Received: 27 December 2017 | Revised: 7 January 2019 | Accepted: 15 January 2019

DOI: 10.1002/joc.6008

International Journal of Climatology RMetS

RESEARCH ARTICLE

Methodological application of quantile mapping to generate precipitation data over Northwest Himalaya

Usha Devi[1] | Manorama S. Shekhar[1] | Gyan P. Singh[2] | Nalamasu N. Rao[1] | Uma S. Bhatt[3]

[1]Snow and Avalanche Study Establishment, Research and Development Centre, Chandigarh, India

[2]Department of Geophysics, Institute of Science, Banaras Hindu University, Varanasi, India

[3]Department of Atmospheric Science, Geophysical Institute, University of Alaska Fairbanks (UAF), Fairbanks, Alaska

Correspondence
Usha Devi, Snow and Avalanche Study Establishment, Research and Development Centre, Sector 37, Chandigarh 160036, India.
Email: usha_v08@rediffmail.com

Funding information
USGS, Grant/Award Number: G17AC00213

Continuous high quality data are critical for weather and climate investigations. Numerous data gaps exist particularly over mountainous regions which limits the ability to construct climatologies and perform trend analysis. This study addresses the issue of sparse precipitation data over Northwest Himalaya (NWH) and fills data voids by applying the quantile mapping (QM) method. QM is applied to observed winter precipitation for a period of 25 years (1991–1992 to 2015–2016) to construct a continuous reliable data set. The first 20 years (1991–1992 to 2010–2011) are used for training and the remaining 5 years (2011–2012 to 2015–2016) are used for validation. In total, 10 stations are available for this study and each one is considered serially as a reference to generate daily precipitation values at the other stations. The mean precipitation of NWH region is constructed by considering the mean of all the stations. Standard statistical measures like root mean square errors, standard deviation, skill score and its decompositions are applied to evaluate the generated datasets. Based on statistical analysis, the Kanzalwan station, located in Great Himalaya range, is one of the best performing reference stations for generating precipitation values over NWH. The statistical measures of this station show the highest skill scores, lowest root mean square error and lowest standard mean errors for all winter months except January. This study provides a successful application of QM to generate precipitation data for climate analysis over the complex terrain of the Himalaya region.

KEYWORDS

Himalaya, precipitation, quantile mapping, sparse data density

1 | INTRODUCTION

Climatological analyses of various meteorological parameters are essential for investigating weather phenomena. However, the reliability of meteorological parameters is affected by the percentage of data in a times series that is missing. There are many reasons for gaps in observed time series such as communication breakdowns, non-response of the observer, technical issues and instrument failure. This problem is amplified when dealing with missing daily precipitation values for stations located over the Himalayan region, since this region has a sparse network and high spatial and temporal variability. The Himalayas provide fresh water for nine rivers in Asia, serving approximately 500 million inhabitants comprising 10% of North India's human population and also providing water for agriculture throughout the year (IPCC 2007). Accurate forecasting is required to anticipate avalanches in order to protect local residents during their day-to-day activities as well as members of the Indian army as they conduct operational activities and plan future deployments. Therefore, improving the precipitation data by filling in gaps will serve these forecasting needs. Scientists generally deal with gaps in time series by either using methods that tolerate missing data, or by limiting analysis to periods of continuously available data in order to do meaningful scientific analysis.

To generate missing precipitation values, various empirical and statistical techniques have been developed over the past few years. The arithmetic averaging (AA) (Linacre 1992) and Inverse distance interpolation (ID) methods (Wei and McGuinness 1973) are examples of empirical methods. Moreover statistical methods such as multiple linear regression (REG) (Tabony 1983; Kim et al. 1984), kriging (Hevesi et al. 1992a, 1992b) and optimal interpolation. However, Ramos-Calzado et al. (2008) demonstrated that AA and linear interpolation (Lowry 1972) method has limited applicability to precipitation, which has high temporal variability. The Handbook of Hydrology (ASCE 1996) recommends using the normal ratio and inverse distance weighting methods for estimating missing data values in a time series. Suhaila et al. (2008) fill missing target station data from neighbouring stations by using the modified inverse distance and normal ratio methods. Silva et al. (2007) produced monthly missing rainfall values at stations in Sri Lanka by applying select statistical methods, from surrounding stations and demonstrated that the inverse distance method was suitable for low terrain. The normal ratio method was found suitable for hilly areas as well as higher altitude zones. Despite modifications to traditional and spatial interpolation methods there are still some limitations. Teegavarapu (2016) discussed the limitations of inverse distance weighting and spatial interpolation methods. To estimate point rainfall, the arithmetic average and inverse-distance methods are not suitable in mountainous regions (Tung 1983).

Previous studies applied various techniques to generate the missing data but found limited applicability in limited areas consistent with numerous studies: Kemp et al. (1983); Eischeid et al. (1995); Degaetano et al. (1995). Simolo et al. (2010) suggest that regression leads to both an over and under-estimation in the number of heavy precipitation events since the probability distribution is not conserved. The Artificial neural network (ANN) technique is also used to generate the missing precipitation values. Coulibaly and Evora (2007) examined six different types of artificial neural network to fill the missing values in daily precipitation and extreme temperatures in Northeastern Canada. They found that the MLP artificial neural network estimated missing values of the precipitation well. Teegavarapu (2007) uses the ANN to a fitted semivariogram model within ordinary kriging and demonstrated that ANN with in-kriging is better than the ordinary kriging. However, it still has limitations in the selection of the semivariogram model, distance intervals, and the computational power (Teegavarapu 2009).

While many studies have been carried out to estimate missing values of rainfall and temperature globally using different methods, few studies (Bhutiyani et al. 2010; Kanda et al. 2017) have estimated missing values over the Northwest Himalayan (NWH) region and the Karakoram region, key regions for Indian national security. Considering the strategic importance of the NWH, an attempt has been made to generate missing precipitation values over the region by using quantile mapping (QM), which employs an empirical cumulative distribution function. Actually, this technique effectively reduces errors by preserving information on the frequency distribution of modelled and observed precipitation data (Lafon et al. 2012). The data and methodology is described in Section 2, results and discussion in Section 3 and the conclusion in Section 4.

2 | DATA AND METHODOLOGY

2.1 | Data

The Indian Snow and Avalanche Study Establishment (SASE) maintain 48 Western Himalaya surface observatories, which collect precipitation data twice daily (at 0830 and 1,730 hr Indian Standard Time [IST]) using a snow stack (a plane surface with a 1 m measuring stick perpendicular to the surface) and a rain gauge. The SASE observatories provided high quality daily observed precipitation data at 10 stations spanning three important Himalayan ranges for the period 1991–1992 to 2015–2016 (25 years) during winter (November–April) for the present study. This study focuses on extreme winter precipitation events that are associated with important synoptic weather systems called western disturbances (WDs). WDs travel from west-to-east across this region and result in snow accumulations that lead to avalanche danger (Rao and Srinivasan 1969). The 10 stations located in the three mountain regions are shown in Figure 1 while mean climatological data are provided in Table 1. The 10 station comprise the following regions: three stations in

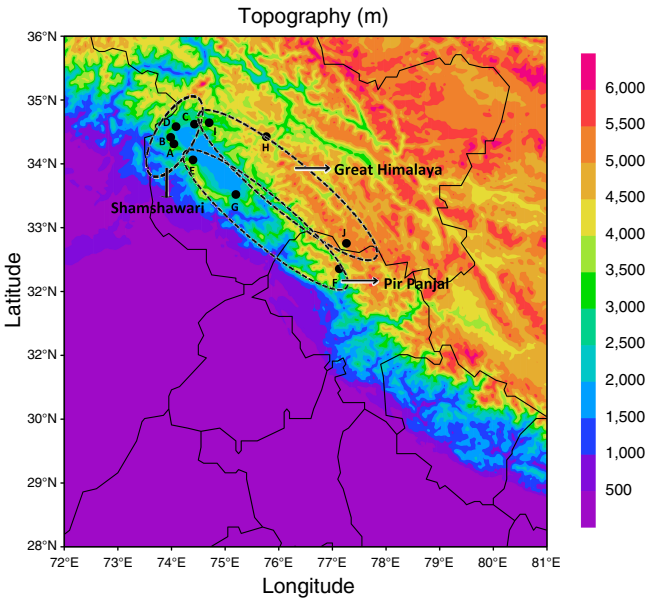

FIGURE 1 Domain representing the study area showing the location of the 10 stations along with topography (m) and names of the mountain ranges in the Northwest Himalaya region. Table 1 provides the station names that corresponds to the capital letters in the map [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Details of stations located in the study domain

Names	Stations code	Altitude (m)	Ranges	Data availability	Average seasonal precipitation (mm)
H-Taj	A	3,080	Shamshawari	1973–2016	1038.6
Stage-2	B	2,650	Shamshawari	1991–2016	1168.4
Z-Gali	C	3,192	Shamshawari	1992–2016	1166.6
Pharkiyan	D	2,960	Shamshawari	1992–2016	1029.8
Gulmarg	E	2,800	Pir Panjal	1991–2016	844.7
Dhundi	F	3,050	Pir Panjal	1989–2016	1161.6
Banihal Top	G	3,250	Pir Panjal	1992–2016	804.0
Drass	H	3,230	Great Himalaya	1991–2016	415.0
Kanzalwan	I	2,440	Great Himalaya	1973–2016	1051.8
Patsio	J	3,800	Great Himalaya	1983–2016	477.2

Pir Panjal, four in Shamshawari and three in the Great Himalaya.

In this study, 25 years of precipitation time series are available without any missing values and for the analysis the data is split into two periods. The first 20 years of data from 1991–1992 to 2010–2011 has been used to develop the algorithm and the remaining 5 years from 2011–2012 to 2015–2016 are used to validate the results. A total of 155 daily values were generated for each winter month during the 5 years validation period.

2.2 | Quantile mapping

Before applying the QM, a Kolmogorov–Smirnov two-sample test (Chakravarti and Laha 1967) was performed on the empirical cumulative distribution functions (ECDFs) between all stations. It revealed that all station time series have similar ECDFs which confirm the capability of the QM technique. The QM approach has been typically used to bias-correct model forecast data with observational data. QM algorithms generally perform better than simple bias correction methods which only preserve mean and variance of the precipitation time series (Teutschbein and Seibert 2012; Chen et al. 2013). In the QM method, the empirical probability distribution of data assumes that there are no gaps in the time series that is used to produce a continuous observed record. Forecasted data generated by numerical weather prediction (NWP) models have their own errors and using these data to fill gaps in observed time series can result in spurious observed precipitation values and may not be very reliable for climate studies. Hence, instead of using the forecasted data, the observed data have been used as reference data to generate the missing precipitation values over Northwest Himalaya. Each station is considered as the reference individually to generate precipitation values at other stations and the combined mean of all these stations (except reference station) produce the mean precipitation over NWH region. This procedure has been adopted for all the months (November–April).

The generated precipitation output is an inverse of cumulative distribution function (CDF) of observed values at the probability which corresponds to the reference CDF at the particular value. Details of this method can be seen in Wood et al. (2002) and Gudmundsson et al. (2012). However, for the sake of completeness, important steps for finding the empirical probability distribution function will be discussed. The empirical probability distribution is obtained by simply fitting a histogram for a given variable and then dividing the frequency of each class by the total number of observations. For this purpose, the number of classes created should be sufficiently large and typically has a minimum number of classes of 5 for short and 20 for long data series to ensure reliable results.

This provides a set of probabilities falling in each class say $P(x_i)$ where the subscript $i = 1\ldots n$, where n is the number of classes.

Mathematically the CDF ($C(x_i)$,) equation is written as

$$C(x_i) = \int f(t)dt, \tag{1}$$

where $f(t)$ is the probability density function.

If the equation is discretized then the function looks like

$$C_i = \sum_{i}^{n} P(x_i), \quad i = 1, 2, 3\ldots\ldots n, \tag{2}$$

where $\sum C_i = 1$. In addition, C_i will provide the fraction of total number of data points below a particular value (the quantile of the particular class). The inverse of the CDF will give the value at a particular probability and is called a quantile function.

Suppose F_o and F_r are CDFs for the observed and reference data sets, respectively. Then for a reference output X the estimated value Y will be as follows.

$$Y = F_o^{-1}(F_r(X)). \tag{3}$$

Here, F^{-1} is an inverse of CDF. Thus, QM is a transformation technique between two CDFs as shown in Figure 2, where Z-Gali (station C) precipitation data is generated by using Kanzalwan (station I) data as a reference data for the month of February.

2.3 | Validation approach

A simple and effective approach to evaluate the process known as split validation has been used. Here, a large

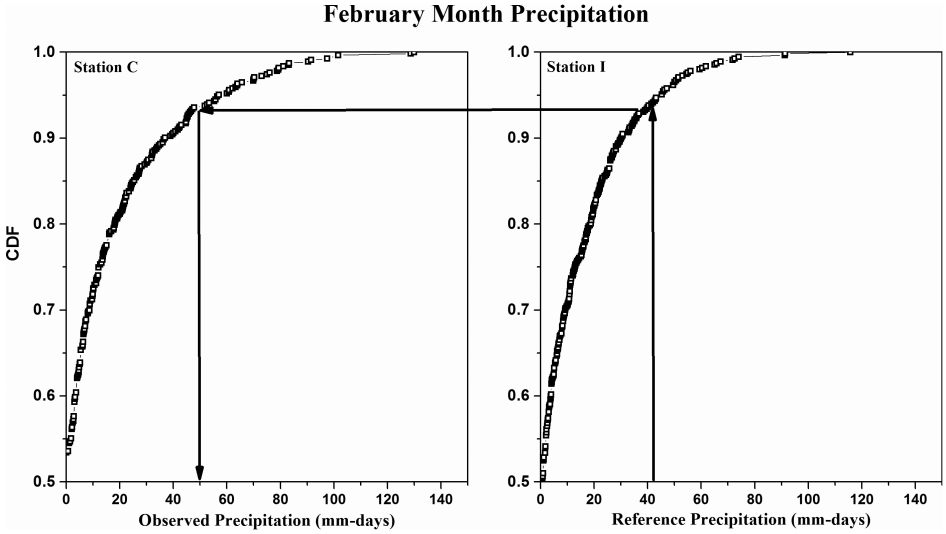

FIGURE 2 Description of QM method with an example for generating precipitation at the station Z-Gali (left panel) using Kanzalwan station as a reference data (right panel) for the month of February

fraction of a dataset is used to develop the algorithm for generating the precipitation values and a smaller fraction of the dataset is kept for verifying the methodology (WMO No.-100 2011). Smaller datasets are generated using the algorithm developed from the larger datasets, and the generated values are compared with observed datasets to evaluate the skill of the algorithm. Most studies evaluate skill score (SS), mean absolute errors (MAE), similarity index (SI), coefficient of efficiency (CE), root mean square error (RMSE) and several other statistical measures because these measures quantify correctness of fit and variability of the generated data compared to observations. In this study, generated daily precipitation time series produced by the QM method are validated by applying RMSE and SS.

The RMSE (also called root-mean-square deviation [RMSD]) represents the fluctuation and error between model generated and observed values. RMSE is a robust measure of accuracy. RMSE values that are less than half of the standard deviation (*SD*) of the observed data are generally considered to be sufficiently low and it is appropriate for model validation (Singh et al. 2004).

$$\mathrm{RMSE} = \sqrt{1/N\sum_{i=1}^{N}\left(X_i^{\mathrm{Gen}} - X_i^{\mathrm{Obs}}\right)^2}, \tag{4}$$

where X_i^{Gen} and X_i^{Obs} represent generated and observed data respectively in Equation (4) and N is the number of total generated data points.

The skill of the generated values provides correctness relative to the observed values. The SS has been calculated by using mean square error (MSE) and *SD* of the observed data as shown in Equation (5). A skill score of 1 represents perfect match between generated and observed and a value of 0 indicates that the generated data has no improvement on climatology (Hashino et al. 2007).

$$\mathrm{SS} = 1 - \frac{\mathrm{MSE}}{\sigma_x^2}. \tag{5}$$

According to the Murphy and Winkler (1992), SS can be further decomposed as:

$$\mathrm{SS} = \rho_{rx}^2 - [\rho_{rx} - (\sigma_r/\sigma_x)]^2 - [(\mu_r - \mu_x)/\sigma_x]^2, \tag{6}$$

where ρ_{rx} represents the correlation between model generated and observed values, σ is the *SD* of generated (subscript *r*) and observed (subscripted *x*) data, μ_r and μ_x are the mean of the generated and observed values, respectively. The terms in right hand side of Equation (6) represents potential skill (PS), slope reliability (SREL), and standardized mean error (SME). Potential skill (PS) is what can be achieved by eliminating conditional and unconditional biases where slope reliability (SREL) is a measure of the conditional bias and SME is a measure of the unconditional bias (Murphy and Epstein 1989).

Taylor (2001) developed a graphical method (Taylor's diagram) for summarizing the statistical comparison between observed and generated data. Luu and Tkalich (2014) have compared the statistics between observations and experiments using Taylor diagrams for the constructed mean sea level time series by establishing a relationship between El Niño–Southern Oscillation (ENSO) and Asian monsoon. A similar statistical approach has been adapted (Taylor diagram shown in Figure 4) in the present study for comparing the precipitation datasets over NWH.

3 | RESULTS AND DISCUSSION

An evaluation of generated and observed precipitation data using QM against competing methods (inverse distance interpolation (ID), normal ratio method (NR) and multiple regression analysis (REG)) is provided in Table 2. The

TABLE 2 Comparison between QM and three other methods

Months	Methods	SS	PS	SR	SME	RMSE
November	ID	0.70	0.77	0.06	0.01	1.72
	NR	**0.79**	**0.83**	**0.04**	**0.01**	**1.44**
	REG	0.67	0.68	0.00	0.01	1.81
	QM	0.74	0.75	0.00	0.00	1.60
December	ID	0.67	0.72	0.06	0.00	3.18
	NR	0.71	0.76	0.05	0.00	2.95
	REG	0.73	0.77	0.04	0.00	2.83
	QM	**0.93**	**0.94**	**0.00**	**0.00**	**1.41**
January	ID	**0.80**	**0.81**	**0.00**	**0.00**	**3.87**
	NR	0.76	0.78	0.02	0.00	4.29
	REG	0.68	0.70	0.01	0.01	4.94
	QM	**0.79**	**0.81**	**0.00**	**0.01**	**3.98**
February	ID	0.70	0.72	0.00	0.01	7.46
	NR	0.76	0.77	0.00	0.01	6.74
	REG	0.76	0.79	0.03	0.00	6.70
	QM	**0.87**	**0.87**	**0.01**	**0.00**	**5.01**
March	ID	0.88	0.90	0.02	0.00	4.08
	NR	**0.91**	**0.91**	**0.00**	**0.00**	**3.61**
	REG	0.77	0.82	0.05	0.00	5.73
	QM	0.89	0.90	0.00	0.01	4.04
April	ID	0.63	0.72	0.09	0.00	4.73
	NR	0.67	0.77	0.10	0.00	4.50
	REG	0.75	0.75	0.00	0.00	3.90
	QM	**0.83**	**0.85**	**0.02**	**0.01**	**3.22**

Note. ID: inverse distance Interpolation; NR: normal ratio method; REG: regression (Significant methods are represented by bold values).

details of these methods are available in Kashani and Dinpashoh (2012). At least two stations are required to use the ID or NR methods otherwise these methods will provide the same values as those of the training station. However, only one station was needed for using the QM methodology to generate data at other stations (Table 2) and additionally, QM produced the most statistically robust results. Therefore, the present study focuses on the QM method. Precipitation data at each of the 10 stations was used to produce data at the other 9 stations. The mean of these 9 station time-series represents data over the whole NWH. The data sets generated by QM for the period 2011–2012 to 2015–2016 (5 years or 155 days for each month, November–April) were validated by using various standard statistics (e.g., *SD*, mean, RMSE, skill score and its decompositions). The MSE Skill Score and its decomposition of precipitation generated for Gulmarg (E), Banihal Top (G) and Kanjalwan (I) displays high skill scores and low errors (SREL and SME) in November (Figure 3). During December, Dhundi (F), Drass (H) and Kanjalwan (I) produced slightly better statistical results than other stations. The stations H-Taj (A), Z-Gali (C) and Pharkiyan (D) display higher skill scores and lower errors compared to the others stations in January. In February, Stage-2 (B), Z-Gali (C), Pharkiyan (D) and Kanjalwan (I) display high skill scores and low errors. While in March, the stations Stage-2 (B) and Kanzalwan (I) had high SS and low error. Similarly, the stations Stage-2 (B), Z-Gali

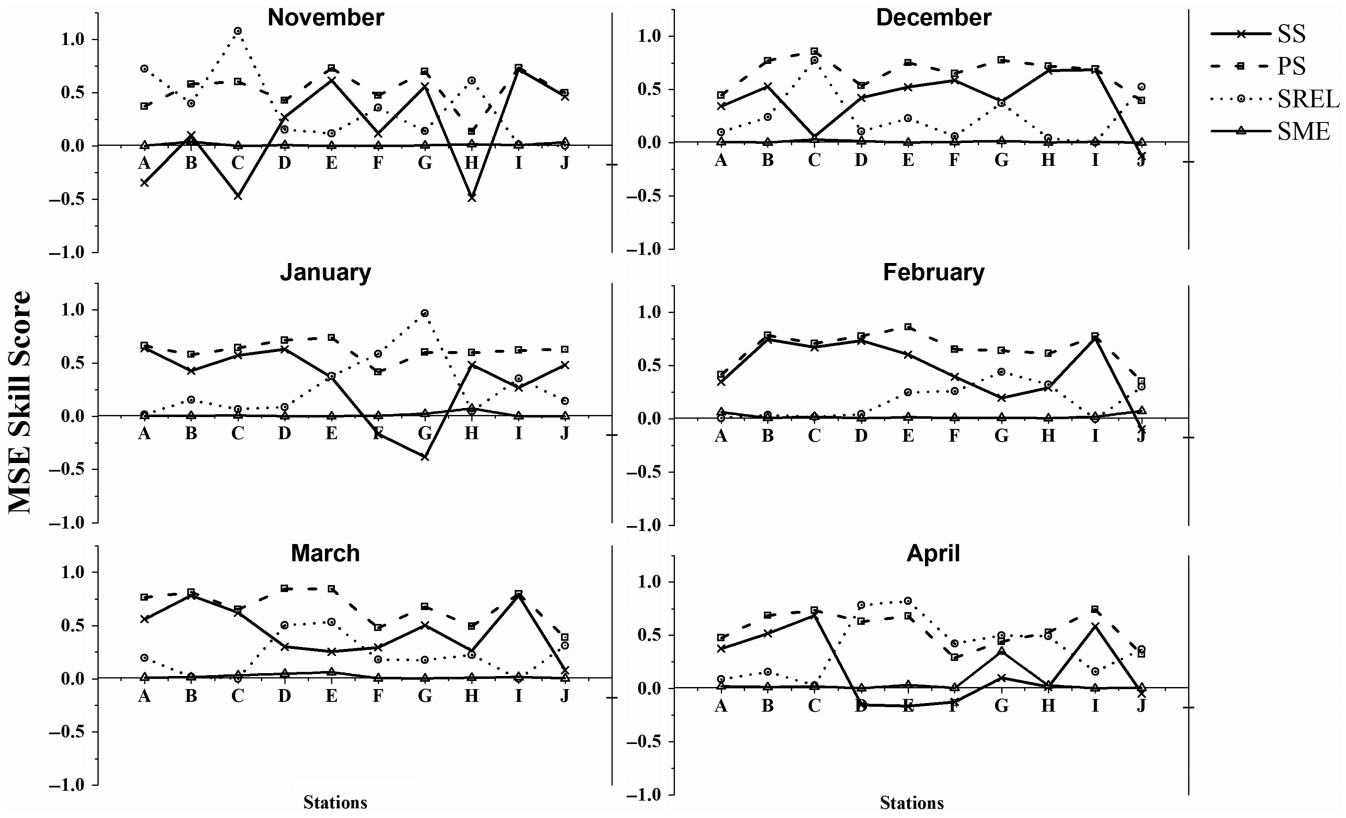

FIGURE 3 Monthly (November to April) MSE skill score (SS) and its decompositions, that is, Potential skill (PS), slope reliability (SREL) and standard mean error (SME) by 10 stations for NWH region

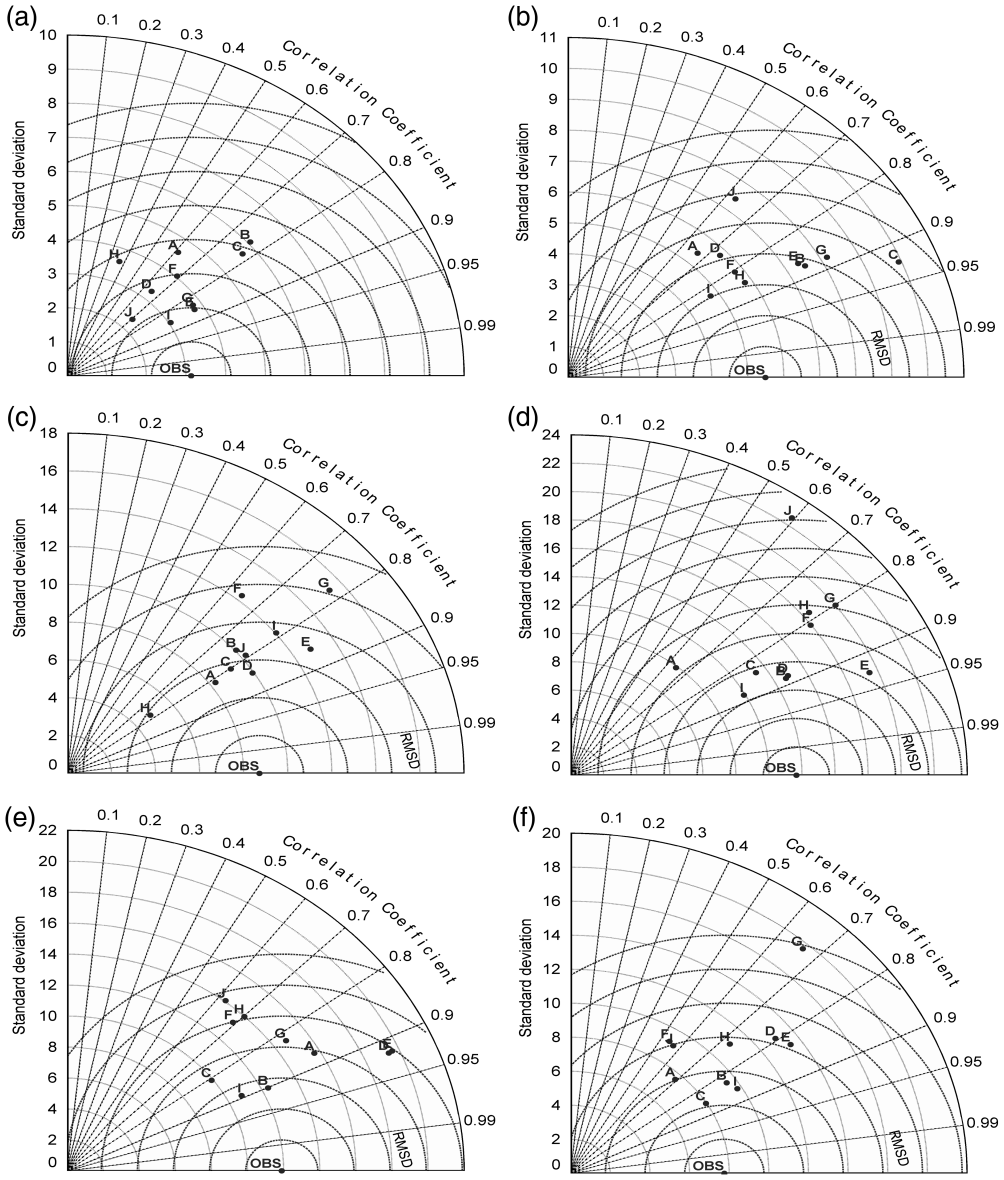

FIGURE 4 Taylors diagram representing the monthly performance of various stations (10) of NWH during (a) November, (b) December, (c) January, (d) February, (e) March and (f) April

(C), and Kanjalwan (I) have better results compared to the other stations over NWH during April.

Kanzalwan (station I) has a high correlation, low RMSE and almost the same *SD* as the observed data in November (Figure 4). During December, observed data have a *SD* of 5.5 mm/day, the stations Drass (H) and Kanzalwan (I) show little deviation from observed data, display low RMSE and high correlation with observations. Compared

TABLE 3 Performance of stations in terms of statistical measures (high SS and low RMSE) represented by tick marks from November to April

Names	Stations code	November	December	January	February	March	April
H-Taj	A			✓			
Stage-2	B				✓	✓	✓
Z-Gali	C			✓	✓		✓
Pharkiyan	D			✓	✓		
Gulmarg	E	✓					
Dhundi	F		✓				
Banihal Top	G	✓					
Drass	H		✓				
Kanzalwan	I	✓	✓		✓	✓	✓
Patsio	J						

TABLE 4 Precipitations thresholds statistics of 10 stations (A–J) for the months of November–April

Months	Precipitation thresholds (mm)	Statistics	OBS	A	B	C	D	E	F	G	H	I	J
November	**<3**	Mean	0.24	0.13	0.08	0.05	0.07	0.12	0.00	0.03	0.07	0.05	0.00
		SD	0.60	0.49	0.36	0.23	0.39	0.47	0.00	0.25	0.45	0.22	0.00
	3–7	Mean	4.30	0.00	4.70	4.57	5.08	4.15	4.16	4.52	4.28	4.82	0.00
		SD	1.09	0.00	1.42	1.18	1.04	1.01	0.91	1.76	0.00	1.17	0.00
	7–12	Mean	8.43	8.14	9.71	0.00	7.33	9.87	9.11	8.81	0.00	9.37	8.66
		SD	0.91	1.15	3.15	0.00	0.21	2.89	1.44	1.81	0.00	0.41	0.00
	12–16	Mean	**12.88**	13.95	0.00	12.27	0.00	13.45	13.21	15.11	13.37	**0.00**	12.81
		SD	**0.41**	0.92	0.00	0.00	0.00	0.89	0.00	0.04	0.00	**0.00**	0.48
	>16	Mean	27.71	46.78	27.11	37.92	26.20	27.31	24.96	27.06	29.38	22.37	16.78
		SD	**0.00**	0.00	14.32	14.77	1.19	7.42	7.16	9.75	6.01	**5.77**	0.00
December	**<5**	Mean	0.30	0.21	0.23	0.23	0.20	0.15	0.11	0.00	0.21	0.41	0.11
		SD	0.82	0.70	0.80	0.85	0.66	0.62	0.67	0.00	0.82	1.15	0.57
	5–10	Mean	6.99	7.53	6.97	6.36	8.60	6.72	7.15	8.35	7.88	7.33	5.83
		SD	1.84	1.13	1.63	0.67	1.67	1.26	1.19	1.47	1.55	1.02	0.35
	10–15	Mean	12.14	12.50	11.26	0.00	14.35	10.87	11.41	13.23	11.99	12.36	12.25
		SD	1.31	2.00	1.83	0.00	0.00	0.64	1.05	1.11	1.38	2.56	1.76
	15–20	Mean	18.19	0.00	18.21	18.42	16.70	0.00	15.69	0.00	17.67	16.63	17.53
		SD	1.13	0.00	0.86	1.36	1.46	0.00	0.64	0.00	1.94	1.34	3.35
	>20	Mean	24.99	29.42	33.68	37.64	36.20	31.64	33.54	29.65	25.01	25.06	35.52
		SD	4.88	10.49	12.71	13.18	5.91	7.31	16.15	8.55	4.59	5.85	7.51
January	**<5**	Mean	0.54	0.40	0.31	0.42	0.37	0.34	0.35	0.04	0.50	0.29	0.29
		SD	1.10	1.08	1.00	1.15	1.11	1.01	0.97	0.43	1.18	0.90	1.02
	5–10	Mean	5.89	6.87	8.10	6.77	7.30	7.25	7.14	7.66	7.52	6.96	7.21
		SD	1.26	1.14	0.89	1.66	1.56	1.22	1.12	1.34	1.34	1.51	1.00
	10–20	Mean	14.36	16.08	16.62	15.76	15.45	14.65	12.80	11.73	14.53	16.52	13.88
		SD	3.48	1.90	2.01	3.17	2.83	2.82	3.09	1.32	3.02	3.46	2.47
	20–30	Mean	24.99	24.19	24.99	25.01	23.73	25.29	23.48	23.96	0.00	23.27	23.16
		SD	3.91	2.77	3.68	2.22	2.77	2.41	2.03	3.25	0.00	2.22	3.27
	>30	Mean	**39.79**	35.91	41.79	44.72	40.92	56.54	61.53	57.43	43.24	**47.14**	48.72
		SD	**5.49**	4.91	13.95	9.98	13.18	29.59	17.73	18.40	0.00	**21.03**	14.06
February	**<5**	Mean	0.70	0.44	0.17	0.34	0.39	0.39	0.30	0.00	0.35	0.54	0.15
		SD	1.12	1.00	0.56	0.72	1.04	1.11	0.98	0.00	1.23	1.14	0.76
	5–10	Mean	5.90	6.88	6.85	7.22	7.57	7.62	7.37	8.00	6.40	8.08	7.34
		SD	1.05	1.61	1.65	1.43	1.84	1.47	0.74	1.20	1.33	1.44	1.68
	10–30	Mean	17.23	17.20	17.23	18.23	19.08	22.05	18.40	18.98	18.16	19.55	18.70
		SD	4.86	7.15	5.66	5.19	6.97	5.43	4.53	6.11	6.22	6.20	5.13
	30–50	Mean	37.59	36.68	37.42	42.15	37.33	36.44	36.30	39.98	36.97	34.39	38.21
		SD	6.17	6.02	6.11	5.31	5.55	3.27	7.40	7.38	5.33	3.14	7.04
	>50	Mean	63.07	51.12	70.24	63.13	61.78	67.77	67.91	71.27	87.46	62.90	71.80
		SD	**10.09**	0.00	8.70	14.72	20.38	13.93	19.84	15.16	40.19	**2.65**	24.24
March	**<5**	Mean	0.70	0.51	0.26	0.59	0.30	0.44	0.31	0.00	0.30	0.51	0.40
		SD	1.18	1.21	0.88	1.24	0.94	1.10	0.93	0.00	1.06	1.07	1.23
	5–10	Mean	6.68	7.63	6.89	7.17	7.82	7.63	7.47	8.99	7.46	7.21	7.46
		SD	1.07	1.61	1.43	1.71	1.61	1.46	1.18	1.10	1.19	1.59	1.19
	10–30	Mean	18.35	17.39	17.22	19.39	20.41	19.45	17.93	17.99	16.80	17.32	17.53
		SD	5.29	5.69	5.49	5.15	5.51	5.93	5.23	4.98	4.70	5.63	4.99
	30–50	Mean	35.77	38.55	42.35	30.83	40.67	41.85	34.34	37.75	37.58	38.06	39.25
		SD	6.78	5.58	6.42	0.00	5.92	7.15	3.26	5.37	6.75	5.81	6.37
	>50	Mean	57.02	61.10	68.11	82.96	92.89	73.33	77.84	62.33	69.59	57.45	77.96
		SD	6.44	15.62	23.75	0.00	25.60	17.19	17.57	7.92	13.46	3.59	16.68

TABLE 4 (Continued)

Months	Precipitation thresholds (mm)	Statistics	OBS	A	B	C	D	E	F	G	H	I	J
April	**<5**	Mean	1.14	0.54	0.59	0.60	0.35	0.75	0.48	0.20	0.04	0.70	0.26
		SD	1.46	1.23	1.07	1.28	0.96	1.26	1.20	0.98	0.42	1.23	1.10
	5–10	Mean	7.08	6.97	7.52	6.73	6.73	7.15	7.67	5.87	7.39	7.69	8.05
		SD	1.41	1.34	1.41	1.32	1.30	1.44	1.11	1.14	1.32	1.35	0.73
	10–20	Mean	15.15	13.88	13.28	13.79	16.06	13.60	16.88	15.51	12.67	15.92	13.79
		SD	2.37	2.47	2.14	2.43	1.60	2.78	2.88	2.32	2.49	2.15	2.60
	20–30	Mean	25.42	25.98	24.81	24.05	24.89	23.77	23.33	23.76	21.89	23.61	26.15
		SD	3.94	3.88	3.80	1.99	2.78	2.32	2.88	3.29	2.05	3.59	1.73
	>30	Mean	**32.91**	36.70	48.18	41.47	58.07	44.59	37.02	50.03	48.73	**38.76**	40.45
		SD	**1.18**	8.02	6.73	5.25	17.09	21.20	9.28	19.86	13.22	**5.92**	10.78

to other stations, H-Taj (A) depicts stronger correlations, lower RMSE and the same *SD* as the observed data in January. For the month of February, the generated data at the stations Stage-2 (B) and Pharkiyan (D) show a slight deviation in *SD* compared to the *SD* of observed data. Whereas at station Kanzalwan (I), the difference in *SD* is slightly higher than observations, but the correlation and RMSE are approximately the same as those of the two stations (B and D). In March the stations Stage-2 (B) and Kanzalwan (I) have a RMSE that is less than half of the *SD* of the observed data and high correlations. However, Stage-2 (B) displays low variation in *SD* compared to Kanzalwan (I). Finally, in April, the stations Stage-2 (B), Z-Gali (C) and Kanzalwan (I) display high correlations and low RMSEs compared to the other stations. The performance of the stations in terms of various statistical measures is summarized in Table 3 and clearly shows that the Kanzalwan (station I) has the best statistical results in all the months except for January.

To quantify biases of the generated data using the QM approach, two statistical measures mean and *SD* have been calculated for 5 precipitation thresholds for each month at different stations (Table 4). Comparing the observed and generated data at Kanzalwan (I) indicates that the mean and *SD* are not well captured for extreme events >12, >30 and >50 mm for November, April and February, respectively. For the month of January, the *SD* of the generated data displays a high deviation for >30 mm events (represented by bold values in table 4). This can be attributed to low skill score in January of the Kanzalwan station. Scatter plots have also been used to assess the performance of QM methodology. Each of the 10 stations was considered individually to serve as the reference station and scatter plots were constructed for all of the months (November to April) to examine the behaviour of very low/zero or very high precipitation events. Figure 5 shows a direct comparison of generated precipitation data for all the stations (A–J) with observed precipitation data. The strength of the statistical relationship is quantified by the adjusted R^2, which varies between 0.13 and 0.85. Note that Patsio (J), located in Great Himalaya range, does not show a strong statistical relationship (low value of adjusted R^2) in any given month. This may be due to its location over high mountains where it is strongly affected by orographic features.

The present study is consistent with the results obtained by Dimri et al. 2008 regarding a precipitation forecast using the k-nearest neighbour method over the Western Himalaya. According to Dimri et al. (2008), the Himalayan regions generally receives different patterns of precipitation depending on the frequency and movement of Western Disturbances eastward and as they cross the Pir-Panjal range, leading to interactions with the other ranges. Due to the close association of WD with precipitation at these stations, the generated value may not perform well for all the other stations. The analysis of this study, which used various statistical measures (e.g., SS, correlations, errors, etc.) suggests that the availability of at least two or three stations which have a high SS and low errors in each month can be used to generate a precipitation time series representative for the whole NWH area. Results over NWH can be improved by considering all the stations that have a good SS and low errors for a given month.

Moreover, the Taylor diagram indicates that the stations A, B, C and I are most suitable to serve as reference stations for generating the mean regional precipitation time series over the NWH. Since the initial data availability at some of the stations begins in 1973 (Table 1), QM can be applied to fill in gaps and generate a longer time series for the entire NWH region. This would result in a complete 30–40 years long time series of data and could be used to define a climatology and for trend analysis over the NWH. This study provides a successful application of the above mentioned method for generating precipitation data sets over the NWH. A next step in this area of research will involve the generation of precipitation data by dividing the stations into different classes based on altitudes and location within mountain ranges of the NWH.

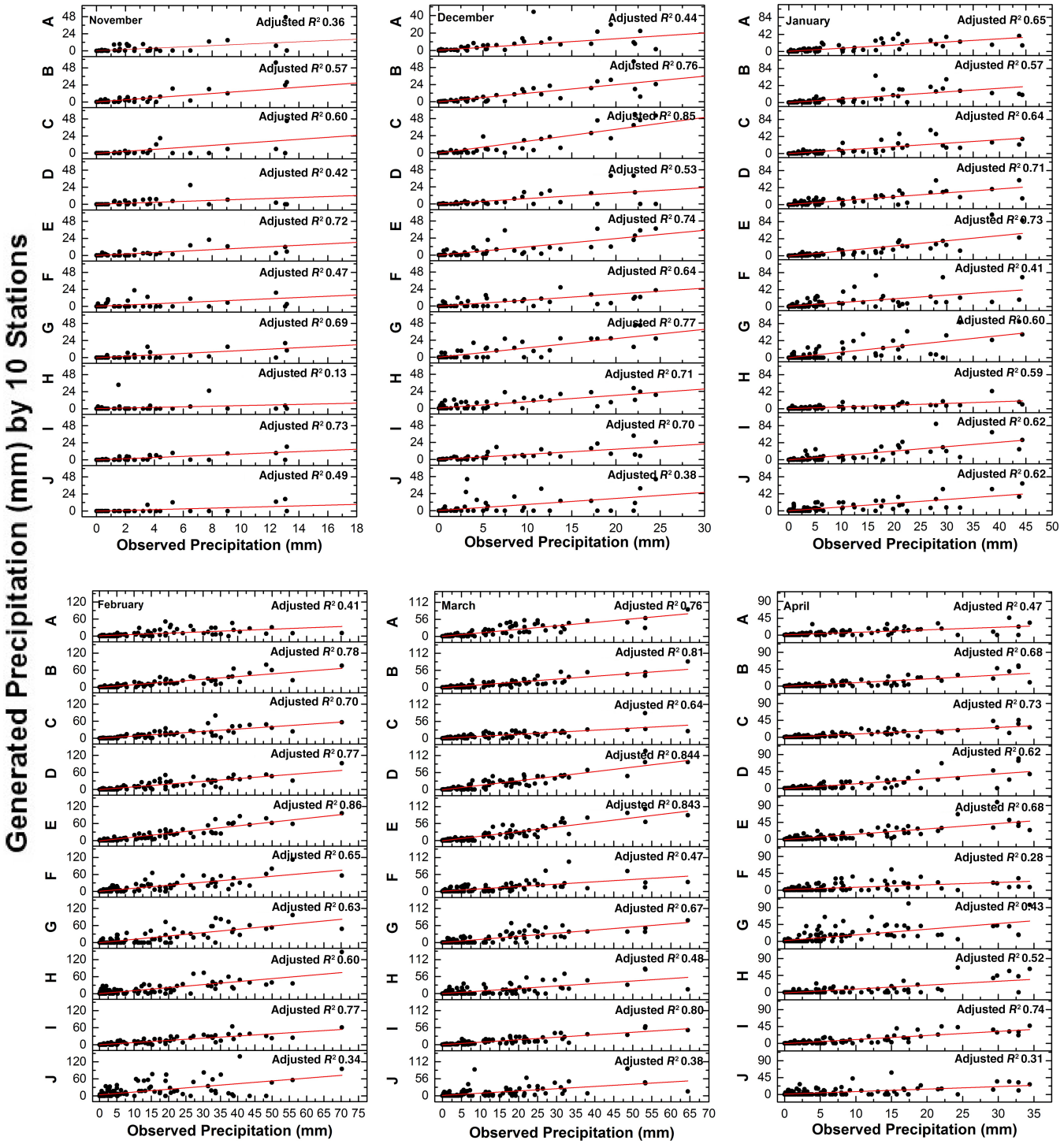

FIGURE 5 Scatter plots of generated precipitation time series (2011–2012 to 2015–2016) versus observed precipitation (mm) from November to April for different stations [Colour figure can be viewed at wileyonlinelibrary.com]

4 | CONCLUSION

There is a large amount of missing data at various stations in the Northwest Himalaya range and due to complex topography these data gaps cannot be easily filled. Of the various methods tested for generating missing precipitation data, QM was the most suitable. The QM approach was successfully applied in this study to fill data gaps in precipitation time series. The suitability of the QM technique in generating precipitation time series over NWH using 10 different stations was evaluated using standard statistical methods and displayed using Taylor diagrams. It is noteworthy that to generate the precipitation time series over the NWH, the Kanzalwan (I) station shows the best statistical measures for generating the missing data. Kanzalwan (I) displays the highest correlation with the observations, high skill score and low RMSE during all the months except January. In fact, Kanzalwan is one of the principal

observatories of our centre and provides long-term high quality data for the winter season. This may in part provide an explanation as to why Kanzalwan shows the best results. However, for January the station H-Taj (station A) has a low RMSE, high correlation and approximately same value of *SD* as the observed data. Using the QM approach, one can extend a continuous time series more than 40 years without any missing values, which would be very useful for climate studies over NWH.

ACKNOWLEDGEMENTS

The work is supported by DRDO funded project and is duly acknowledged. Authors are thankful to Director, SASE for constant encouragement and providing necessary support and guidance for this study. U.S.B. was supported by the Alaska Climate Science Adaptation Center through a Cooperative Agreement G17AC00213 from the USGS. We are also thankful to reviewers for their valuable and constructive comments which help us in improving our manuscript.

ORCID

Usha Devi https://orcid.org/0000-0002-5663-4622

REFERENCES

ASCE. (1996) *Hydrology Handbook*, 2nd edition. New York, NY: American Society of Civil Engineers (ASCE).

Bhutiyani, M.R., Kale, V.S. and Pawar, N.J. (2010) Climate change and the precipitation variations in the northwestern Himalaya: 1866–2006. *International Journal of Climatology*, 30, 535–548.

Chakravarti, I.M. and Laha, R.G. (1967) *Handbook of Methods of Applied Statistics*. New York, NY: John Wiley & Sons, p. 460.

Chen, J., Brissette, F.P., Chaumont, D. and Braun, M. (2013) Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America. *Water Resources Research*, 49, 4187–4205. https://doi.org/10.1002/wrcr.20331.

Coulibaly, P. and Evora, N.D. (2007) Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology*, 341, 27–41.

Degaetano, A.T., Eggleston, K. and Knapp, W.W. (1995) A method to estimate missing maximum and minimum temperature observations. *Journal of Applied Meteorology*, 34, 371–380.

Dimri, A.P., Joshi, P. and Ganju, A. (2008) Precipitation forecast over western Himalayas using k-nearest neighbour method. *International Journal of Climatology*, 28, 1921–1931.

Eischeid, J.K., Baker, C.B., Karl, T.R. and Diaz, H.F. (1995) The quality control of long-term climatological data using objective data analysis. *Journal of Applied Meteorology*, 34, 2787–2795.

Gudmundsson, L., Bremnes, J.B., Haugen, J.E. and Skaugen, E.T. (2012) Technical note: downscaling RCM precipitation to the station scale using quantile mapping—a comparison of methods. *Hydrology and Earth System Sciences Discussions*, 9, 6185–6201. https://doi.org/10.5194/hessd-9-6185.

Hashino, T., Bradley, A.A. and Schwartz, S.S. (2007) Evaluation of bias-correction methods for ensemble stream flow volume forecasts. *Hydrology and Earth System Sciences*, 11, 939–950.

Hevesi, J.A., Flint, A.L. and Istok, J.D. (1992a) Precipitation estimation in mountainous terrain using multivariate geostatistics Part 2. Isohyetal maps. *Journal of Applied Meteorology*, 31, 677–688.

Hevesi, J.A., Istok, J.D. and Flin, A.L. (1992b) Precipitation estimation in mountainous terrain using multivariate geostatistics Part 1. Structural analysis. *Journal of Applied Meteorology*, 31, 661–675.

IPCC. (2007) Synthesis Report. In: Pachauri, R. K. and Reisinger, A. (Eds) Climate Change 2007: Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland: IPCC, p. 104.

Kanda, N., Negi, H.S., Rishi, M.S. and Shekhar, M.S. (2017) Performance of various techniques in estimating missing climatological data over snow-bound mountainous areas of Karakoram Himalaya. *Meteorological Applications*, 25, 337–349. https://doi.org/10.1002/met.1699.

Kashani, M.H. and Dinpashoh, Y. (2012) Evaluation of efficiency of different estimation methods for missing climatological data. *Stochastic Environmental Research and Risk Assessment*, 26, 59–71.

Kemp, W.P., Burnell, D.G., Everson, D.O. and Thomson, A.J. (1983) Estimation missing daily maximum and minimum temperature. *Journal of Climate and Applied Meteorology*, 22, 1587–1593.

Kim, J.W., Chang, J.T., Baker, N.L., Wilks, D.S. and Gates, W.L. (1984) The statistical problem of climate inversion: determination of the relationship between local and large-scale climate. *Monthly Weather Review*, 112, 2069–2077.

Lafon, T., Dadson, S., Buys, G. and Prudhomme, C. (2012) Bias correction of daily precipitation simulated by a regional climate model: a comparison of methods. *International Journal of Climatology*, 33, 1367–1381. https://doi.org/10.1002/joc.3518.

Linacre, E. (1992) *Climate Data and Resources. A Reference and Guide*. London and New York: Routledge.

Lowry, W.P. (1972) *Compendium of Lecture Notes in Climatology for Class IV Meteorological Personnel, WMO-No. 327*. Geneva: Secretariat of the World Meteorological Organization.

Luu, Q.H. and Tkalich, P. (2014) Reconstruction of gappy mean sea level data. *Indian Journal of Geo-Marine Sciences*, 43(7), 1316–1321.

Murphy, A.H. and Epstein, E.S. (1989) Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, 117, 572–581.

Murphy, A.H. and Winkler, R.L. (1992) Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7, 435–455.

Rao YP, Srinivasan V. 1969. *Discussion of Typical Synoptic Weather Situation: Winter Western Disturbances and their Associated Features. Indian Meteorological Department: Forecasting Manual Part III*. New Delhi: India.

Ramos-Calzado, P., G´omez-Camacho, J., P´erez-Bernal, F. and Pita-L´opez, M. F. (2008) A novel approach to precipitation series completion in climatological datasets: application to Andalusia. *International Journal of Climatology*, 28, 1525–1534. https://doi.org/10.1002/joc.1657.

Silva, D.R.P., Dayawansa, N.D.K. and Ratnasiri, M.D. (2007) A comparison of methods used in estimating missing rainfall data. *The Journal of Agricultural Science*, 3(2), 101–108.

Simolo, C., Brunetti, M., Maugeri, M. and Nanni, T. (2010) Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *International Journal of Climatology*, 30, 1564–1576.

Singh, J, Knapp, H.V., and Demissie, M. (2004). *Hydrologic Modeling of the Iroquois River Watershed Using HSPF and SWAT. Illinois Department of Natural Resources and the Illinois State Geological Survey.* Illinois State Water Survey Contract Report 2004-08. Available at: http://www.isws.illinois.edu/pubdoc/CR/ISWSCR2004-08.pdf.

Suhaila, J., Sayang, M.D. and Jemain, A.A. (2008) Revised spatial weighting methods for estimation of missing rainfall data. *Asia-Pacific Journal of Atmospheric Sciences*, 44(2), 93–104.

Tabony, R.C. (1983) The estimation of missing climatological data. *Journal of Climatology*, 3, 297–314.

Taylor, K.E. (2001) Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, 106(7), 7183–7192.

Teegavarapu, R.S.V. (2007) Use of universal function approximation in variance-dependent interpolation technique: an application in hydrology. *Journal of Hydrology*, 332, 16–29.

Teegavarapu, R.S.V. (2009) Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. *Journal of Hydroinformatics*, 11(2), 133–146.

Teegavarapu, R.S.V. (2016) Spatial and temporal estimation and analysis of precipitation. In: Singh, V.P. (Ed.) *Handbook of Applied Hydrology*. New York, NY: McGraw Hill.

Teutschbein, C. and Seibert, J. (2012) Bias correction of regional climate model simulations for hydrological climate-change impact studies: review and evaluation of different methods. *Journal of Hydrology*, 456–457, 12–29. https://doi.org/10.1016/j.jhydrol.2012.05.052.

Tung, Y.K. (1983) Point rainfall estimation for a mountainous region. *Journal of Hydraulic Engineering ASCE*, 109(10), 1386–1393.

Wei, T.C., and McGuinness, J.L. (1973). *Reciprocal distance squared method, a computer technique for estimating area precipitation.* U.S. Agricultural Research Service, North Central Region, OH. Technical Report ARS-Nc-8.

Wood, A.W., Maurer, E.P., Kumar, A. and Lettenmaier, D.P. (2002) Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research*, 107, 4429. https://doi.org/10.1029/2001JD000659.

World Meteorological Organization. (2011) *Guide to Climatological Practices (WMO-No. 100)*. WMO: Geneva.

How to cite this article: Devi U, Shekhar MS, Singh GP, Rao NN, Bhatt US. Methodological application of quantile mapping to generate precipitation data over Northwest Himalaya. *Int J Climatol*. 2019; 1–11. https://doi.org/10.1002/joc.6008