Prediction of climate variables by comparing the *k*-nearest neighbor method and MIROC5 outputs in an arid environment

Hamid Reza Golkar Hamzee Yazd[1], Nasrin Salehnia[2,*], Sohrab Kolsoumi[2], Gerrit Hoogenboom[3]

[1]Ferdows Branch, Islamic Azad University, PO Box 9771-848664, Ferdows, Iran
[2]Ferdowsi University of Mashhad, PO Box 9177-949207, Mashhad, Iran
[3]Institute for Sustainable Food Systems, University of Florida, PO Box 110570, Gainesville, Florida, USA

ABSTRACT: The goal of this study was to compare the ability of the *k*-nearest neighbors (*k*-NN) approach and the downscaled output from the MIROC5 model for generating daily precipitation (mm) and daily maximum and minimum temperature (T_{max} and T_{min}; °C) for an arid environment. For this study, data from the easternmost province of Iran, South Khorasan, were used for the period 1986 to 2015. We also used an ensemble method to decrease the uncertainty of the *k*-NN approach. Although, based on an initial evaluation, MIROC5 had better results, we also used the output results of *k*-NN alongside the MIROC5 data to generate future weather data for the period 2018 to 2047. Nash-Sutcliffe efficiency (NSE) between MIROC5 estimates and observed monthly T_{max} ranged from 0.86 to 0.92, and from 0.89 to 0.93 for T_{min} over the evaluation period (2006–2015). *k*-NN performed less well, with NSE between *k*-NN estimates and observed T_{max} ranging from 0.54 to 0.64, and from 0.75 to 0.78 for T_{min}. The MIROC5 simulated precipitation was close to observed historical values (−0.06 < NSE < 0.07), but the *k*-NN simulated precipitation was less accurate (−0.36 < NSE < −0.14). For the studied arid regions, the *k*-NN precipitation results compared poorly to the MIROC5 downscaling results. MIROC5 predicts increases in monthly T_{min} and T_{max} in summer and autumn and decreases in winter and spring, and decreases in winter monthly precipitation under RCP4.5 over the 2018–2047 period of this study. This study showed that the *k*-NN method should be expected to have inaccurate results for generating future data in comparison to the outputs of the MIROC5 model for arid environments.

KEY WORDS: RCP4.5 · Statistical downscaling · Delta method · Ensemble · LARS-WG · Lut Desert

1. INTRODUCTION

The changing climate over recent decades has resulted in extensive effects on society and natural systems. Anthropogenic greenhouse gas emissions have risen since the pre-industrial age, driven mainly by an increase in population and economic growth (IPCC 2014). Both precipitation and temperature averages and variation are directly affected by these increased emissions. Furthermore, there is a strong linkage between the changes in daily weather conditions and hazardous events such as droughts, wildfires, floods, and other phenomena (Salehnia et al. 2018). Monitoring and assessment of these changes will play a vital role in making robust decisions about water allocation for use in agriculture, water resource management, and management of climate risk (Kharin et al. 2007). Evaluation of future climate change will increase our knowledge and assist in developing adaptive management programs to limit impacts on agriculture (Ullah al. 2018). General circulation models (GCMs) and their outputs are one

*Corresponding author: salehnia61@gmail.com

of the most useful tools for estimating future climates and measuring their changes (Mandal et al. 2016, Faiz et al. 2018). Alternatively, non-parametric weather generators such as *k*-nearest neighbors (*k*-NN) have also been used for the generation of future data (Eum et al. 2010, King et al. 2015).

One of the weaknesses of the GCM projections is their coarse spatial scale (>100 km^2). GCM outputs cannot capture the impacts of climate change at a local scale, and this native resolution limits the representation of mesoscale processes. Therefore, downscaling is used to convert low spatial resolution GCM output to high spatial resolution climate variables. There are 2 general approaches for downscaling: statistical downscaling (SD) and dynamical downscaling (DD). DD requires initial boundary conditions and additional details to create local-scale predictions. Therefore, it is time-consuming, costly, and prone to error (Maurer & Hidalgo 2008, Fita et al. 2017).

SD utilizes the statistical relationship between a local station weather data (the predictand) and the GCM data output (the predictors). The many iterations required by the SD method are easily run by a simple computer with one core, so it is comparatively computationally inexpensive and efficient (Hellström et al. 2001, Cavazos & Hewitson 2005, Boé et al. 2007). SD is broadly categorized into linear methods, stochastic weather generators, and weather pattern-based approaches (Semenov & Barrow 1997, Hanssen-Bauer et al. 2005, Vrac et al. 2007). Linear SD methods include various sub-methods such as simple and multiple linear regression, the delta method, and canonical correlation analysis (CCA). The stochastic weather generator SD category is exemplified by the Long Ashton Research Station Weather Generator (LARS-WG) (Semenov & Barrow 1997), the WGEN (Weather GENerator) model, MarkSim GCM and *k*-NN. The weather pattern-based SD category includes the analog method, cluster analysis, and the artificial neural network (ANN) (Wu et al. 2012). Many studies have compared a large number of SD methods during the past two decades, including investigation of the impact of climate change on hydrology and water resources (Fowler et al. 2007, Chen et al. 2011a,b), producing weather data such as precipitation and temperature, and generating extreme values of climate variables.

Many recent studies have compared various SD methods to determine the best approach for a specific region. For example, Wilby & Wigley (1997) compared the performance of 6 statistical downscaling methods (e.g. regression-based methods, ANN, and weather generators) for downscaling daily precipitation in

North America. Their results showed that regression-based methods provided better assessments. Bürger et al. (2012) compared 5 statistical downscaling methods for temperature and precipitation extremes in western Canada and found that the weather pattern-based approach performed best. For British Columbia, Canada, SD methods have been used (Mandal et al. 2016). Several studies have compared SD approaches for Europe, including Haylock et al. (2006) for northwest and southeast England, Piani et al. (2010) across Europe, Boé et al. (2007) for northwestern France, Maraun (2013) for central northern Germany, and Schmidli et al. (2007) for the European Alps.

A promising nonparametric technique for generating weather data is the *k*-NN resampling approach. Forecasting weather data through analogue approaches has been applied in several studies, including Lorenz (1969), Barnett & Preisendorfer (1978), and Shabbar & Knox (1986). Van den Dool (1994) assessed these methods in his research over the USA. Young (1994) employed *k*-NN to generate temperature and precipitation values for Tucson and Safford, Arizona, USA. Lall & Sharma (1996) assessed and further developed this method. The delta method bias correction approach is a downscaling method that has been used for assessing extreme rainfall estimation (Sunyer et al. 2012), estimation of future discharge (Lenderink et al. 2007), and hydrological impact studies (Fowler et al. 2007, Dessu & Melesse 2013).

One of the main questions raised here for generating future data is to determine whether the *k*-NN method or CMIP5 downscaled data is more precise for generating future weather data. The main objectives of the present study were to: (1) improve the result of *k*-NN through the ensemble method and investigate its use for generating weather variables in arid regions, (2) compare the relative performance of the *k*-NN method with MIROC5 downscaled (through the delta method) outputs for generating weather variables in arid environments, and (3) develop a user-friendly tool for applying *k*-NN ensemble outputs.

2. MATERIALS AND METHODS

2.1. Study area and climate data

The study area is located in the easternmost province of Iran, South Khorasan. It lies between 30° 35′ and 34° 14′ N latitude and 57° 46′ and 60° 57′ E longitude, an arid location with an area of 151 193 km^2.

The maximum annual temperature is 44°C, and the lowest recorded temperature is −5.21°C. Because Southern Khorasan is located in the desert climate, rivers are seasonal. The Lut Desert, one of the driest and hottest deserts in the world, is located in the immediate vicinity of this province and affects its climate. The average annual precipitation in South Khorasan is 134 mm and the average annual temperature is 17.5°C. Historical daily weather data of 4 different meteorological station locations (Birjand, Ferdows, Tabas, and Nehbandan) were collected from 1986 to 2015. These data include maximum and minimum temperature (°C), precipitation (mm), wind speed (m s^{-1}), humidity (%), and sunshine hours (h). The prediction period was 2018–2047 and the evaluation period was 2006–2015. The geographical location of the stations is shown in Fig. 1, and the physiographic details of locations are presented in Table 1.

2.2. Statistical downscaling method and the MIROC5 model

We used the outputs of the GCM Model for Interdisciplinary Research on Climate (MIROC5; Watanabe et al. 2010) using Representative Concentration Pathways (RCP) 4.5. We applied the delta method for statistical downscaling on MIROC5 under RCP4.5. The characteristics of the MIROC5 model are presented in Table 2. The MIROC5 model has been widely used (Fettweis et al. 2013, Gaetani & Mohino 2013, Harding et al. 2013, Hsu et al. 2013, Chen & Frauenfeld 2014). We choose the RCP4.5 scenario because many applications are based on this scenario (Hsu et al. 2013). RCP4.5 represents a 'medium' scenario under radiative forcing of 4.5 W m^{-2} by 2100. The CO_2-equivalent concentration in the year 2100 for RCP4.5 is 538 ppm, the emissions in RCP4.5 peak around 2040 (Meinshausen et al. 2011) and then
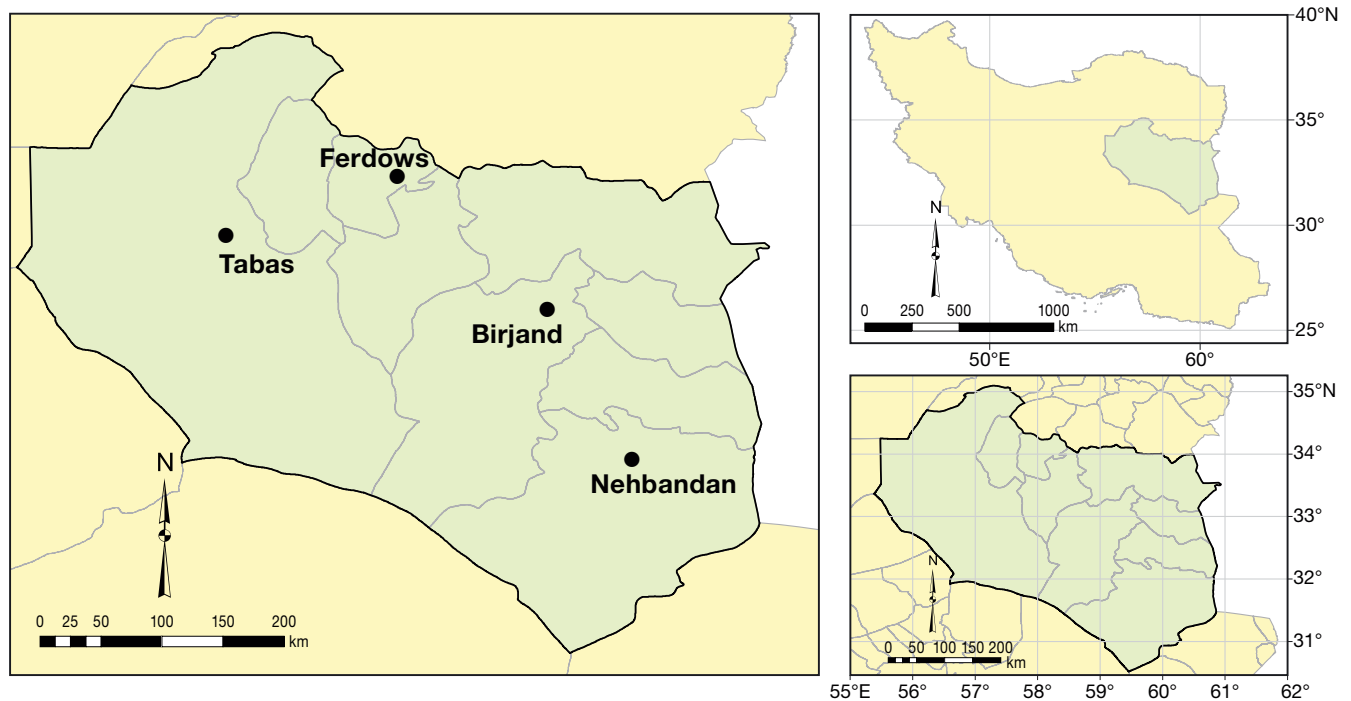

Fig. 1. Geographical location of Birjand, Ferdows, Nehbandan, and Tabas stations within South Khorasan, Iran

Table 1. Physiographic details of study locations and weather data during the observation period (1986–2015), on a mean annual scale. T_{max}: maximum temperature; T_{min}: minimum temperature

Location	Latitude (°N)	Longitude (°E)	Elevation (m)	Average T_{max} (°C)	Average T_{min} (°C)	Average total precipitation (mm)	Average no. wet days
Birjand	32°52′	59°12′	1491	24.6	8.5	165	38
Ferdows	34°10′	58°10′	1293	24.5	10.5	142	37
Nehbandan	36°16′	58°48′	1213	27.2	13.1	119	27
Tabas	33°36′	56°55′	711	29.1	16.4	87	26

Table 2. The characteristics of the MIROC5 model

Modeling center	Spatial resolution	Horizontal resolution (latitude × longitude) (deg)
Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology, Japan	128 × 256	1.4062 × 1.4008

decline. As we have projected data for ~30 yr into the future (2018–2047), we decided to apply RCP4.5 for the evaluation of reliable outputs.

The delta method was selected because it is the most widely used method for downscaling GCM outputs (Maraun et al. 2010, Themeßl et al. 2011). This method was used to derive downscaled weather data, computing the ratio between averaged observation and GCM-simulated climate and multiplying this ratio by the GCM data. It is calculated as a separate equation for precipitation and temperature (Eqs. 1 & 2):

$$P_{SD}^{Delta} = P_{GCMrcp} \times \frac{\overline{P}_{Obs}}{\overline{P}_{GCMhist}} \qquad (1)$$

$$T_{SD}^{Delta} = T_{GCMrcp} + \left(\overline{T}_{Obs} - \overline{T}_{GCMhist} \right) \qquad (2)$$

where P_{SD}^{Delta} and T_{SD}^{Delta} are downscaled data of precipitation and temperature, respectively. $\overline{P}_{Obs}$ denotes the mean observed (obs) precipitation and $\overline{P}_{GCM}$ refers to GCM precipitation data over a future period. The subscripts GCMrcp and GCMhist represent the GCM outputs under the RCP scenario and the GCM historical outputs, respectively. The subscript Obs represents the observation values. In Eq. (2), all subscripts are the same as in Eq. (1) but for temperature instead of precipitation.

2.3. The *k*-NN method

The *k*-NN method is based on selecting a specified number of days similar in characteristics to the day of interest from the historical record. In other words, *k*-NN has a resampling strategy for generating data on the basis of 'similarity' to the historical period of weather data. The vector of input data consists of *p* variables across each station for each day of the historical record. We selected data from the 4 stations. The choice of the nearest neighbors number (*k*) is essential to ensure good performance of the method (Sharif & Burn 2006).

We ran the *k*-NN method with different numbers of weather variables and we have added this option to software that has been developed as part of this pro-

ject. For each run, the results of the neighbor's vector were maintained. Then the efficiency criteria were applied to all variables and the vector of the most desirable neighbor was selected. We used the *k*-NN method for daily maximum temperature (°C), minimum temperature (°C), and precipitation (mm) data over 4 stations for the period 1986–2015, and applied the vectors to maximum temperature (°C), minimum temperature (°C), precipitation (mm), wind speed (m s⁻¹), humidity (%), and sun hours (h) on a daily scale.

Suppose that the daily historic weather vector includes *p* variables. For example, for *p* = 3, the variables T_{min}, T_{max}, and Pr refer to minimum temperature (°C), maximum temperature (°C), and precipitation (mm), respectively. In this study, the number of stations is equal to 4. Consider that the simulation begins on day *t*, which is considered here to be 1 January. The procedure cycles through various steps to obtain the weather data for day *t* + 1. The process continues for all 365 d of a year and is repeated to produce data for 30 yr (in the present study). For running the *k*-NN method, there are sequential steps, which are listed as follows, according to the variables that were involved in our research:

(1) Create a feature vector for the weather variables (*p* = 3) and for each station and each day of the historical period, 1986–2015, as follows:

$$X_t = [x_{1,t} \ x_{2,t} \ \ldots x_{p,t}] \qquad (3)$$

where X_t refers to the vector of weather variables for day *t* and the selected station. $x_{1,t}$ is the value of weather variable 1 (e.g. precipitation) at time step *t* for the selected station.

(2) Develop a vector for all potential neighbors, called the matrix of neighbors, the length of which is *L*. To compute *L*, a new concept must be defined: *w* refers to a window of chosen width centered on the feature day. We set this value as a 14-d window with a 7-d lag and a 7-d lead based on Yates et al. (2003). For example, we can place a 14-d window centered on 10 January (i.e. 3 January to 17 January), and then select all of the historical day pairs inside the window, over the 30 yr. Now, the data block of

potential neighbors from which to resample consists of $L = (w + 1) \times N - 1$ d. So, in the present study, $L = (14 + 1) \times 30 - 1 = 449$. Rajagopalan & Lall (1999) and Yates et al. (2003) recommended the use of a heuristic method for choosing k, according to which $k = \sqrt{L}$, so the k nearest neighbors compute as $k = \sqrt{L}$, and thus $k = \sqrt{449} = 21$.

With these numbers, the weather on the first day t comprising all $p = 3$ variables at each station is randomly chosen from the set of all 1 January values for $N = 30$ yr during the observation period. The procedure cycles through the following steps to select one of the nearest neighbors to show the weather for day $t + 1$ of the simulation period.

(3) Determine Mahalanobis distances between the mean vector of the current day's weather, $\overline{X}_t$, and the mean vector $\overline{X}_i$ for day i, where $i = 1, …, L$. The Mahalanobis distance can be determined as follows:

$$d_i = \sqrt{\left(\overline{X}_t - \overline{X}_i\right) \times C_t^{-1} \times \left(\overline{X}_t - \overline{X}_i\right)^T} \tag{4}$$

where T refers to the transpose operation and C_t^{-1} is the inverse of the covariance of the neighbors' matrix.

(4) Sort the Mahalanobis distances in ascending order and maintain the first k nearest neighbors. Then, the probability distribution should compute to achieve higher weights for the closer neighbors. Determine weights for each of the j neighbors according to the cumulative probability function:

$$p_i = \sum_{i=1}^{j} \left(\frac{1/j}{\sum_{i=1}^{K} 1/i} \right) \tag{5}$$

(5) The weather on the feature day in the simulation period (i.e. the day $t + 1$) is selected among the k nearest neighbors using Eq. (5). Then, generate a random number (r) as $r \subset (0,1)$, and if $p_1 < r < p_k$, then the day j for which r is closest to p_j is selected. If $r \leq p_1$, the day corresponding to d_1 (Eq. 4) is selected, and if $r \geq p_k$, then the day corresponding to d_k is selected. Finally, steps 3 to 5 are repeated to produce synthetic data for the required number of years. In this study, these steps are repeated to generate 30 yr in the first evaluation; for the second evaluation during 2006–2015, the k-NN method is performed again.

The k-NN method — as a stochastic and nonparametric technique — has remarkable advantages: despite it having several steps to run, the steps are not complicated, and they are simple and easy to implement; in addition, there is no normality assumption for the residuals. However, this approach only works well with a large sample size (Bannayan & Hoogenboom 2008).

Further details on the k-NN model can be found in Sharif & Burn (2006). The stochastic processes in weather generators create random outputs and the climate might not be represented adequately if only one outcome is used. Since the k-NN approach is a stochastic weather generator, we used an ensemble method to decrease the uncertainty of the outputs. We applied 'weighted averaging' for the ensemble method; weighted averaging obtains the combined output by averaging the outputs of an individual run of the k-NN method with different weights implying different importance. Specifically, weighted averaging gives the combined output $H(x)$ as Eq. (6):

$$H(x) = \sum_{i=1}^{T} w_i \times h_i(x) \tag{6}$$

where w_i is the weight for the run h_i, and the weights w_i are usually assumed to be constrained by Eq. (7):

$$w_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{T} w_i = 1 \tag{7}$$

where T is the number of iterations or runs of the k-NN method. For further details of the calculation, see Zhou (2012).

2.4. Efficiency criteria

To assess the relative performance of the downscaling and k-NN methods against observation data, we used root mean square error (RMSE), Pearson's correlation coefficient (r), the Nash-Sutcliffe coefficient of efficiency (NSE), and mean absolute error (MAE) on a monthly scale, computed as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(OBS_i - SIM_i)^2}{n}} \tag{8}$$

$$r = \frac{\sum_{i=1}^{n}\left(OBS_i - \overline{OBS}\right) \times \sum_{i=1}^{n}\left(SIM_i - \overline{SIM}\right)}{\sqrt{\sum_{i=1}^{n}\left(OBS_i - \overline{OBS}\right)^2 \times \sum_{i=1}^{n}\left(SIM_i - \overline{SIM}\right)^2}} \tag{9}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(OBS_i - SIM_i)^2}{\sum_{i=1}^{n}\left(OBS_i - \overline{OBS}\right)^2} \tag{10}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|(SIM_i - OBS_i)|}{n} \tag{11}$$

where SIM and OBS refer to 'simulated or predicted data' and 'observation data', respectively; n is the

total number of pairs of simulated and observed data; i is the ith value of the simulated and observed data; and $\overline{SIM}$ and $\overline{OBS}$ are the mean values of simulated and observed data, respectively.

RMSE evaluates the average error magnitude between simulated and observed data. Pearson's correlation was used to determine the agreement between simulated and observed data. MAE measures the average magnitude of the errors in a set of predictions but is less sensitive to extreme values than RMSE. NSE was used to quantify how well the plot of observed versus simulated data fits the 1:1 line. For a perfect model, NSE is 1.

2.5. Software

In this project, the k-NN-WG software package (AgriMetSoft 2017; https://agrimetsoft.com/KNN-WG.aspx) was developed using C#. In this package, users can generate weather variables through k-NN and compare outputs with those of other methods in a user-friendly interface. The tool provides a suite of efficiency criteria methods for the evaluation process. Furthermore, the users can plot the baseline data graph versus k-NN's outputs. Further information is available in the Supplement at www.int-res.com/articles/suppl/c077p099_supp.pdf.

3. RESULTS AND DISCUSSION

In this study, the ability of the k-NN method versus MIROC5 downscaled data under RCP4.5 for generating weather data was compared with observation data. Then, the methods were assessed with a common base of observation data to check their performance through efficiency criteria. Finally, the outcomes of the 2 methods were compared in terms of their utility in predicting weather data for future decades. The weather data characteristics of data generated by the 2 methods were compared with historical precipitation, minimum temperature, and maximum temperature data on a daily time scale, and the results are presented on a monthly scale.

3.1. Analyzing k-NN results

In the first step, k-NN was run with different weather variables, including maximum and minimum temperature (°C), precipitation (mm), wind speed (m s^{-1}), humidity (%), and sunshine hours (h),

to determine the best composition for the neighbor's vector. Statistical indices, including RMSE, MAE, r, and NSE, were calculated to compare the k-NN outputs with the weather data from 4 synoptic stations. These results are presented in Table 3 on a monthly time scale. For T_{min} and T_{max} at all stations, the vector of neighbors with 6 variables was the best selection. For example, the best result of efficiency criteria of T_{max} between station data and k-NN outputs was for the Birjand station, with RMSE of 5.70°C mo^{-1}, MAE of 3.18°C mo^{-1}, r of 0.81, and NSE of 0.58 across all 6 variables. There was a high correlation between observed station data and k-NN data for T_{max} (r = 0.71 to 0.83) and T_{min} (r = 0.74 to 0.90) for all locations, while precipitation showed a much lower correlation (r = 0.15 to 0.40; Table 3). Precipitation amounts on a given day are notably highly variable when compared to similar days in past years due to the arid climate, so analyzing and generating precipitation amount can be difficult. Table 3 shows that for precipitation no combination of neighbor's vectors across the 4 stations was acceptable. However, the 6 variables of the neighbor's vector were selected to generate precipitation predictions.

3.2. Comparison of simulation and observation outputs in the evaluation process

k-NN and MIROC5 downscaled outputs were plotted against observed T_{max}, T_{min}, and precipitation for the observation period. k-NN and MIROC5 had very similar results for T_{max} and T_{min} values from different study stations for the evaluation period from 2006 to 2015 (Figs. 2 and 3). Clearly, both methods are capable of closely tracking monthly variations in T_{max} and T_{min}. The statistics in Table 4 show the comparison between the baseline record and the simulated record of weather variables, and reveal that MIROC5 is slightly better than k-NN for both temperature variables. For the 4 stations, MIROC5 T_{max} data versus observed T_{max} (at monthly time scale) had an NSE > 0.86, RMSE < 3.36°C mo^{-1}, and MAE < 2.87°C mo^{-1}. For MIROC5 output T_{min} versus observed T_{min}, NSE ranged from 0.89 to 0.93; for k-NN, the NSE ranged from 0.75 to 0.78. This agrees well with previous studies that have shown that k-NN has the potential to accurately generate temperature data (Yates et al. 2003). Statistical criteria such as MAE show similar patterns, implying that MIROC5 has better results than k-NN for T_{max} and T_{min}. Therefore, MIROC5 is a better choice than k-NN for future temperature predictions in the study region (Table 4); however, we

Table 3. Results of different efficiency criteria, used to create the matrix of different neighbors of weather variables, between historical station data and the *k*-NN method for the period 1986–2015 on a daily scale. T_{max}: maximum temperature (°C); T_{min}: minimum temperature (°C); Pr: precipitation (mm); Wd: wind speed (m s^{-1}); Hm: humidity (%); Sh: sun hours (h). RMSE: root mean square error; MAE: mean absolute error; r: Pearson's correlation coefficient; NSE: Nash-Sutcliffe coefficient of efficiency

Data	Simulated variable	Birjand				Ferdows				Tabas				Nehbandan			
		RMSE	MAE	r	NSE	RMSE	MAE	r	NSE	RMSE	MAE	r	NSE	RMSE	MAE	r	NSE
T_{min}	T_{max}, T_{min}, Pr	5.32	4.41	0.74	0.53	5.23	4.09	0.78	0.58	5.43	3.86	0.83	0.66	5.39	4.29	0.81	0.63
	T_{max}, T_{min}, Pr, Sh	5.03	4.08	0.80	0.58	4.88	3.76	0.83	0.63	5.15	3.55	0.86	0.68	5.01	3.76	0.85	0.68
	T_{max}, T_{min}, Pr, Hm	4.38	3.37	0.84	0.68	4.58	3.39	0.84	0.68	4.82	3.16	0.87	0.72	4.57	3.27	0.88	0.73
	T_{max}, T_{min}, Pr, Wd	4.66	3.70	0.82	0.64	4.40	3.18	0.85	0.70	4.93	3.29	0.86	0.71	4.57	3.27	0.88	0.73
	T_{max}, T_{min}, Pr, Sh, Hm	4.21	3.26	0.87	0.70	4.29	3.08	0.87	0.72	4.70	2.95	0.88	0.74	4.48	3.17	0.89	0.74
	T_{max}, T_{min}, Pr, Sh, Wd	4.26	3.26	0.88	0.70	4.18	3.03	0.88	0.73	4.71	2.94	0.88	0.73	4.48	3.14	0.89	0.74
	T_{max}, T_{min}, Pr, Hm, Wd	4.14	3.15	0.87	0.71	4.14	2.90	0.88	0.74	4.60	2.80	0.88	0.75	4.42	3.03	0.89	0.75
	T_{max}, T_{min}, Pr, Sh, Hm, Wd	3.88	2.87	0.90	0.75	3.92	2.64	0.89	0.76	4.51	2.66	0.89	0.76	4.19	2.76	0.90	0.78
T_{max}	T_{max}, T_{min}, Pr	6.64	4.60	0.71	0.43	6.86	4.86	0.74	0.48	7.47	4.81	0.75	0.47	6.84	4.28	0.72	0.45
	T_{max}, T_{min}, Pr, Sh	6.36	4.17	0.74	0.48	6.34	4.13	0.78	0.56	7.15	4.29	0.77	0.52	6.59	3.99	0.75	0.49
	T_{max}, T_{min}, Pr, Hm	6.18	3.89	0.76	0.51	6.30	4.10	0.79	0.56	6.97	4.01	0.78	0.54	6.48	3.78	0.75	0.50
	T_{max}, T_{min}, Pr, Wd	6.35	4.12	0.75	0.48	6.10	3.82	0.80	0.59	7.12	4.29	0.77	0.56	6.38	3.61	0.77	0.52
	T_{max}, T_{min}, Pr, Sh, Hm	5.85	3.42	0.80	0.56	6.01	3.62	0.81	0.60	6.85	3.74	0.79	0.56	6.31	3.50	0.78	0.53
	T_{max}, T_{min}, Pr, Sh, Wd	6.06	3.71	0.77	0.53	5.86	3.33	0.82	0.62	6.85	3.79	0.79	0.56	6.27	3.48	0.79	0.53
	T_{max}, T_{min}, Pr, Hm, Wd	5.94	3.52	0.80	0.55	5.92	3.47	0.82	0.61	6.84	3.77	0.79	0.56	6.33	3.56	0.78	0.53
	T_{max}, T_{min}, Pr, Sh, Hm, Wd	5.70	3.18	0.81	0.58	5.74	3.14	0.83	0.64	6.77	3.64	0.80	0.57	6.25	3.37	0.78	0.54
Pr	T_{max}, T_{min}, Pr	18.0	10.0	0.06	−0.44	15.8	8.6	0.17	−0.38	10.0	5.3	0.05	−0.36	17.7	7.9	0.17	−0.24
	T_{max}, T_{min}, Pr, Sh	18.0	10.0	0.08	−0.45	16.0	8.9	0.11	−0.43	10.0	5.3	0.12	−0.37	17.7	8.2	0.08	−0.24
	T_{max}, T_{min}, Pr, Hm	17.7	9.84	0.16	−0.40	16.3	8.9	0.03	−0.45	10.0	5.3	0.07	−0.36	17.7	7.9	0.30	−0.16
	T_{max}, T_{min}, Pr, Wd	18.0	10.0	0.12	−0.44	16.0	8.9	0.13	−0.40	10.0	5.3	0.14	−0.37	17.5	7.9	0.35	−0.21
	T_{max}, T_{min}, Pr, Sh, Hm	18.0	10.0	0.04	−0.44	16.0	8.9	0.06	−0.42	10.0	5.3	0.05	−0.36	17.7	7.9	0.19	−0.24
	T_{max}, T_{min}, Pr, Sh, Wd	18.0	10.0	0.18	−0.44	16.0	8.9	0.08	−0.41	10.0	5.3	0.08	−0.36	17.7	8.1	0.13	−0.24
	T_{max}, T_{min}, Pr, Hm, Wd	17.5	9.60	0.14	−0.40	16.30	8.9	0.07	−0.42	10.0	5.3	0.17	−0.36	17.7	7.9	0.12	−0.22
	T_{max}, T_{min}, Pr, Sh, Hm, Wd	18.0	10.0	0.24	−0.36	16.0	8.9	0.19	−0.37	10.0	5.3	0.24	−0.37	17.0	7.5	0.32	−0.14

used both for projecting weather data for future decades.

Precipitation prediction accuracy for the 4 locations was also assessed (Fig. 4). These outputs show a much wider spread between the model's outputs and observed data in the evaluation period, with much lower correlations for precipitation than for temperature. This is in part due to the stochastic nature of precipitation in arid regions, with many zero values in its daily time series. This limitation has been noted before (e.g. Richardson 1981, Dibike & Coulibaly 2005, Sirangelo et al. 2007). Our case study location is characterized by an arid climate, and it has been affected by climate change, drought, and water scarcity. This should be considered when evaluating the performance of the *k*-NN method. The MIROC5 precipitation outputs underestimated monthly precipitation for some months such as December and overestimated amounts for April (Fig. 4, Table 4). For the remaining months, the MIROC5-simulated precipitation is very close to the observed historical values. The *k*-NN model was unable to reproduce the historical precipitation values, with only a weak correlation between simulated *k*-NN precipitation and observed values; on a monthly time scale, r ranged from 0.19 to 0.32, while NSE ranged from −0.14 to −0.37, RMSE ranged from 10 to 18 mm mo^{-1}, and MAE ranged from 5.3 to 10 mm mo^{-1}.

3.3. Projecting future weather variables using the MIROC5 downscaled data and k-NN

We projected future climate for 30 yr including 2018 to 2047. The period of 1986–2015 was used as the baseline period. The box plots between station data and
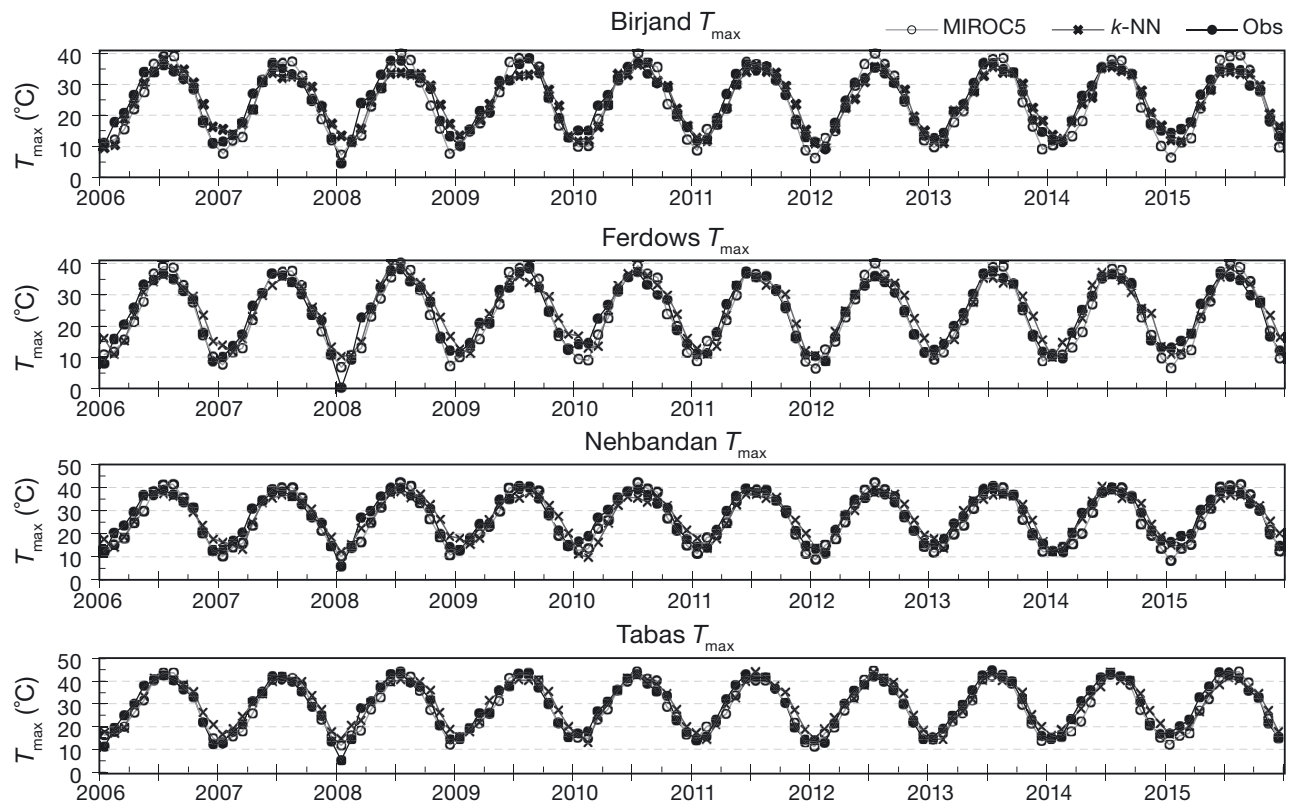
Fig. 2. Comparison between k-NN, MIROC5 and observation (Obs) data for maximum temperature (T_{max}) from 2006 to 2015 on a monthly scale
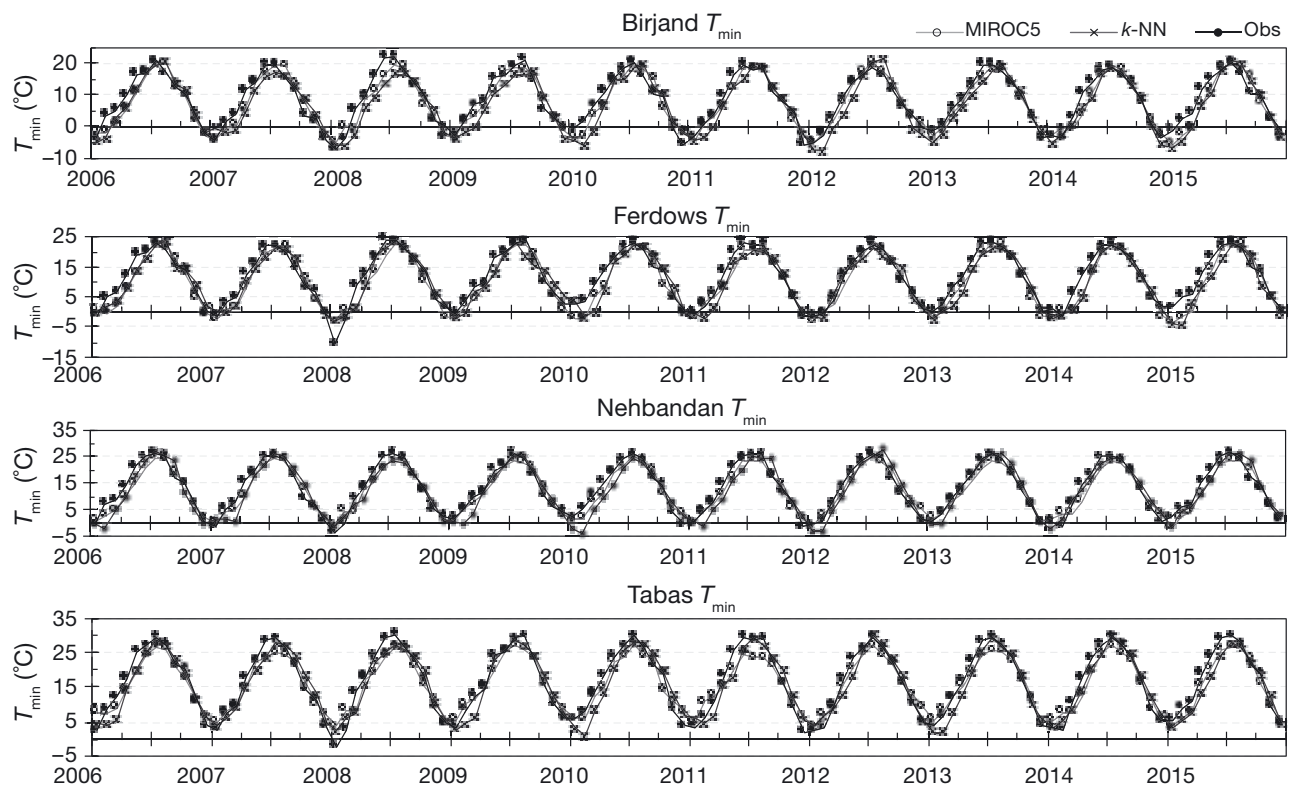

Fig. 3. Comparison between k-NN, MIROC5 and observation (Obs) data for minimum temperature (T_{min}) from 2006 to 2015 on a monthly scale

Table 4. Statistical results between *k*-NN and MIROC5 with observation data over the evaluation period (2006–2015) on a monthly scale. T_{max}: maximum temperature; T_{min}: minimum temperature; RMSE: root mean square error; MAE: mean absolute error; NSE: Nash-Sutcliffe coefficient of efficiency

		T_{max}			T_{min}			Precipitation		
		RMSE	MAE	NSE	RMSE	MAE	NSE	RMSE	MAE	NSE
Birjand	MIROC5	3.36	2.87	0.86	2.59	2.13	0.90	14.7	9.26	0.03
	k-NN	5.70	3.18	0.58	3.88	2.87	0.75	18.0	10.0	−0.36
Ferdows	MIROC5	3.08	2.58	0.90	2.77	2.32	0.89	13.1	8.42	0.00
	k-NN	5.74	3.14	0.64	3.92	2.64	0.76	16.0	8.90	−0.37
Nehbandan	MIROC5	3.31	2.85	0.87	2.50	2.18	0.93	15.7	9.54	0.07
	k-NN	6.25	3.37	0.54	4.19	2.76	0.78	17.0	7.50	−0.14
Tabas	MIROC5	2.98	2.48	0.92	2.84	2.48	0.91	8.92	5.71	−0.06
	k-NN	6.77	3.64	0.57	4.51	2.66	0.76	10.0	5.30	−0.37
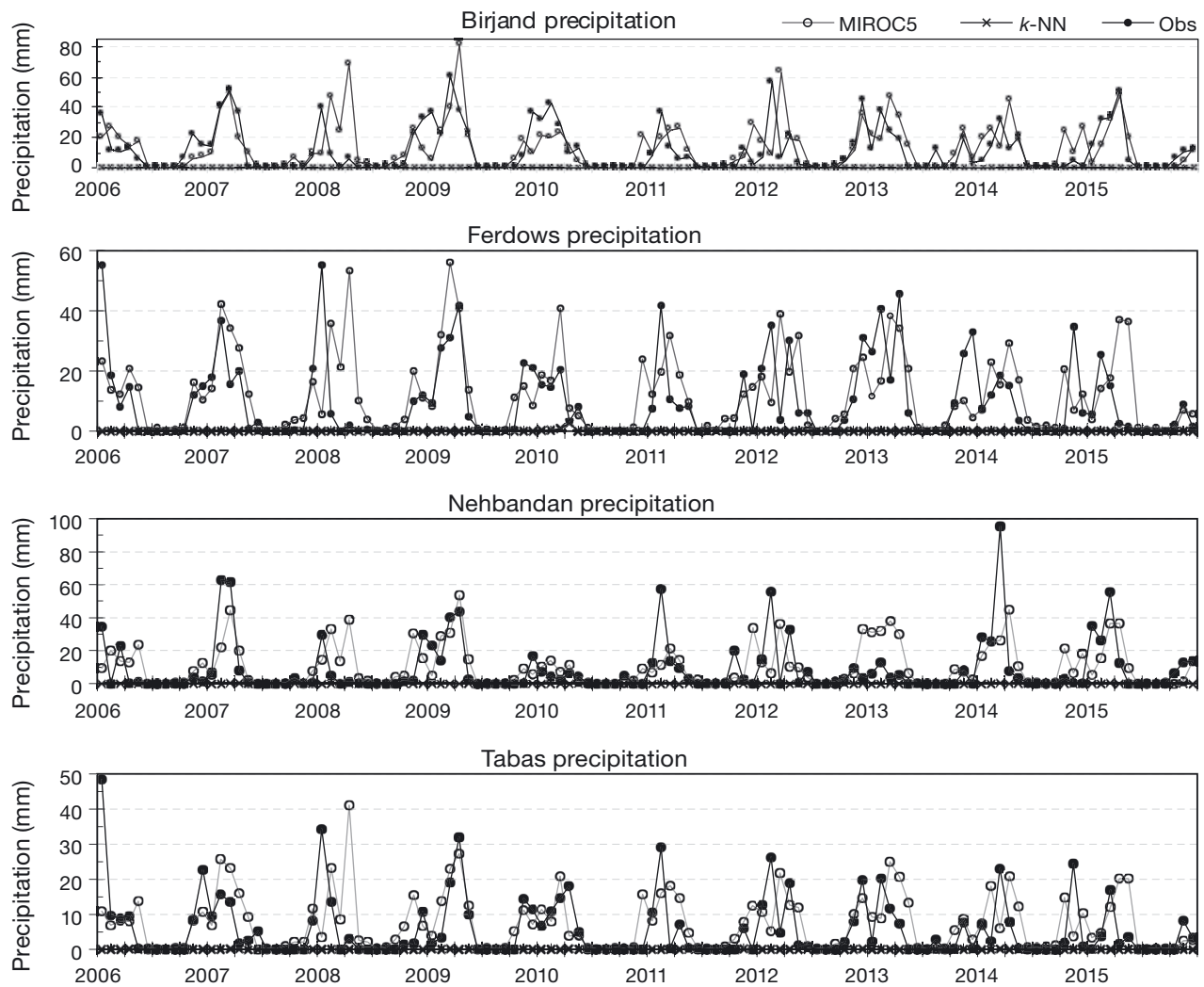
Fig. 4. Comparison between *k*-NN, MIROC5 and observation (Obs) data for precipitation from 2006 to 2015 on a monthly scale

MIROC5 downscaled data under RCP4.5 (Figs. 5–7) show the median and variation in the estimates of T_{min}, T_{max}, and precipitation. Monthly T_{min} is expected to increase during the summer and autumn, and decrease during the winter and spring for Birjand, Ferdows, and Nehbandan (Fig. 5) during the projection period. For Tabas, the driest location (Table 1), the future monthly T_{min} is expected to increase dur-
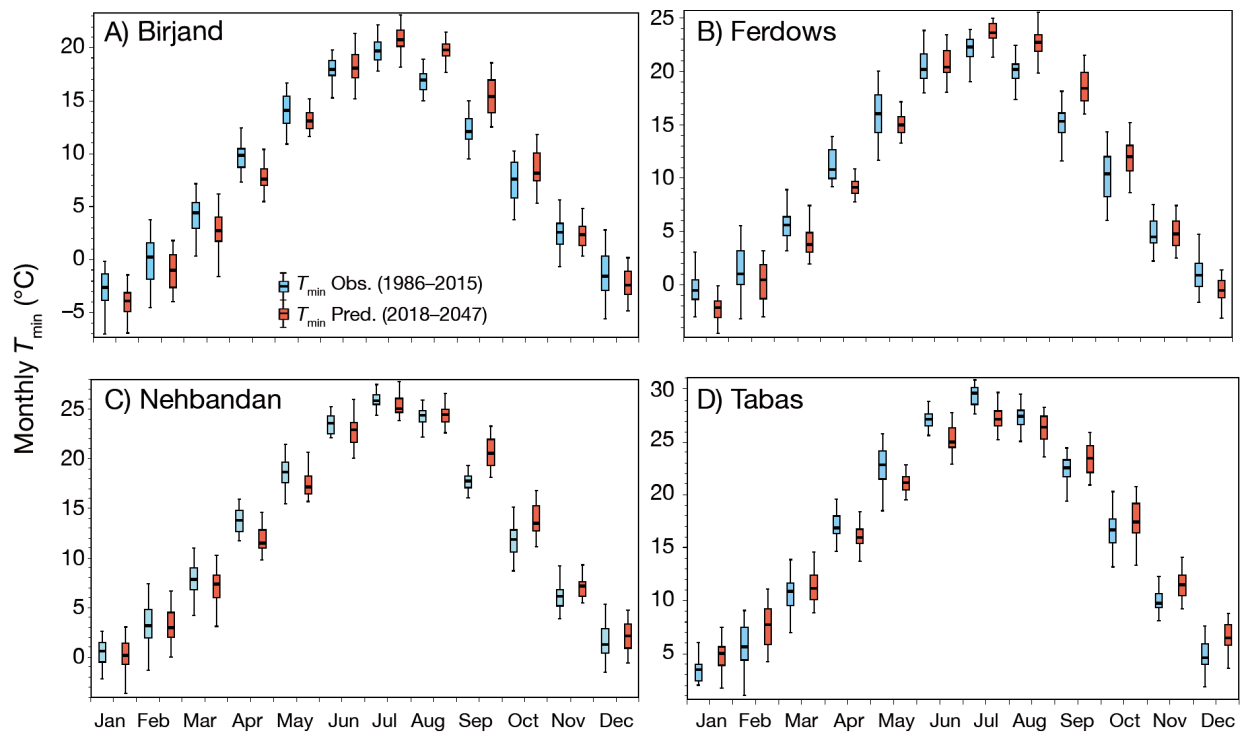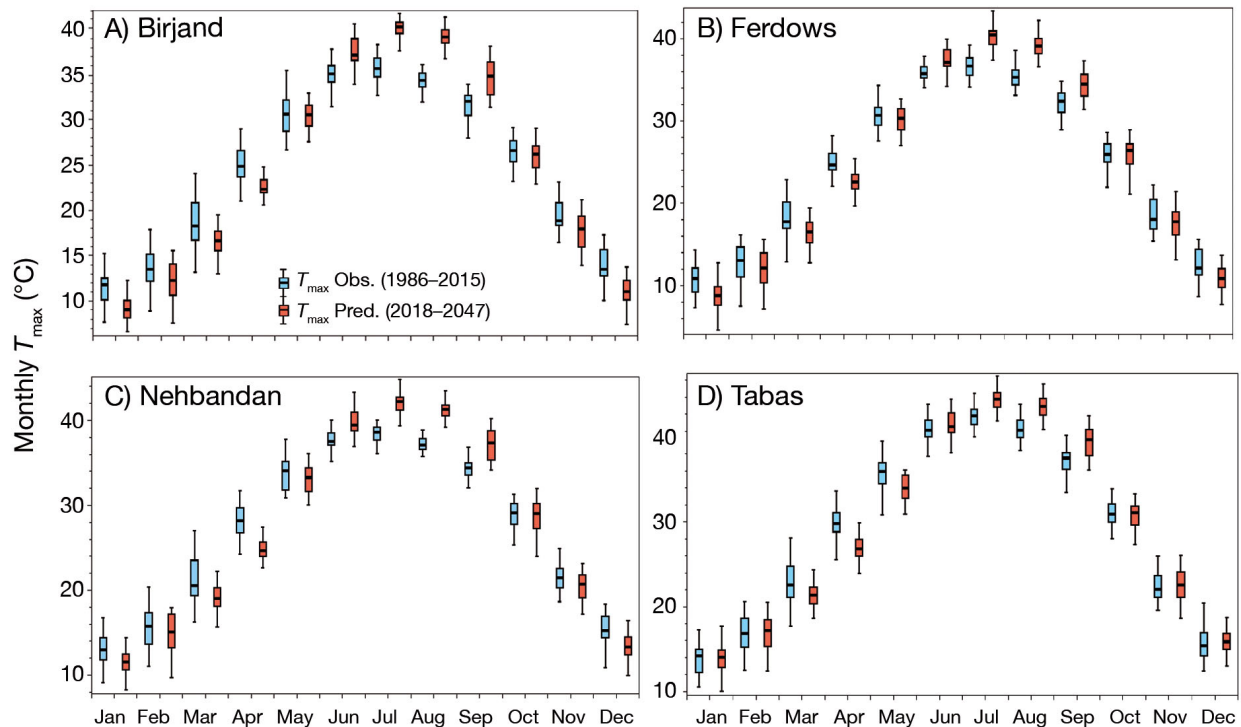
Fig. 5. Boxplots of monthly minimum temperature (T_{min}; °C) from observation (Obs.) versus prediction (Pred.) periods for (A) Birjand, (B) Ferdows, (C) Nehbandan, and (D) Tabas stations. The horizontal line in the middle of the box represents the median, the upper edge of the box represents the 75th percentile (upper quartile), while the lower edge is the 25th percentile (lower quartile). The boxes extend between the 25th to the 75th percentiles (interquartile range), and the whiskers show the 5th and 95th percentiles; points are values outside this range

Fig. 6. Boxplots of monthly maximum temperature (T_{max}; °C) from observation (Obs.) versus prediction (Pred.) periods for (A) Birjand, (B) Ferdows, (C) Nehbandan, and (D) Tabas stations. See Fig. 5 for boxplot definitions
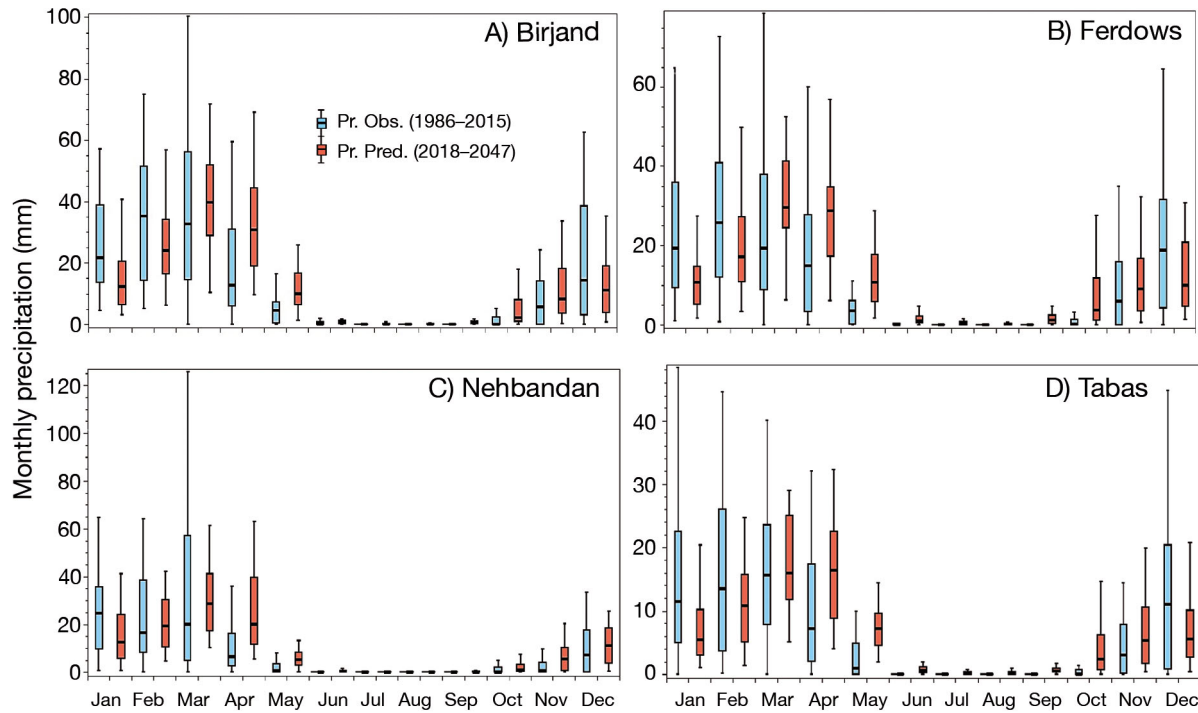
Fig. 7. Boxplots of total monthly precipitation (mm) from observation (Pr. Obs.) versus prediction (Pr. Pred.) periods from (A) Birjand, (B) Ferdows, (C) Nehbandan, and (D) Tabas stations. See Fig. 5 for boxplot definitions

ing the winter and autumn, but decrease during the summer and spring. For Tabas, the maximum monthly T_{min} is predicted to be 29.5°C for July compared to 30.8°C for the baseline period. The change in maximum monthly T_{min} for Birjand, Ferdows, and Nehbandan for the future period relative to baseline is expected to be −0.93°C, −2.5°C, and +0.42°C, respectively.

The T_{max} projections from MIROC5 show trends similar to those of T_{min} (Fig. 6), except that Tabas' trends mirror those of the other 3 sites. For all 4 stations, the monthly T_{max} is expected to decrease for the winter and spring, and increase for summer and autumn. For Tabas, the maximum monthly T_{max} is expected to increase to 47°C for July in the future, compared to 45°C for the baseline period. The changes of the maximum monthly T_{max} for Birjand, Ferdows, and Nehbandan in the future are projected to equal −4.5°C, −4°C, and −4.5°C, respectively.

Under RCP4.5, the downscaled MIROC5 projections of monthly precipitation show a general reduction in winter precipitation and increase in autumn precipitation (Fig. 7). A slight increase is projected for early summer (June and July) and late spring (April and May). In Birjand, the maximum monthly precipitation for the historical period was 127 mm,

and in the future, this was projected to be 83 mm in April, a dramatic decline of 43 mm. Similar declines of 41 mm for Ferdows, 72 mm for Nehbandan, and 44 mm for Tabas were projected.

The precipitation data generated by *k*-NN show a noticeable reduction in comparison to the past 30 yr over the 4 locations (Fig. 8). As we discussed in Section 3.2, we can state that in comparison to MIROC5 under RCP4.5, the *k*-NN method does not have enough ability to generate precipitation data over arid regions. The monthly average maximum temperature is illustrated in Fig. 9; it can be noted that, on average, the range of monthly T_{max} of the *k*-NN method during the future period (2018−2047) is generally lower than that during the observation period (1986−2015). As well, the *k*-NN method showed that the minimum T_{max} is higher than the observed T_{max}, whereas the maximum T_{max} would be lower than observation data, for all 4 stations. Generally, according to the *k*-NN method, monthly average T_{min} did not change considerably during future decades in comparison to the past 30 yr (Fig. 10). The changes generated by *k*-NN are slightly lower than the observation period for T_{min}.

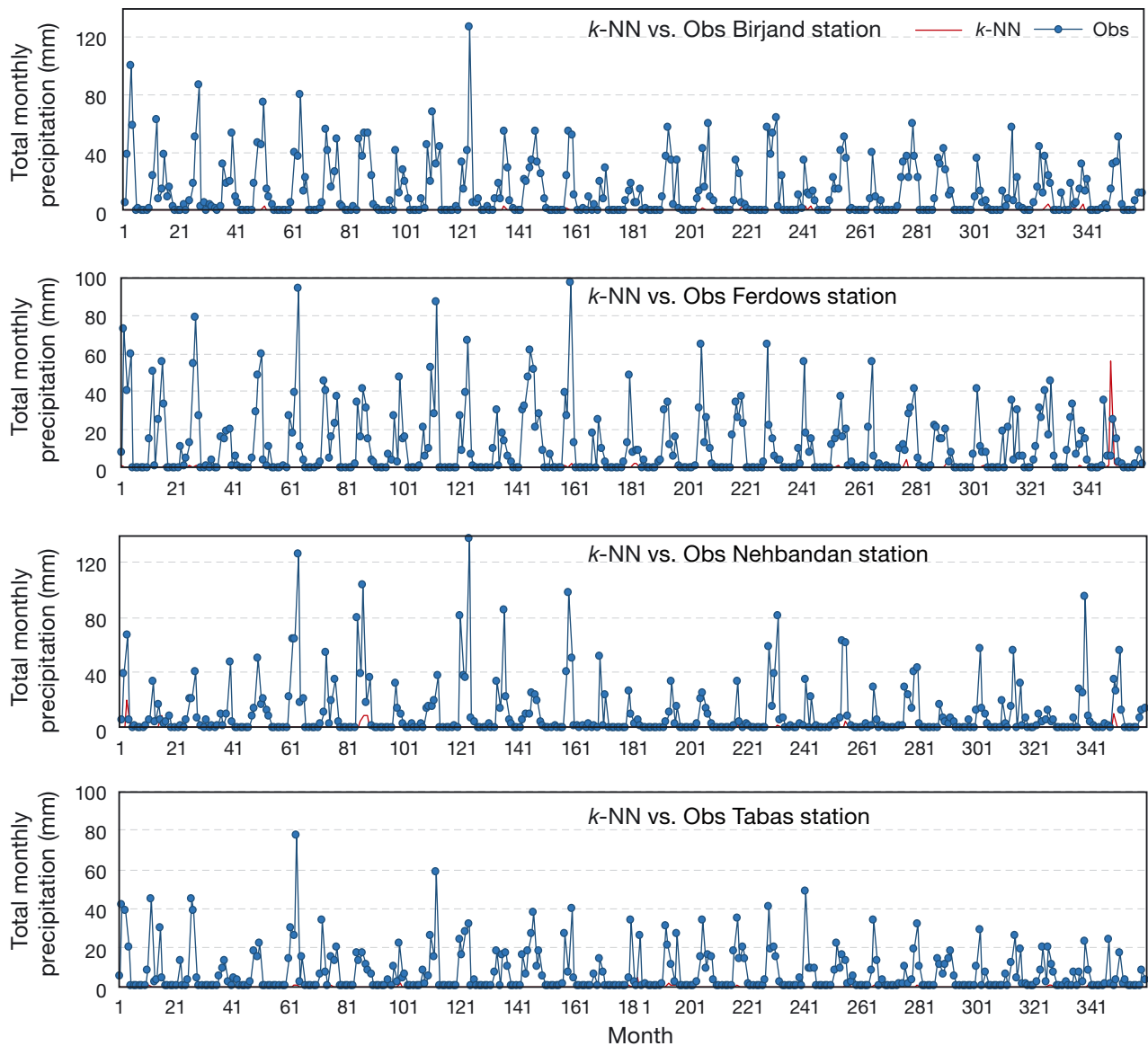This finding demonstrates that the temperature values during future decades are not the same

Fig. 8. Total monthly precipitation (mm) based on observation (Obs) (1986–2015) versus prediction periods (2018–2047) using the
k-NN method (360 mo)

through the MIROC5 downscaled data under RCP4.5
and the *k*-NN approach. Also, the projected precipi-
tation amounts are not the same using the 2 methods;
however, the results of MIROC5 revealed reliable
outcomes in comparison to the *k*-NN method, as the
efficiency criteria showed.

4. CONCLUSIONS

The use of projected weather data is important
when attempting to account for weather and climate
changes in future decades, and this has a significant

effect on various meteorological projects, such as
water allocation, hydrological modeling, hazardous
events, and other environmental research. This study
compared MIROC5 with *k*-NN methods for generat-
ing future weather data projections (2018–2047) rel-
ative to baseline (1986–2015) over arid regions. A
user-friendly software program was presented, for
easy comparison of the *k*-NN method with other
models such as CMIP5. We presented a new method
for decreasing uncertainty in *k*-NN through an
ensemble method. To select the best composition of
neighbor's vectors, *k*-NN was run with a different
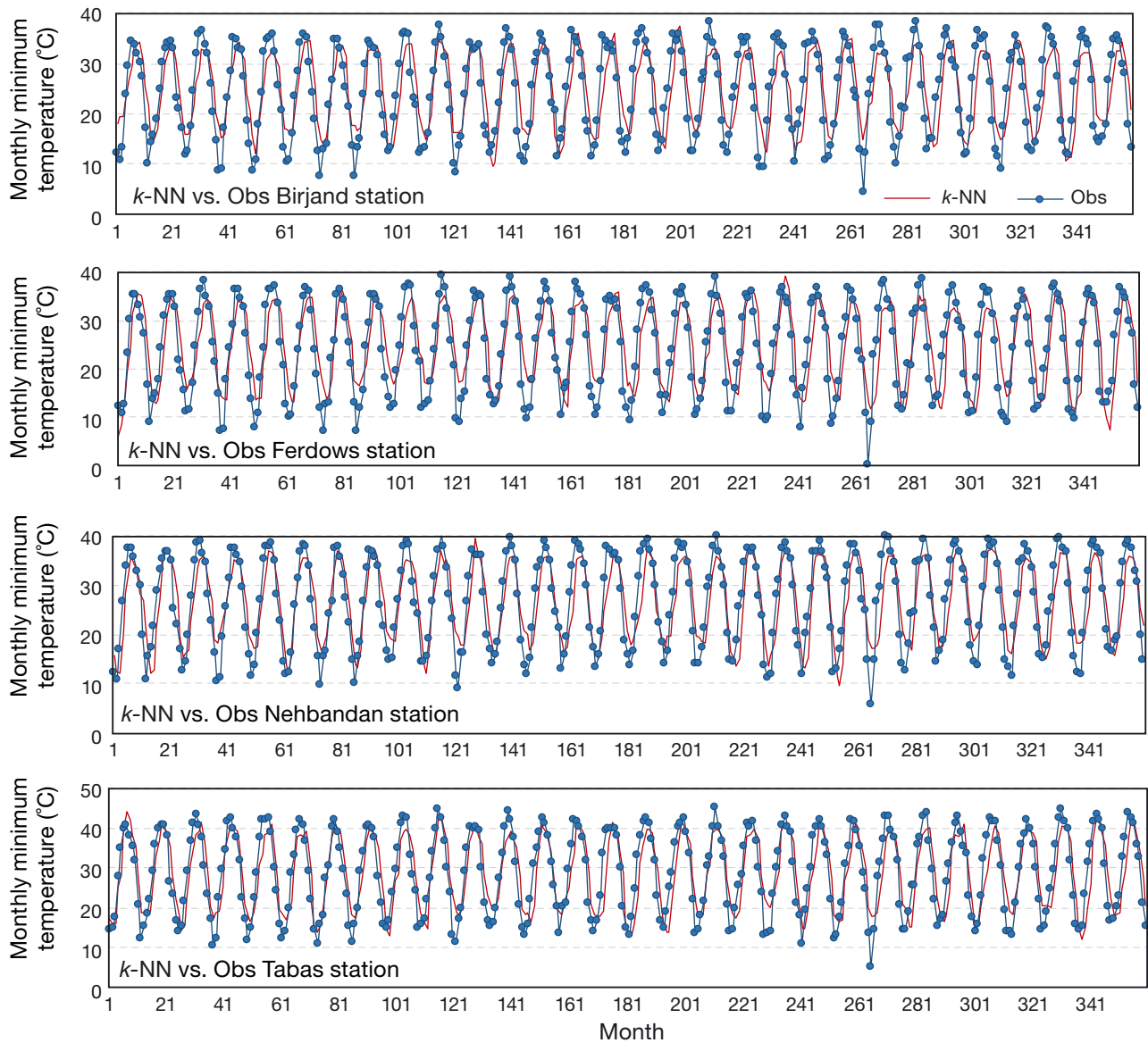number of weather variables. The vector of neigh-

Fig. 9. Monthly maximum temperature (°C) based on observation (Obs) (1986–2015) versus prediction periods (2018–2047) using the *k*-NN method (360 mo)

bors with 6 variables was the best selection. MIROC5 downscaled data performed better than *k*-NN over the evaluation period, but both methods were used for future projections. Our results predict increases in monthly T_{max} and T_{min} in summer and autumn and decreases in winter and spring for all locations. The downscaled projections of monthly precipitation show general reductions in winter.

Finally, according to the presented results in this study, although the *k*-NN weather generator is one of the most popular methods used in various studies — especially for assessing the vulnerability of basins to floods and droughts under climate change, manage-

ment of water allocations, and hydrological concepts — we reveal that the outputs of the *k*-NN method are less suitable in comparison to MIROC5 downscaled future data over arid regions, especially for precipitation. In this regard, we recommend that other researchers use these 2 methods in different climate regions across the globe to obtain more definitive and meaningful results.
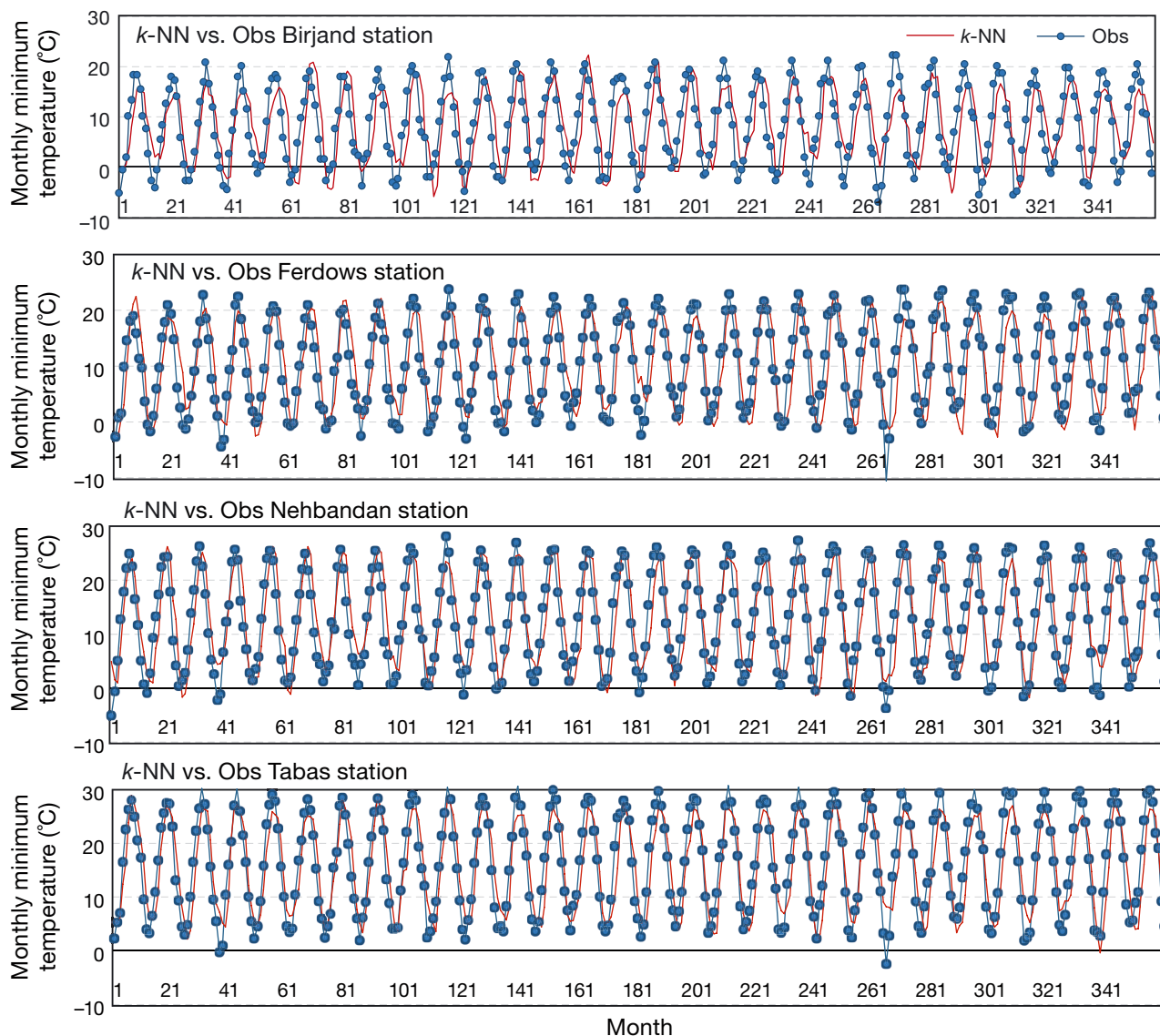
Fig. 10. Monthly minimum temperature (°C) based on observation (Obs) (1986–2015) versus prediction periods (2018–2047) using the *k*-NN method (360 mo)

LITERATURE CITED

AgriMetSoft (2017) KNN Weather Generator software. https://agrimetsoft.com/KNN-WG.aspx

Bannayan M, Hoogenboom G (2008) Weather analogue: a tool for real-time prediction of daily weather data realizations based on a modified *k*-nearest neighbor approach. Environ Model Softw 23:703–713

Barnett TP, Preisendorfer RW (1978) Multifield analog prediction of short-term climate fluctuations using a climate state vector. J Atmos Sci 35:1771–1787

Boé J, Terray L, Habets F, Martin E (2007) Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. Int J Climatol 27: 1643–1655

Bürger G, Murdock TQ, Werner AT, Sobie SR, Cannon AJ (2012) Downscaling extremes—an intercomparison of

multiple statistical methods for present climate. J Clim 25:4366–4388

Cavazos T, Hewitson BC (2005) Performance of NCEP–NCAR reanalysis variables in statistical downscaling of daily precipitation. Clim Res 28:95–107

Chen L, Frauenfeld OW (2014) A comprehensive evaluation of precipitation simulations over China based on CMIP5 multimodel ensemble projections. J Geophys Res Atmos 119:5767–5786

Chen J, Brissette FP, Leconte R (2011a) Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. J Hydrol 401:190–202

Chen C, Haerter JO, Hagemann S, Piani C (2011b) On the contribution of statistical bias correction to the uncertainty in the projected hydrological cycle. Geophys Res Lett 38:L20403

Dessu SB, Melesse AM (2013) Impact and uncertainties of

climate change on the hydrology of the Mara River basin, Kenya/Tanzania. Hydrol Process 27:2973–2986

Dibike YB, Coulibaly P (2005) Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models. J Hydrol 307:145–163

Eum HI, Simonovic SP, Kim YO (2010) Climate change impact assessment using k-nearest neighbor weather generator: case study of the Nakdong River Basin in Korea. J Hydrol Eng 15:772–785

Faiz MA, Liu D, Fu Q, Wrzesiński D and others (2018) Extreme precipitation and drought monitoring in northeastern China using general circulation models and pan evaporation-based drought indices. Clim Res 74:231–250

Fettweis X, Franco B, Tedesco M, van Angelen JH, Lenaerts JTM, van den Broeke MR, Gallee H (2013) Estimating the Greenland ice sheet surface mass balance contribution to future sea level rise using the regional atmospheric climate model MAR. Cryosphere 7:469–489

Fita L, Evans JP, Argüeso D, King A, Liu Y (2017) Evaluation of the regional climate response in Australia to large-scale climate models in the historical NARCliM simulations. Clim Dyn 49:2815–2829

Fowler HJ, Blenkinsop S, Tebaldi C (2007) Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int J Climatol 27:1547–1578

Gaetani M, Mohino E (2013) Decadal prediction of the Sahelian precipitation in CMIP5 simulations. J Clim 26: 7708–7719

Hanssen-Bauer I, Achberger C, Benestad RE, Chen D, Førland EJ (2005) Statistical downscaling of climate scenarios over Scandinavia. Clim Res 29:255–268

Harding KJ, Snyder PK, Liess S (2013) Use of dynamical downscaling to improve the simulation of Central U.S. warm season precipitation in CMIP5 models. J Geophys Res 118:522–536

Haylock MR, Cawley GC, Harpham C, Wilby RL, Goodess CM (2006) Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. Int J Climatol 26: 1397–1415

Hellström C, Chen D, Achberger C, Räisänen J (2001) Comparison of climate change scenarios for Sweden based on statistical and dynamical downscaling of monthly precipitation. Clim Res 19:45–55

Hsu P, Li T, Murakami H, Kitoh A (2013) Future change of the global monsoon revealed from 19 CMIP5 models. J Geophys Res 118:1247–1260

IPCC (2014) Climate change, synthesis report. In: Core Writing Team, Pachauri RK, Meyer LA (eds) Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva

Kharin VV, Zwiers FW, Zhang X, Hegerl GC (2007) Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. J Clim 20:1419–1444

King LM, McLeod AI, Simonovic SP (2015) Improved weather generator algorithm for multisite simulation of precipitation and temperature. J Am Water Resour Assoc 51:1305–1320

Lall U, Sharma A (1996) A nearest neighbor bootstrap for time series resampling. Water Resour Res 32:679–693

Lenderink G, Buishand A, van Deursen W (2007) Estimates of future discharges of the river Rhine using two scenario methodologies: direct versus delta approach. Hydrol Earth Syst Sci 11:1145–1159

Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. J Atmos Sci 26:636–646

Mandal S, Srivastav RK, Simonovic SP (2016) Use of beta regression for statistical downscaling of precipitation in the Campbell River basin, British Columbia, Canada. J Hydrol 538:49–62

Maraun D (2013) Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. J Clim 26: 2137–2143

Maraun D, Wetterhall F, Ireson AM, Chandler RE and others (2010) Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. Rev Geophys 48:RG3003

Maurer EP, Hidalgo HG (2008) Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. Hydrol Earth Syst Sci 12: 551–563

Meinshausen M, Smith SJ, Calvin K, Daniel JS and others (2011) The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. Clim Change 109:213–241

Piani C, Haerter JO, Coppola E (2010) Statistical bias correction for daily precipitation in regional climate models over Europe. Theor Appl Climatol 99:187–192

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

Richardson CW (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour Res 17:182–190

Salehnia N, Zare H, Kolsoumi S, Bannayan M (2018) Predictive value of Keetch-Byram Drought Index for cereal yields in a semi-arid environment. Theor Appl Climatol 134:1005–1014

Schmidli J, Goodess CM, Frei C, Haylock MR, Hundecha Y, Ribalaygua J, Schmith T (2007) Statistical and dynamical downscaling of precipitation: an evaluation and comparison of scenarios for the European Alps. J Geophys Res 112:D04105

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

Shabbar A, Knox JL (1986) Monthly prediction by the analogue method. Proceedings of the 1st WMO workshop on the diagnosis and prediction of monthly and seasonal atmospheric variations over the globe. Ser 6. Vol II, Tech. doc. WMO/TD 87. World Meteorological Organization, Geneva, Switzerland, p 672–681

Sharif M, Burn DH (2006) Simulating climate change scenarios using an improved K-nearest neighbor model. J Hydrol 325:179–196

Sirangelo B, Versace P, De Luca DL (2007) Rainfall nowcasting by at site stochastic model P.R.A.I.S.E. Hydrol Earth Syst Sci 11:1341–1351

Sunyer MA, Madsen H, Ang PH (2012) A comparison of different regional climate models and statistical downscaling methods for extreme rainfall estimation under climate change. Atmos Res 103:119–128

Themeßl MJ, Gobiet A, Leuprecht A (2011) Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. Int J Climatol 31:1530–1544

Ullah A, Salehnia N, Kolsoumi S, Ahmad A, Khaliq T (2018) Prediction of effective climate change indicators using statistical downscaling approach and impact assessment on pearl millet (*Pennisetum glaucum* L.) yield through genetic algorithm in Punjab, Pakistan. Ecol Indic 90: 569–576

Van den Dool HM (1994) Searching for analogues, how long must we wait? Tellus 46:314–324

Vrac M, Stein M, Hayhoe K (2007) Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. Clim Res 34:169–184

Watanabe M, Suzuki T, O'ishi R, Komuro Y and others (2010) Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity. J Clim 23:6312–6335

Wilby RL, Wigley TML (1997) Downscaling general circula-

tion model output: a review of methods and limitations. Prog Phys Geogr 21:530–548

Wu W, Liu Y, Ge M, Rostkier-Edelstein D and others (2012) Statistical downscaling of climate forecast system seasonal predictions for the Southeastern Mediterranean. Atmos Res 118:346–356

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

Young KC (1994) A multivariate chain model for simulating climatic parameters with daily data. J Appl Meteorol 33: 661–671

Zhou ZH (2012) Ensemble methods: foundations and algorithms. CRC Press, Boca Raton, FL