CHAPTER 2

Study Area and Data

Chapter 2

Study Area and Data

Contents

2.1 Study area

The north-east region of India, hereafter mentioned as NER (plate 2.1), is one of the richest biodiversity zones of the world, comprising of eight states of India, namely, Assam, Arunachal Pradesh, Nagaland, Manipur, Mizoram, Tripura, Meghalaya, and Sikkim. This region shares the international boarders of India with Bhutan and Nepal to the north, Myanmar to the east and Bangladesh to the southwest. Spread between 22° to 29°25/ N (29.4°N) and 89°42/E (89.7°E) to 97°25/ E (97.4°N), the NER is a transition zone between India and southeast Asia [1], and it provides habitat to diverse biota with high level of endemism. The importance of this region lies in its uniqueness as it hosts the 2nd largest global biodiversity hotspot- Indo-Burma Hotspot [2], World's highest rainfall zone- Mawsinram and Cherrapunji (https://www.guinnessworldrecords.com/world-records/greatest-monthly-rainfall-) and considered for richness in diverse rice germplasm [3].

From the physiographical point of view, the NER can be divided into three regions, namely, the eastern Himalaya, its extensions to the east and south-east, and the Brahmaputra-Barak-Imphal valley plains. Apart from these plain areas, there lies comparatively smaller flatlands scattered in the hilly terrains of Meghalaya and Tripura.

Humid subtropical climate conditions (overall humidity>=80%) predominantly exhibits in the NER. The region faces humid, hot summers and comparatively dry and mild winters except in Sikkim and parts of Arunachal Pradesh that faces alpine climatic conditions characterized by cold, snow-covered winters and mild summer seasons. The region is influenced by south-east monsoon in summer seasons. In general, the monsoon winds generated at the Bay of Bengal shifts eastward during monsoon and brings adequate rainfall to this region, thus making the region significant in agricultural productivity [4]. Parts of NER receives the highest amount of rainfall in India due flourished to the south-west monsoon mainly in summer. Mawsinram and Cherrapunji of Meghalaya are such regions, known for holding the records of the wettest places in world.

The NER is also known for the agricultural contribution to Indian economy. It is noteworthy that India's Jute cultivation is mainly concentrated in the NER regions (https://farmer.gov.in/cropstaticsjute.aspx), and Assam is ranked the third highest raw jute (jute and mesta) producing state after West Bengal and Bihar. It contributes to a total of 7.93% of all India's share in this sector of cash crop [5]. The NER also contributes a countable amount of production of spices to India's export sector [6]. NER is regarded as one of the highest ginger producing areas in world [7]. Economic plant diversity is also very high [8].

We have selected 5 station locations in the NER for our study, depending on data availability and quality. These locations are- Cherrapunji (CHR), Dibrugarh (DBR), Guwahati (GHY), Kailashahar (KSH) and Tulihal (TUL). The details of these locations are presented in table 2.1. Sikkim is excluded entirely from our analyses as the sub-division Gengatic West-Bengal (GWB) covering this state of NER contains area-wise larger parts of West Bengal as compared to Sikkim. Thus, we have kept in view of the characteristics of NER and made sure to include at least one representative location from each of the physiographical groups as discussed earlier in this section only (Table 2.1).

Selected Location	**WMO station ID**	**State**	**Geographical location**		**Physiographical area where it belongs to**	**Annual Average Climate***			
			Lat(°N)	**Lon(°E)**		**T_{max} (°C)**	**T_{min} (°C)**	**RH (%)**	**RF_{total} (mm)**
CHR	42515	Meghalaya	25.2702	91.7323	Eastern Himalayan extension (EH)	21.1	13.8	75	11856.8
DBR	42314	Assam	27.4728	94.9120	Brahmaputra valley (BV)	28.1	18.8	80	2609.0
GHY	42410	Assam	26.1445	91.7362	Brahmaputra valley (BV)	29.5	19.8	81	1751.8
KSH	42618	Tripura	24.3268	92.0126	Eastern Himalayan extension (EH)	30.5	20.0	83	2616.2
TUL	42623	Manipur	24.6738	94.0300	Imphal valley (IV)	27.3	14.9	78	1436.7

Table 2. 1 Selected study areas of NER, WMO= World Meteorological Organisation; Annual Average Climate*= 30 years climate normal (1981-2000) as per the climatological tables of observatories in India issued by the Director General of Meteorology, New Delhi; Here T_{max}= maximum temperature, T_{min}= minimum temperature, RH(%)= Relative humidity in percentage, RF_{total}= total rainfall*

The locations of the selected stations are displayed in Plate 2.1. The average annual rainfall rates (1981–2010 climate normal) in the NER are presented as a contour map (Plate 2.1, bottom) for a better clarity on the study areas.

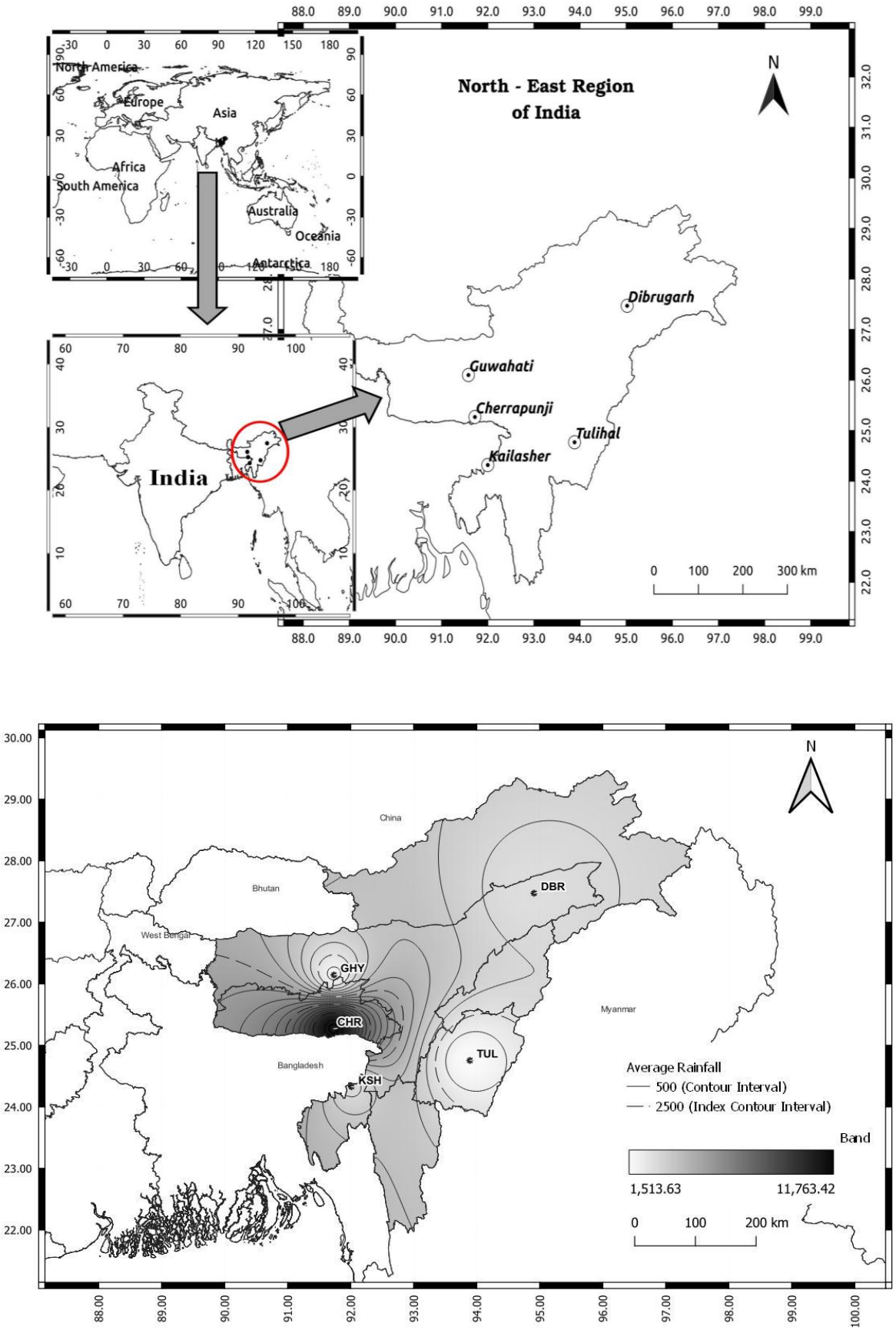

Plate 2. 1 The locations of the selected study areas (top) and contour map displaying the distribution of average annual rainfall across the study areas (bottom)

2.2 Data

Daily time series data on rainfall (mm), temperature (maximum and minimum, ^{o}C), sea level pressure (hPA) and wind speed (ms^{-1}) were used in the present study. The data used in this thesis work was acquired from India Meteorological Department (IMD), Pune, for a period of 49 years (1969-2017) as per availability and adequacy. The raw daily data for selected five locations were further on segregated into monthly and seasonal data as per the requirement in this thesis work. The regulations set by World Meteorological Organisation [9] were followed while considering the monthly sum (in case of rainfall data) or monthly mean (in case of temperature maximum and minimum).

2.2.1 Treating the missing values

Missing values are known to create a significant number of errors in data analysis. Hence, they were treated very carefully. Deletion of missing values from the data column is avoided, as a) we were dealing with time series data, and b) deletion causes loss of important information from data. Also, mean filling is not applied, as this may disturb the distribution pattern of the original data. A proper choice of imputation is even based on the type of missing data pattern and mechanisms that concerns the relationship between missingness and the values of variables in the dataset. Therefore, in order of knowing the missing data pattern in our selected data, we took help of aggregation plot, discussed in the next section. The imputation requirement in the data is dealt afterwards as per the choice of missing data mechanisms.

Aggregation plot

The aggregation plot is a useful tool to address questions associated with missing data observations, e.g., in identification of the variables containing missing data values, the missing data counts in the observations, as well as the patterns of missingness in the variable of interest. There may exist typical combinations of variables with a greater amount of missing value count, that can easily be illustrated with the help of the aggregation plots [10]. We've taken help of R's VIM package [11], that provides interface for customized plotting as a visualization tool of data structure in details. Here, the proportion of missing values by total observation can be counted for both univariate as well as multivariate data (alternatively, absolute frequencies can be shown instead of proportions, and is visualized with the help of a bar plot. The aggregation plot, based

on this missing count, plots all different combinations of missing and non-missing observations present in the observed data. Various color palettes are available that can be accessed to assign the missing and non-missing data patters. Additionally, the frequencies of various combinations are represented by a small bar plot at the extreme right-hand side of the plot panel.

The aggregation plot on the right-hand side of Figure 2.1 displays all different combinations of missing and non-missing observations (vertical axis) present in the daily data for different meteorological variables we acquired for the selected locations of NER (horizontal axis). Here, the black color indicates missingness, while the gray color represents non-missing responses in the data. The frequencies of these different combinations are presented in the smaller bar plot (to the right side of the aggregation plot). Thus, each rectangular block of gray/black color represents the combinations of non-missing/missing data with their corresponding frequencies per total observation. As for example, in case of daily rainfall we can see that a larger proportion (0.75) of non-missing values per total data is present as a common pattern in all the stations. From the aggregation plot it can also be concluded that the amount of missing data, which is also very low in number, shows missing at random pattern (MAR) across all the sites (the commonality in most of the missing data pattern is negligible, as the proportion of these values to total data is very less, ~<0.05). Likewise, the temperature series (both maximum and minimum) contains very low number of missing values (<20%) among all the selected meteorological variables. Among other meteorological time series, sea level pressure, relative humidity and wind speed contained <35% missing data. In the next section therefore, we elaborate the missing data mechanisms and choice for imputation.

Missing data: mechanisms and choice of estimation

There are three important missing data mechanisms involved in the generation of missing values: missing at random (MAR; missingness depends only on the observed data values), missing not at random (MNAR; missingness dependent on the missing data values) and missing completely at random (MCAR; missingness independent of the values of data, missing or observed) mechanism [12]. In order to assess the adequacy of the assumptions, minute review of data generation process is necessary, beyond which data elimination process is adopted. Many imputation methods, including multiple imputation, assume MCAR or MAR. However, MCAR is the most

difficult mechanism to prove as it requires that missingness to be independent of the observed or other missing values. Conversely, MAR is less strict, as it allows for

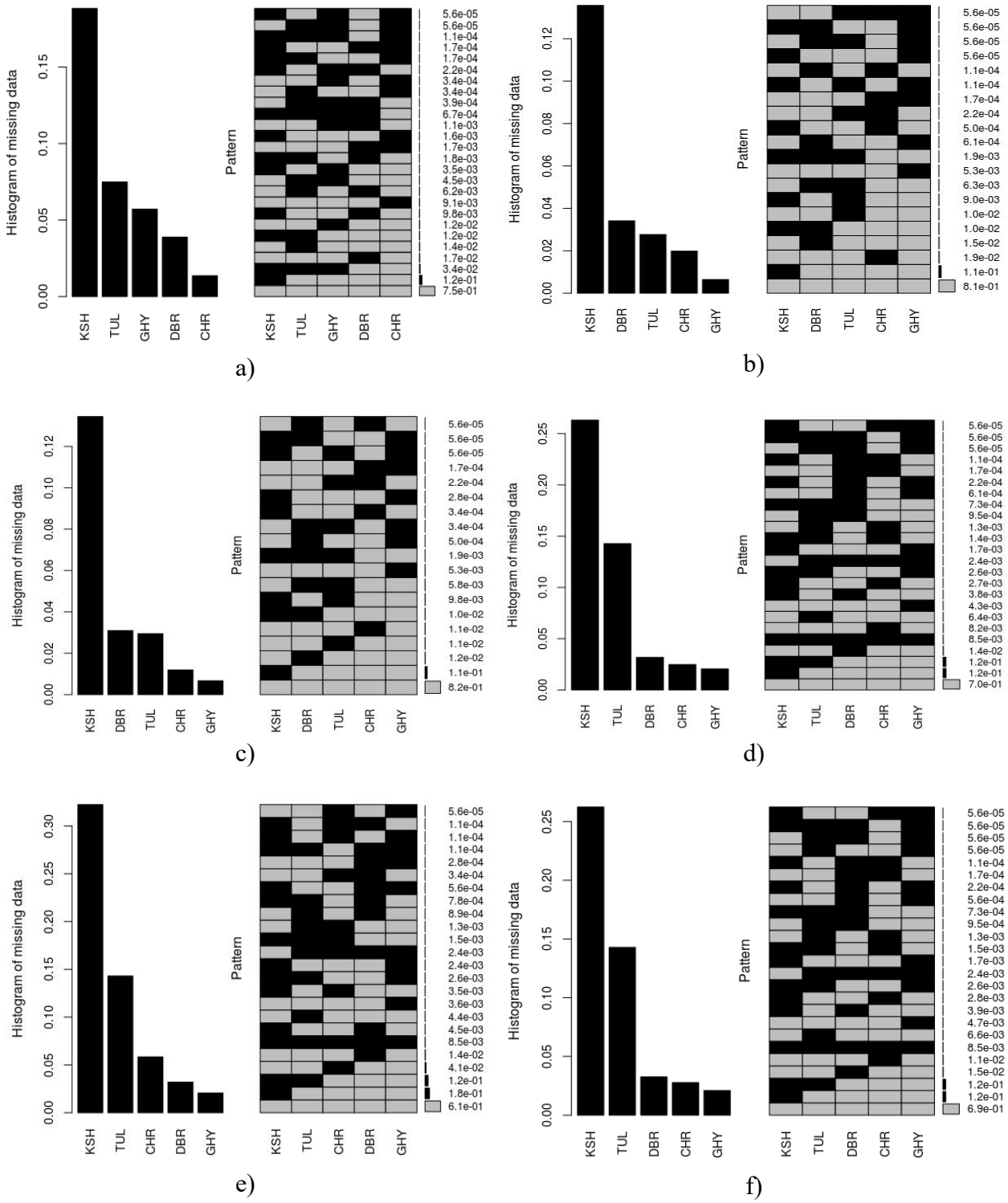

Figure 2. 1Aggregation plots of a) Rainfall, b) maximum temperature, c) minimum temperature, d) relative humidity, e) sea level pressure and f) wind speed of the selected study areas of NER.

missingness to be dependent on observed values, at the same time expecting the missingness to be independent from other missing responses. The missingness pattern is assumed to be MCAR unless there is strong evidence against it. If such is not possible, then it is believed that the pattern is MAR. Likewise, if not MAR, then MNAR has to be assumed.

In case of missing value imputation concerning rainfall data, many techniques are adopted by the researchers. It is of special mention that according to WMO guideline, 2017, a user may adopt appropriate methodologies for estimated data by carefully making proper investigation, considering available data sources, climatic characteristics, and geography of a region they consider. Hence, as an alternate method we applied multiple imputation (MI) procedure, developed by Rubin (1987) [13], for the imputation of missing values in the data. R's MICE package was used for this purpose.

In layman's term, MI is the method for imputing more than one value for each missing item, to allow appropriate assessment of uncertainties involved in imputation (in this regard, differs from single imputation). This technique not only reduces the increase in variance to negligible levels, but also provides valid standard errors that take into account of imputation uncertainty [12]. It largely corrects the disadvantage of imputing draws from the predictive distribution, i.e., the loss of precision. Multiple imputed data sets are generated based on the joint distribution and serve a wide variety of analytical purposes [14].

Quality control measures must be taken prior to data processing in case if the data is subjected to any sort of modification. In our case, the original data containing missing values were subjected to multiple imputation by chained equations (MICE). Although among the five selected stations, the station containing the highest no. of missing count (%) is very less, i.e., 0.18824384 (KSH), yet the accuracy of imputation was taken care of, as a part of maintaining the data quality. For this purpose, we took help of quantile-quantile (Q-Q) plots (Figure 2.1), elaborated in detail, in the next section. Descriptive statistics is also presented to have a view of the standard deviation so that to know the spread of variance inherent in the data.

2.2.2 Q-Q plot

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help in assessing if a set of data plausibly came from some theoretical distribution, or to compare two datasets [15]. While comparing two distributions, the quantiles of one distribution is matched with the same quantiles of the other one. Thus, a Q-Q plot is a scatterplot created by plotting two sets of quantiles against one-another. The slope and intercept estimate the scale and location. The points plotted are always in non-decreasing order while viewing from left to right. If both sets of quantiles come from the same distribution, the points form a straight line, passing through the origin at an angle of 45°. The points forming the Q-Q plot, if concentrates around a straight line with a slope not necessarily at a 45° angle, and intercept not necessarily zero, then it is assumed that the location and scale family is correctly specified up to an unspecified location and scale [16]. In this way, Q-Q plots are of great help in knowing the differences in scale and location of two distributions. Besides, it also helps in the identification of outliers.

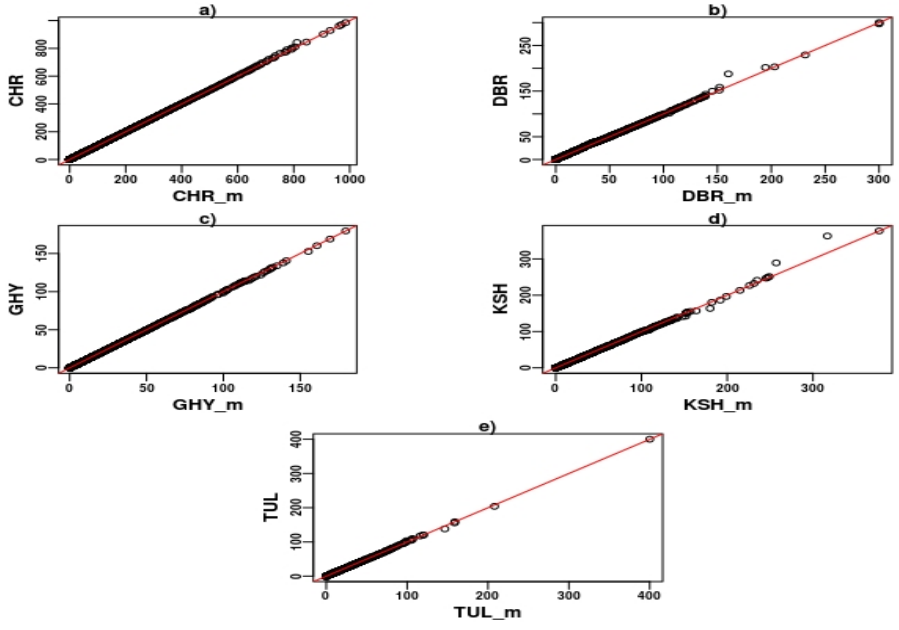

Figure 2. 2 A representative Q-Q plot for Rainfall time series (original vs. imputed)

Here, the quality of the estimated data values (or, simply the imputed data) were to be evaluated as a quality control measure for the data we had to use for our analysis. Hence, with the help of Q-Q plots, we tried to assess the quality of the imputed data by comparing the original datasets containing missing values with the new datasets

containing imputed values. In Figure 2.3 from top left to bottom, we have five Q-Q plots, plotted for rainfall data, for each of the selected stations, viz. CHR, DBR, GHY, KSH and TUL, as a representative Q-Q plot. In each of these Q-Q plots, the x-axis is represented by the quantiles of the imputed daily data for a particular station (CHR_m/ DBR_m/ GHY_m/ KSH_m/ TUL_m) and the y-axis is represented by the original daily data for that station of interest (CHR/ DBR/ GHY/ KSH/ TUL). As evident from the plots, the points follow a straight line, more or less in all the sites, passing through the origin (y=x) in all our studied stations, implying both original and imputed data to be belonging to the same distribution. The same inference could be drawn from the Q-Q plots of other meteorological variables also (see Appendix).

2.2.3 Descriptive statistics

Parameters	Sites	Min.	Q1	Median	Mean	Q3	Max.	SD
Rainfall	CHR	0.00	0.00	0.00	31.68	25.20	985.50	77.26
	DBR	0.00	0.00	0.00	7.32	7.00	300.10	15.88
	GHY	0.00	0.00	0.00	4.88	2.80	179.60	12.65
	KSH	0.00	0.00	0.00	7.49	6.20	377.20	18.00
	TUL	0.00	0.00	0.00	4.00	3.30	400.00	10.15
MaxT	CHR	9.20	19.30	21.50	21.25	23.50	31.10	3.19
	DBR	13.60	25.30	28.60	28.28	31.50	39.80	4.20
	GHY	15.00	26.80	30.20	29.65	32.60	40.60	3.85
	KSH	15.00	28.50	31.30	30.67	33.10	38.90	3.25
	TUL	12.00	24.80	28.00	27.30	30.10	36.10	3.58
MinT	CHR	-1.00	10.20	14.90	13.84	17.70	23.40	4.38
	DBR	1.00	14.00	19.70	18.64	23.90	31.50	5.82
	GHY	4.90	14.40	20.90	19.55	24.70	29.50	5.69
	KSH	4.500	15.200	21.700	19.820	24.600	29.500	5.59
	TUL	-2.700	9.200	16.600	14.920	21.000	27.500	6.60
RH	CHR	40.00	62.00	81.00	76.78	96.00	100.00	20.29
	DBR	30.00	71.00	81.00	80.30	91.00	100.00	12.16
	GHY	25.00	76.00	83.00	81.51	89.00	100.00	10.52
	KSH	33.00	78.00	83.00	82.88	89.00	100.00	8.37
	TUL	27.00	71.00	79.00	78.06	87.00	100.00	12.09
SLP	CHR	1016.00	1473.00	1502.00	1498.00	1525.00	2490.00	37.34
	DBR	989.20	1006.00	1011.40	1011.10	1016.20	1026.80	6.08
	GHY	990.20	1005.30	1010.60	1010.40	1015.60	1212.40	6.39
	KSH	988.70	1006.50	1011.40	1010.90	1015.60	1023.60	5.61
	TUL	991.20	1006.20	1011.50	1011.60	1016.90	1077.20	6.43
WS	CHR	0.00	0.00	4.00	4.87	6.00	106.00	6.70
	DBR	0.00	0.00	8.00	6.85	10.00	77.00	5.32
	GHY	0.00	0.00	0.00	3.69	6.00	80.00	4.93
	KSH	0.00	0.00	4.00	4.86	8.00	120.00	5.45
	TUL	0.00	0.00	0.00	2.11	4.00	118.00	4.02

Table 2. 2 Summary of the datasets (daily time series)

Summary of a dataset is represented by descriptive statistics that explains descriptive measures to explain the properties of that dataset. In this section we have included the

measures of central tendency (mean, median, 1st and 2nd quartile range) and the measures of dispersion (standard deviation). A higher value of range can be seen in case of rainfall across all selected sites. Table 2.2 depicts the summary of the datasets used in this study.

2.3 References

[1] Das, S.D. Ethnic and Cultural Ties between Northeast India and China: Insights from the Past. *International Research Journal of Social Sciences*, 4(1), 44-47, 2015.

[2] Myers N., Mittermeier, R. A., Mittermeier, C. G., de Fonseca, G. A. B. and Kent, J. Biodiversity hotspots for conservation priorities. *Nature*, 403:853-858, 2000.

[3] Ngachan, S. V., Mohanty, A. K and Pattanayak, A. Status Paper on Rice in North East India. Rice Knowledge Management Portal (RKMP), Directorate of Rice Research, Rajendranagar, Hyderabad 500030.

[4] Dikshit, K.R., Dikshit, J.K. Weather and climate of North-East India. In: *North-East India: Land, People and Economy*. pages 149–173, 2014.

[5] Pocket book of agricultural statistics, 2017; Govt. of India, Directorate of Economics & Statistics New Delhi, 2017.

[6] Agricultural statistics at a glance, Govt. of India, Directorate of economics and statistics, 2017.

[7] Rahman, H., Karuppaiyan, R., Kishore, K. and Denzongpa, R. Traditional practices of ginger cultivation in northeast India. *Indian Journal of Traditional Knowledge*, 8:23-28, 2009

[8] Roy, A., Kumar, D. S., Tripathi, A.K., Uttam, S.N. and Barman, H.K. Biodiversity in northeast India and their conservation. *Progressive Agriculture*, 15(2):182-189, 2015.

[9] WMO Guidelines on the Calculation of Climate Normals, World Meteorological Organization, WMO-No. 1203, 2017.

[10] Templ, M., Alfons, A. and Filzmoser, P. Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification,* 6:29–47, 2012. DOI: 10.1007/s11634-011-0102-y.

[11] Kowarik, A. and Templ, M. Imputation with the R package VIM. *Journal of Statistical Software,* 74:1–16, 2016.

[12] Little, R. and Rubin, D. *Statistical analysis with missing data*, 2nd edition., Wiley, Hoboken. ISBN:0-47118386-52002, 2002.

[13] Rubin, D. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

[14] Cheng, X., Cook, D. and Hofmann, H. Visually Exploring Missing Values in Multivariable Data Using a Graphical User Interface. *Journal of Statistical Software*, 68(6):1-22, 2015.

[15] Marden, J. I. Positions and QQ Plots. *Statistical Science*, 19(4):606-614, 2004.

[16] Das, B., and Resnick, S.I. Q-Q plots, random sets and data from a heavy tailed distribution. *Stochastic Models,* 24:103–132, 2008.