} (°C)			
			Mean		SD		Mean		SD		Mean		SD		Mean		SD	
			Obs	Pre	Obs	Pre	Obs	Pre	Obs	Pre	Obs	Pre	Obs	Pre	Obs	Pre	Obs	Pre
USA	Alabama	Baldwin	16.0	16.5	6.1	6.2	6.2	5.7	19.0	14.2	27.0	28.1	6.1	7.3	16.6	15.8	6.8	7.9
Burkina Faso	Bale	Boromo	19.4	20.0	4.8	3.5	4.0	2.9	11.5	7.0	34.2	33.9	3.5	3.8	21.1	21.2	3.2	3.1
USA	Alabama	Covington	16.6	16.9	6.8	5.7	6.1	6.1	19.2	18.9	27.4	28.2	6.1	7.4	13.4	15.5	7.5	7.3
USA	California	Davis	19.2	18.3	8.6	9.1	1.3	1.0	6.2	3.9	27.5	26.4	8.1	9.7	10.5	10.3	4.2	5.1
Burkina Faso	Mouhoun	Dedougou	21.3	21.3	5.2	3.2	3.6	3.5	8.9	10.9	34.2	35.3	3.4	3.5	22.0	21.8	3.1	2.9
USA	Alabama	Limestone	16.5	16.6	6.5	6.5	4.2	3.6	11.5	9.9	25.3	24.6	7.5	9.3	12.1	12.0	7.8	8.7
Iran	Khorasan	Mashhad	*	*	*	*	0.4	0.6	2.1	2.3	25.7	26.0	9.6	10.0	11.5	12.3	7.9	7.6
USA	Michigan	Michigan	14.9	16.5	8.9	9.2	2.6	2.3	6.9	6.7	20.5	18.8	11.3	10.2	8.5	6.9	9.1	9.3
USA	Georgia	Midville	18.4	17.0	7.0	6.5	2.3	3.3	7.9	10.6	27.6	27.0	6.9	7.3	14.8	13.9	7.4	7.8
USA	Florida	Monroe	19.7	19.3	5.5	5.6	2.5	4.6	7.4	18.8	29.3	29.4	2.9	3.9	23.9	22.4	3.6	4.0
USA	Florida	Nassau	17.2	16.7	6.6	6.8	4.7	4.7	11.7	11.9	27.0	29.0	5.2	6.0	18.5	18.7	5.7	6.1
USA	Georgia	Plains	18.5	17.5	7.5	6.2	3.7	2.6	11.9	7.0	27.8	27.0	6.9	7.6	14.9	13.4	7.1	7.9
USA	Florida	Sumter	17.2	17.9	6.5	6.2	4.7	4.6	12.8	12.1	28.9	29.3	4.4	5.0	16.9	15.3	6.1	5.9
USA	Alabama	Tallapoosa	17.2	16.9	6.3	6.4	3.9	4.2	10.5	15.5	26.6	27.6	6.7	6.9	12.8	14.4	7.3	7.2
USA	Georgia	Tifton	18.3	17.0	6.8	5.7	2.5	3.2	7.8	8.8	27.7	27.7	6.2	8.3	15.9	15.5	6.8	7.6
UK	Harpenden	Rothamsted	11.6	10.7	8.0	7.5	2.0	2.1	4.5	4.7	16.4	15.5	6.0	5.0	8.4	7.0	4.4	4.0

T_{max}: maximum temperature, T_{min}: minimum temperature, *: not available.

the prediction of solar radiation and maximum and minimum temperature was reasonably matched with the observed data, precipitation was slightly overestimated especially for July when compared to December. However, across all sites the MASE, except for minimum temperature, was lower in December when compared to July. The overestimation of precipitation in July was for the sites where the variation of predicted annual precipitation (e.g. standard deviation) was lower compared to the target year for the same site. Further analysis of data showed those sites with higher MSD values for precipitation contained heavy daily precipitation for a certain day, including Boromo with 119.1 mm on day of year 113, Baldwin with 228 mm on day of year 250, and Covington with 189.5 mm on day of year 140. Our results showed that those sites with lower MSD values do not contain any daily precipitation that was higher than 40 mm in both the target year and the historical data.

The comparison of the *k*-mean approach with the full observed data in the feature vector of the target year with the second approach that contained only part of observed weather data is encouraging for further work. The *k*-mean approach showed promising for sites where there is no high variation in observed data and can provide a promising estimate of future weather realizations for these types of sites. The *k*-mean approach could be considered a potential approach for estimating missing weather data as well. However, when only part of the observed data is available, prediction of unrecorded weather data requires the second approach. Further work is needed to improve the performance of the second approach employed in the original *k*-NN algorithm. Applying different weight factors for different weather variables or using a larger time scale than daily data should be considered for further analysis, especially for precipitation.

4. Conclusion

In this paper, we describe the development of a tool that predicts daily weather data realizations consisting of solar radiation, maximum and minimum temperature, and precipitation using a modified *k*-nearest neighbor methodology. The tool was evaluated across 16 sites in the USA, Europe, Africa, and Asia. Employing the *k*-mean approach for *k*-NN was promising for the prediction of the weather variables that were included in this study, even for precipitation when there was no heavy precipitation for any given day. Using the second approach, the comparison of observed and predicted data showed that it was possible to predict the unrecorded daily weather data with reasonable accuracy. This program can be used as a standalone tool or can be incorporated into any decision support system for various applications that require realizations of future daily weather data. In addition to observed weather data for the target year, the tool requires observed historical weather data, both in daily format. Obviously, a larger number of historical years of weather data would benefit the accuracy of the model simulation by providing a better chance to find the best matching year.

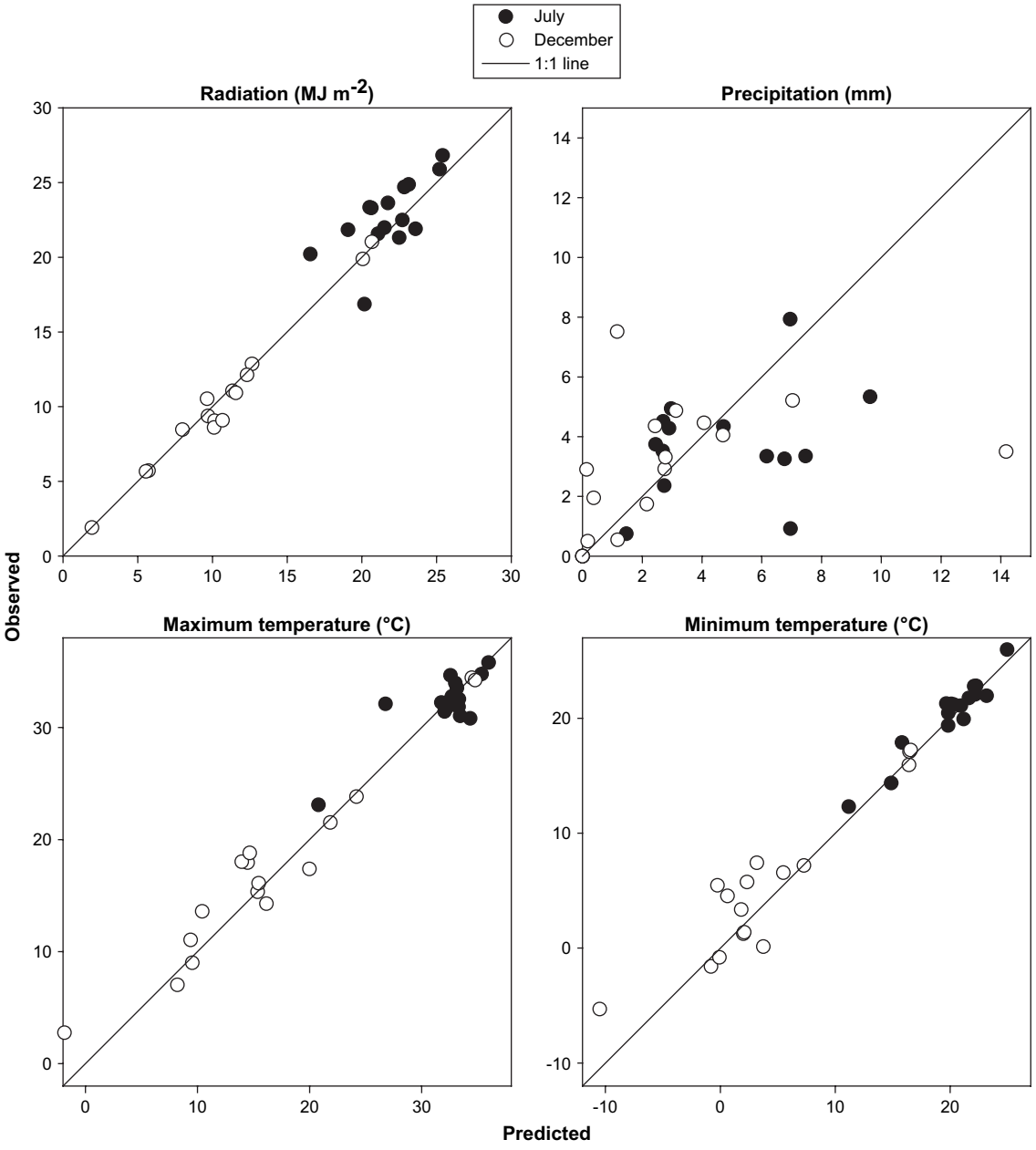

Fig. 6. Mean monthly observed and predicted solar radiation, precipitation, and maximum and minimum temperature across all sites for July and December.

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–Climate Program Office (NOAA–CPO) and the 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. We also thank from Mr. Hamed Shariat Yazdi for help on programing codes.

References

Armstrong, J.S., 2001. Evaluating forecasting methods. In: Armstrong, J.S. (Ed.), Principles of Forecasting: a Handbook for Researchers and Practitioners. Kluwer Academic Publishers, Norwell, MA, pp. 171–192 (Chapter 14).

Bannayan, M., Hoogenboom, G. Daily weather sequence prediction realization using the non-parametric nearest-neighbor re-sampling technique. International Journal of Climatology, in press.

Bannayan, M., Crout, N.M.J., 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., Konnen, G.P., 2006. Application of nearest-neighbor resampling for homogenizing temperature records on a daily to sub-daily level. International Journal Climatology 26, 75–89.

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

Buishand, T.A., 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, G.A., Nihan, N.L., 1991. Nonparametric regression and short-term freeway traffic forecasting. Journal of Transportation Engineering, ASCE 117 (2), 178–188.

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

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

Goel, S., Dash, S.K., 2007. Response of model simulated weather parameters to round-off-errors on different systems. Environmental Modelling and Software 22, 1164–1174.

Hartkamp, A.D., White, J.W., 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., 2000. Contribution of agrometeorology to the simulation of crop production and its applications. Agricultural and Forest Meteorology 103 (1–2), 137–157.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688.

Jones, P.G., Thornton, P.K., 2000. Marksim: software to generate daily weather data for Latin America and Africa. Agronomy Journal 92, 445–453.

Jones, J.W., Hansen, J.W., Royce, F.S., Messina, C.D., 2000. Potential benefits of climate forecasting to agriculture. Agricultural, Ecosystem and Environment 82, 169–184.

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

Kilsby, C.G., Jones, P.D., Burton, A., Ford, A.C., Fowler, H.J., Harpham, C., James, P., Smith, A., Wilby, R.L., 2007. A daily weather generator for use in climate change studies. Environmental Modelling and Software 22, 1705–1719.

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.

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

Nicks, A.D., Harp, J.F., 1980. Stochastic generation of temperature and solar radiation data. Journal of Hydrology 48, 1–7.

Oram, P.A., 1989. Sensitivity of agricultural production to climate change, an update. In: Climate and Food Security. IRRI, Manila, Philippines, pp. 25–44.

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

Rackso, P., Szeidi, L., Semenov, M., 1991. A serial approach to local stochastic weather models. Ecological Modelling 57, 27–41.

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

Semenov, M.A., Brooks, R.J., Barrow, E.M., Richardson, C.W., 1998. Comparison of WGEN and LARS-WG stochastic weather generators for diverse climates. Climate Research 10, 95–107.

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

Soltani, A., Hoogenboom, G., 2003. Minimum data requirements for parameter estimation of the stochastic weather generators WGEN and SIMMETEO. Climate Research 25 (2), 109–119.

Todini, E., 2000. Real-time flood forecasting: operational experience and recent advances. In: Marsalek, J., et al. (Eds.), Flood Issues in Contemporary Water Management. Kluwer Academic Publisher, The Netherlands, pp. 261–270.

Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), 1998. Understanding Options for Agricultural Production. Systems Approaches for Sustainable Agricultural Development. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Wu, W., Xing, E.P., Myers, C., Mian, I.S., Bissell, M.J., 2005. Evaluation of normalization methods for cDNA microarray data by *k*-NN classification. Bioinformatics 6, 191.

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, K.C., 1994. A multivariate chain model for simulating climatic parameters from daily data. Journal of Applied Meteorology 33, 661–671.