Journal of
Earth Science & Climatic Change

Research Article Open Access

# Reconstruction of a Long Reliable Daily Rainfall dataset for the Northeast India (NEI) for Extreme Rainfall Studies

**Rahul Mahanta*, Prolay Saha, P V Rajesh, Sudipta Nandy, Yasmin Zahan, Anupam Mahanta**

*Interdisciplinary Climate Research Centre, Department of Physics, Cotton University, Guwahati-781001, Assam, India*

## Abstract

The North East India (NEI) is an IUCN (International Union for Conservation of Nature) biodiversity hot spot. A region known for its highest annual rainfall in the world together with the unique topography and mighty Brahmaputra, makes the region vulnerable to climate change induced hydrological disasters and biodiversity loss. For building resilience to extreme rainfall events, food security and biodiversity management, dependable and consistent estimates of trend and modes of variability based on over 100 years of daily rainfall are critical. However, the region is poorly sampled by continuous rain gauges and in a region of large spatial variability of the mean rainfall, approximately 10 stations with such data are highly inadequate for estimating extreme event statistics. We were successful in developing a quality controlled daily rainfall data collection on a set of 24 well-distributed fixed stations around the region in order to improve this condition. This technical note describes combining conventional weather station records with rain-gauge records from a number of sources like privately owned tea estates to create a continuous daily rainfall record from 1 January 1920 to 31st December 2009 for the north-eastern region of India. Remaining data gaps are less than 3% of the total data in each station. With the goal of improving estimates of long-term changes in climate variability over NEI, every attempt has been made to reconstruct data gaps. The NEI final rebuilt data set is ideally adapted to estimating both long-term trend and multi-decadal variability of rainfall over the region.

## Introduction

Long-term reliable rainfall measurement is required for a variety of applications such as past or future climate assessments, water resource management, and infrastructure development, among others. In order to understand climate change, such digitised long-term, high-quality climatological daily rainfall data are critical [1]. They're also useful for verifying climate models and assessing trends and extreme rainfall occurrences over longer periods of time. In addition, the dataset can help with better climate risk management and future climate projections. This technical note presents the first extensive rainfall reconstruction in the North East India (here after NEI). To construct daily rainfall data, we employed an assortment of both published and unpublished sources pertaining to the region, spanning 1920-2009.In the backdrop of a warming global environment, this data set is also important to disentangle contributions from anthropogenic and natural drivers to the trend and periodicities of the normal rainfall together with intense rainfall events over the region.

Unlike other parts of India that enjoy good data coverage, NEI is a data-sparse region. Although some India Meteorological Department (IMD) daily rainfall databases exist, the density of rainfall data both in temporal and spatial scales is rather poor in NEI. There are several reasons for the absence of a dense, long-term and reliable daily rainfall data from IMD over NEI. The most crucial reason for this data crisis is the remoteness of the region with low population density. Further, the region has passed through extended periods of regional conflicts and environmental disasters that made regular observations difficult. Also, in some stations it was observed that the location of the observatories is changed resulting in fragmented or conflicting data series within the same locale. In the hand written records, we observed human error occurring during the process of observation, and in the recording. Changes in the surrounding environment also contribute to the errors in the existing dataset.

As such, availability of an adequate length and high-quality instrumental records of rainfall data for NEI continues to hinder our ability to carry out a robust evaluation of rainfall variability. To better understand, identify, anticipate, and respond to changes in rainfall variability related to climate change, such rigorous assessments are required. Although India can boast of a rich legacy of about 150 years of rainfall data, the region of NEI remains poorly sampled spatially as well as temporally. Additionally, some of the available rainfall records (with their space-time resolution) do not conform to the standards required for supporting climate assessment with high degree of confidence.

In NEI floods, erosion and landslides represent the main natural hazards with human and economic effects and all three hazards are meticulously associated with the intensity, frequency and period of rainfall events. For developing superior projections of future rainfall variability, a better understanding of the dynamical behavior of NEI's past rainfall variability is crucial. Hence a reliable estimate of the existence of any trend together with multi-decadal variability of regional mean rainfall and that of the extreme events are critical for delineating impact of climate change over the region. Although the Climatological mean rainfall in NEI has large spatial variability, number of rain gauge stations with continuous data is very small [2,3]. Also, year-to-year variation of stations contributing to the mean introduces large artificial variability. To address these challenges and provide a reliable database for extreme value analysis, we need a thorough understanding of historical rainfall fluctuations with high temporal precision and over long time scales throughout the whole NEI to evaluate climate model simulations.

***Corresponding author:** Mahanta R, Interdisciplinary Climate Research Centre, Department of Physics, Cotton University, Guwahati-781001, Assam, India; E-mail: rahulmahanta@gmail.com

**Received** August 20, 2021; **Accepted** September 06, 2021; **Published** September 15, 2021

**Citation:** Mahanta R, Saha P, Rajesh PV, Nandy S, Zahan Y, et al. (2021) Reconstruction of a Long Reliable Daily Rainfall dataset for the Northeast India (NEI) for Extreme Rainfall Studies. J Earth Sci Clim Change 12: 580.

Such an understanding necessitates long-term, high-quality daily rainfall data; whose present readiness and accessibility is irregular in both space and time in the NEI. Therefore, the recovery of instrumental records of daily rainfall from printed manuscripts and other sources is of great importance for climatological studies. Further the data recovery also allows us to examine the leading modes of rainfall variability in NEI. The basic objective of our work is to make historical rainfall data in NEI from the different sources digitally accessible, with the highest possible time resolution and quality.

Since the formation of India Meteorology Department (IMD), there has been significant interest in the rainfall over NEI, particularly by the Tea Gardens. Since 2nd half of 1800 over 100 rain-gauge sites have operated in NEI. But sadly, the length of rainfall record is short for most observation sites. The data records are frequently incomplete and, in some situations, erroneous. As such, existing studies on rainfall distribution over NEI have been grounded on incomplete data (both in terms of station density and period of records). As a result, the rainfall distribution patterns over NEI have been grossly oversimplified or investigated incorrectly, oblivious to the actual rainfall gradients.

This research aims to assess the available data and then use the acceptable data to create a 90-year rainfall database for chosen NEI locations. The generated database includes long-term rainfall climatology as well as a regional annual rainfall distribution map. This technical note gives a short description of the process of cataloguing, recording, reconstructing and digitizing rainfall records from original data recorded in paper from different sources to develop a daily rainfall dataset for NEI. This dataset is expected to generate a new baseline for NEI that will allow researchers and policymakers to analyze and address impacts and risks due to intense/extreme rainfall, in ways and over time periods that was not possible earlier. In addition, using the reconstructed database we present a rainfall characterization of the region, which is going to be complemented in other future statistical studies.

## Rainfall Data

We define a wet day as one on which a quantifiable amount of rain equal to 0.3 mm or greater has been recorded at any station, as described by Soman and Krishna Kumar [4]. India Meteorology Department (IMD) rain gauge stations in India reckons daily rainfall as the amount of rainfall collected from 0300Z (0830LST) of the previous day to 0300Z of the concerned day, where the rainfall is measured correct to 0.1 mm. To indicate 24 hour rainfall total the term 'rain intensity' is used in this study.

It's worth noting that the rainfall data utilised in this study are 24-hour accumulations for the time ending at 0830 LST on each day, not 24-hour totals like those acquired from a continuously recording rain gauge. This means that if 100 mm of rain fell between 1500 LST and 0830 LST on day 2 and 200 mm fell between 0830 LST and mid-day on day 2, the observational dataset would record daily rainfall data of 100 and 200 mm for two days instead of a 24-hour figure of 300 mm. As a result, the measured rainfall maxima for some places may be underestimated.

## Current Status

Even after 150 years of rainfall observation in India, the network of meteorology observation stations remains sparse. We found that relatively uninterrupted daily data in the region is available on about 15 stations for about 32 years after 1975, using which we carried out studies of extreme events and heavy rainfall events over NEI [3]. Prior to 1975, the number of stations with data in a given year decreases. Particularly, number of common stations with continuous data is even smaller, not representative of the region. Rajeevan [5] developed a database for daily rainfall in NEI, but the spatial density is too low (07 observatories in NEI) to enable adequate spatial analysis. Thus, due effort to create a reliable and long (~100 years) daily data set in the NEI has been lacking so far. This new dataset produced comprises daily rainfall time series over 24 well distributed stations that have operated in NEI for 90-years which will be extended to 100 years shortly. (See Figure 1a for locations of the stations). From 1975 to 2009 it involved collecting and creating the time series in the 9 additional stations than the 15 reported in Goswami et al. [3]. Prior to 1975 going back to 1920, it involved filling up the extended gaps in the 15 stations together with collecting and creating the time series for the 9 additional stations going back to1920. While there are a few data gaps in the rebuilt time series, every attempt has been taken to replace them in so that estimates of long-term changes in rainfall variability in the region can be improved. This work was undertaken in 2006 under the supervision of Prof. B.N. Goswami, CAOS, IISc, Banglore. Starting in 2006 the data collection work continued till 2017. First we identified the concerned establishments/authorities collecting/archiving meteorological data, followed by the extraction and archiving of the required data. A consistent standard of the data achieved after quality controlling of the data. Daily rainfall records spanning over 4000 years (not a 4000-year time-series) have been collected and used to generate daily rainfall time-series dating back to the 1920s.

## Background on data availability

The longest gridded daily rainfall (for over 100 years) constructed over the whole of India [6], while was based on approximately 2000 stations reasonably well distributed over rest of India, contained only 10 stations over the NEI. Due to the orography and high spatial variability of the mean rainfall in the region (Figure 1b), such a small number of stations are inadequate for constructing the mean over the season and missing data even in one of these stations could introduce artificial variability of the mean. For this reason, long records of monthly and seasonal means (starting from 1871) constructed by Parthasarathy et al. [7] based on 306 fixed stations well distributed over the country, that included 12 stations over the NEI, making this data set most reliable for studying low-frequency variability of Indian monsoon rainfall. It is rather interesting to note that they did not include Cherrapunji in their list of stations to construct the mean. The seasonal mean rainfall at Cherrapunji being over 30% higher than the mean over the rest of the stations, missing data in Cherrapunji even for one or two years could give rise to serious spurious variability. However, the 12 stations underestimate the seasonal mean and more importantly are grossly inadequate for estimating the number of extreme rainfall events during the rainy season over the region.

In India, historical instrumental observations of past rainfall data are in the reference libraries and archives of India Meteorology Department(IMD), Central Water Commission(CWC), state government administrations, military authorities, oil – natural gas exploration companies, JESUIT library, national data center and private tea gardens. After the establishment of British rule in this part of South Asia, following an order in 1839 by Colonial Secretary of the British Empire Lord Glenelg, all administrators and public functionaries across South Asia had to preserve chronicles of the "state of the weather" [8]. After this in NEI also regular meteorological records were kept that was to be included in the district Gazetteers annual reports which regional administrators sent to the British administrative headquarters in India. As such, in this part of India also regular and meticulous instrumental rainfall data recording started in latter half of the 19th century, and

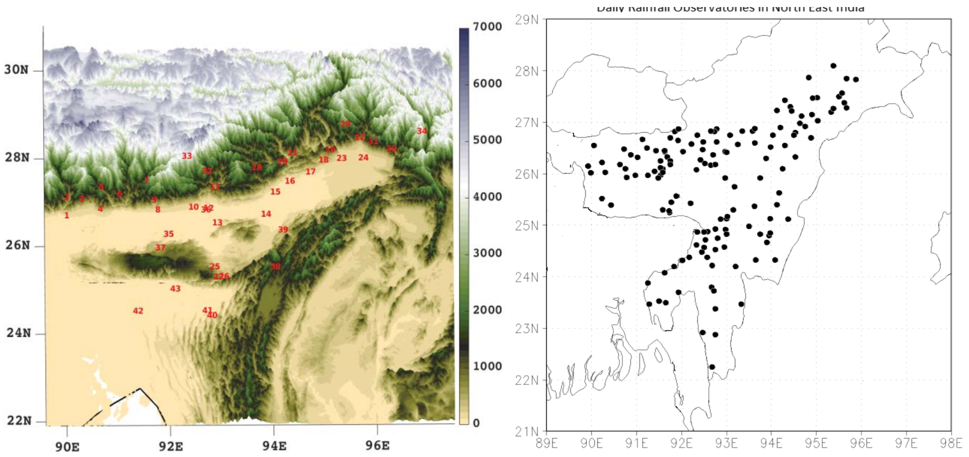


**Figure 1:** (a) Topographic map of NEI with stations used as baseline (b) Total number of Rain gauge stations used for the reconstruction rainfall data over NEI after first screening.

was made possible by the arrival of British with their technologically advance instruments. Thus, Meteorological observations have been performed at a number of locations in NEI in the past, and numerous institutions, particularly tea gardens, retain archives with handwritten rainfall records dating back to the late 1800's.

Apart from IMD and tea gardens, the record of rainfall and temperature, exists in many documentary sources that include private diaries, periodicals and journals and log books of medical and scientific research found in missionary hospitals in the region, JESUIT libraries, newspapers and personal diaries. It must be mentioned that Doctors working for colonial administrations in order to investigate the links between climate and disease, recorded precipitation and temperature in tropical and monsoonal regions [9]. Since its founding in 1540, the Society of Jesus (the "Jesuits") has maintained an interest in natural sciences [10]. For meteorological and climatological research the observations and the results of Jesuits scientific traditions have been of great interest [11].

Over many parts of NEI in the period after 1950 (Figure 2), observational data has not been preserved regularly over time in some of the stations and in some other stations the preservation and archiving of documents has not been prioritised by the relevant government departments. This problem is particularly heightened in NEI where any systematic station-by-station method of preserving meteorological data is lacking and where extended periods of social conflicts have threatened the survival of historic weather records. Data management has only been practised for a few decades. As a result, many observational rainfall data in various sections of the region are fragmented, scattered, or sometimes non-existent. While IMD statistics were used as the starting point, there were numerous data gaps. Many of the IMD cooperative rain gauges and other rain gauges failed to fulfil the IMD data deadlines, and hence the late data was not captured by the IMD databases. Much of the Tea Garden data, for example, did not even make it into the IMD reports. These had to be retrieved from original sources, which were frequently old (but unpublished) official records. The authors extracted these rainfall data from available "tea garden" records. Most of the IMD data after 1980 was available in Regional Meteorological Centre (RMC), IMD, Guwahati.

## Data sources

To fill up the gaps, only data from IMD and cooperating IMD approved Government and Private agencies were used for reasons of equipment standardisation and calibration, as well as data reliability. The dearth of data in the NEI is not due to a lack of measurements, as meteorological observations have been recorded by former colonial endeavours from the mid-nineteenth century. Even though hills, rivers and jungle dominate NEI topography, colonial administrators deployed individual station net-works within most of the populated parts of NEI and the observed weather data were published in periodical reports. Numerous sources of documentation (documentary sources, observatories, reports, etc.,) had to be used in order to recover these data. These sources included: Indian Meteorological Department; Tocklai Experimental Research Station; Assam Agricultural University; Assam state Electricity Board; Meghalaya State Electricity Board; Various Hydro Electrical Power projects of NEI; State Agriculture Departments of seven states in NEI, Tea gardens that are member of Assam Branch-Indian Tea Association, Assam Tea Planters Association and Bharotiya Cha Parisad and The Society of Jesus (the "Jesuits"). Because tea and lumber production were the main drivers of the colonial economy, and most cultivable land was allocated to them, tea plantation records make up a significant part of NEI's documented heritage. The state of Assam has more than 750 major Tea Estates. Most of these Tea Estates have been functional for more than 100 years now and many of these Tea Estates have been recording and documenting Rainfall & Temperature Data since the time they began operations. Hand written records were collected from the tea Estates, collated and digitized for the years 1920-2009).

The drop in number stations correspond to years with natural calamity, war, social movements, e.g., 1950 major earthquake in NEI, 1962 China War along Tibetan Border, 1970s Language movement and Bangladesh independence war, 1980s Assam foreigner's Movement followed by other ethnic tensions. The initial database included 326 daily rainfall observatories in the research area, with activity dating back to the IMD's inception (1875-till date). The rainfall data from the IMD database, on the other hand, was extremely fragmented and unreliable, resembling many observatories in the same location but covering different time periods.

Consequently the original IMD rainfall databases were highly variable both in terms of record lengths of data and data gap durations for different stations. The data collection periods for the majority of observatories were very short (125 had between 1-5 years of complete records, another 50 observatories had rainfall records from 6-30 years, 89 observing stations had 30-60 years, 30 stations had 60-75 years and only 23 stations had more than 75years of data). Some of the chosen

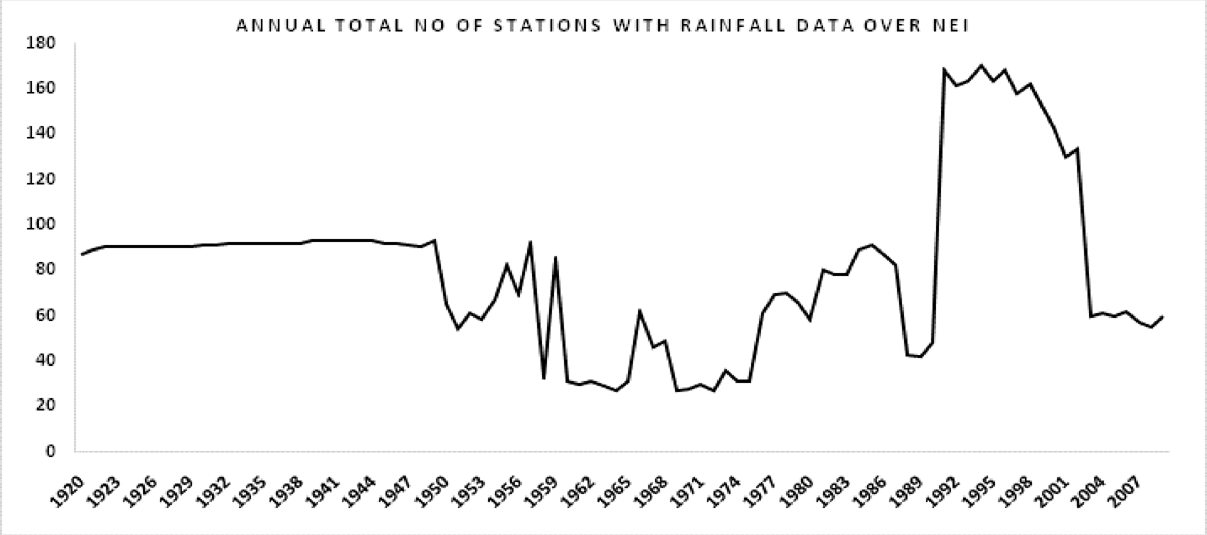


**Figure 2:** Available number of stations in NEI with daily rainfall available for each year from 1920-2009.

rainfall stations with data records spanning shorter periods were also included, done mainly to capture data in cases where rainfall stations were replaced by new stations whose location is little different (but within the period under consideration). In the end, we had 24 stations spread across Northeast India, with an average height of 345.25 metres and elevations ranging from 16 metres to more than 2000 metres above mean sea level.

## Methodology

The method comprised of two phases:

**(a) Reconstruction missing data**

The first phase involved reconstructing rainfall time series in order to create a continuous long-term series by merging short-duration rainfall series from adjacent observatories. Auxiliary information obtained from observatories in close proximity was also considered while filling data gaps.

**(b) Quality Control**

The second phase entailed a quality control examination of the reconstructed rainfall series, with the goal of identifying and replacing any records in the database that were inconsistent or dubious (e.g., negative rainfall, extreme rainfall events, some zero values, and records that differed markedly from values recorded in neighbouring observatories).

To allow for future post-processing, a special effort was made to document all of the available metadata. As a result, we accumulated data acquisition information and the history of each observatory (metadata) during this process, including changes in location, measurement circumstances, observers, and observation periods. This metadata was extremely helpful in determining the quality of a series and identifying potential problems, but it was commonly lacking in raw climate databases obtained from IMD. The regional variability of numerous measures defining daily rainfall features (mainly mean, percentile, and standard deviation) was compared. The final database had better spatial coherence, according to the findings. Figure 1a depicts the location of the IMD rainfall stations over NEIN. The final dataset includes rainfall stations that are spatially well scattered all through the NEI, with the exception being the high altitude region in north bordering Arunachal Pradesh and south Tibet, without any official observatory for rainfall data. There is a higher concentration of rainfall stations in and around major urban areas of NEI.

## Station selection criteria

The daily rainfall data from 24 rain gauge stations scattered pretty well to represent the spatial variability of rainfall over the NEI for the period 1920 to 2009 is the backbone of our investigation, as described in the preceding section. It may be noted that we also do not include the Cherrapunji in our list as we found that the data at the station around 1950 are missing for years at a time that could introduce serious artificial variability. Unfortunately, there is no alternate station nearby (say, within 30 km) from which one could recover any missing data. It may be noted that the high rainfall region in the central easternmost part is in Myanmar and no station data with long-term record was available to us from this region. There is potential for adding one station over the high rainfall region in the westernmost sector. However, we could not recover data for the entire period for technical reasons, but hope to add such a station in future. The Meteorological observatories all over India were established by the India Meteorology Department starting from its inception in 1875 and upgraded periodically. After independence (1947) many State governments augmented the number of rain gauge stations significantly through State Agriculture Departments and Disaster Management Authorities. Recently, over NEI there has been renewed attention and initiatives from a number of different agencies, like, Indian Space Research Organisation (ISRO), Central water Commission (CWC), Agriculture departments and most notably by few researchers from Japan [12]. However, such high-quality high-frequency rainfall observations with increasing station density are only available for two decades at most. We chose the 24 IMD sites since our goal is to assess long-term trend and multi-decadal variability. We focused on creating substantially unbroken daily rainfall time series at all 24 locations from 1920 to 2009.

In order to select appropriate daily rainfall series for reconstruction, we selected some criteria, which are function of the data gaps, temporal duration and the period covered. The rainfall series were separated into three clusters.

- Cluster A: Consist of rainfall series (23) with more than 75 years daily data with less than 10 years of missing data.
- Cluster B: Consist of rainfall series (39) with data available for 50 to 75 years, or data available for more than 75 years but missing data exceeding 10 years.
- Cluster C: Data series with a time span of less than 20 years were deemed too short to be useful for reconstruction. This collection of rainfall series is extremely useful since it may supplement long-term data from adjacent observatories that had stopped collecting data for a time. 89 rainfall series complemented this criterion. The residual stations had data for 1-5 years. This set of observatories (labelled C in the dataset) was put aside for reassembling data from other observatories.
- The stations from cluster A and B were used in the reconstruction process.

## Filling data gaps in each series

We don't want to extrapolate from stations further than 20-30 km (approximately the scale of intense rainfall occurrences) from the missing station because of the substantial particular variability of mean rainfall. If there were, no real observations in the neighborhood of any station, it would be impossible to fill the missing data. One of the primary objectives of this reconstruction process is to select or improve an existing procedure to fill the gaps in daily rainfall data over NEI. Many researchers have noticed that for rainfall reconstruction, it is more consistent to utilise algorithms based on data reported at adjacent observatories [13,14].

Three to five years of rainfall data were randomly destroyed in order to develop ways for reconstructing daily rainfall data for the target stations. These years were regarded as times in which rainfall data was unavailable. The missing (deleted) daily rainfall data for each year were then calculated for each station using daily rainfall data available at three to five surrounding stations. This was done for each of the three techniques. Following that, three error statistics were used to compare the estimated data to the real observations.

To fill data gaps in rainfall series of target stations, following Hasana, et al. [15], the following assumptions were made:

### (i) Similar Statistical properties

It is assumed that missing rainfall amounts to have comparable statistical properties to the available rainfall data for any station. Existence of large volume of rainfall data made it possible to deduct statistically the parameters from the existing times to the station with missing data. Table 1 shows the various statistics of 1st and 2nd halves of rainfall series for 24 stations, supporting our assumption that the statistical features of missing rainfall measurements are identical to those of data from available periods.

### (ii) Station Correlation

Between neighbouring stations there exists a spatial correlation for both rainfall incidence and quantity if the distance between the stations is not large. This assumption is highlighted by Figure 3 that point toward a rather negative association between rainfall amounts and distances between stations.

For each target station with missing rainfall data, clusters of two to five rain gauging stations were selected. Based on the observations of nearby stations, the daily rainfall data of target stations were estimated using the chosen techniques. For each target station only the nearest and well spread out observatories were selected. For each cluster (target

| Station | Abbreviation | Mean Rainfall per day (mm) | | Standard Dev(mm) | | 95th percentile(mm) | | 99th percentile(mm) | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1920-1965 | 1966-2009 | 1920-1965 | 1966-2009 | 1920-1965 | 1966-2009 | 1920-1965 | 1966-2009 |
| **Agartala** | AGT | 9.5 | 9.4 | 19.6 | 20.1 | 51.8 | 57.3 | 110.5 | 108.5 |
| **Aijal** | AZL | 5.9 | 5.7 | 12.9 | 12.7 | 26.6 | 25.1 | 53.6 | 59.3 |
| **Chaparmuh** | CMK | 10.1 | 12.2 | 16.8 | 20.3 | 28.3 | 29.2 | 61 | 56.7 |
| **Dhubri** | DHB | 7.1 | 9.8 | 19.9 | 20.3 | 33 | 35.6 | 69.3 | 76.1 |
| **Dibrugarh** | DBR | 7.6 | 6.9 | 16.1 | 17.6 | 55.9 | 52.7 | 163.8 | 148.2 |
| **Digboi** | DIG | 6.5 | 7.8 | 13.7 | 14.1 | 20.8 | 22.1 | 50 | 54.2 |
| **Guwahati** | GHT | 4.5 | 4.7 | 11.8 | 12.8 | 33.3 | 31.7 | 80.0 | 84.7 |
| **Goalpara** | GLP | 6.7 | 6.1 | 18.0 | 19.8 | 33.77 | 34.9 | 63.3 | 58.8 |
| **Gohpur** | GHP | 8.6 | 9.3 | 23.0 | 22.3 | 53.3 | 50.6 | 97.8 | 101.5 |
| **Golaghat** | GLT | 5.1 | 4.9 | 11.9 | 12.4 | 39.4 | 39 | 88.4 | 92.6 |
| **Halflong** | HFL | 6.3 | 7.2 | 16.3 | 18.2 | 42.7 | 40.8 | 105.7 | 98.3 |
| **Imphal** | IMP | 3.8 | 4.8 | 9.2 | 8.5 | 31.2 | 34.7 | 67.1 | 63.2 |
| **Jorhat** | JRT | 6.2 | 5.7 | 13.9 | 12.2 | 45.32 | 41.8 | 119.4 | 124.3 |
| **Kailasahar** | KSH | 11.5 | 11.7 | 20.8 | 21.9 | 68.6 | 57 | 137.0 | 144.7 |
| **Kohima** | KOH | 5.3 | 4.1 | 11.2 | 12.9 | 74.81 | 68.4 | 166.1 | 158.0 |
| **Lumding** | LMD | 3.5 | 3.9 | 10.1 | 9.3 | 29.2 | 32.6 | 59.2 | 64.8 |
| **Mazbat** | MZB | 5.4 | 6.6 | 13.4 | 11.8 | 42.2 | 40.8 | 100.6 | 108.3 |
| **Lakhimpur** | NLP | 9.6 | 7.3 | 21.5 | 20.6 | 26.9 | 29.5 | 57.2 | 61.6 |
| **Pasighat** | PGT | 12.3 | 11.2 | 28.8 | 29.3 | 40.6 | 45.2 | 74.4 | 64.8 |
| **Shillong** | SHL | 6.8 | 6.2 | 18.0 | 19.8 | 21.6 | 26.7 | 45.2 | 48.7 |
| **Silchar** | SLC | 14.1 | 12.8 | 30.9 | 34.3 | 33.5 | 30.2 | 80.3 | 87.5 |
| **Tangla** | TNG | 7.4 | 7.7 | 23.3 | 24.9 | 32.3 | 31.5 | 60.5 | 53.4 |
| **Tezpur** | TZP | 5.1 | 5.8 | 12.2 | 14.7 | 36.8 | 33.4 | 71.4 | 68.8 |
| **Tura** | TUR | 9.1 | 9.8 | 23.0 | 24.6 | 68.5 | 64.3 | 118.4 | 124.6 |

**Table 1:** Various statistics of first and second halves of the data series.

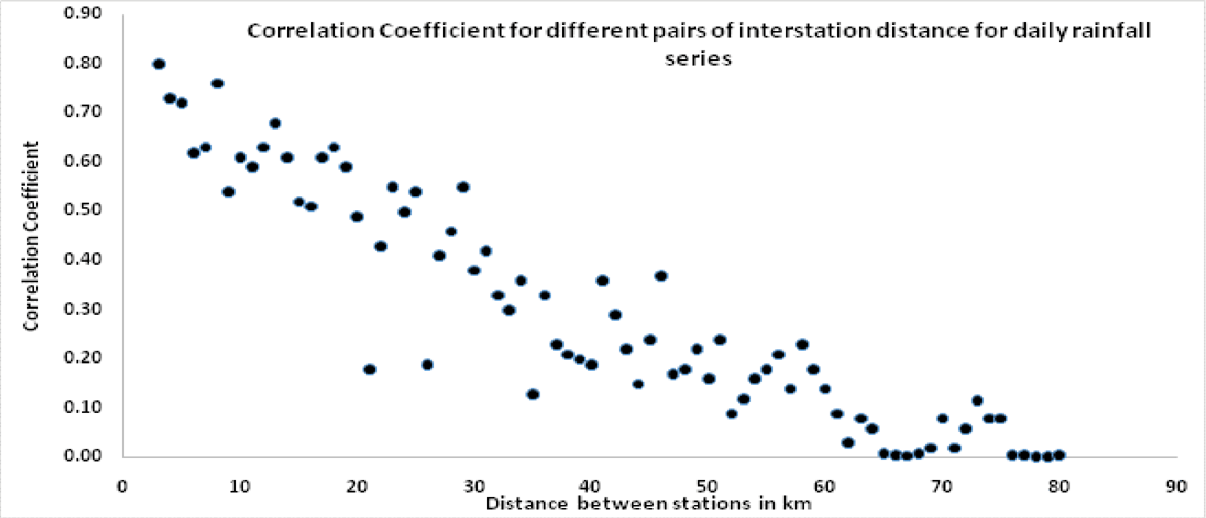


**Figure 3:** Scatter plot showing the relation between distance between rainfall stations and their correlation. The values are average for each distance gap among all stations.

and nearest neighbors) a time period was considered when all the stations in the cluster had 100% data. Then the correlation coefficient (CC) for rainfall series of each target station with the selected nearest neighbors were calculated. Based on correlation coefficient value for rainfall estimation the neighboring stations were ranked and only stations with CC greater than 0.5 were considered for reconstructing an estimation of missing rainfall data.

## Estimation of missing rainfall data

In this study for filling rainfall gaps attention was given to methods involving the data from nearby stations. For our purpose three stand methods were tested, namely, the nearest neighbor, inverse distance weighted interpolation, and normal ratio method. The three procedure used in the estimation of missing data are all deterministic methods and are given below.

## Arithmetic Mean (AM) Method

In this method missing rainfall data is derived by arithmetically averaging data from the nearest weather stations to the station of interest, provided the normal annual rainfall at the neighboring stations are inside the range of 10% of the normal annual rainfall at the target station [16,17]. AM method presume equal weights from all nearby rain gauge stations and estimate the missing amount by taking the mean rainfall, using equation (1), from the neighboring stations [18].

$$V_{est} = \frac{\sum_{i=1}^{n} v_i}{n} \quad (1)$$

In eqn 1, $V_{est}$ is the assessed or estimated amount of the missing rainfall data, $v_i$ represents the rainfall amount at the ith nearest weather station and n is the number of the nearest stations for our target station.

## Normal Ratio (NR) Method

If any nearby gauges have normal annual precipitation exceeding 10% of the target rain gauge, the Normal Ratio approach, first proposed by Paulhus and Kohler in 1952 and later updated by Young in 1992, is utilised.

For calculating the missing rainfall amount, NRM considers stations with significant weight located around the target station. The equation is:

$$P_o = \frac{\sum_{i=1}^{n} (P_i * W_i)}{\sum_{i=1}^{n} W_i} \quad (2)$$

Where Wi is the weight of the ith nearest observatory and can be estimated as:

$$P_o = \frac{\sum_{i=1}^{n} P_i * (\frac{1}{d_i^2})}{\sum_{i=1}^{n} (\frac{1}{d_i^2})} \quad (3)$$

Where $d_i$ is the distance between the target station (with missing data) and the ith nearest observatory

### Inverse Distance Weighting (IDW) Method

Using the observed values at other stations, the inverse distance weighting approach guesses the missing value of an observation, Vest [17,18].

$$V_{est} \frac{\sum_{i=1}^{n} V_i d_i^{-k}}{\sum_{i=1}^{n} d_i^{-k}} \quad (4)$$

Where $V_{est}$ is the estimated rainfall amount; n gives the number of stations; $V_i$ the rainfall at station i, d the distance between target station and the ith station, and k the power of distance, referred to as a friction distance ranging between 0.5 and 2. Weights for each sample are inversely proportional to their distance from the calculated point in this method [19].

## Performance criteria

Artificial data gaps were produced by arbitrarily eliminating 5 years of accessible observations from the A and B observatories to compare the three strategies. The nearest neighbour, inverse distance weighted interpolation, and normal ratio methods were used to examine a total of 140 rainfall series with 28760 artificial daily data gaps. After applying above three methods to these data, the most appropriate method was chosen by determining the root mean square error (RMSE), Maximum Absolute Error (MAE) and correlation coefficient (CC) of the reconstructed gaps. The MAE and RMSE are indicators of the extent of severe mistakes, while the CC shows the interpolation's relative correctness [20].

**Root mean square error (RMSE)**

In meteorology, air quality, and climate research investigations, the root mean square error (RMSE) has been employed as a standard statistical tool to analyse model performance. In meteorology and other geosciences research, a common statistical metric called root mean square error (RMSE) is frequently used to assess the effectiveness of model findings [21-24]. It denotes the disparity between what a model predicts and what is actually observed in the field. The approach with the lowest RMSE value is designated as the best method for the research field. The RMSE has a range of 0 to +∞. The RMSE is defined as follows:

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n}(P_{obs} - P_{est})^2}}{\sqrt{n}} \tag{5}$$

## Mean absolute errors (MAE)

The average absolute error (MAE) is a commonly used metric in model evaluations. It was recommended by Willmott and Robeson. It calculates the amount of estimation error. The approach with the lowest MAE value is considered to be the best for the study region. MAE is a scale that ranges from 0 to +∞ [25]. The MAE is defined as follows.

$$MAE = \frac{\sum_{i=1}^{n}|P_{obs} - P_{est}|}{n} \tag{6}$$

## Correlation Coefficient

Correlation Coefficient ($r_{pearson}$) gives an indication of the strength of the association among observed and estimated rainfall series, i.e. provides evidence about the appropriateness of the estimating procedure. Thus this coefficient indicates whether the estimates will be high or low when the observed or actual series is high or low.

$$r_{pearson} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{7}$$

## Quality control of rainfall data

For reconstruction of such a large dataset, the quality of the rainfall data used is of considerable concern. As a result, a significant amount of effort was spent on data quality checking. There were a few obvious inaccuracies that were easily identified and eliminated from the data set. However, there were several data values that made determining the credibility of the observations very impossible. As a result, before constructing the daily databases, the data was subjected to some basic quality assurance checks, such as deleting any negative values and all unusually large values that could have resulted from human error, such as missing a decimal point.

An obvious error in translation from journal/register to digital format includes shifting a decimal point by mistake or negative rainfall records. Another problem was a missing negative sign generally assigned to a missing data as -999. The negative sign is sometimes missed in digitization making it look like real observation of 999 mm. These are identified and manually corrected. Another error that enters in some stations is due to difference in units of rainfall records maintained by two different sources. For example, the IMD stores the rainfall records in tenths of mm while some TGs record them in mm. When we see an exact factor of 10 jumps at a merging point, we suspect that it is the cause of the problem. Then original journals are consulted and data corrected in the time series. Individual station data time series were examined throughout time for possible abnormalities. The entire rainfall station dataset was analysed, with stations having more than 70% complete records being used. This method increases the number of stations accessible for analysis on each day while also increasing the spatial rainfall data for the region. Rejecting multi-day rainfall records (which usually happens on a Monday) and physically inspecting the heaviest rainfall days to ensure that the extreme values were consistent with the surrounding records for that day were also quality control methods.

All rainfall occurrences with 24-hour rainfall at a single station above 150 mm were further evaluated for probable mistakes, as 150 mm was defined as an exceptional rainfall event and reflects the 99th percentile of daily maximum rainfall at the majority of NEIN stations. When rainfall at any NEIN station surpassed 150 mm on a given day, the rainfall data at other NEIN sites were also examined. The value was recognised as correct if there were additional stations reporting considerable rainfall on that day or if there was a large percentage of rainfall stations over NEIN reporting rainfall.

As a result, the second group of mistakes deleted from the data were all the instances where a station recorded extremely heavy rainfall for multiple days in a row despite no evidence of this from nearby stations. From time to time, it has been noticed that accumulated rainfall (lasting more than one day) is not recognised as such in the raw data set. When there was missing data for one or more days followed by a day with extremely high rainfall, this was relatively easy to spot in the data set. Only after a comparison with rainfall from nearby stations was this high rainfall value discarded.

The final mistake check was to compare the occurrences that remained after the previous checks were eliminated with other meteorological data and journals, such as the IMD publications, IITM publications and archived website (Academic and Research Institutes) data. However, it's probable that the daily rainfall dataset created for this study contains some flaws, resulting in some discrepancies in the results. However, because our concentration is on extreme/heavy rainfall occurrences, the impact would be restricted [26].

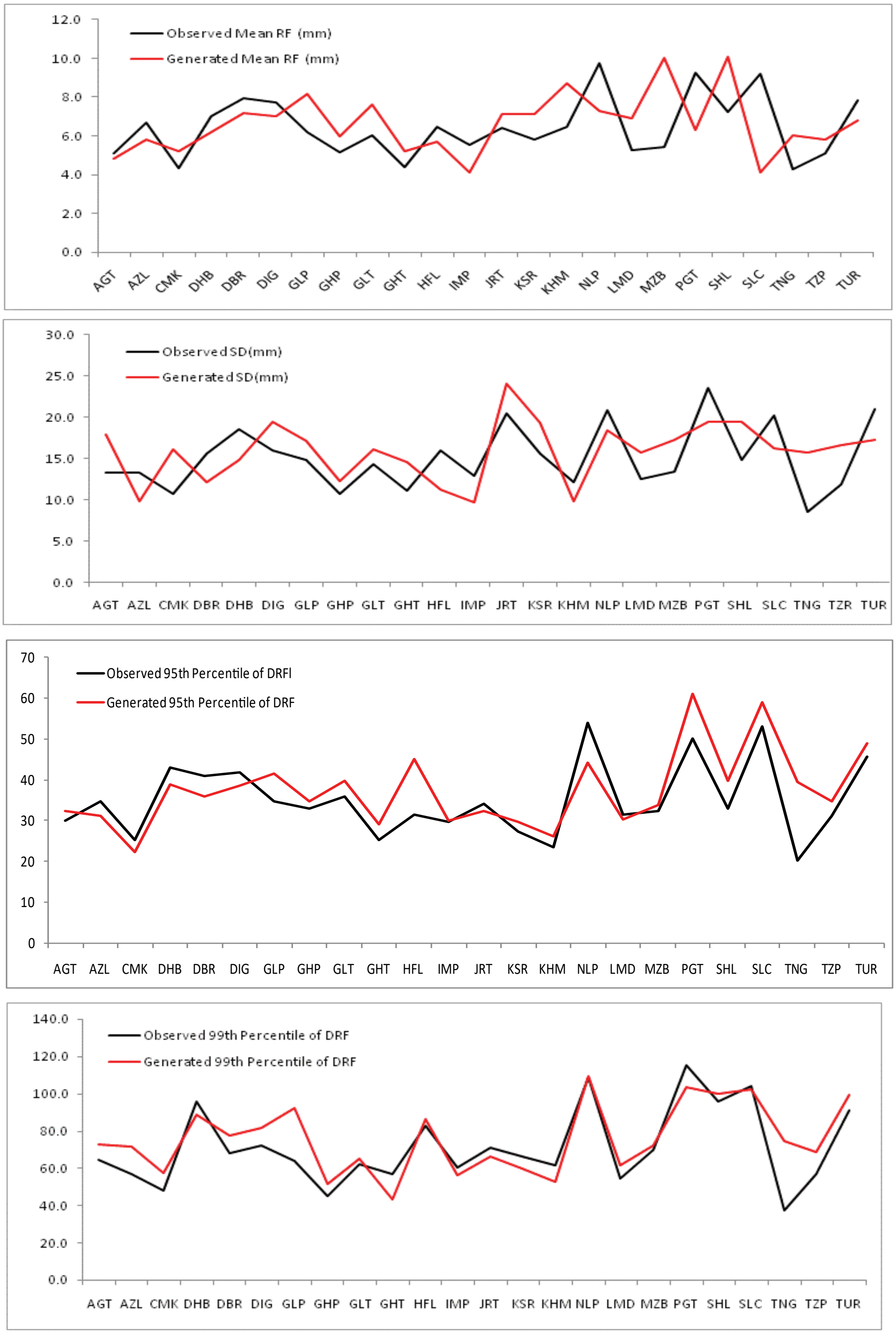
## Results and Discussion

Table 1 shows that the mean rainfall in our target stations varies from 3.5 mm/day to about 15 mm/day indicating that our choice of stations in the region are well distributed accounting for both low and high rainfall regions of NEI. Among the stations used for this reconstruction we have 49 percent Plain stations(at a height of 16m to 200 m above msl), 25 percent foothill stations(200-500 m above msl) and 29 percent hill stations(>500 m above msl). The average and maximum distances from the targeted station to the closest stations are 9.57 km and 65.29 km, respectively (i.e. the interpolation is done with rainfall data from stations up to 65.29 km from the targeted station).

With data from closer stations, better interpolation and data matching can be obtained (if available). For any of the stations, data generation based on a single neighbouring station was ineffective.

The method to reconstruct data differs for different stations. Using line plots presents the various statistics, such as the mean, standard deviation, 95th percentile, 99th percentile and percentage of days with no rainfall of observed and reconstructed datasets.

Figure 4 (a)-(e) Line plots comparing mean, standard deviation, 95th percentile, 99th percentile and percentage of days with no rainfall for observed and reconstructed Daily Rainfall Series.



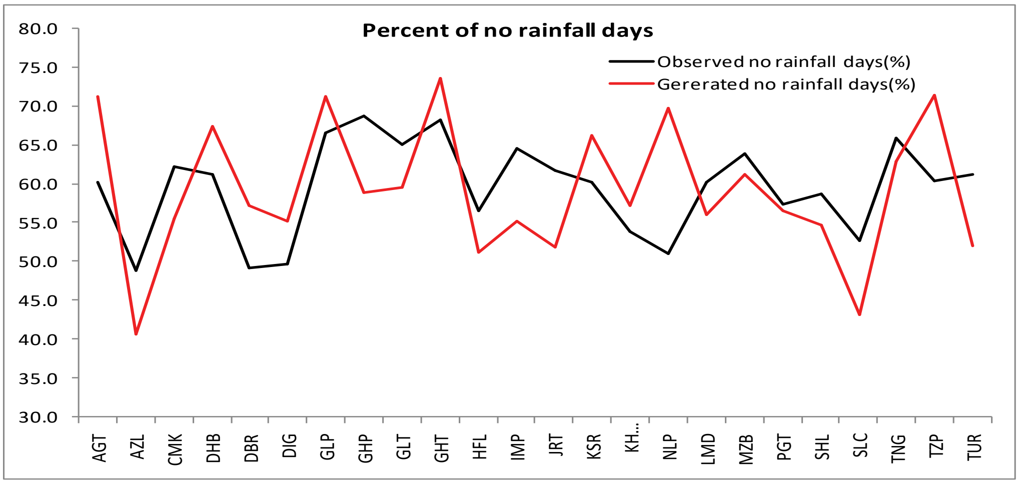


**Figure 4 (a-e):** Line plots comparing mean, standard deviation, 95th percentile, 99th percentile and percentage of days with no rainfall for observed and reconstructed Daily Rainfall Series.

## Percentage of rainfall stations for different methods used for reconstruction

In about 38 percent stations, AMM is the most preferred method for filling missing rainfall data. For another 33 percent stations NRM reconstructed the rainfall series with minimum error. IDWM reconstructed the series with least error in 29 percent stations, all lying in the hills of the region.

### Plains

More than 55 percent stations in the plains preferred AMM for gap filling followed by NRM for 36 percent stations and IDWM was favored for remaining 9 percent plain stations.

### Foothills

Among foothill stations 66 percent stations prefers NRM, with AMM and IDWWN favored by 20 percent and 12 percent stations respectively.

### Hills

Hill stations in general preferred IDWM with 71 percent stations. AMM and NRM sharing 6 percent and 18 percent of the remaining hill stations. For 5 percent hills station, the preferred method for filling missing data was both NRM and IDWM [27].

## Preferred Method of data reconstruction:

### AMM

- In about 38 percent stations, AMM is the most preferred method for filling missing rainfall data, i.e. in these stations, AMM generates rainfall series with numberof rainfall days very close to those observed. AMM capture better the data statistics compared to the NRM for the plain stations interpolated datasets.
- The recreated data via the AMM approach has extremely comparable qualities to the observed series (data) in terms of mean and standard deviation for the stations, Guwahati, Jorhat, Lakhimpur, Dibrugarh, Gohpur and Pasighat all plain stations lying along Brahmaputra valley (along with high CC).

### NRM

- For 33 percent stations NRM reconstructed the rainfall series with minimum error. The NRM generates data that is fairly comparable to actual data in terms of the mean and proportion of rainy days.
- The NMR technique reconstructs data with highly comparable qualities to observed data in terms of mean and proportion of days with rain, primarily in valley and foothill stations. In the Foothill stations, however, NRM slightly overestimate the 95th percentile while underestimating the 99th percentile.
- For the plain stations using the NRM, we find the mean rainfall to deviate a little from the observed value, but it significantly changes the number of heavy rainfall days from observed dataset.
- The NRM approach is extremely reliant on the correlation between the target and nearby stations, and it is only employed if the correlation coefficient is greater than 0.5.

### IDWM

- The IDW technique gave respectable predictions for the hill stations. IDWM reconstructed the series with least error in 29 percent stations, all lying in the hills of the region.
- For IDWM when applied to foothill and plain stations, the interpolated data shows a mean that differ significantly, together with a significant decrease in the probability of extremely heavy ( >244 mm per day) rainfall days from observed dataset. Thus, IDWM method considerably underestimates the variability in daily rainfall in these stations. However, the IDW method gave consistent results for the hill stations in the region.
- Few valley stations favour both NRM and AMM for filling gaps in rainfall series.

None of the algorithms can accurately forecast the zeros of the missing days [28].

## Performance Metrics

The overall performance of all three error statistics must be used to determine the optimal strategy for each station.

- With error method RMSE about 35 percent stations preferred AMM, 39 percent favored NRM and 26 percent followed IDWM.
- MAE preferred AMM for 43 percent stations followed by 30 percent for NRM and 26 percent station favored IDWM.
- CC favored AMM for 35 percent stations, NRM for 39 percent stations and 26 percent IDWM.

Compared to the stations at Plains and foothill stations, the hill stations have relatively lower CC value and higher MAE and RMSE values. Furthermore, on the highlands, the distances between neighbouring stations are also rather long.

All the hill stations (above 1000 m altitude) recorded minimum MAE for IDW method. For all three methods employed to generate missing rainfall values, RMSE is very high for the stations in hills. However using the smallest values of MAE and RMSE for selecting the most suitable method might lead to improper results, as these values tend to be small for small rainfall amounts. Lack of fitting may also be attributed to data matching on the basis of surrounding stations with inadequate daily time scale correlation between rainfalls of neighbouring stations [29] (Table 2).

## Choice of Methods to fill data gaps for individual stations

The best strategy for filling in the missing rainfall gaps was determined using data from nearby observatories and evaluated using three distinct procedures. Table 3 shows the most appropriate approach for each station after comparing estimated data with real observations using three error statistics (Table 4,5).

| Station | RMSE(mm) | | | CC(R) | | | MAE(mm) | | |
|---|---|---|---|---|---|---|---|---|---|
| | AMM | NRM | IDWM | AMM | NRM | IDWM | AMM | NRM | IDWM |
| Golaghat | 11.6 | 5.7 | 11.6 | 0.64 | 0.92 | 0.65 | 5.1 | 2.2 | 5.2 |
| Jorhat | 8.5 | 10.3 | 15.6 | 0.70 | 0.50 | 0.54 | 3.6 | 5.9 | 6.1 |
| Chaparmukh | 7.4 | 10.1 | 16.5 | 0.78 | 0.58 | 0.63 | 2.6 | 4.3 | 4.9 |
| Lumding | 9.2 | 13.1 | 13.3 | 0.71 | 0.57 | 0.53 | 3.7 | 6.0 | 6.2 |
| Halflong | 24.6 | 18.6 | 12.9 | 0.53 | 0.45 | 0.86 | 11.8 | 8.8 | 4.9 |
| Goalpara | 21.9 | 13.9 | 18.8 | 0.51 | 0.61 | 0.49 | 9.7 | 6.8 | 7.9 |
| Tangla | 19.0 | 11.3 | 23.1 | 0.35 | 0.62 | 0.54 | 8.3 | 5.2 | 11.1 |
| Mazbat | 7.8 | 10.2 | 15.4 | 0.84 | 0.55 | 0.59 | 3.0 | 6.4 | 7.0 |
| Silchar | 12.1 | 10.9 | 14.9 | 0.78 | 0.75 | 0.69 | 6.6 | 6.0 | 7.9 |
| Shillong | 13.4 | 15.2 | 8..4 | 0.76 | 0.79 | 0.87 | 5.4 | 6.7 | 4.1 |
| Agartala | 7.9 | 12.9 | 17.0 | 0.39 | 0.58 | 0.32 | 6.7 | 5.2 | 7.7 |
| Imphal | 10.3 | 10.5 | 4.9 | 0.63 | 0.62 | 0.92 | 4.5 | 4.9 | 1.9 |
| Aijal | 15.2 | 23.8 | 8.4 | 0.49 | 0.56 | 0.73 | 7.1 | 12.2 | 4.7 |
| Dibrugarh | 6.8 | 16.4 | 18.5 | 0.90 | 0.49 | 0.50 | 2.2 | 7.6 | 7.3 |
| Tezpur | 12.9 | 10.0 | 12.9 | 0.26 | 0.33 | 0.56 | 5.8 | 4.8 | 5.4 |
| Pasighat | 9.7 | 13.3 | 14.3 | 0.89 | 0.47 | 0.60 | 5.6 | 5.1 | 9.0 |
| Lakhimpur | 11.7 | 23.3 | 20.5 | 0.61 | 0.53 | 0.51 | 7.2 | 10.6 | 9.3 |
| Digboi | 17.3 | 14.6 | 12.8 | 0.52 | 0.51 | 0.63 | 7.9 | 7.0 | 5.4 |
| Tura | 28.4 | 26.7 | 16.0 | 0.49 | 0.15 | 0.68 | 9.6 | 10.3 | 7.6 |
| Guwahati | 7.6 | 12.8 | 13.5 | 0.59 | 0.30 | 0.54 | 5.0 | 5.8 | 6.0 |
| Dhubri | 18.6 | 10.7 | 23.5 | 0.43 | 0.69 | 0.62 | 7.2 | 5.3 | 10.7 |
| Kohima | 14.4 | 14.4 | 6.1 | 0.52 | 0.49 | 0.54 | 7.2 | 7.3 | 5.7 |
| Gohpur | 11.0 | 7.7 | 14.1 | 0.89 | 0.56 | 0.55 | 4.5 | 7.0 | 7.4 |

**Table 2:** Error statistics for each method at the 24 stations.

RMSE-Root Mean Square Error; CC-Correlation Coefficient, MAE: Mean absolute errors.

| Method adopted | Stations |
|---|---|
| AMM | Chaparmukh, Lumding, Mazbat, Dibrugarh, Lakhimpur, Guwahati, Gohpur, Pasighat, Silchar |
| NRM | Golaghat, Jorhat, Goalpara, Tangla, Agartala, Dhubri, Kailasahar, Tezpur, Pasighat, Silchar |
| IDWM | Halflong, Shillong, Imphal, Aijal, Digboi, Tura, Kohima, Tezpur |

**Table 3:** Method adopted to fill data gaps for individual stations.

| Test Stats | RMSE | MAE | CC |
|---|---|---|---|
| AMM | 14.3 | 6.1 | 0.78 |
| NRM | 15.3 | 6.6 | 0.57 |
| DWM | 15.1 | 6.7 | 0.68 |

**Table 4:** Average value for test Statistics (mm).

| Error Stats | Range Error |
|---|---|
| RMSE(mm) | 4.9-22.7 |
| CC | 0.50-0.92 |
| MAE(mm) | 1.9-9.1 |

**Table 5:** Range of error statistics.

### Station Correlation Coefficient (CC) between the reconstructed and observed daily rainfall series

The correlation coefficient between reconstructed and observed daily rainfall series for individual stations range from 0.53 to 0.92 (Figure 4c) indicating fairly good relationships between the reconstructed and observed datasets, with highest correlation, seen for Pasighat and Jorhat has the lowest value of correlation coefficient. Significant high correlation between the data sets is seen over western and eastern part of Assam, west Arunachal Pradesh bordering Bhutan and Tripura valley in the south. In general stations at higher altitude have comparatively lower value of CC. Most foothill stations in the region show strong CC between the observed and reconstructed datasets (Figure 5).

### Station Correlation Coefficient (CC) between the reconstructed and observed daily rainfall series

The correlation coefficient between reconstructed and observed daily rainfall series for individual stations range from 0.53 to 0.92 (Figure 4c) indicating fairly good relationships between the reconstructed and observed datasets, with highest correlation, seen for Pasighat and Jorhat has the lowest value of correlation coefficient. Significant high correlation between the data sets is seen over western and eastern part of Assam, west Arunachal Pradesh bordering Bhutan and Tripura valley in the south. In general stations at higher altitude have comparatively lower value of CC. Most foothill stations in the region show strong CC between the observed and reconstructed datasets (Figure 5 and 6).

### Comparison with other datasets

Figure 7 shows that IITM data set for Northeast India Homogenous region and the present dataset show almost similar inter annual and inter decadal variation. Their differences in long term mean values are consistent with difference in number of stations going into the mean. However, the IMD gridded dataset shows a different inter-decadal variation along with a much higher mean rainfall. This difference may be due to the inclusion of the Cherrapunji observatory in the IMD dataset. Both our dataset and IITM dataset, without the inclusion

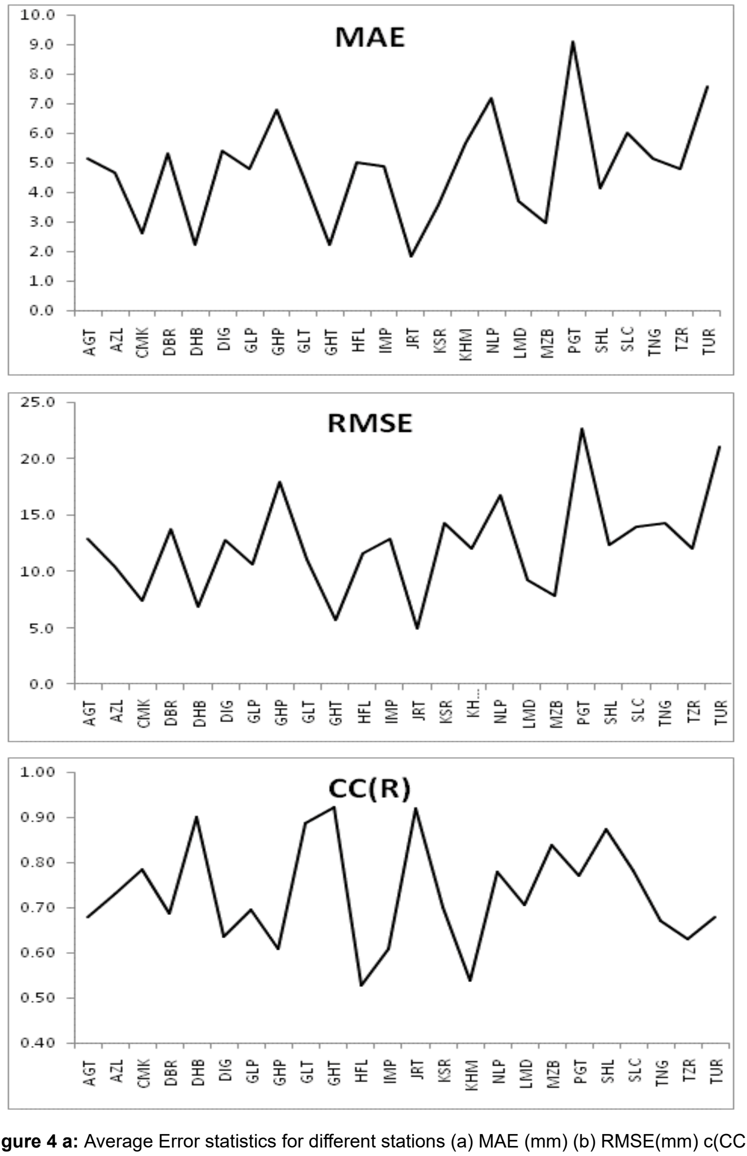


**Figure 4 a:** Average Error statistics for different stations (a) MAE (mm) (b) RMSE(mm) c(CC).

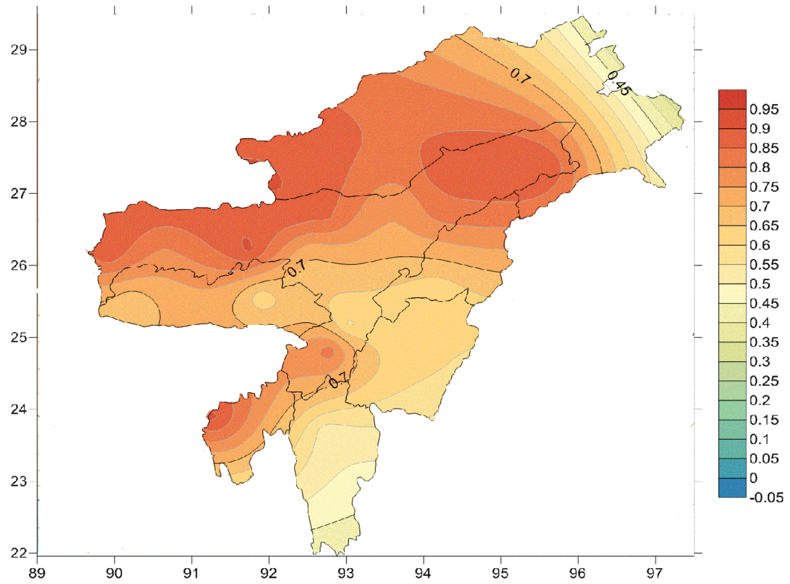


**Figure 5:** Spatial distribution of correlation coefficient values for reconstructed and observed rainfall series over NEI.

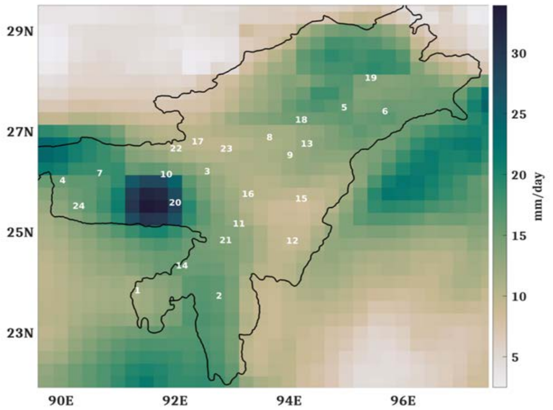
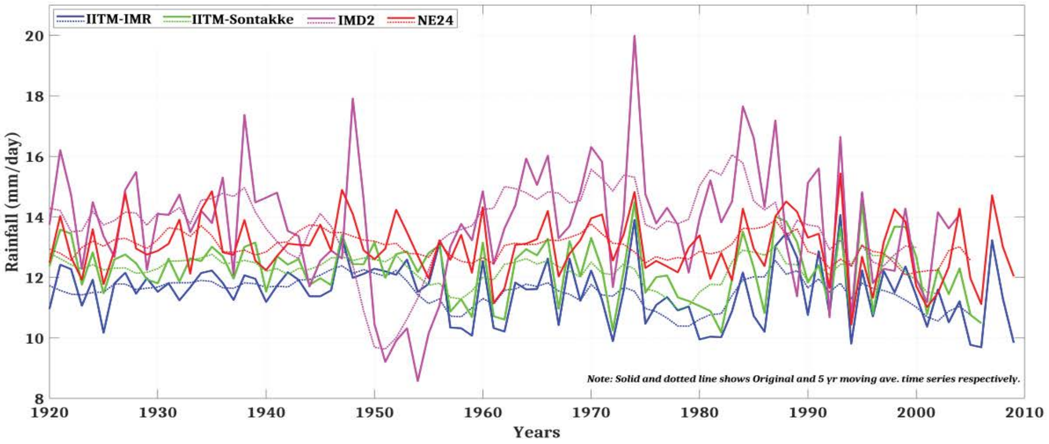


**Figure 5:** Rainfall climatology of NEI for JJAS period extracted from TRMM Data -3B43 Precipitation (0.25o X0.25o), 1998-2013 along with the distribution of stations chosen for reconstruction.



*Note: Solid and dotted line shows Original and 5 yr moving ave. time series respectively.*

**Figure 7:** Comparison plot of Seasonal (JJAS) mean rainfall over the NEIR from three different data sets. (a) Gridded dataset from 1920-2006, (blue) of (Rajeevan et al. 2008)(b) 24-NE station data set (red) from 1920-2009 and (c)IITM generated dataset for homogeneous region from 1920-2009 (purple) from the (Parthasarathy et al. 1994).

of Cherrapunji give a much lower value for the mean rainfall. This suggests that NEI sampling design is an important aspect of collecting rainfall (meteorological) data for scientifically sound decision-making. When generating reliable rainfall data, it's important to examine the data's representativeness in relation to the study's goal [30].

## Conclusions

In the absence of a sufficiently long reliable time series of rainfall in the region, the primary objective has been to reconstruct a reliable and sufficiently long time series of daily rainfall over the region. Consequently, for studying climate change impact on rainfall in general and extreme rainfall in particular, we have reconstructed daily rainfall at 24 spatially well distributed stations over NEI from 1920-2009. For a daily rainfall database for the period 1920-2009 across NEI, we looked at three potential strategies for filling gaps in records. Based on Root Mean Square Error (RMSE), MAE, and Correlation Coefficient, the data recreated for the target stations in NEI were compared to real observations recorded (CC). Both AMM and NRM provide good forecasts for target stations with a high correlation coefficient with neighbouring stations, according to the study's findings. For stations having comparatively low correlation coefficients with surrounding stations, IDM and NRM outperformed all other approaches for guessing missing rainfall amounts.

For 29 percent stations the preferred method is IDWM, 33 percent stations gap filling was done with NRM and about 37.5 percent station were filled using AMM. The IDWM do not capture well the extremely heavy rainfall events. None of the algorithms can accurately forecast the zeros of the missing days.

The new data set for the period 1920-2009, reconstructed for a fixed number of 24 stations, well distributed among the low to high rainfall region with less than 3% missing value will be a much more reliable and stable data set over NEI. It can give a better understanding of rainfall variability over this region. Because the dataset is fixed, so this data set is less biased dataset, which is necessary for the study of rainfall. We have excluded the station of Cherapunjee with its very high mean rainfall compared to other stations, a few missing data of which could give rise to false trend and present biased statistical results. This long term fixed stations RF data can be used to study the inter-decadal variation of RF, analysis of extreme events, spells etc., and can provide many insights about rainfall variability over NEI.

## Discussion

When our analysis demonstrated that a decreasing tendency of both mean and extreme events identified in an earlier study [3] is an artefact of that study's short data length of 32 years, the relevance of generating the larger time series of NEIR became instantly apparent. Furthermore, because climate records contain multi-decadal variability, at least a hundred years of data are necessary to detect any anthropogenic signal in climate time series. For more than 30 years, there had been no good data on seasonal mean rainfall and the frequency of occurrence of extreme events in the NEI.

With a major data mining exercise, we constructed such a quality controlled and reliable data set based on daily rainfall records on 24 well distributed stations over the region for the period 1920 to 2010 (~90 years). The efforts made in construction of this data set for 90+ years has made it possible to derive this new valuable insight that in the background of rather strong natural multi-decadal and centennial variability of the NEIR, influence of anthropogenic forcing, if any, is impossible to decipher.

Using this unique data set Zahan et. al. 2021 examined the trends of seasonal mean rainfall, extreme events as well as those of wet and dry spells, as well as inter-annual, decadal and multi-decadal modes of variability. The results show that in the seasonal mean rainfall, a long term trend is conspicuous by absence but dominated by a major multi-decadal variability strongly coherent with the global multi-decadal mode of variability with ~ 65 year period with the AMO as the primary pacemaker. Their study concluded that the modulation of the NEIR by the natural variability (e.g., the global multi-decadal mode with ~65 year period) is stronger than the anthropogenic response(of greenhouse gas forcing and the aerosol forcing) making it challenging to isolate the impact of climate change in the NEI.

### Acknowledgements

We thank Prof. B. N. Goswami, former Director of IITM for motivating and providing valuable suggestions for reconstruction of this data set; Prof. Arup k Sarma, IIT, Guwahati; Dulal Charkarvorty, IMD, Delhi; Late R C Kachari, Mr Amar Bhuyan, Mr Jayanta Sarma and Mr. Deva Kanta Handique, RMC, IMD, Guwahati; Dr. Prasanta Neog of Assam Agriculture University; Dr. Walter Fernandes of Northeast Social Science Research Centre, Guwahati, for introducing to the vast data with JESUITS; Assam State Archive, Dispur; Assam Historical and Antiquarian Study Centre, Guwahati, Assam state Media Archive, Guwahati; National Data Center, Pune; Assam Branch of Indian Tea Association, Guwahati; Assam tea Planters' Association, Jorhat; Tocklai experimental Research Centre, Jorhat; Oil and Natural Gas Corporation, Sibsagar; Oil India Limited, Digboi; NEEPCO, Shillong, ICAR, Barapani.

### References

1. Brunet M, Jones P (2011) Data Rescue initiatives: bringing historical climate data into the 21st century. Clim Res 47: 29-40.
2. Mahanta R, Sarma D, Choudhury A (2013) Heavy rainfall occurrences in northeast India. Int J Climatology 33: 1456-1469.
3. Goswami BB, Mukhopadhyay P, Mahanta R, Goswami BN, (2010) Multiscale interaction with topography and extreme rainfall events in the northeast Indian region. J Geophys Res Atmos p: 115.
4. Sonam MK, Kumar KK (1990) some aspects of daily rainfall distribution over India during the south-west monsoon season. Int J Climatology 10: 299-311.
5. Kothawale DR, Rajeevan M (2017) monthly, seasonal, Annual Rainfall Time Series for All-India. Homogenous Reg, Meteorological Sub-Divisions Pp: 1871-2016.
6. Rajeevan M, Bhate, J, Jaswal AK (2008) Analysis of variability and trends of extreme rainfall events over India using 104 years of gridded daily rainfall data. Geophys Res lett 35.
7. Hubbard KG (1994) spatial variability of daily weather variables in the high plains of the USA. Agric For Meteorol 68: 29-4.
8. Grove R (1998) The East India Company, the Raj and the El Nino: the critical role played by colonial scientists in establishing the mechanisms of global climate teleconnections 1770–1930. Nature and the Orient. The environmental history of south and southeast Asia Pp: 123-154.
9. Endfield GH, Morris C (2012) Cultural spaces of climate. Clim Change 113: 1-4.
10. Udías A (1996) Jesuits' contribution to meteorology. Bull Am Meteorol Soc 77: 2307-2316.
11. Udías A (2003) Searching the heavens and the Earth: the history of Jesuit observatories. Berlin: Kluwer Academic Publishers 369.
12. Murata F, Terao T, Fujinami H, Hayashi T, Asada H, et al. (2017) Dominant Synoptic disturbance in the extreme rainfall at Cherrapunji, Noreast India, based on 104 years of rainfall data (1902-2005). J Clim 30: 8237-8251.
13. Paulhus JLH, Kohler MA (1952) Interpolation of missing precipitation records. Mon Weather Rev 80: 129–133.
14. Eischeid JK, Pasteris PA, Diaz HF, Plantico MS, Lott NJ (2000) Creating a serially complete, national daily time series of temperature and precipitation for the western United States. J Appl Meteorol 39: 1580-1591.
15. Xia Y, Fabian P, Stohl A, Winterhalter M (1999) Forest climatology: Estimation of missing values for Bavaria. Germany Agric For Meteorol 96: 131-144.

16. Bennett ND, Newham LT, Croke BF, Jakeman AJ (2007) Patching and disaccumulation of rainfall data for hydrological modelling. InInt. Congress on Modelling and Simulation (MODSIM 2007), Modelling and Simulation Society of Australia and New Zealand Inc., New Zealand Pp: 2520-2526.

17. Tabios GQ, Salas JD (1985) A comparative analysis of techniques for spatial interpolation of precipitation. Water Resource Bull 21: 365-380.

18. Teegavarapu RS, Chandramouli V (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol 312: 191-206.

19. McKeen SA, Wilczak J, Grell G, Djalalova I, Peckham S, et al. (2005) Assessment of an ensemble of seven real time ozone forecasts over eastern North America during the summer of 2004. J Geophys Res p: 110.

20. Savage NH, Agnew P, Davis LS, Ordóñez C, Thorpe R, et al. (2013) Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation. Geosci Model Dev 6: 353-372.

21. Chai T, Kim HC, Lee P, Tong D, Pan L, et al. (2010) Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and $NO_2$ measurements. Geosci Model Dev 6: 1831–1850.

22. Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci Model Dev 7: 1247-1250.

23. Kashani MH, Dinpashoh Y (2012) Evaluation of efficiency of different estimation methods for missing climatological data. Stoch Environ Res Risk Assess 26: 59–71.

24. Campozano L, Sanchez E, Aviles A, Samaniego E, (2014) Evaluation of Infilling Methods for Time Series of Daily Precipitation and Temperature: The Case of the Ecuadorian Andes. MASKANA 5: 101-115.

25. Chen FW, Liu CW (2012) Estimation of the Spatial Rainfall Distribution using Inverse Distance Weighting (IDW) in the Middleof Taiwan. Paddy Water Environ 10: 209-22.

26. De Silva R, Dayawansa N, Ratnasiri M, (2007) A Comparison of Methods Used in Estimating Missing Rainfall Data. The J Agric Sci 3: 101-108.

27. Hasan M, Croke B (2013) Filling Gaps in Daily Rainfall Data: A Statistical Approach. In 20th Int Congr on Modelling Simul. Adelaide, Australia.

28. Moodie DW, Catchpole AJW (1976) ‘Valid climatological data from historical sources by content analysis’. Science, 2: 3- 51.

29. Ranade A, Singh N, Singh HN, Sontakke NA (2008) on variability of hydrological wet season, seasonal rainfall and rainwater potential of the river basins of India (1813-2006). J Hydrol Res Dev 23: 79-108.

30. Willmott CJ, Matsuura K, Robeson SM (2009) Ambiguities inherent in sums-of-squares-based error statistics. Atmos Environ 43: 749-752.