Conversion of CruTS 3.23 data and evaluation of precipitation and temperature variables in a local scale

Sinan Jasim Hadi[1,*], and *Mustafa* TOMBUL[1]

[1]Department of Civil Engineering, Faculty of Engineering, Anadolu University Eskisehir, Turkey.

Abstract. The precipitation and temperature data obtained from the last version 3.23 of Climate Research Unit Time Series (CruTS) data are evaluated in a local scale. The data obtained for the period 1901-2014 consists of ten climate variables: Precipitation, Mean temperature, Diurnal Temperature Range, Wet-day Frequency, Vapour Pressure, Cloud Cover, Maximum temperatures, Minimum temperatures, frost day frequency, and potential evapotranspiration. These data converted to ASCII format which is usable in GIS environments. Daily precipitation and temperature data of 9 station located in Seyhan basin in Turkey for the period 1973-2000 used for the assessment after converting them to monthly total precipitation and average temperature. The assessment implemented using several statistical measurements: correlation, RMSE, mean difference, median difference, and Mann-Whitney test. The study found that CruTS has high agreement with the observed data in the meaning of the correlation and low RMSE in all stations. The result of mean, median, and Mann-Whitney tests indicate no significant difference between CruTS and the observed data in all station with the exception of two stations; Tomarza station has significant values in the temperature values in all tests; Goksun station has significance difference in the mean test in the two climate variables.

1 Introduction

Climate Research Unit Time Series (CruTS) was named based on the name of the producer of these data i.e. the Climate Research Unit, University of East Anglia, Norwich, UK. Initially these data developed by [1] for the period 1901- 1996, updated by [2] for the period 1901 - 2002, and finally updated one more time by [3] for the period 1901-2009. The data are still being added to the collection and corrected.

The last version 3.23 of CruTS datasets consists of 0.5º latitude-longitude grid cells covering the global with the exception of Antarctica region. The datasets provide monthly values of ten climate variables. Six of these variables are independent: Precipitation (pre), Mean temperature (tmp), Diurnal Temperature Range (dtr), Wet-day Frequency (wet), Vapour Pressure (vap), and Cloud Cover (cld). The remain four (i.e. Maximum

*Correspondingauthor: sinan.jasim@yahoo.com

temperatures (tmx), Minimum temperatures (tmn), frost day frequency (frs), and potential evapotranspiration (pet)) are calculated and estimated using the independent variables [4].

CruTS is routinely updated using several resources: CLIMAT monthly data for about 2400 station, for the last 4-5 years, produced by the collaboration between World Meteorological Organization (WMO) and the US National Oceanographic and Atmospheric Administration (NOAA), Monthly Climatic Data for the World (MCDW) produced by UK National Climatic Data Center (NCDC) for incorporating in WMO by 200 stations, World Weather Records (WWR) decadal data for about 1700 stations during 1991-2000 published as exchange between the archive center of NCDC and the National Meteorological Services (NMSs). The aforementioned sources provide a systematic incorporation in the updating process.

The importance of CruTS datasets has been increasing in Geosciences researches with a focus on the climate change studies due to two reasons: the long period that covering from 1901 – 2014 that give the ability for studying the trend and the change in the climate, and the high spatial resolution which helps in local not only regional studies. CruTS with its different versions have been widely used in many researches and in many applications such as trend analysis, climate change, desertification and many others. For example, but not limited to [5-10].

In order to use this data for applications, they need to be validated in different scales. [3] created the CruTS 3.10 data using Climate anomaly method (CAM) [11] and validated them in a sub-continental scales by using: temperature dataset developed by University of Delaware (UDEL) based on GHCN-M [12] for the period 1901-2008, and precipitation data developed by Global Precipitation Climatology Center (GPCC) for the period 1901-2009. They found a tight agreement between the temperature data of CruTS and UDEL.CruTS and GPCC precipitation data have high agreement also, but not as tight as temperature data.

The main objective of this study is to evaluate the last version of CruTS 3.23 released 1 September 2015 in a local scale.The dataset to be compared with the observed monthly total precipitation and mean temperature data of 9 stations located in Seyhan basin, Turkey for the period 1973 – 2000. Additionally, the CruTS 3.23 dataset for the period 1901 – 2014 for all the variables converted to GIS-based ASCII files to make them usable in GIS environments.

2 Data collection

CruTS data are originally available at Centre for Environmental Data Archival of the British Atmospheric Data Center (BADC) which can be reached by (http://www.badc.nerc.ac.uk). The data provided in two format: ASCII and NetCDF. In the same time, a temporal or spatial subset of the datasets can be obtained with CSV file format from CEDA OGC WEB SERVICES/ Web Processing Service reached by (http://wps-web1.ceda.ac.uk/).

The Consortium for Spatial Information, CSI, which is the CGIAR spatial scientists community (http://www.cgiar-csi.org/) processed the CruTS 3.10 data which covers the period 1901 to 2009 and converted them to ASCII in order to facilitate their use in GIS environment.

In this study, CruTS 3.23 data for the period 1901 – 2014 released 1 September 2015 evaluated by comparing them with the observed data. The data set collected from CEDA OGC WEB SERVICES/ Web Processing Service with CSV format. The collected data set converted to ASCII format to provide a full set contains the entire available variables and period.

Observations of 9 stations were collected from DMI (DevletMeteorolojiIsleri); the Tukrish General Directory of Meteorology. These stations located at three sub-basins called

(coded): 1822, 1801, and 1805 as part of Seyhan basin located in the southwest of Turkey. The area is taking part from four Turkey's provinces: Adana, Kayseri, Kahramanmaras, and Sivas, as seen in Figure 1.

The data were collected for the period from 1/1/1973 to 31/12/2000. For every station, the collected data were total daily precipitation and average daily temperature. These data converted to monthly total precipitation and monthly average temperature to be compared, after that, with the CruTS data.

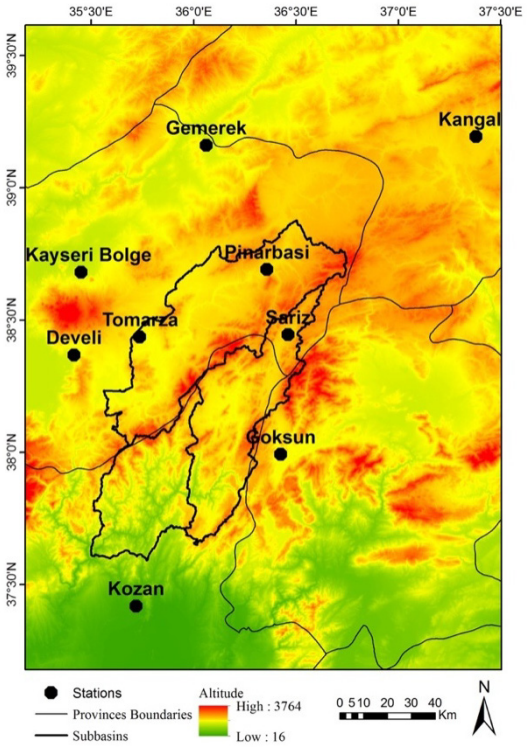

Fig. 1. Study area with the 9 stations.

3 Methodology

The obtained CruTS data of ten climate variables for the period 1901 -2014 have csv file format as one file for every month. MATLAB was used for reading files and converting them to ASCII after adding the required text to those files. The obtained ASCII files can be then easily used in any GIS software for further analysis.

The obtained files of the precipitation and temperature were then imported to R (i.e. Statistical Programming Language) for implementing the main objective of this study. CruTS data has a period from 1901 -2014 while the observed data 1973 -2000, therefore subsets of the CruTS data were extracted. The values of the CruTS ASCII files cells where the stations are located were extracted that 9 values obtained for each month which represented by one ASCII file.

The strength of the relationship between the values extracted from CruTS data and the observed data were evaluated by calculating the correlation coefficient which is given as:

$$r_{X,Y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad (1)$$

Where x_i is the value of the CruTS data for month i, $y_i$is the observed value of the same month, $\bar{x}$ is the mean of the CruTS data, $\bar{y}$ is the mean of the observed data, and N the number of data which represents months in this case. The result of the correlation for every station and for the two climate parameters; precipitation and temperature shown in table 1.

In order to measure the error between the two datasets, Root Mean Squared Error (RMSE) was calculated for each station using:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \quad (2)$$

The result of RMSE listed in table 1.

The significance of the difference between the mean of the CruTS data μx and the mean of observed data μy was tested using a hypothesis test:

Ho: μx = μy

Ha: μx ≠ μy

The test conducted for each station using t student's test:

$$t = \frac{\bar{x} - \bar{y}}{SE} \quad (3)$$

The standard error SE is calculated by:

$$SE = \sqrt{\frac{{s_x}^2}{N_x} + \frac{{s_y}^2}{N_y}} \quad (4)$$

Where s_x and s_y are the standard deviation of the CruTS and Observed data respectively [13, 14].

After calculating the t statistic, P – value calculated for every station in order to compare them with the significance level.

Some arguments saying that mean should not be used in case of the data are not normally distributed unless the number of observations large enoughwhich is the case here in the precipitation data. Although the samples used here can be considered as sufficiently large, the median test was implemented.

In addition, Non-Parametric test named Mann-Whitney (known as Rank-Sum test and Wilcoxon Rank-Sum test) was used. This test is used for checking whether one group has higher or lower values than a second group. The two-sided hypothesis of this test is:

Ho: Prob[x > y] = 0.5

Ha: Prob[x > y] ≠ 0.5

Where x the first group data (i.e. CruTS data) and y the second group data (i.e. the observed data). In words, if the null hypothesis rejected it means that the data of two groups are different and vice versa. To save some space, the details of median and Mann-Whitney test are not mentioned here; for detailed description see [14-17]. The P-values of mean, median, and Mann-Whitney test are shown in table 2.

4 Results and discussion

The main objective of this study is to evaluate the CruTS 3.23 data by comparing them with the observed data of 9 stations located at the southwest of Turkey. To achieve this objective and to make the CruTS data (i.e. obtained with CSV file) usable in several software, especially GIS softwares, the files of the monthly 10 climate variables mentioned earlier were converted to ASCII. An example of the obtained files (i.e. the first month January of 1973 of precipitation) was added to ArcGIS and symbolized shown in figure 2.

In order to ease the access, the ASCII files of the ten climate variables for the period 1901-2014 are available for download as a compressed file for every variable from (https://goo.gl/vFGNdp) or all the variables in one compressed file from (https://goo.gl/TXfxZm).

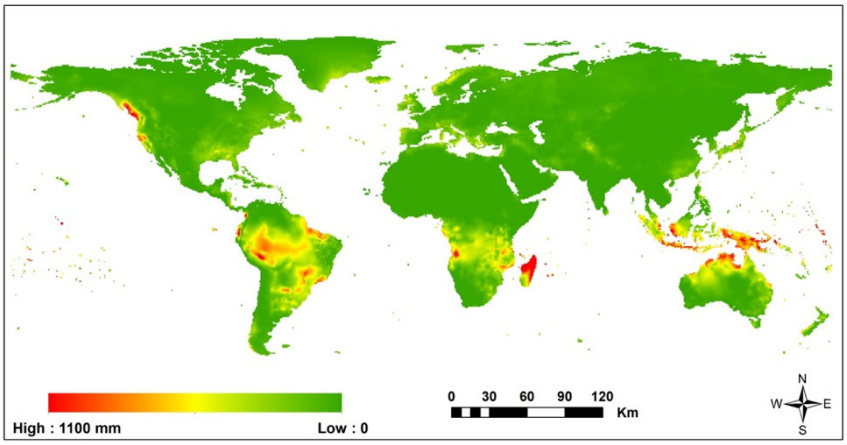

Fig. 2. Global Precipitation map of January 1973 obtained from CruTS 3.23 data.

Precipitation and Temperature variables for the period 1973-2000 were extracted from the CruTS data in order to have the same monthly data collected from the DMI for the 9 station. By having subsetting the original data, consistent data are obtained that can be used for comparison.

Box plots were created for precipitation and temperature (i.e. figure 3 and 4 respectively) illustrating the data of every station from the two sources CruTS and observation side by side. Looking at figure 3, Kozan station has the highest values of the total monthly precipitation while other stations are close to each other. In all stations, data coming from CruTS and observed data are close to each other in terms of the 0.25, 0.5 (i.e. median), and 0.75 quantiles with an exception of Goksun station in which the observed data has obvious higher values than CruTS data. Tomarza station shows a significant difference in the median between the two data sources. Based on the time series plot of the precipitation (i.e. figure 5), there is an obvious mismatch between the two data sources; highest in Tomarza station, a bit lower in Goksun station while other stations matching better.

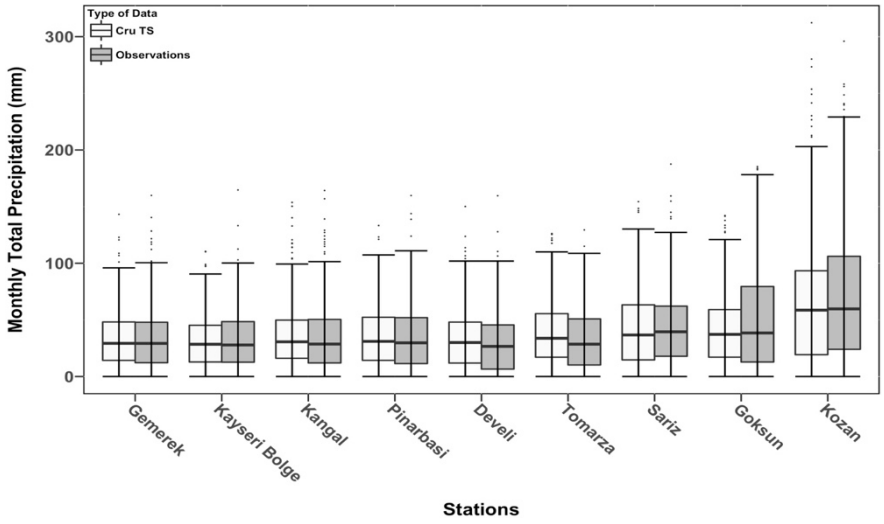

Fig. 3. CruTS and observed precipitation data boxplot.

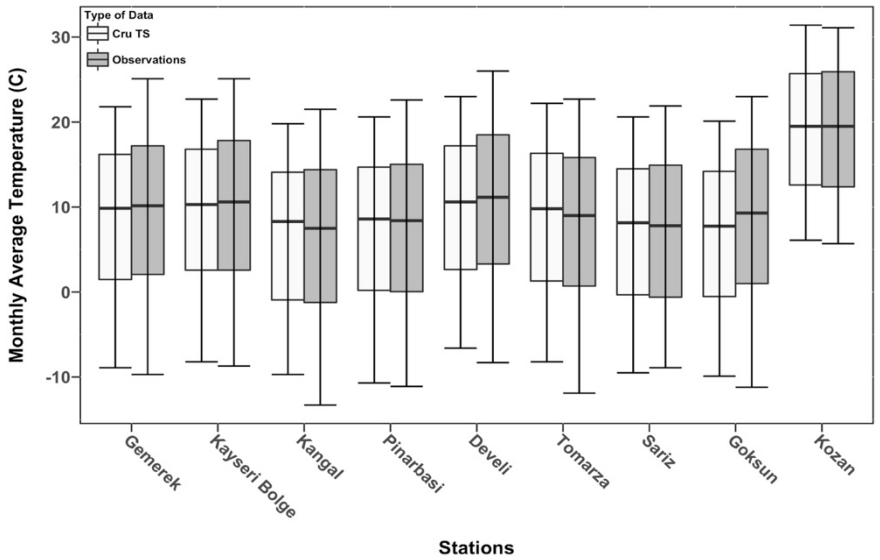

Fig. 4. CruTS and observed temperature data boxplot.

Based on the temperature box plot figure 4, Kozan station has, also, the highest temperature values in comparison with other stations which are not much different from each other. Similarly, to the precipitation, Goksun station has the highest difference between CruTS and observed data where as other stations have slight close differences. According to the temperature time series plot figure 6, the temperature in comparison with precipitation has changed, generally, smoothly. In addition, all stations show a high matching except Goksun station where there is a slight mismatch between CruTS and observed data. According to figure 1 which shows the Digital Elevation Model (DEM), Kozan station is located at a low altitude in comparison with other station which are located

in higher altitude and that could be the reason that this station has highest values in both precipitation and temperature.

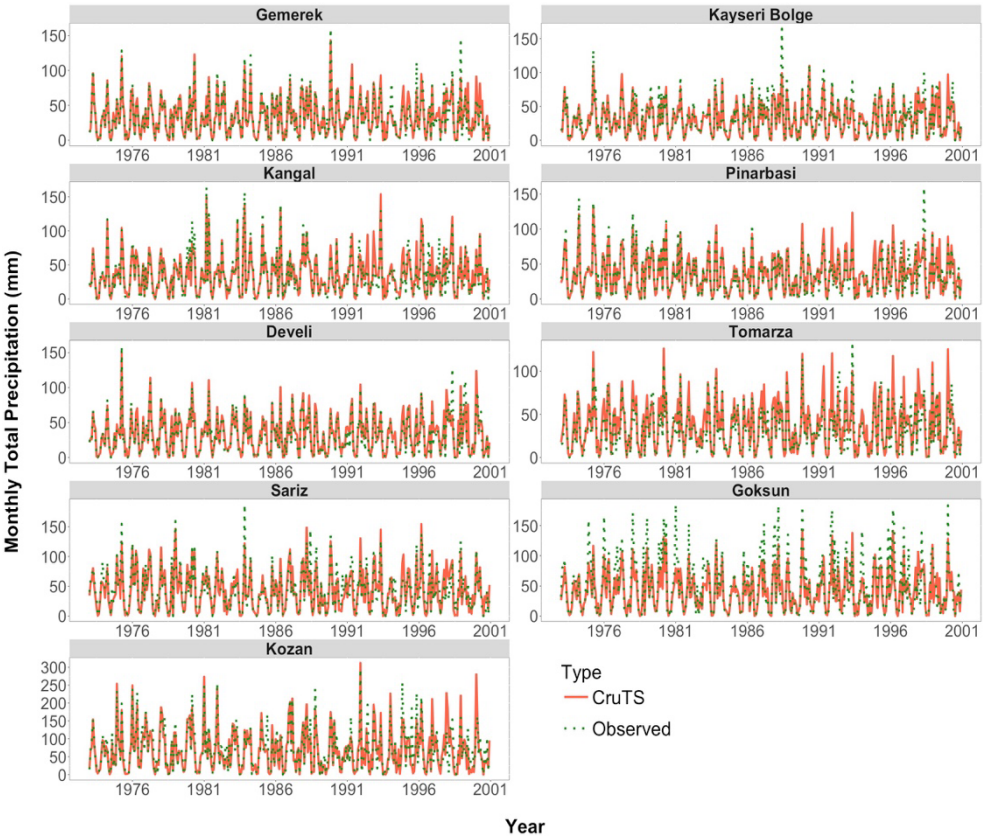

Fig. 5. Time series plot of the monthly total precipitationobtainedfromCruTS and Observed data for the period 1973 - 2000.

Observed precipitation and temperature data were used for the comparison with the CruTS 3.23 in the meaning of several statistical measurements. Correlations listed in table 1 showed very high agreement for all stations in the temperature data ranges from 0.989 for Kangal station to 0.996 for KayseriBolge station. From the same table, RMSE which measures the error between the two datasets, illustrates that Kozan station has the lowest while Goksun station has the highest error between the CruTS 3.23 and observed temperature data which agrees with what have been resulted from figure 4.

In general, the agreement between the precipitation data was lower than temperature data. The precipitation correlations lowest value is 0.87 for Kozan station while the highest value is 0.942 for Gemerek station. RMSE results for precipitation, also, showed higher error values. The highest error belongs to Kozan station while Gemerek station has the lowest. This lower agreement in the precipitation could be resulted from the nature of creating this data as these data created by interpolating a number of stations. The temperature data has more homogeneous distribution than precipitation around the interpolated area regardless the used method.

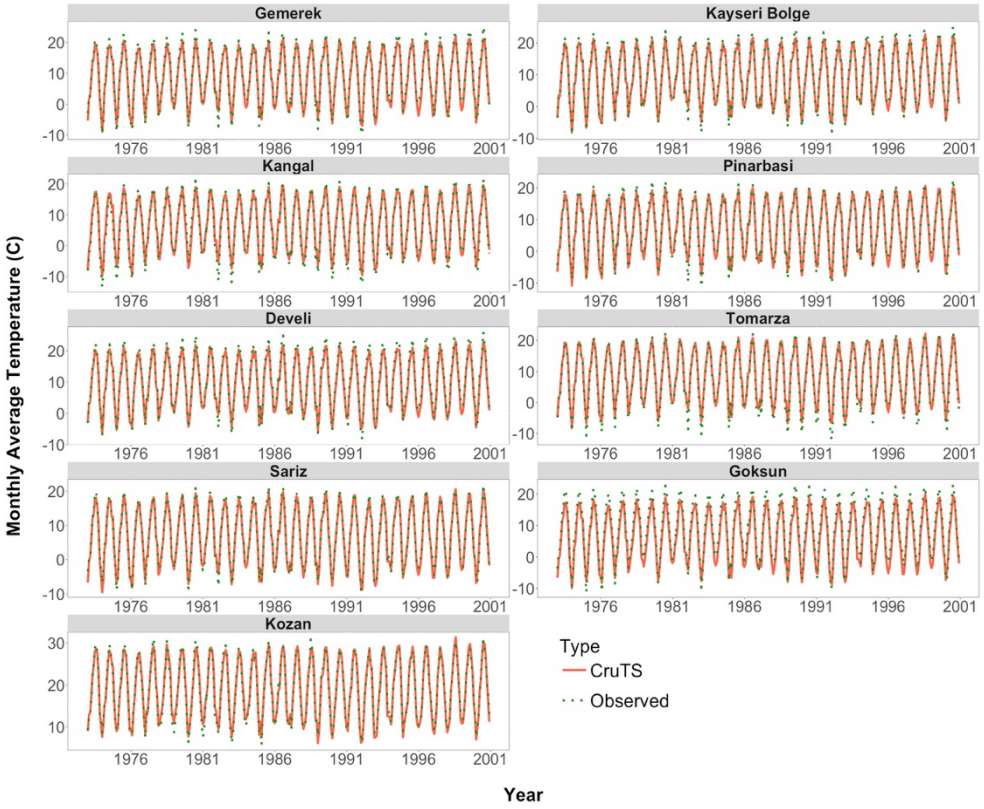

Fig. 6. Time series plot of the monthly average temperature obtained from CruTS and Observed data for the period 1973 - 2000.

The P values of testing the difference between the means of two sources data for two variables and every station listed in table 2. The result shows no significant values in most of the stations which means fail to reject the null hypothesis that the there is no difference between the means with several exceptions; Tomarza station has a significant precipitation mean difference at the significance level $\alpha = 0.05$ and not significant at$\alpha = 0.01$ where P values is 0.018, Goksun station is significant in both precipitation and temperature at both significance levels with values 0.005 and 0.003 respectively.

The median P value results listed in the same table indicate that two variables in all stations have no significant difference in the mean with the exception of the precipitation variable in Tomarza station which is significant at the significance level $\alpha = 0.05$ and not significant at$\alpha = 0.01$ with the value 0.029. Surprisingly, temperature variable in Kozan station has a P value = 1 which indicates the median are same in the two data sources and that agrees with very high value of this station's mean result which is 0.949.

The Non – Parametric Mann-Whitney test p values are also listed in table 2. These values indicate fail to reject the null hypothesis (i.e., no difference between the two groups) as no significance difference for the two variables in all station except two stations i.e., Tomarza and Goksun stations. Precipitation variable is significant in Tomarza station at the significance levels $\alpha = 0.05$ only which agrees with the two mean and median tests. Goksun station has a significant value in the temperature variable at the two significance levels $\alpha =$ 0.01 and 0.05 which matches the mean test.

Table 1. The correlation and MSRE between CruTS and observed data for precipitation and Temperature.

Station Name	Precipitation		Temperature	
	Correlation	RMSE(mm)	Correlation	RMSE (C)
Gemerek	0.947	9.007	0.992	1.383
Kayseri Bolge	0.916	10.866	0.996	1.178
Kangal	0.924	11.694	0.989	1.562
Pinarbasi	0.920	11.288	0.991	1.217
Develi	0.929	10.618	0.995	1.409
Tomarza	0.891	13.663	0.992	1.506
Sariz	0.885	16.320	0.994	0.948
Goksun	0.870	25.387	0.992	2.414
Kozan	0.870	30.326	0.992	0.930

Table 2. The P value of the mean difference, median difference and Mann-Whitney tests.

Station Name	Mean		Median		Mann-Whitney test	
	Precipitation	Temperature	Precipitation	Temperature	Precipitation	Temperature
	P value					
Gemerek	0.582	0.324	0.977	0.693	0.343	0.222
Kayseri Bolge	0.397	0.452	0.862	0.762	0.793	0.338
Kangal	0.524	0.580	0.464	0.340	0.257	0.718
Pinarbasi	0.518	0.813	0.462	0.873	0.355	0.721
Develi	0.130	0.161	0.083	0.507	0.060	0.102
Tomarza	0.018*	0.265	0.029*	0.490	0.021*	0.365
Sariz	0.548	0.743	0.289	0.806	0.594	0.677
Goksun	0.005**	0.003**	0.823	0.139	0.291	0.001**
Kozan	0.614	0.949	0.911	1.000	0.618	0.980

* significant at $\alpha = 0.05$.
* *significant at $\alpha = 0.01$.

In general, CruTS 3.23 data has a tight agreement with the observed data. Most of the used 9 stations has; high correlations; low errors; and no significance in the differences tests. The result agrees with the result of [3] who found close-fitting matches in a sub-continental scale during their updating of CruTS 3.10 data. As an exception, Goksun station has a statistical significance in the difference in several test which agrees with the low correlation and high error in comparison with other stations. However, although Kozan station has the lowest correlation (sharing the same value with Goksun) and highest error in the precipitation variable, it has no significance in any of the tests.

5 Conclusion

This study aims to evaluate the CruTS 3.23 data for only precipitation and temperature variables in a local scale. To have suitable format of the data, files of monthly data of ten climate variables; precipitation, mean temperature, diurnal temperature range, minimum and maximum temperature, vapour pressure, cloud cover, rain days, frost days and potential evapotranspiration for the period January 1901 to December 2014 was converted to GIS-based ASCII formats.

The local scale comparison results indicate that the two sources of data have a tight agreement in both precipitation and temperature, however, the agreement of temperature variable is higher.

The three tests (i.e., Mean difference, Median difference, and Mann-Whitney) results show no significance difference between the two data sets in all stations with an exception of two stations; Tomarza and Goksun stations. Tomarza station has significant difference only in the precipitation variable in the three test while Goksun station has significant difference only in the mean test of the two variables. Kozan station shows the opposite of these two stations as, for example, it has a difference in the median reaches to zero. This difference among stations originated from the nature of the CruTS data that the different versions were created by interpolating a number of station around the world.

This research was funded by Anadolu University under the following project: BAP-000261. The authors would like to thank (DMİ) DevletMeterolojiİşleri (General Department of Meteorology - Ministry of Forest and Water Affairs) for providing the data to complete this study.

References

1. M. New, M. Hulme, and P. Jones, "Representing twentieth-century space-time climate variability. Part II: Development of 1901-96 monthly grids of terrestrial surface climate," J Climate. **13**, 13, pp. 2217-2238, (2000).
2. T. D. Mitchell and P. D. Jones, "An improved method of constructing a database of monthly climate observations and associated high-resolution grids," Int J Climatol. **25**, 6, pp. 693-712, (2005).
3. I. Harris, P. D. Jones, T. J. Osborn, and D. H. Lister, "Updated high-resolution grids of monthly climatic observations - the CRUTS3.10 Dataset," Int J Climatol. **34**, 3, pp. 623-642, (2014).
4. P. D. Jones and I. Harris. (2008, 10/08/2016). *Climatic Research Unit (CRU) time-series datasets of variations in climate with variations in other phenomena.* Available: http://catalogue.ceda.ac.uk/uuid/3f8944800cc48e1cbc29a5ee12d8542d
5. R. A. Garcia, M. Cabeza, C. Rahbek, and M. B. Araújo, "Multiple dimensions of climate change and their implications for biodiversity," Science. **344**, 6183, (2014).
6. U. Hellden and C. Tottrup, "Regional desertification: A global synthesis," Global Planet Change. **64**, 3-4, pp. 169-176, (2008).
7. H. Hoff, M. Falkenmark, D. Gerten, L. Gordon, L. Karlberg, and J. Rockstrom, "Greening the global water system," Journal of Hydrology. **384**, 3-4, pp. 177-186, (2010).
8. M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel, "World map of the Köppen-Geiger climate classification updated," Meteorologische Zeitschrift. **15**, 3, (2006).
9. B. Poulter, D. Frank, P. Ciais, R. B. Myneni, N. Andela, J. Bi*, et al.*, "Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle," Nature. **509**, 7502, (2014).
10. A. K. Taxak, A. R. Murumkar, and D. S. Arya, "Long term spatial and temporal rainfall trends and homogeneity analysis in Wainganga basin, Central India," Weather and Climate Extremes. **4**, pp. 50-61, (2014).
11. T. C. Peterson, T. R. Karl, P. F. Jamason, R. Knight, and D. R. Easterling, "First difference method: Maximizing station density for the calculation of long-term global temperature change," Journal of Geophysical Research: Atmospheres. **103**, D20, pp. 25967-25974, (1998).
12. T. C. Peterson, R. Vose, R. Schmoyer, and V. Razuvaëv, "Global historical climatology network (GHCN) quality control of monthly temperature data," Int J Climatol. **18**, 11, pp. 1169-1179, (1998).
13. D. M. Diez, C. D. Barr, and M. C. etinkaya-Rundel, *OpenIntro Statistics*: Duke University, (2015).

14. D. R. Helsel and R. M. Hirsch, "Statistical Methods in Water Resources," in *Techniques of Water Resources Investigations*. vol. Book 4, ed: U.S. Geological Survey, p. 522, (2002).
15. M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*, Third Edition ed.: Wiley Online, (2014).
16. J. Hájek, Z. Šidák, and P. K. Sen, *Theory of Rank Tests*, Second Edition ed.: Academic Press, (1999).
17. R. H. McCuen, *Modeling Hydrologic Change: Statistical Methods*, 1 edition ed.: CRC Press, (2002).