J. Geogr. Sci. 2016, 26(12): 1661-1674
DOI: 10.1007/s11442-016-1351-7

Quality control and homogenization of daily meteorological data in the trans-boundary region of the Jhelum River basin

Rashid MAHMOOD, JIA Shaofeng

Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Science and Natural Resources Research, CAS, Beijing 100101, China

Abstract: Many studies such as climate variability, climate change, trend analysis, hydrological designs, agriculture decision-making etc. require long-term homogeneous datasets. Since homogeneous climate data is not available for climate analysis in Pakistan and India, the present study emphases on an extensive quality control and homogenization of daily maximum temperature, minimum temperature and precipitation data in the Jhelum River basin, Pakistan and India. A combination of different quality control methods and relative homogeneity tests were applied to achieve the objective of the study. To check the improvement after homogenization, correlation coefficients between the test and reference series calculated before and after the homogenization process were compared with each other. It was found that about 0.59%, 0.78% and 0.023% of the total data values are detected as outliers in maximum temperature, minimum temperature and precipitation data, respectively. About 32% of maximum temperature, 50% of minimum temperature and 7% of precipitation time series were inhomogeneous, in the Jhelum River basin. After the quality control and homogenization, 1% to 11% improvement was observed in the infected climate variables. This study concludes that precipitation daily time series are fairly homogeneous, except two stations (Naran and Gulmarg), and of a good quality. However, maximum and minimum temperature datasets require an extensive quality control and homogeneity check before using them into climate analysis in the Jhelum River basin.

Keywords: quality control; homogenization; daily meteorological data; Jhelum River basin; Pakistan

1 Introduction

High quality and homogeneous long-term data series are essential in climate research, especially in climate change studies, which are used to assess climate variability and historical climate trends of mean and extreme climate events. However, most of the long climatic series not only have outliers and missing values but also are inhomogeneous (Cao and Yan, 2012; Trewin, 2013). Homogeneous climate time series are those where the variations are

Received: 2015-07-22 **Accepted:** 2015-10-29
Foundation: National Natural Sciences Foundation of China, No.41471463; President's International Fellowship Initiative CAS
Author: Rashid Mahmood, E-mail: rashi1254@gmail.com; Jia Shaofeng, E-mail: jiasf@igsnrr.ac.cn

caused solely due to variation in climate and not due to non-climatic factors. The potential non-climatic factors are changes in instruments, changes in surroundings, relocation of monitoring stations, changes in observation methods etc. (Li-Juan and Zhong-Wei, 2012; Štěpánek *et al*., 2013). These factors may hide true signals of climate variability and climate change, leading towards some wrong conclusions of climate and hydrological studies (Costa and Soares, 2009). These are discussed in more details in Peterson *et al*. (1998), Aguilar *et al*. (2003) and Trewin (2010).

The climatic series that span from decades to centuries are rarely free of irregularities, errors and missing values. Although some specific inhomogeneous sites have only a marginal effect on the observed climate trends at the global scale, they can have substantial impact at the local or regional scale (Trewin, 2013). Thus, it is essential to produce homogeneous and quality controlled climate records before using them in climate analysis (Costa and Soares, 2009).

Several techniques such as Buishand range test (Buishand, 1982), Krukal-Wallis test (Kruskal, 1952; Kruskal and Wallis, 1952), Mann-Kendal test (Mann, 1945; Kendall, 1975), Multiple Analysis of Series for Homogenization (MASH) (Szentimrey, 1999), Pettit test (Pettitt, 1979), Regression-Based methods (Easterling and Peterson, 1995; Vincent, 1998), Standard Normal Homogeneity Test (SNHT) (Alexandersson, 1986) etc. have been developed for detection of irregularities on a site and their adjustment.

There are two main groups of homogeneity testing techniques; 'absolute' and 'relative'. In the first group, the statistical tests are applied on each time series separately. In the relative methods, the statistical tests are applied on the difference of test series (time series under consideration) and reference series—created from some highly correlated stations in the region. Although both approaches are useful and valid to detect an inhomogeneity, the relative approach is more reliable than the absolute because it also considers the changes on the neighbor stations in the region (Peterson *et al.*, 1998).

In homogenization, first, inhomogeneities are identified in a time series by using some techniques and then these irregularities are adjusted to make the site homogeneous (Trewin, 2013). Although several techniques are available, no single procedure is recommended.

Thus, the following four steps are commonly used to detect and adjust an inhomogeneous site: 1) basic quality control and metadata analysis, 2) reference series creation, 3) inhomogeneity detection and 4) adjustment for the compensation of inhomogeneity (Costa and Soares, 2009).

Many countries such as Australia (Trewin, 2013), Spain (Vicente-Serrano *et al*., 2010), Croatia (Zahradníček *et al*., 2014), Czech Republic (Štěpánek *et al*., 2009) and China (Feng *et al*., 2004) have developed homogenized meteorological datasets for climate analysis. However, in Pakistan and India, no quality controlled and homogeneous datasets are available for climate research. Thus, in the present study, quality controlled and homogeneous daily maximum temperature, minimum temperature and precipitation datasets are developed for the Jhelum River basin, Pakistan and India. This will provide an unprecedented resource for climate and climate change research in Pakistan. Station characteristics and data sources are described in Section 2. In Section 3, quality control and homogenization techniques are outlined. The main results and discussion are described in Section 4 and conclusions in Section 5.

2 Study area and data description

The upper Jhelum River basin is located in the north of Pakistan and spans between 33°–35°N and 73°–75.62°E, as shown in Figure 1. This is the second biggest tributary of the Indus River basin. The Jhelum basin has a drainage area of 33,342 km^2, with an elevation ranging from 200 to 6248 m. The whole basin drains into the Mangla Reservoir, the second largest reservoir in Pakistan, which was construction in 1967. The primary function of this reservoir is to provide water for irrigation of 6 million ha of land and to produce electricity as byproduct. The installed capacity of the reservoir is 1000 MW, which is 6% of the installed capacity of the country's power production (Archer and Fowler, 2008; Mahmood and Babel, 2013).

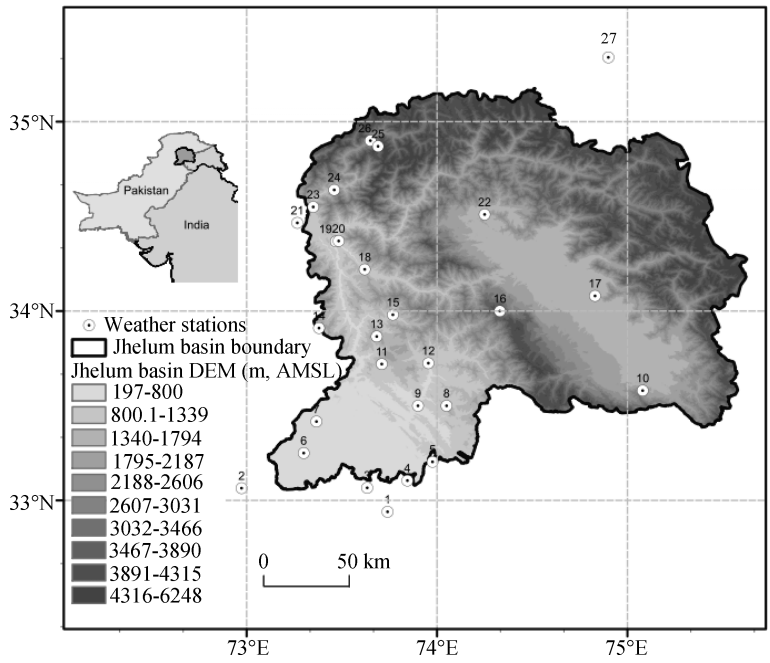

Figure 1 Location of the study area and geographic distribution of weather stations

Observed daily historical data of maximum temperature (22 weather stations), minimum temperature (22) and precipitation (27) were collected from Pakistan Meteorological Department (PMD), the Water and Power Development Authority of Pakistan (WAPDA) and the Indian Meteorological Department (IMD). The daily data of Gulmarg, Kupwara, Qazigund and Srinagar weather stations were obtained from IMD. The PMD provided climate data of Astore, Balakot, Garidopatta, Kotli, Muzaffarabad, Murree and Jhelum climate stations, and the remaining data was collected from WAPDA. Most of the precipitation series have data periods from 1961–2009. However, most of the temperature time series range from 1971–2009. The geographic distribution of these stations is shown in Figure 1. This shows that most of the stations are located in the eastern parts of the basin and on lower altitudes. The basic information about the stations such as location, mean distance between the stations, mean altitudinal differences between the stations, available data period and missing data of each station are given in Table 1.

Table 1 Geographic and basic information about the climate stations available in the Jhelum River basin

SR	Station	Latitude (°)	Longitude (°)	Altitude (m. AMSL)	Precipitation		Temperature		
					Period	(% missing)	Period	Tx (% missing)	Tn (% missing)
1	Jhelum	32.94	73.74	287	1970–2009	5.39	1980–2009	20.92	20.97
2	Dhudial	33.06	72.97	518	1983–1997	18.8	1983–1997	1.58	8.85
3	Mangla	33.07	73.63	283	1961–2009	0.01	1962–2009	5.52	7.34
4	Jarikass	33.10	73.84	295	1992–2009	0.00	N/A	N/A	N/A
5	Tandar	33.20	73.98	671	1990–2009	0.62	N/A	N/A	N/A
6	Gujar khan	33.25	73.30	457	1961–2009	4.11	1969–2008	6.53	7.98
7	Kallar	33.42	73.37	518	1961–2009	0.00	1980–2009	3.36	1.77
8	Khandar	33.50	74.05	1,067	1961–2009	4.08	N/A	N/A	N/A
9	Kotli	33.50	73.90	614	1961–2009	2.27	1961–2009	2.66	2.78
10	Quazigund	33.58	75.08	1,690	1962–2009	7.15	1969–2009	9.29	8.53
11	Palandri	33.72	73.71	1,402	1962–2009	0.00	1962–2009	3.50	4.11
12	Sehar kokuta	33.73	73.95	914	1961–2009	5.11	N/A	N/A	N/A
13	Rawalakot	33.87	73.68	1,676	1961–2009	7.80	1970–2009	7.69	7.97
14	Murree	33.91	73.38	2,213	1970–2009	0.23	1970–2009	2.89	2.65
15	Bagh	33.98	73.77	1,067	1961–2009	7.79	1966–2009	9.74	9.86
16	Gulmarg	34.00	74.33	2,705	1961–2009	9.16	1969–2009	22.01	15.76
17	Srinagar	34.08	74.83	1,587	1961–2009	6.55	1961–2009	7.08	0.01
18	Gharidopatta	34.22	73.62	814	1961–2009	4.93	1961–2009	6.92	5.94
19	Domel	34.37	73.47	711	1962–2009	0.00	1963–2008	3.96	1.81
20	Muzaffarabad	34.37	73.48	702	1961–2009	3.55	1961–2009	0.46	0.61
21	Shinkiari	34.47	73.27	991	1961–1996	1.02	1983–1996	16.08	36.60
22	Kupwara	34.51	74.25	1,609	1961–2009	15.38	1977–2009	16.12	15.06
23	Balakot	34.55	73.35	995	1961–2009	0.37	1961–2009	5.01	7.10
24	Shogran	34.64	73.46	2,895	1998–2009	23.11	N/A	N/A	N/A
25	Saif-ul-maluk	34.87	73.69	3,200	1996–2009	1.69	1998–2009	13.80	1.72
26	Naran	34.90	73.65	2,362	1961–2009	8.13	1962–2009	20.10	20.32
27	Astore	35.34	74.90	2,168	1961–2009	0.28	1971–2009	4.09	4.00

Tx max temperature, *Tn* min temperature, *N/A* data not available

3 Methodology

3.1 Quality control

It is the primary emphasis of quality control to treat with outliers before application of any homogenization approach, which can mislead homogenization results (González-Rouco *et al*., 2001; Štěpánek *et al*., 2013; Zahradníček *et al*., 2014). There is a lack of generally recommended methodology for quality control of meteorological data. Thus, in the present study, a combination of different methods applied in Feng *et al*. (2004), Štěpánek *et al*. (2013) and Zahradníček *et al.* (2014) was used to identify erroneous data resulting from observation sources and digitization.

3.1.1 Extreme value check

In this method, daily values of a variable such as temperature are compared with the global and/or local historically observed extreme values of this variable. The data values which are greater than the highest and less than the lowest observed values of a variable are considered as erroneous values. These values are adjusted or removed from the data for subsequent quality control (Feng *et al*., 2004). In the present study, local temperature and precipitation extremes (PMD, 2014; Atta Ur and Shaw, 2015) were compared with the daily records to check outliers in the data series. These local extreme values for temperature and precipitation are presented in Table 2.

Table 2 Local extremes of Tx, Tn and Pr

Variable	High Extreme	Low Extreme	Source
Tx (°C)	53.5	–24.1	(PMD, 2014; Atta Ur and Shaw, 2015)
Tn (°C)	53.5	–24.1	
Pr (mm)	668	–	(PMD, 2014)

3.1.2 Internal consistency check

Reek *et al*. (1992) concluded that the errors in the data series are mostly due to digitizing, unit difference, typos, different way of data reporting etc. So, they developed eight rules to check the erroneous data in meteorological time series. In the present study, the following three rules were used to check the daily time series, as used in Feng *et al*. (2004): 1) *internal consistency* detects the errors such as Tx is lower than Tn, 2) *Flat-liner check* recognizes the same data values for at least seven consecutive days (not applied to zero values of Pr) and 3) excessive diurnal temperature range (Tx–Tn>53.5°) is used to detect extraordinary large daily temperature range (Tx–Tn). Since no highest diurnal temperature range is found in the literature for Pakistan, a value of 53.5°—the highest maximum temperature in Pakistan—was used as the highest diurnal temperature range in the present study. If data values exceed the range of 53.5°, the values are identified as outliers.

3.1.3 Temporal outlier check

The above methods can detect some obvious errors in the data series. However, they cannot identify the errors such as where a data value is significantly different from the previous or the following value in the same time series (Feng *et al.*, 2004). To detect these kinds of outliers, Tukey's method, known as Inter Quartile Range (IQR) method, developed by Tukey

(1997) was used in the present study, as in González-Rouco *et al.*, (2001), Štěpánek *et al*. (2013) and Zahradníček *et al*. (2014) to detect the outliers in the climatic datasets. There are three main steps to detect outliers: 1) to find out the inter quartile range (IQR)—which is the difference between the first quartile (Q1) and the third quartile (Q3); 2) to calculate lower and upper extremes—the lower and upper extremes are calculated by subtracting 1.5×IQR from Q1 and adding 1.5×IQR into Q3, respectively; 3) values beyond these limits are considered to be possible outliers. If an IQR-coefficient of 3 is used, instead of 1.5, to calculate the upper and lower limits, then the values beyond these limits are considered to be the most probable outliers. This method is less sensitive to extreme values than the methods such as Z-score and Standard deviation method which use mean and/or standard deviation to detect outliers. This method has more resistant against outliers because quartiles are used in this method (Tukey, 1977; González-Rouco *et al.*, 2001; Seo, 2006). In the present study, this method was applied on the differences of the test (specific station) and reference series (discussed in the next section) for the detection of erroneous data. In this study, IQR coefficient of 2 was used to give more assurance about outliers.

3.1.4 Spatial outlier check

This method is used to detect those outliers which are not detected by the previously mentioned methods. This method detects outliers by comparing test station values with neighbor stations' values (Feng *et al*., 2004). Since no single method is generally recommended to deal with spatial outliers (Štěpánek *et al*., 2013), a combination of several methods was applied in the present study to identify outliers as done in Štěpánek *et al*. (2013) and Zahradníček *et al*. (2014). In this study, ProclimDB software developed by Štěpánek *et al*. (2010) was used for this purpose. This is a fully automated software for quality control of climate data. In this, several methods are available to detect spatial outliers. Among them, the following methods were used in the present study:

1) *Pairwise comparison method*. In this method, series of differences between test and neighbor stations are created and standardized. Cumulative density funtions (CDFs) for each difference series is calculated. If the average CDF exceeds the critical value (0.95), that value is considered as outlier. It means if the difference between the values at test and neighbor stations is statistically siginificant (α=0.05), the values of the test stations are considered as outliers.

2) *Inter quartile range method*. In this method, limits (higher and lower) are calculated from the neighbor stations and applied to the test series to find out the outliers. In the present study, a value of 2 (Tukey's coefficient) was used during calculation of limits.

3) *Technical series method*. A technical (theoretical) series is created from neighbor stations by means of some statistical methods for spatial data (e.g., kriging and IDW). Then, this series is compared with the test series at a significance level of 0.5.

In the present study, five highly correlated neighbor stations, as discussed below, were used to create theoretical (technical) series for calculating limits for IQR method and for pairwise comparison method.

3.1.5 Creation of reference series

A change in a climatic time series which may be considered as an inhomogeneity in a dataset, but it may also be a result of a change in local or regional climate (Peterson *et al*., 1998).

Several techniques have been introduced to overcome this kind of problems. Most of them use data from some highly correlated nearby stations in the region to establish a new time series (called as reference series) as a descriptor of regional climate. In the present study, a technique used in Zahradníček *et al*. (2014) and Štěpánek *et al*. (2013) was used to create reference series for each variable on each site. According to them, the first step is to select neighbour stations. These stations can be selected either by distances or by correlations. Correlation coefficients can be calculated either from raw station data or first order differences. In the present study, five highly correlated neighbor stations were selected, with the distance restricted to 150 km and altitude difference of 600 m. Then, the datasets of these highly correlated stations were standardized with the mean and standard deviation. At the end, Inverse Distance Weighting (IWD) method, equation 1, was used to take average of five selected standardized neighbors to create reference series.

$$\overline{x} = \frac{\sum_{i=1}^{n} \frac{1}{d_i^p} \times y_i}{\sum_{i=1}^{n} \frac{1}{d_i^p}} \tag{1}$$

where $\overline{x}$ is the reference series; y_i neighbor station; d is the distance between the test and neighbor stations; n is the number of neighbor stations; p is the power of distance—the higher the value of p, the greater the weights for the nearest neighbor station. In this study, a power of 0.5 and 1, as recommended in the manual of ProclimDB (Processing of Climatological Database) software (Štěpánek, 2010), was used to create reference series for temperature (Tx and Tn) and precipitation, respectively.

3.2 Homogenization

The presence of inhomogeneities is a common problem in climate time series. Most of these are related to abrupt changes in average values but also appears as changes in the trend of time series. These irregularities in climate data can deceive the actual results and lead to some wrong conclusions (Vicente-Serrano *et al*., 2010). Thus, to assess some meaningful climate analysis, the climate data must be homogeneous (Štěpánek *et al*., 2009). An ideal way to deal with such irregularities is to examine the station's metadata—that records the historical information about station such as relocation of station, instrument change, type of instruments used etc. After detection of inhomogeneity through the metadata, the temporal variation of the inhomogeneous dataset from the station can be compared with the variation of the neighbor station or regional climate variation. However, most of the time, a complete metadata is not available for all stations in the region. Thus, some alternative subjective and objective methods are used to check the homogeneity (Feng *et al*., 2004). These methods are reviewed comprehensively in Peterson *et al*. (1998) and Costa and Soares (2009). These methods generally used the following steps during homogenization: 1) creation of reference series for comparison with the test series for relative homogenization, 2) application of statistical test to detect irregularities and 3) homogenization—adjustments to compensate with inhomogeneities and imputation of missing data. Since each statistical test renders results with some degree of uncertainty because of noise in the time series (Zahradníček *et al*., 2014), a combination of different tests is considered to be more effective to uncover data inhomogeneity. Thus, in this study, three relative homogeneity tests were applied for homo-

geneity check: SNHT (Alexandersson, 1986), Maronna and Yohai Bivariate test (Maronna and Yohai, 1978; Potter, 1981) and Easterling & Peterson test (Easterling and Peterson, 1995). Reference series for each test station were created from five highly correlated neighbor stations. These series can be divided into a duration of 40 years, with an overlap of 10 years if the series are of long period, e.g., more than 70 years. This is recommended for SNHT test to perform properly. Since there is a lack of methods to detect the inhomogeneities directly from the daily time series (Vicente-Serrano *et al*., 2010), these tests were applied on the monthly, seasonal and annual time series, the same as in Feng *et al*. (2004), Vicente-Serrano *et al*. (2010), Zahradníček *et al*. (2014) and Štěpánek *et al*. (2013). This approach is commonly used for inhomogeneity detection.

In ProclimDB, the main criterion for the identification of a year of breakpoint (abrupt change) is the probability of detection (PD) of a given year. This is the ratio of total detected breakpoints for a given year from all tests to the all theoretically possible breakpoints from all tests. PD values exceeding 10% and 20% (recommended by Štěpánek *et al*., 2010) are used to identify potential inhomogeneities in precipitation and temperature, respectively. The same values were used for the present study. Before taking the final decision about breakpoints, these breakpoints were also examined graphically to reduce any uncertainty.

The inhomogeneous series were corrected on a daily scale. The daily adjustments were calculated based on the reference series and smoothed by low pass filter because this better reflects the physical properties of time series (Figure 1). A 15-year data on both sides of breakpoint was used during the calculation of adjustments.

In adjustment calculation, first the difference series between the test and reference series are calculated before and after the breakpoint. Then, the adjustments are calculated by subtracting the difference series before breakpoint and the difference series after breakpoint. These adjustments can be smoothed by low pass filter, high pass filter or moving average (Štěpánek *et al*., 2013).

For validation of homogenization, correlation coefficients between test and reference series are calculated for each month before application of adjustments and after application of adjustments. Then, these correlations are compared with each other. If there is an increase in change in correlation coefficients, the adjustments are accepted (Zahradníček *et al*., 2014). The same was done in the present study.

The presence of missing data in climate time series is a common problem which must be considered when dealing statistically with the climate data. It can mislead the results and even prevent important analysis of the considered variable from being carried out. Currently, several statistical techniques have been developed to overcome this problem. They span from some simple methods, such as using a mean value, to some very sophisticated techniques, such as multiple imputation. However, their application depends mainly on the percentage of missing data. It is suggested that if percentage of missing values is not greater than 5, any method can be used. However, if percentage of missing values is greater than 5, some sophisticated methods such as regression and multiple imputation methods must be applied (Lo Presti *et al*., 2010). In the present study, a multiple imputation method, predictive mean matching (Heitjan and Little, 1991), was used to deal with missing data because most of the stations available for this study have missing data greater than 5%. On some stations, the missing percentage is even greater than 15% (Table 1). This is a

semi-parametric approach which is similar to regression method except that these missing values are imputed randomly. This method ensures that the imputed values are plausible. It may perform better than regression if the normality assumption is violated (Horton and Lipsitz, 2001).

4 Results and discussion

4.1 Spatial correlation

For homogenization and quality control of climate data, it is essential to get information about spatial correlations among climate stations. Figure 2 shows average correlation coefficient of each climate station with all the other stations in the study area. The correlations were calculated for each variable (Tx, Tn and Pr) between stations, on daily time series. In case of Tx, the highest correlation (0.95) was observed on Kallar, Kotli, Muzaffarabad and Srinagar and the lowest (0.79) on Bagh. In case of Tn, the highest correlation (0.96) was found on Mangla, Srinagar, Domel and Muzaffarabad and the lowest (0.85) on Dhudial. Among Pr stations, Domel had the highest correlation of 0.62, and Gulmarg had the lowest correlation of 0.2. The spatial correlations for Tx and Tn were much stronger than Pr.

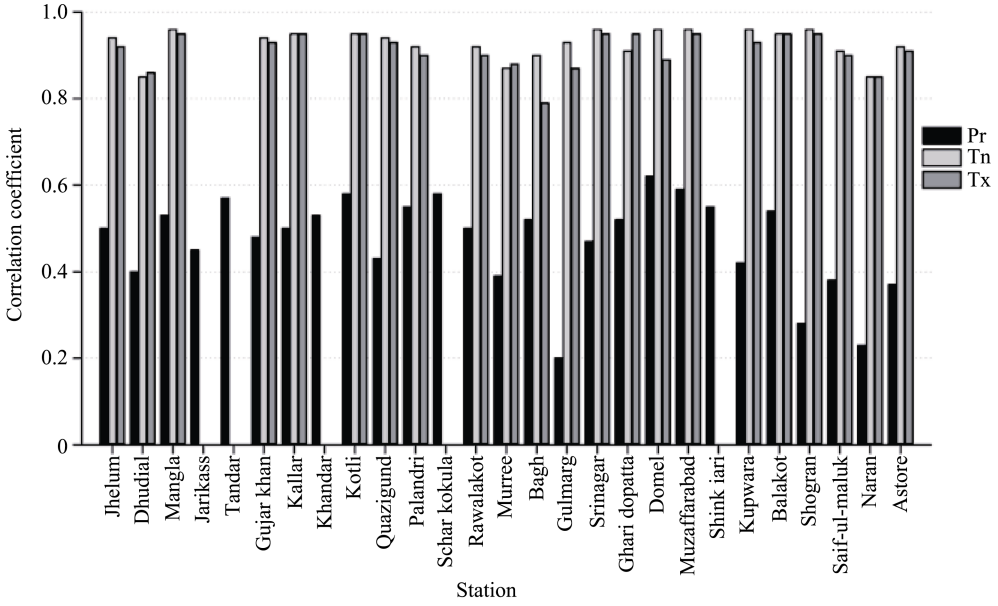

Figure 2 Average correlation coefficients between weather stations in the Jhelum River basin

Figure 3 shows average correlations with respect to distance between climate stations in case of Tx, Tn and Pr. Highly correlated stations were observed within 40 km. As distance exceeded 40 km, correlations decreased quickly in case of Pr. On the other hand, in case of temperature, decreasing rate was very small. As expected, larger distances showed lower correlations between stations. As distance increased, the correlations decreased more quickly in precipitation as compared to temperature.

4.2 Quality control of daily data

In the present study, an extensive methodology comprising four checks (high/low extremes,

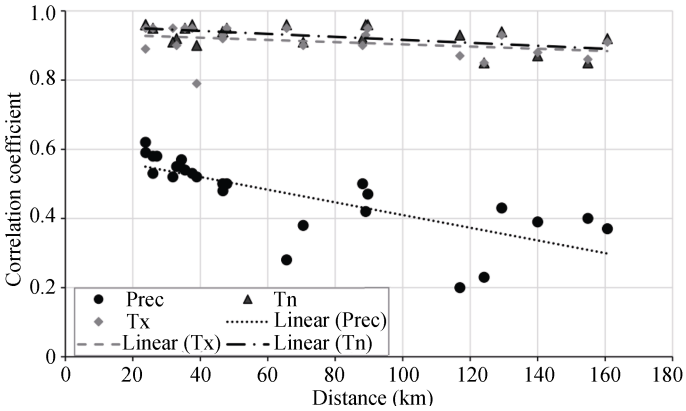

Figure 3 Variation in correlation coefficients with respect to distance between weather stations in the Jhelum River basin

internal consistency, temporal, and spatial outliers' checks) were used to detect outliers. Table 3 shows percentage of outliers detected by each quality control method in Tx, Tn and Pr. These are the total outliers detected on all climate stations, in three variables (Tx, Tn and Pr). Very few values 6 (0.0019%), 24 (0.0076%) and 4 (0.001%) were identified by high/low extreme check in Tx, Tn and Pr, respectively. These values were adjusted manually by examining the neighbor station values as well as previous and following values around the infected value. A total of 846 (0.211%) were recognized as errors in Tx and Tn time series during the internal constancy check. Among them, 354 (0.0558%) errors were identified by *Tx lower than Tn rule*, and no error was detected by *excessive diurnal range check*. These errors were corrected by taking the average of five neighbor stations, and previous and following values of the infected value. Flat-liner check (same seven consecutive values) detected about 216 (0.0677%) values in Tx and 276 (0.0874%) in Tn time series. In this case, all values were removed from the datasets except the first value of each group of same consecutive values, the same as in Feng *et al*. (2004).

Tukey's method detected 768 (0.241%) and 892 (0.283%) temporal outliers in Tx and Tn, respectively, and 730 (0.229%) and 1095 (0.347%) errors were identified by spatial outliers method. The errors detected by temporal and spatial outliers check were removed from the datasets before homogenization process. Table 3 shows that the most infected variable is Tn,

Table 3 Percentages of erroneous data in Tx, Tn and Pr time series during quality control in the Jhelum River basin

Method		Tx (%)	Tn (%)	Pr (%)
Total number of values processed		319100	315794	418499
High/Low extremes		0.0019	0.0076	0.0010
Internal consistency	*Tx lower than Tn*	0.0279	0.0279	–
	Flat-liner	0.0677	0.0874	0.0000
	Excessive diurnal temperature range	0.0000	0.0000	–
Temporal outliers		0.2407	0.2825	–
Spatial outliers		0.2288	0.3467	0.0222
Total		0.5669	0.7521	0.0232

and the less effected variable is Pr. Since, according to the best of our knowledge, no studies are reported on quality control in the Jhelum basin, these results were compared with Feng *et al*. (2004) conducted in China. It was found that Tx and Tn are more problematic than Chinese stations. Nonetheless, precipitation data is of high quality, which is comparable with China precipitation data.

4.3 Homogenization

Table 4 describes the number of stations having inhomogeneous datasets, the number of inhomogeneities in climate variables, and the years of inhomogeneities. A total of 23 inhomogeneities (2 in Pr time series, 9 in Tx, and 12 in Tn) and datasets of 20 stations (2 in Pr time series, 7 in Tx, and 11 in Tn) were identified as inhomogeneous. This means 28% of the climatic series (Tx, Tn and Pr) were found as inhomogeneous. Most of the inhomogeneous stations were found in Tx and Tn data series, with 7 (32%) and 11 (50%) stations, respectively. Only 2 (7%) of the Pr data series were detected as inhomogeneous. Some example of inhomogeneous station and breakpoints detected are show in Figure 4. Since no studies about homogenization are found in Pakistan, these results were compared with some other studies such as Feng *et al*. (2004) conducted in China, Štěpánek *et al*. (2013) in Czech Republic and Zahradníček *et al*. (2014) in Croatia. In Feng *et al*. (2004), Štěpánek *et al*. (2013) and Zahradníček *et al*. (2014), a percentage of inhomogeous stations was 37%, 42% and 23%, respectively. However, they conducted homogenization on more climate variables than this study.

When some data series are detected as inhomogeneous, that data then become questionable or invalid for climate change, climate variability and trend analysis (Feng *et al*., 2004). Since overall 28% of the stations are inhomogeneous, it is necessary to remove inhomogeneities from data series to make it useful for climate analysis in the Jhelum River basin. So,

Table 4 Inhomogeneous stations and number of breakpoints in Tx, Tn and Pr in the Jhelum River basin

SR	Station	Year of inhomogeneities		
		Tx	Tn	Pr
1	Bagh	1970	1970, 2004	
2	Balakot		1979	
3	Dhudial	1989	1989	
4	Domel	1969, 1989	1969	
5	Gulmarg		1987	1968
6	Gharidopatta	1969		
7	Kotli	1981, 1995	1970	
8	Kupwara		1985	
9	Mangla		1972	
10	Murree	1989	1989	
11	Muzaffarabad	1969	1969	
12	Naran	1989		1988
13	Rawalakot		1990	
	Stations having inhomogeneities	7	11	2
	Stations having inhomogeneities (%)	31.8	50.0	7.4

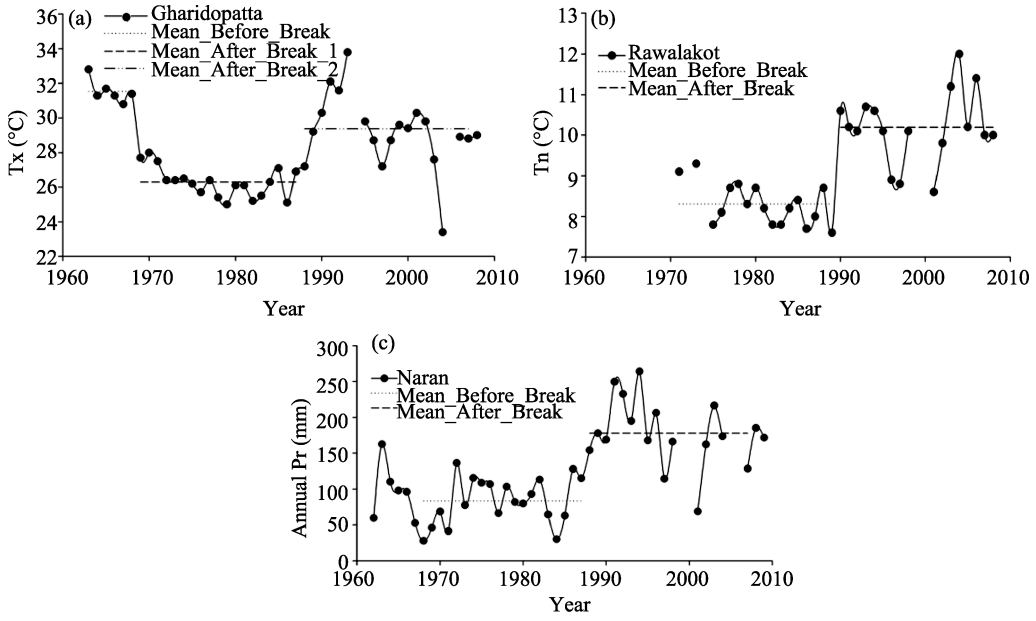

Figure 4 Detected inhomogeneities (a) in Tx on Gharidopatta weather station, (b) in Tn on Rawalakot and (c) in Pr on Naran, in the Jhelum River basin

correction (adjustments) were calculated for each inhomogeneous climate station from daily reference series and then adjusted with the infected raw data to compensate the breakpoints. An example of adjustments calculated for the breakpoint in 1990 for Tn of Rawalakot station is shown in Figure 5.

Figure 6 shows changes in average correlation coefficients (CC) calculated between the test and reference series before and after homogenization, on the infected climate stations. These changes were calculated for each climate station and for each climate variable (Tx, Tn and Pr), and then the monthly average change in CC was taken for all infected climate station. Increasing (positive) change in CC means improvement after homogenization, and decreasing (negative) change means no improvement in data series. Almost all months showed positive change except September (in case of Pr), October (Tx), and November (Pr). The improvement is ranged from 2% to 11% in Tx, from 1% to 8% in Tn and 0.1% to 3% in case of Pr, as shown in

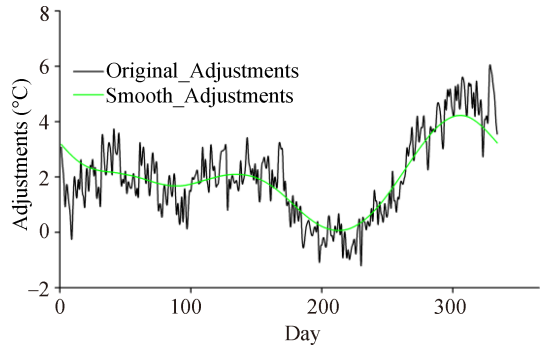

Figure 5 Daily adjustments for the identified breakpoint in 1990 in Tn of Rawalakot station (shown in Figure 4b)

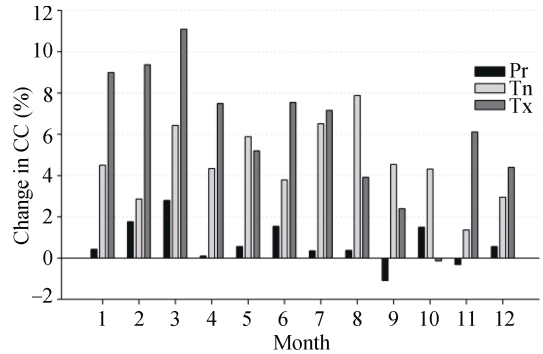

Figure 6 Change in correlation coefficients (CC) between test and reference series before and after homogenization

Figure 6. The maximum improvement was observed in the month of March (in case of Tx), August (Tn) and March (Pr). On the whole, after homogenization, the climate time series are improved and can be used for further climate analysis.

5 Conclusions

In the present study, daily climate data (maximum temperature, minimum temperature and precipitation) of the Jhelum River basin was extensively quality controlled by applying a combination of different methods (high/low extreme check, internal consistency check, temporal and spatial outlier check). Then, inhomogeneities were detected by a combination of relative homogeneity methods (Standard Normal Homogeneity Test (SNHT), Bivariate test and Easterling & Peterson test) and adjusted by applying correction factors calculated from the reference series on daily basis.

During quality control, 0.59%, 0.78% and 0.023% of the total data values were detected as outliers in maximum temperature, minimum temperature and precipitation time series, respectively. During homogenization, maximum temperature series of 32%, minimum temperature series of 50% and precipitation series of 7% were identified as inhomogeneous, in the Jhelum River basin. After homogenization, the infected series were improved by 1% to 11%.

It was concluded that the precipitation daily time series are fairly homogeneous, except two stations (Naran and Gulmarg), and of a good quality. However, the maximum and minimum temperature datasets require an extensive quality control and homogeneity check before using them in climate analysis, especially in climate variability, climate change and trend analysis. The homogenized dataset will be used to assess the impact of climate change on the water resources of the Jhelum River basin in further studies.

Acknowledgements

The authors would like to acknowledge the Pakistan Meteorological Department, the Water and Power Development Authority of Pakistan and the Indian Meteorological Department for providing important and valuable data for this research.

References

Aguilar E, Auer I, Brunet M *et al*., 2003. Guidelines on Climate Metadata and Homogenization WMO/TD No.1186. W. M. Organization: Geneva, 52 pp.

Alexandersson H, 1986. A homogeneity test applied to precipitation data. *Journal of Climatology*, 6: 661–675. doi: 10.1002/joc.3370060607.

Archer D R, Fowler H J, 2008. Using meteorological data to forecast seasonal runoff on the River Jhelum, Pakistan. *Journal of Hydrology*, 361: 10–23. doi: http://dx.doi.org/10.1016/j.jhydrol.2008.07.017.

Atta Ur R, Shaw R, 2015. Disaster and climate change education in Pakistan. In: Rahman A U, Khan A N, Shaw R (eds.) Disaster Risk Reduction Approaches in Pakistan. Japan: Springer, 315–335.

Buishand T A, 1982. Some methods for testing the homogeneity of rainfall records. *Journal of Hydrology*, 58: 11–27. doi: http://dx.doi.org/10.1016/0022-1694(82)90066-X.

Cao L J, Yan Z W, 2012. Progress in research on homogenization of climate data. *Adv. Clim. Change Res.*, 3: 59–67. doi: 10.3724/SP.J.1248.2012.00059.

Costa A, Soares A, 2009. Homogenization of climate data: Review and new perspectives using geostatistics. *Math. Geosci.*, 41: 291–305. doi: 10.1007/s11004-008-9203-3.

Easterling D R, Peterson T C, 1995. A new method for detecting undocumented discontinuities in climatological time series. *International Journal of Climatology*, 15: 369–377. doi: 10.1002/joc.3370150403.

Feng S, Hu Q, Qian W, 2004. Quality control of daily meteorological data in China, 1951–2000: A new dataset. *International Journal of Climatology*, 24: 853–870. doi: 10.1002/joc.1047.

González-Rouco J F, Jiménez J L, Quesada V *et al.*, 2001. Quality control and homogeneity of precipitation data in the southwest of Europe. *Journal of Climate*, 14: 964–978. doi: 10.1175/1520-0442(2001)014<0964: QCAHOP>2.0.CO; 2.

Heitjan D, Little R, 1991. Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society. Series C* (*Applied Statistics*), 40: 13–29. doi: 10.2307/2347902.

Horton N J, Lipsitz S R, 2001. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55: 244–254. doi: 10.2307/2685809.

Kendall M G, 1975. Rank Correlation Methods (Charles Griffin). London: Oxford University Press.

Kruskal W H, 1952. A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, 23: 525–540. doi: 10.2307/2236578.

Kruskal W H, Wallis W A, 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47: 583–621. doi: 10.2307/2280779.

Lo Presti R, Barca E, Passarella G, 2010. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environ. Monit. Assess.*, 160: 1–22. doi: 10.1007/s10661-008-0653-3.

Mahmood R, Babel M S, 2013. Evaluation of SDSM developed by annual and monthly sub-models for downscaling temperature and precipitation in the Jhelum basin, Pakistan and India. *Theor. Appl. Climatol.*, 113: 27–44. doi: 10.1007/s00704-012-0765-0.

Mann H B, 1945. Nonparametric tests against trend. *Econometrica*, 13: 245–259. doi: 10.2307/1907187.

Maronna R, Yohai V J, 1978. A bivariate test for the detection of a systematic change in mean. *Journal of the American Statistical Association*, 73: 640–645. doi: 10.2307/2286616.

Peterson T C, Easterling D R, Karl T R *et al.*, 1998. Homogeneity adjustments of in situ atmospheric climate data: A review. *International Journal of Climatology*, 18: 1493–1517. doi: 10.1002/(SICI)1097-0088(19981115) 18:13<1493::AID-JOC329>3.0.CO;2-T.

Pettitt A N, 1979. A non-parametric approach to the change-point problem. *Applied Statistics*, 28: 126–135. doi: 10.2307/2346729.

PMD, cited 2015: Extreme Events Reports. [Available online at http://www.pmd.gov.pk/journal/extreme- event-slist.htm.]

Potter K W, 1981. Illustration of a new test for detecting a shift in mean in precipitation series. *Monthly Weather Review*, 109: 2040–2045. doi: 10.1175/1520-0493(1981)109<2040:IOANTF>2.0.CO;2.

Seo S, 2006. A review and comparison of methods for detecting outliers in univariate data sets University of Pittsburgh 59 pp.

Štěpánek P, cited 2015: ProClimDB – Software for Processing Climatological Datasets. Available online at http://www.climahom.eu/ProcData.html.

Štěpánek P, Zahradníček P, Skalák P, 2009. Data quality control and homogenization of air temperature and precipitation series in the area of the Czech Republic in the period 1961–2007. *Advances in Science and Research*, 3: 23–26. doi: 10.5194/asr-3-23-2009.

Štěpánek P, Zahradníček P, Farda A, 2013. Experiences with data quality control and homogenization of daily records of various meteorological elements in the Czech Republic in the period 1961–2010. *Quarterly Journal of the Hungarian Meteorological Service*, 117: 1–158.

Szentimrey T, 1999: Multiple analysis of series for homogenization (MASH). Proceedings of the Second Seminar for Homogenization of Surface Climatological Data, Budapest, Hungary, 27–46.

Trewin B, 2013. A daily homogenized temperature data set for Australia. *International Journal of Climatology*, 33: 1510–1529. doi: 10.1002/joc.3530.

Tukey J W, 1977. Exploratory Data Analysis. Pearson.

Vicente-Serrano S M, Beguería S, López-Moreno J I *et al*., 2010. A complete daily precipitation database for northeast Spain: Reconstruction, quality control, and homogeneity. *International Journal of Climatology*, 30: 1146–1163. doi: 10.1002/joc.1850.

Vincent L A, 1998. A technique for the identification of inhomogeneities in Canadian temperature series. *Journal of Climate*, 11: 1094–1104. doi: 10.1175/1520-0442(1998)011<1094:ATFTIO>2.0.CO;2.

Zahradníček P, Rasol D, Cindrić K *et al.*, 2014. Homogenization of monthly precipitation time series in Croatia. *International Journal of Climatology*, 34: 3671–3682. doi: 10.1002/joc.3934.