To be submitted to Monthly Weather Review

A Novel Method for Interpolating Station Rainfall Data using a Stochastic Lattice Model

Boualem Khouider*

Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada.

C. T. Sabeerali, R. S. Ajayamohan, V. Praveen

Center for Prototype Climate Modelling, New York University, Abu Dhabi, UAE.

Andrew J. Majda

Department of Mathematics & Center for Atmosphere and Ocean Sciences,

Courant Institute of Mathematical Sciences, New York University, USA.

Center for Prototype Climate Modelling, New York University, Abu Dhabi, UAE.

D. S. Pai

Climate Services Division, India Meteorological Department, Pune, India.

M. Rajeevan

Ministry of Earth Sciences, New Delhi, India.

**Corresponding author address:* Boualem Khouider, Department of Mathematics and Statistics, University of Victoria, 3800 Finnerty Rd. (Ring Rd.), Victoria, BC, V8P 5C2, Canada.

E-mail: khouider@uvic.ca

ABSTRACT

Rain gauge data are routinely recorded and used around the world. However, their sparsity and inhomogeneity make them inadequate for climate model calibration and many other climate change studies. Various algorithms and interpolation techniques have been developed over the years to obtain an adequately distributed dataset. Objective interpolation methods such as the inverse-distance weighting (IDW) are the most widely used and have been employed to produce some of the most used gridded rainfall datasets (e.g. the India Meteorological Department gridded rainfall). Unfortunately, the skill of these techniques becomes very limited to non existent in areas located far away from existing recording stations. This is problematic as many areas of the world are lacking an adequate rain gauge coverage throughout the recording history. Here, we introduce a new probabilistic interpolation method in an attempt to address this issue. The new algorithm employs a multitype particle interacting stochastic lattice model which assigns a binned rainfall value, from an arbitrary number of bins, to each lattice site or grid cell, with a certain probability according to the rainfall amounts assigned to neighbouring sites and a background climatological rainfall distribution, drawn from the available data. Grid points containing recording stations are not affected and are being used as "boundary" input conditions by the stochastic model. The new stochastic model is successfully tested and validated against two standard gridded rainfall datasets, over the Indian land mass.

1. Introduction

Rainfall is one of the most important meteorological parameters on which the lives and the well beings of many living organisms and especially humans depend. The spatial and temporal variability of rainfall is directly linked to the socio-economic development of people in the tropical continents. Real time monitoring of the precipitation on a daily basis is required for planning of various activities like agriculture, construction, travel, and consequently many of the local industries.

To study the dynamics of precipitation variability and to make an assessment of its future variability, a gridded data product from the widely distributed observation stations is essential. Besides, the availability of such a product, on various time scales (hourly to monthly) is imperative to assessing water resources in mountains, arid regions, and river basins. Gridded rainfall data is also required for hydrological and high resolution climate models. Many modelling groups try to understand the characteristics of precipitation using general circulation models. The underlying models need to be verified using the observed gridded datasets in order to improve their performance and prediction skills. The daily observed precipitation is also required to monitor and forecast the subseasonal variability such as monsoon intraseasonal oscillations (MISO) and Madden Julian Oscillations (Sabeerali et al. 2017).

Despite the progress in estimating the precipitation from satellite, the rain gauge observations has a critical role in generating gridded precipitation data over the land areas (Xie and Arkin 1996) and thereby studying spatial and temporal variability of precipitation and its long term trend. Rain gauge data are routinely recorded over the Indian subcontinent and it has the longest recording period than the satellite observations, which make them an ideal source to estimate the precipitation quantitatively and to assess changes in precipitation variability on different time scales. The rain gauge observations are the direct point measurement of precipitation, whereas the satellite estimates and model predictions of precipitation is indirect in nature. Moreover, over the land it is still difficult to estimate accurate precipitation using satellite and hence the satellite estimated precipitation need to be verified or calibrated using the gauge based gridded rainfall data (Xie and Arkin 1995).

Giving the importance of gauge based precipitation data, significant progress has been made to develop various algorithms and techniques to construct gridded datasets from unevenly distributed observational station networks. There are several global or regional gridded precipitation datasets that are available to use for modelling, forecasting, or analysis purposes (Rajeevan et al. 2006; Rajeevan and Bhate 2009; Pai et al. 2014; Xie and Arkin 1997; Huffman et al. 1997; Chen et al. 2002; Gruber et al. 2000; Yatagai et al. 2012; Adler et al. 2003; Xie et al. 1996). These datasets however differ substanitally in terms of their spatial resolution, temporal resolution or the type of techniques used to interpolate the rain gauge data to the regular grid. The most popular gridded rainfall data sets like the Climate Prediction Center Merged Analysis of Precipitation (CMAP; Xie and Arkin (1997)) and the Global Precipitation Climatology Project (GPCP; Adler et al. (2003); Huffman et al. (1997)) are prepared by merging satellite and rain gauge data. The daily gridded precipitation product under the Asian Precipitation Highly Resolved Observational Data Integration Towards Evaluation of Water Resources (APHRODITE) project (Yatagai et al. 2012), covering the whole Asia, and India Meteorological Department (IMD) gridded data (Rajeevan et al. 2006; Pai et al. 2014), covering the Indian subcontinent, are purely rain gauge based products. All these products, irrespective of whether they are merged or gauge based, employ somewhat similar techniques (Shepard 1968; Willmott et al. 1985) for interpolating station rainfall data into regular grid. Despite the abundance of gridded products, the pertaining analyses do not provide estimates of the precipitation variability and the impact of man-made climate change with reasonable accuracy everywhere, and there exists a large difference in the estimated precipitation distributions among different datasets (Yatagai et al. 2005). In a previous study, Xie et al. (1996) have reported that precipitation analysis is not really sensitive to the algorithms used in regions of dense network of rain gauge stations whereas the bias is likely to exist over the regions of sparse networks of gauge observations when spatial inhomogeneities in precipitation exist. Hence, the performance of all these algorithms primarily depends on the density of the rain gauge network and the spatial variability of precipitation.

The algorithms used to interpolate unevenly distributed rainfall gauge data into a regular (usually rectangular) grid are commonly known as objective analysis (OA) methods. OA techniques are often classified into empirical or functional and statistical methods. The empirical or functional techniques provide a functional distribution of rainfall on the regular spatial grid, at a given point in time, using a weighted interpolation of the available station data with weights that are typically inversely proportional to the distance of the stations to the grid point under consideration. The most common statistical technique is due to Gandin (1965). Gandin's method assumes that the rainfall rate at a given grid point is the wighted sum of all station data within a prescribed radius of influence region. The weights attributed to each station are optimized by minimizing the expected interpolation error at the stations, which requires the knowledge of the station variances and covariances Bussieres and Hogg (1989). This method, thus called the optimal interpolation (OI) technique, uses the extra-global information, namely the rainfall variability, instead of simply using the localized station values only.

It is important to note at this point that in each one of these OA techniques, a radius of influence beyond which the algorithm is not applicable is preset to maximize accuracy, and any grid point whose closest data station is beyond this distance is assigned a missing data code (Bussieres and Hogg 1989). Bussieres and Hogg (1989) found an optimal radius of influence, for the four techniques they assessed, of about 40 km, for their particular network of pseudo-gauge data, but they choose to set it to about 110 km for all methods to avoid missing data points on their prescribed grid of 0.05×0.05 degree resolution.

To construct the best possible gridded rainfall products, comparative studies of many different OA techniques are routinely conducted. For instance, (Bussieres and Hogg 1989) compared the empirical or functional OA algorithms of Barnes (1973), Shepard (1968), and Cressman (1959), and the OI method of Gandin (1965) using an unevenly distributed network of pseudo-rainfall station data based on radar observations while Chen et al. (2008) compared the last three algorithms based on real-quality controlled 16,000 rain-gauge station data. Both studies found that Gandin's OI statistical technique is superior to the others but it is often closely followed by Shepard's method. However, Shepard's method is much easier to implement and perhaps it is for this reason only that the aforementioned APHRODITE and IMD datasets, that will be used in this study, are based mainly on Shepard's OA algorithm.

The accuracy of rainfall data depends critically on the interpolation technique and hence the choice of the algorithm is important. Unfortunately the skill of the existing gauge based gridded products are very limited in the data sparse regions. Large errors in the analysis are likely to occur over areas with large spatial variability in precipitation and poor station coverage gauge network (Xie et al. 1996). For example, extremely large rainfall rates are reported occasionally over some individual stations. However, they are unlikely representative of their surrounding areas.

This is problematic as many of the regions in the world still lack an adequate number of rain gauge networks throughout the recording history. Here, we propose a new probabilistic interpolation technique, using a stochastic lattice model (SLM) to grid a network of station rainfall data over India and validate it against the aforementioned APHRODITE and IMD datasets that are based on Shepard's OA technique. The SLM is somewhat a variant of the stochastic multicloud model will local interactions of Khouider (2014) (see also Khouider et al. (2010)) for organized tropical convection. It is based on the concept of particle interacting systems on a lattice, where particles occupying lattice sites or cells randomly switch states according to prescribed probability rules depending on the way the lattice sites interact which each other and on an external potential representing the environmental state. In the present context, the SLM technique uses the global climatological information, namely, the rainfall rate distribution, to stochastically propagate the station gauge values to neighbouring points on the given regular grid. In this sense, the proposed method is closer to the statistical method of Gandin (1965) but instead of minimizing the expected errors it actually samples an estimated probability density at each grid point conditional on the station data and the climatological rain rate distribution. The main motivational question is to assess whether such a stochastically based OA is capable of performing better in regions of sparse observations. In this sense, this study introduces a new concept in station rainfall data analysis that can be extended to global rainfall station data interpolation and especially back in time when the coverage was limited.

While the existing IMD gridded rainfall data is based on a dense network of 6955 stations, here the new SLM algorithm employs only 1830 stations on purpose. To have a fair comparison, we also use Shepard's OA algorithm on the same 1830 stations both with and without a radius of influence.

The paper is organized as follows. Section 2 describes the station data used, the regular grid used to interpolate it, the new SLM algorithm and its parameter calibration, and an overview of Shepard's method. The five rainfall data products, including the high resolution IMD data set, the APHRODITE data set, and the newly produced low station density interpolation data, based on the SLM and Shepard's method with and without radius of influence restriction, are analysis and compared to each other in Section 3. In particular, we first provide a localized assessment of the SML versus Shepard's method by comparing the associated rain event distributions at various locations against those of actual observations. Then we follow up with direct comparisons of the seasonal rain fall climatologies and daily rain fall estimates, statistical metrics such root mean square error, absolute relative error, and cross correlation maps of high resolution IMD dataset versus each one of the four remaining products. The section is concluded with the analysis of the interannual and daily rainfall variabilities corresponding to the five products. Finally, a summary of the results and a few concluding remarks are given in Section 4.

2. Data and Algorithm

a. The Indian rain gauge station data

The Indian subcontinent posses one of the oldest networks of rain-gauge-data in the word. A brief history of the Indian rain gauge data collection and its archival can be found in Walker (1910) and Parthasarathy and Mooley (1978). The first gridded precipitation product for the Indian region is constructed by Hartmann and Michelsen (1989) for the period 1901-1970. The variability of Indian summer monsoon has been routinely studied using this dataset (Hartmann and Michelsen 1989; Krishnamurthy and Shukla 2000, 2007, 2008). A series of studies were conducted, more recently, by the India Meteorological Department (IMD) scientists to quality control the wide network of rain gauge station data in India and to generate a gridded data set that represents the rainfall characteristics in a realistic manner (Rajeevan et al. 2005; Rajeevan and Bhate 2008; Rajeevan et al. 2006; Rajeevan and Bhate 2009; Pai et al. 2014). Although the number of stations and the spatial resolution of the gridded product varied, the algorithm used in these studies was based on the aforementioned Shepard scheme.

We collected a long term record (more than 100 years) of quality controlled daily station rainfall data over the Indian subcontinent from the National Data Centre, IMD, Pune, India. These station data are daily 24 hour accumulated rainfall ending 0300 UTC. For pedagogical reasons, rainfall data of only 1380 stations, spanning across the Indian subcontinent were used to test the new algorithm developed here. We specifically, choose the data used to generate the gridded rainfall data in the Rajeevan et al. (2008) study, which we refer to, here, as the IMD1380 data product. However, the new method developed here is assessed against the IMD high resolution gridded data in Rajeevan and Bhate (2009), which is based on a much denser network of 6955 stations and used here as a high standard reference. This data product will be referred to as IMD6955.

As already stated, the specific question asked here is whether such a scheme can improve the precipitation estimate over grids with poor rain gauge coverage. As indicated in Figure 1a., the 1830 stations are distributed unevenly over the Indian subcontinent with fewer stations over the northeast region and eastern coast of India while the network is relatively dense over the central India and southern peninsular region. Besides, not every rain gauge station has an acceptable precipitation record every day.

We note in particular that because the radius of influence constraint associated with Shepard's method, the IMD1380 leaves large areas of the Indian continent grid with missing data, especially in the readily mentioned low station density regions. We thus decided to expand the application of Shepard's interpolation scheme to data points beyond the predefined radius of influence.The resulting data product is referred to, here, as the IMD1380-relaxedR. We note however, the issue could have been addressed by simply increasing the values of the radius of influence until the whole grid is fully covered as Bussieres and Hogg (1989) did but our results indicate that within the radius of influence, the IMD1380 and IMD1380-relaxedR products are hardly different from one another.

The probability of occurrence of rainfall events used all existing stations, in India from 1910 to 1970, is shown in Figure 2b together with a power law fit. The rainfall distribution over the Indian subcontinent seems to follow the fitted power law. The maximum probability of rainfall occur in the range of 0-100 mm day^{-1} and then the probability decreases rapidly with the intensity of rainfall (Figure 2b).

As already stated not all stations have recorded good quality rainfall data every data. The 130 stations used in this study have a minimum of 70% data availability during the analysis period 1951-1970. However, the data density is not uniform over the Indian subcontinent. While the gauge network over southern India and northwest and central India is dense, it is scattered over northeast and eastern coastal region (Figure 1a). Note that in this data source, there are no stations reported with precipitation over Jammu and Kashmir.

b. The stochastic model on a triangular lattice for rainfall data interpolation

1) TRIANGULATION, MASK, BINNING, AND BACKGROUND DISTRIBUTION

To better accommodate the complexity of the continental boundaries of the Indian peninsula, we adopt a triangular configuration for the stochastic lattice model. The Indian subcontinent is divided in to M triangular mesh elements as shown in Figure 1b. In our analysis, we consider $M = 11921$ which is approximately equivalent to 0.25° spatial resolution.

At any given time, t, spanning the period of interest, a given triangle $I, I = 1, 2, \cdots, M$, on the triangulation lattice may or may not contain station data. Station data will be present at site or cell I if there are stations inside the triangle and if some of these stations have recorded quality-control-acceptable measurements. In such case, the average of all these station values is computed an assigned, as an observation value, and the corresponding triangle or cell j, which is considered as an observation cell. All other cells are meant to be filled in by the OA procedure.

To illustrate, Figure 2a shows day to day variation of the number cells containing stations with recorded rainfall data, from 1910 to 2003. The data density is satisfactory and more or less uniform till 1995. Out of total 11921 triangular cells over the Indian subcontinent, on an average around 1200 cells with rainfall is available. However, during the recent times there is a drop in the number of cells recorded with rainfall data (Fig 2a). In this study, we restricted our analysis to the period from 1951 to 1970, for homogeneity.

For convenience, we introduce the binary function, defined on the lattice as

$$\mathscr{M}_t(I) = \begin{cases} 1, & \text{if there is station data in cell } I \text{ at time } t \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

for $I = 1, 2, \cdots, M$, which serves as a mask defining the lattice points with station data and those without any station data, at any given time t. Comparing the number of triangles $M = 11921$ to the number of cells with recored data in Figure 2a, which is limited from above by the total number of stations used, 1380, there is at least 88% of lattice cells that are attributed the values $\mathscr{M}_t = 0$, at any given time. It is the job of the interpolation method to fill up those gaps.

The new stochastic lattice model (SLM), introduced here, is based on the concept of multi-type particle interacting systems (Khouider 2014), which define an order parameter, denoted by σ, that takes one of the discrete values from 0 to $N-1$, at each one of the lattice sites and makes random jumps from one discrete state to another depending on prescribed probabilistic rules, based on the states of the nearest neighbours. In the present study, the station rainfall data are binned into N rain rates, corresponding to the N states of the SLM. To better accommodate the distribution of the recorded rainfall, we adopt a piecewise-uniform binning strategy. Various bin configurations have been tested, with a total size ranging from $N = 51$ to $N = 137$. Our results indicate that the finer the bin sizes are the more accurate the interpolated rain fall is. However, the finer bins are associated with larger bin sizes and as such the computational time increases with the increased accuracy. As a compromise between accuracy and computational efficiency, we adopt the configuration with $N = 137$ illustrated in Table 1, as our standard case. The results of our model calibration with respect to the bin size are reported below for completeness.

The choice of the bin configuration, is partly motivated by the background or climatological rainfall rate distribution reported in Figure 2b. To accommodate the SLM implementation, this distribution was binned accordingly. The resulting coarsened distribution, denoted by $\rho_j, j = 0,1,2,\cdots,N-1$, is obtained by further assigning the probability of occurrence of rain rates, based on the full IMD dataset spanning from 1951 to 2004, corresponding to each SLM bin,

$$\rho_j = \frac{\text{number of rainfall events with a rain rate within bin } j}{\text{total number of rainfall records}}. \quad (2)$$

The bin resolution is thus set to be higher in regions where the rainfall rate distribution varies the most, resulting in the configuration in Table 1.

2) THE JUMP PROCESS AND MARKOV SAMPLING

One can think of the previously defined lattice as containing particles. Different numbers of particles are contained at different sites. At any given time t, each lattice site is either occupied by a certain number of particles, corresponding to a rainfall bin number or none, if there is no rainfall. More precisely, we consider the order parameter

$$\sigma_I^t = j, \quad j = 0,1,.....,N-1 \quad (3)$$

on a given lattice site I, $I = 1,2,\cdots,M$, and at any given time t, according to whether there is a rain event within the bin $j, j = 0,1,2,\cdots,N-1$, in that cell at that time t. Let R_j be the rain rate associated with bin j, $j = 0,1,2,.....N-1$. In the jargon of particle interacting systems, a realization of the order parameter σ^t on the lattice is called a configuration. The size of the configuration space, formed by all possible such configurations, increases exponentially with the number of lattice cells M. It is given by N^M where N is the number of discrete states.

Particles interacting systems in a heat bath, with infinite external energy supply, assume the Gibbs canonical distribution,

$$G(\sigma) = \frac{1}{Z}\exp[-\beta H(\sigma)], \tag{4}$$

as their equilibrium measure (Liggett 1999; Thompson 1972), where H is the Hamiltonian energy which includes the energy associated with the way the lattice sites interact with each other and and external energy source, and Z is a normalization constant known as the partition function. Here, we view rainfall rates as particles of such a system that respond to weather conditions as random deviations from the climatology represented by the distribution ρ_j in (2). The interpolation problem becomes then one of finding the best possible Hamiltonian H or distribution G given the station data. We assume that H takes the form

$$H(\sigma) = -\frac{1}{2}\sum_I \sum_{I'} J(\sigma_I, \sigma_{I'}) + \sum_I h(\sigma_I),$$

where J is the internal interaction potential between neighbouring sites and h is the external energy potential. The specific form of J, which is not necessary at this stage, will be given through the definition of the energy differences, between nearest configurations, when designing our sampling methodology, which takes into account the knowledge of the rainfall climatology and instantaneous station data at lattice sites with $\mathscr{M}_t(I) = 1$. The sampling strategy is given next.

For practical reasons, we use the Markov Chain Monte Carlo sampling method based on Arrhenius Dynamics (Thompson 1972), where for any fixed physical time, the order parameter σ^t is viewed as a Markov process that makes random transitions at random lattice sites, over a long enough period of pseudo-time, t, until it reaches a statistical equilibrium, whose distribution is the Gibbs measure conditional on the climatology and the instantaneous station data.

Next, we introduce the Hamiltonian energy differences at each lattice site, including where station data is available, based on the nearest neighbour interaction potential J (Khouider 2014). We define

$$\triangle_+^I \tilde{H}(\sigma) = J_0[\max_{I'}(|R(\sigma_I+1)-R(\sigma_{I'})|) - \max_{I'}(|R(\sigma_I)-R(\sigma_{I'})|)] + h(\sigma_I+1) - h(\sigma_I)$$
$$\triangle_-^I \tilde{H}(\sigma) = J_0[\max_{I'}(|R(\sigma_I-1)-R(\sigma_{I'})|) - \max_{I'}(|R(\sigma_I)-R(\sigma_{I'})|)] - h(\sigma_I+1) + h(\sigma_I) \quad (5)$$

as the Hamiltonian energy differences between a state with a given configuration σ and the two closest possible states where the rainfall at site I jumps either to the next bin up or to the next bin down. Here $J_0 > 0$ represents the strength of local interactions and is considered as an interpolation parameter and $R(x)$ is the rainfall rate, R_x, associated with bin x, $0 \leq x \leq N$. Our tests indicate that the optimal J_0 value depends on the number of bins, N, and $J_0 = 1.05$ seems to be the ideal choice when $N = 137$. Increasing J_0 diminishes the weight of the prior climatological equilibrium distribution, which is set so as to replicate the influence of the external potential h (Khouider 2014) as specified below.

To guarantee convergence to the proper equilibrium distribution, the jump rates of the Markov process, σ^t, from a given configuration σ to its two closest "neighbours" in the configuration space, are given by

$$C_+^{I,j}(\sigma) = [1-\mathscr{M}(I)]\tilde{C}_+^I e^{-\triangle_+{}^I H(\sigma)/2} + \frac{\mathscr{M}(I)}{\tau}[\max(e^{-\alpha(\sigma_I-\sigma_I^\star)}, 1.0) - 1.0]$$
$$C_-^{I,j}(\sigma) = [1-\mathscr{M}(I)]\tilde{C}_-^I e^{(-1/2)\triangle_-{}^I H(\sigma)} + \frac{\mathscr{M}(I)}{\tau}[\max(e^{\alpha(\sigma_I-\sigma_I^\star)}, 1.0) - 1.0] \quad (6)$$

Here α and τ are positive parameters that are specified in Table 2 together with the other model parameters while $\mathscr{M}$ is the binary mask function in (1) and $0 \leq \sigma_I^* \leq N-1$ is a fixed bin index corresponding to the observed rainfall data at the given cell I, if available.

The background rates $\tilde{C}^j_+$ and $\tilde{C}^j_-$ on the other hand are defined based on the climatological rainfall distribution in (2). We set

$$\tilde{C}^j_+ = \frac{1}{\tau}\frac{\rho_{j+1}}{\rho_j}, \;\; j = 0,1,2,...,N-1$$
$$\tilde{C}^j_- = (1/\tau), \;\; j = 1,2,...,N, \tag{7}$$

which is equivalent to defining the external potential h so that $\rho_j = e^{h(j)}$.

This completes the formal definition of a Markov jump process according to which, the order parameter σ^t_I can jump up by one unit or jump down by one unit with transition probabilities depending on whether its neighbours have more or less particles and the prescribed background climatology. We have

$$\mathrm{Prob}\{\sigma_I^{t+\triangle t} = \sigma^t_I + 1\} = C^I_+(\sigma^t)\triangle t + o(\triangle t)$$
$$\mathrm{Prob}\{\sigma_I^{t+\triangle t} = \sigma^t_I - 1\} = C^I_-(\sigma^t)\triangle t + o(\triangle t),$$
$$\mathrm{Prob}\{\sigma_I^{t+\triangle t} = \sigma^t_I\} = 1 - [C^I_+(\sigma^t) + C^I_-(\sigma^t)]\triangle t + o(\triangle t),$$

for small time increment $\triangle t$, $\triangle t/\tau \ll 1$, of the pseudo-time t, used to iterate the process to equilibrium.

The definition of the transition rates in (6) and (7) ensures that the underlying Markov process is in "partial detailed balance" with respect to the Gibbs measure in (4) and as such the probability distribution of the stochastic process σ_t converges to $G(\sigma)$ in the long run (Khouider 2014). Therefore, according to the MCMC theory, the time series of the process σ_t can be used to sample $G(\sigma)$, conditional to the station data, and thus to provide probabilistic estimates or interpolates for the rainfall rates at lattice sites where observations are not available.

The dependence of the transition rates in (6) on the mask function $\mathscr{M}$ is such that the convergence of the process to the observed values σ^*_I occurs on an exponentially faster time scale, at all lattice sites with station data, independently on the background climatology distribution and on the state of the neighbouring sites; σ_t becomes quickly (almost) deterministic at those locations. The rate of this convergence is set by the parameter α which bears the large value $\alpha = 4$. The station values are then used to update the values of its neighbouring cells, which then transmit the information to their own neighbours are so on. The process goes back and forth until statistical convergence. Our tests indicate that fixing the values to $\sigma_I^t = \sigma_I^*$ at the cells with observation data lead to the same results by also results in a less smooth convergence of the process.

To implement the MCMC procedure, we adopt Gillespie's exact algorithm as done in Khouider (2014). Accordingly, we introduce the total transition rate, contributed from all grid cells

$$S_R = \sum_I (C_+^I(\sigma) + C_-^I(\sigma)). \tag{8}$$

Also, to avoid the occurrence of unphysical values of σ, we enforce the "boundary conditions",

$$R_-^I(\sigma_I) = 0, \text{ if } \sigma_I = 0 \text{ and } \; R_+^I(\sigma_I) = 0, \text{ if } \sigma_I = N,$$

at each lattice cell $I = 1, 2, \cdots, M$.

In a few words, Gillespie's exact sampling algorithm can be summarized as follows. Let $T_0 > 0$ be a fixed peuso-time measured in the units of the algorithm's time scale τ, chosen to be large enough. Given an initial guess distribution σ_I^0,

1. Read the station day at the given physical time (day of the year between 1951 and 1970 for us) and set $T = T_0$.

2. Compute the up and down transition rates C_+^I and C_-^I using (6) at every cell $I, I = 1, 2, \cdots, M$ and compute the total rate S_R using (8)

3. Draw a uniform random number U between 0 and 1 and set $s = -(1/(S_R))\log(U)$

4. If $s \leq T$, make a single transition at a random site I in the following way.

(a) Renumber the rates $C_+^I(\sigma)$ and $C_-^I(\sigma)$ from 1 to $2M$, say, $C_1 = C_+^1$, $C_2 = C_+^2, \cdots, C_M = C_+^M$, $C_{M+1} = C_-^1$, $C_{M+2} = C_-^2, \cdots, C_{2M} = C_-^M$. Compute the probabilities $P_k = C_k/S_R$ and their cumulative sums $S_k = \sum_{l=1}^k P_l$, $k = 1,2,...,2M$.

(b) Draw a second random number U^1, uniformly between 0 and 1 and independent of U, and find the first k_0 such that $S_{k0} \geq U^1$ and perform the transition associated with C_{k_0}:

$$\sigma_I = \begin{cases} \sigma_I + 1 & \text{if } C_{k_0} = R_+^I \\ \sigma_I - 1 & \text{if } C_{k_0} = R_-^I \\ \sigma_I, & \text{otherwise.} \end{cases}$$

(c) Set $T = T - s$ and go back to Step 2.

5. If $s > T$ stop.

We note that one and only one site is affected at each iteration of the Markov process. Thus, only the transition rates, $C_\pm^I$, corresponding to that site and its immediate neighbours need to be recalculated, every time Step 1 is called again.

When dealing with an observation time series of rainfall like it is the case here, the converged values at the previous time can be used as the initial guess for the present physical time.

To facilitate comparison with existing data products, namely the IMD6955 and APHRODITE datasets, the unstructured triangular cell output is converted to point values at grid points with regular lat-lon grid ($0.25^\circ \times 0.25^\circ$) using the bilinear interpolation. This newly gridded dataset is named as the CPCM1380 data product, in reference to the Center for Prototype Climate Model at NYU Abu Dhabi where this research was conducted and the 1380 rain gauge stations used. Given that the triangular and rectangular grids have the same resolutions of 0.25°, it is expected that the error inducted by this grid conversion is minimal compared to the errors induced by the original objective analysis of inferring the lattice rainfall data from the rain gauge data.

c. Convergence of the MCMC time-series and sensitivity to parameters of the SLM scheme

As already mentioned, the MCMC algorithm consists in running an ergodic Markov process to equilibrium, whose equilibrium distribution is the one one wishes to sample, and use the converged pseudo-time series to draw samples for that distribution. To ensure that the MCMC runs in our SLM scheme have been satisfactorily run to convergence, we monitored the Markov chains at several grid points and time instances, and set the iteration pseudo-time accordingly. The results from this exercise led us to choose a conservative iteration time $T_0 = 24$ hours. For the sake of illustration, we plot in Figure 3 the MCMC pseudo-time series corresponding to the lattice point with lat-lon coordinates 28N, 80.75E and the day 19-Jul-1951, for 6 different bin sizes. As we can see from Figure 3, after a transient period of up to $\tilde{3}$ hours (10,000 pseudo-time steps), the chains enter a statistical steady state where they fluctuate up and down within their stochastic variability range. As can be surmised from Figure 3 both the length of the transient period and the width of the variability range depend strongly on the bin number. As expected the transient period is longer for the larger number of bins (137) while the variability range is shorter for the larger bin number. Notice, however, despite these discrepancies, the converged values seem to oscillate around fairly the same rainfall limit. In our preliminary tests presented here, we took the average over the last 20% of each chain as the interpolated rainfall value at the corresponding lattice cell. To take full advantage of the stochastic nature of the scheme, the associated variances can be also recorded to provide some measure of uncertainty in the interpolated data. This will be done in the future.

Preliminary tests indicated that the scheme is most sensible to the values of J_0 and the number of bins N. In Table 3, we report the root mean square errors (RMSE) between the interpolated and regridded rainfall data based on the SLM scheme, CPCM1380, and the high resolution IMD6955 dataset for various values of J_0 and bin size N, integrated over the totality of the structured grid for the monsoon season JJAS 1951 . As we can see from this table, for a fixed J_0 the RMSE typically increases with decreasing bin size while its variation with respect to J_0 is more subtle. For a fixed bin size, the RMSE seems to increase both when J_0 is increased and when J_0 is decreased and suggests the prevalence of a sweet spot somewhere in between. According to Table 3, $J_0 = 1.05$ and $N = 137$ seem to be an optimal choice in terms of minimizing the RMSE in comparison to the high resolution IMD6955 dataset. It is worth noting that in the process, we have also calculated the correlation coefficient between the CPCM1380 and the IMD6955 data products of JJAS 1951, for the parameters in Table 3. Our results indicate that the correlation coefficient hardly changes, regardless of the value of J_0 or the bin number N. It varies between 0.94 and 0.95 for all the parameter pairs recorded in Table 3, which suggest that the scheme is robust and can eventually be trusted even at coarse bin configurations. It is in particular at the higher 0.95 value when $J_0 = 1.05$ and $N = 137$. This is the main reason why this value of J_0 is chosen to be our default value instead of simply $J_0 = 1.1$, which appears to have the same smallest RMSE value of 1.09 mm day^{-1}.

d. Shepard weighted interpolation method and its relaxation

As already mentioned, the SLM interpolation technique is assessed in comparison to the high resolution (0.25x0.25) rainfall product IMD6955 which is obtained using the inverse distance weighted interpolation method of Shepard (1968) based on data collected by 6955 rain gauge stations (Pai et al. 2014). Since we choose to use much less stations to test the SLM technique, namely, because we wanted to test its performance on a coarse station network, we also apply Shepard's technique to these 1380 stations to reproduce in situ the IMD1380 product for a fair validation of the SLM method.

In Shepard's method, the interpolated values at a grid node are computed from a weighted sum of the neighbourhood observations. Following the previous studies (Rajeevan et al. 2006; Pai et al. 2014), we considered a limited number of neighbouring points (minimum 1 and maximum 4) within a search distance (radius of influence) of 1.5° around the grid node where we want to compute the interpolated values.

Consider the grid point P_i, the inverse distance based weighting interpolation method is defined as follows. Let d_i denote the distance from P_i to the nearest rain gauge station. If $d_i = 0$, then the station data is used directly and no interpolation is required, otherwise, the rainfall rate at P_i is given by

$$R_i := f(P_i) = \frac{\sum_s W_i^s Z_s}{\sum_s W_i^s},$$

where the summation is taken over all stations with available data at the given time, Z_s is the observed rainfall rate at station s, and W_i^s is the associated weight which depends on the inverse of the distance, d_i^s, of P_i from the location of Station s modulo a shadowing factor to mitigate over-representation due to many stations from the same direction. In particular, a radius of influence D_x is prescribed and the weights are set by mathematical formulas depending on wether $d_i^s \leq D_x/3$ or $D_x/3 < d_i^s \leq D_x$ and $W_i^s = 0$ is $d_i^s > D_x$. The interested reader is referred to Rajeevan et al. (2006) and Pai et al. (2014) for details.

3. Results

Following the aforementioned previous studies, here also we used a radius of influence of $D_x = 1.5°$ as already mentioned. We termed this product as the IMD1380 station product. Since we used less number of stations (1380 stations) as opposed to 6955 stations used in Pai et al. (2014), a lot of missing values are noted in the final gridded product as opposed to Pai et al. (2014). To provide a fair test for the SLM technique, we decided to push Shepard's method beyond its limits and have uplifted/relaxed the radius of influence restriction and reproduced a full coverage gridded rainfall data for the Indian continent based on the same 1380 stations. We termed this data as the

IMD1380-relaxedR product. The area within the search radius D_x is termed as inside radius of influence (inside Rinf) domain and the area outside the search radius D_x is termed as the outside radius of influence (outside Rinf) domain while the entire area which includes both inside and outside the radius of influence areas is termed as the global domain.

In the following analysis, we compared various statistical metrics of CPCM, IMD and APRODITE gridded datasets. In addition to the traditional root mean square error, and correlation estimates, deviations between the various data products are estimated according to the following equation, which is namely, the accumulated relative error. If R^1 and R^2 represent the rainfall rates corresponding to the data products 1 and 2, respectively, then their difference is estimated by the quantity

$$N_{12} = \sum_x \sum_t \frac{2|R^1(x,t) - R^2(x,t)|}{R^1(x,t) + R^2(x,t)}. \tag{9}$$

Here x is the generic spatial location of all rectangular grid points and t spans over all days of the analysis period from 1951 and 1970. However we will begin in Section 3a. by looking at how well the SLM and Shepard's schemes represent the local rainfall event distributions in comparison to the observed gauge data.

The SLM and the relaxed Shepard's algorithms are run and the interpolated datasets or products, CPCM1380 and IMD1380-relaxedR, respectively, on the $0.25^\circ \times 0.25^\circ$ are constructed for the 20 years period, 1951-1970, using the procedures outlined above. Here, we report the results of the comparative tests of these products, against each other and against the high resolution IMD6955, IMD1380, and the APHRODITE products. Notice that because rainfall is very rare to non-existent during the dry winter months, all the analysis-comparative tests presented below are restricted to the summer months of June-September (JJAS), coinciding with the Indian summer monsoon.

a. Validation tests: Local rainfall distribution skill

First, we assess how well the new SLM and the Shepard technique reproduce the observed local rainfall intensity probability density functions (PDFs). Following Chen et al. (2008), we have selected 8 validation points over the Indian landmass and the daily precipitation from all the stations in a 2° square around each validation point are withdrawn from the dataset. These square correspond to boxed regions shown in Figure 1a. With two boxes (A and B) along the west coast and two along the east coast (G and H) of the southern tip, and four boxes (C, D, E, and F) distributed along the east-west extend of Northern India, the network of validation points spans a variety of physical conditions both in terms of the meteorology and in terms of the rain gauge station density in the corresponding neighbourhoods. The validation point locations are representative of the complexity of the Indian rain gauge dataset in both respects.

The SLM and the Shepard algorithms are performed using the gauge data from the remaining stations to define the precipitation values at the locations of the withdrawn stations. The PDF of precipitation intensity is computed by aggregating the values of precipitation of all withdrawn station locations in each box around each validation point, leading to one localized PDF for each validation point and for each algorithm. The estimated PDFs are compared to the corresponding PDF of the withdrawn station observed precipitation (i.e, instead of using inferred data we now use the actual station data) to assess the accuracy of the two algorithms in reproducing the precipitation intensity distribution at the given locations. The results are summarized in Figure 4 where the bar diagrams corresponding to the two methods and to the station data are compared against each other.

As can be surmised from Figure 4, the PDF estimates are given in terms of rainfall events falling into the 6 bins

$$R < 1,\ 1 \leq R < 6,\ 6 \leq R < 11,\ 11 \leq R < 16,\ 16 \leq R < 21,\ 21 \leq R,$$

where R is the rainfall rate, expressed in mm day^{-1}, averaged over all station locations in each of the boxes in Figure 1a. In general, the PDF of precipitation intensity at each validation point is dominated by weak to no rain events ($R < 1$ mm day^{-1}). However, as can be seen in Figure 4, the frequency of occurrence of such low to no rain events varies strongly between the validation points. In terms of the station data (red bars), it goes from as high as 80% at the North East validation point C to less than 40% at the South East point B located at the northern tip of the Western Ghats mountain range (Figure 1a).

According to Figure 4, except for the two validation points A and B, the no rain events are better represented in the LMS algorithm (yellow bars) compared to Shepard's method (blue bars). These two validation points are located over the windward side of Western Ghats, where we get torrential rain during the monsoon season (Seasonal mean rainfall over these locations is larger than 25 mm day $^{-1}$). Over these two validation points the rainfall intensity is mainly controlled by the orography. The number of stations reporting the precipitation are also large on these locations (Number stations: 26 at validation point A and 25 at validation point B).

At every validation point, the light precipitation events (within the range $1 < R < 16$) are better represented by the SLM method compared to Shepard's method. The moderate and heavy precipitation events ($R > 16$ mm day $^{-1}$) are also well represented by the SLM method except for two validation points (Figure 4e and g). At the validation point E, the SLM method overestimates the moderate and heavy precipitation events compared to the observed station precipitation whereas at validation point G (Figure 4g), the moderate and heavy precipitation events are underestimated by the SML. The validation point E is located within the monsoon trough region where we get heavy rainfall during the passage of monsoon depression/low pressure systems. The number of stations is also very large at this validation point (31 stations). Whereas the point G is located on the eastern coast of India where normally the monsoon depression/low pressure systems first hit on land. However, around this validation point the number of stations reporting precipitation data is comparatively less (17 stations).

As seen in Figure 4, except for aforementioned four occurrences, the SLM method provides a much better representation of the rainfall PDF at these validation points . Shepard's method has the tendency to overestimate the frequency of the light rain events ($1 < R < 16$) and underestimates the moderate to strong rain events ($R > 16$).

To better see this, the agreggated PDF of precipitation intensity at all station locations over the eight 2 $^\circ$square boxes are given in Figure 4i. As expected, the PDF of station precipitation is largely dominated by the no rain events which has frequency of occurrence 58%, while the probability of heavy rainfall ($R \geq 21$) is 12.5%. The Shepard method underestimate the frequency of no-rain events and heavy rain events (blue bars) whereas it overestimates the frequency of occurrence of light precipitation events ($1 \leq R < 16$). The frequency of no-rain events in Shepard's method is 44%; it is 25% less than that of the station precipitation. In all the categories of rain events the SLM method outperforms the Shepard method (yellow bars). The frequency of no-rain event in the SLM method is 57% which is comparable to the station precipitation. Similarly the frequency of occurrence of light rainfall events ($1 \leq R < 16$) and moderate or heavy rainfall ($R \geq 16$) in the SLM method is also comparable to the station precipitation. Figure 4i, may seem to indicate that Shepard's is as good as the SLM in estimating moderate rain events within the range $16 \leq R < 21$ but looking back at the local panels this is clearly due to cancelations of errors some of which is also inevitably true for the SLM results, though to a much lesser extent.

b. Seasonal mean and daily rainfall direct comparisons

Figure 5 compares the JJAS mean 20-year climatology obtained from the CPCM1380 (c) and the IMD1380 (d) gridded rainfall datasets against those corresponding to the two existing rainfall products, namely, the high resolution IMD6955 (b) and APHRODITE (a). The JJAS climatology corresponding to IMD1380-relaxed is shown on panel (e). We note that data from all the 1380 stations is used to produces the CPCM1380, IMD1380, and IMD1380-relaxedR datasets.

Compared to the high resolution product IMD6955, the heavy precipitation over the windward side of Western Ghats and the copious rainfall over Central India are well captured in all the datasets including the new CPCM1830 dataset (Figure 5c). Even with the significantly reduced number of stations, CPCM1380 (Figure 5d) is in good agreement with the high resolution IMD6955 and APHRODITE gridded rainfall products all over the Indian continent while IMD1380 in Figure 5d misses large areas, namely the northern and northeastern tips of India, because of the lack of station coverage. The IMD1380-relaxedR climatology on the other hand shows significant biases especially over those mentioned areas.

The seasonal rainfall averaged over the Indian subcontinent of all the five products are contrasted in Table 4. The seasonal precipitation of APHRODITE is the smallest among all the precipitation products consistent with the maps in Figure 5. Seasonal rainfall of CPCM1380 and IMD6955 are almost identical, whereas that of IMD1830 underestimates the mean rainfall by 10 mm day^{-1} and IMD 1380 RelaxedR overestimates it by about almost 60 mm day^{-1}, compared to IMD6955. This suggest again that gridded rainfall data is method dependent and that the stations density is less important is one is interested only in the climatological regional mean values.

The 20 year averaged JJAS seasonal mean rainfall differences between the IMD6955 dataset and the other rainfall products are presented in Figure 6. Consistent with Table 4, the difference between APHRODITE and IMD6955 datasets shows negative values in most areas of the Indian subcontinent, especially over Central India and the northeast region and a narrow band of positive values along the Western coast (Figure 6a). This implies that APHRODITE estimates less precipitation over most areas of the Indian subcontinent including Central India, the northeast region and the high orographic precipitation along the Western Coast than the IMD6955 dataset. On the other hand, the difference between CPCM1380 and IMD6955 data is positive along Central India and the Western Coast (Figure 6b). However, over the data sparse regions of the northeast and the eastern coast of India, CPCM1380 estimates less precipitation than the IMD6955 dataset (Figure 6b). This may have contributed to error cancelation when we computed the seasonal mean climatology in Table 4.

The difference between the IMD1380 and IMD6955 datasets, on the other hand, does not show a significant difference along the central plains of India and southern peninsula (differences are mostly between 1 mm/day over the country) except Western coast where IMD1380 slightly overestimates the orographic precipitation with respect to the IMD6955 dataset (Figure 6c and d). Thus, once again, not only the number of stations used but also the methodology is important when it comes to OA of rain gauge data. Nonetheless, the errors are within 1 to 2 mm day^{-1} in most places which is within 3 to 6% on average.

When taking into account the fact that the SLM method used to produce the CPCM1830 dataset is based on rainfall binning with bin sizes of 2 mm day^{-1} and larger, according to Table 1, errors in the range of 1 to even 3 mm days^{-1} are expected and are deemed acceptable. As shown in Table 3, decreasing the number of bins decreases, though slowly, the RMSE relative to the high resolution IMD6955 product but unfortunately increasing further the bin number is computationally prohibitive and we refrain from pursuing this at this stage of this research. The goal here is to demonstrate that the SLM OA may offer a reliable method that can be applied in regions of sparse station data, especially when one is interested only in the gross features of the rainfall statistics. Besides of not discriminating grid points that are far away from available data stations, the other attractive feature of this method resides in the fact that it is a stochastic method that in effect incorporates some uncertainly into the interpolated data.

To assess how good the new SLM scheme captures the inter-annual variability of precipitation at each grid point, compared to existing data products, we plot in Figure 7 the standard deviation of seasonal mean rainfall as it varies from year to year, for the five data products. All datasets exhibit large standard deviation over the regions that receive large amounts of precipitation, during each monsoon season. For example, the western coast of the Indian peninsula, the central Indian plains, and Northeast India reveal large standard deviations during the boreal summer monsoon season. In comparison to the IMD6955 high resolution dataset, the overall pattern of standard deviation is fairly well captured in all rainfall products (Figure 7), except for the IMD1380-relaxedR (Figure 7e) which has clear issues in the low station density area in the Northern and Northeastern tips of India. However, in APHRODITE the standard deviation over the central India and northeast India is weak compared to both IMD and CPCM1380 products. While the APHRODITE and IMD1380 datasets (Figure 7a,d) underestimate the highly scattered and high values of standard deviation displayed by the high resolution IMD6955 product over central India, the CPCM1380 product shows a fairly similar pattern as IMD6955 though without some exageration (Figure 7b,c).

Finally, the daily averaged precipitation for a single day (01-July-1960) of the five products are compared against each other in Figure 8. The precipitation for this day is mainly concentrated over the Western Ghats and the eastern coastal regions of India. The precipitation is well organized in these two regions whereas it is more or less scattered over the central Indian plains. All the gridded rainfall products reasonably capture this pattern of precipitation with a maximum of 30-40 mm day^{-1} over the eastern coast and Western Ghats. However, APHRODITE precipitation variability is relatively smooth (Figure 8a) especially over the Eastern coast when compared to IMD and CPCM1380 gridded rainfall products. The high resolution IMD6955 dataset shows higher spatial details than the other gridded datasets. The IMD rainfall produced using the 1380 stations underestimates the 01-July-1960 rainfall over the eastern coast of India (Figure 8d,e). Note that the east coast rain gauge network is relatively sparse in the source (Figure **??**). In spite of this sparse network, the rainfall as by CPCM1380 (Figure 8c) is comparable to that of IMD6955 dataset (Figure 8b), on this particular day.

These direct comparisons show that the SLM method is a reliable interpolation method that can be confidently used, especially when the station data is sparse, both for capturing the global mean rainfall as well as its local distribution and variability, in time and in space.

c. Statistical metrics

We show in Figure 9 the maps of the root mean square error (RMSE) of seasonal mean precipitation at each grid point to measure the differences between the different data products, relative to the reference- high resolution IMD6955 dataset. The RMSE is always large over the data sparse and complex topography regions. In all the cases the maximum uncertainty is over the northeastern region and Western Ghats. Generally, the RMSE is minimum over the low elevation plains such as central India. However, compared to APHRODITE and IMD1380 datasets, the CPCM 1380 dataset shows slightly large RMSE of seasonal mean precipitation with respect to IMD6955 high resolution datasets especially over the northeast region, Western Ghats and low plains of central India Figure 9b. This is expected from the CPCM1830 product because of the combination of the stochasticity of the SLM method and the coarseness of the bin size used to implement it. Nonetheless, the RMSE displayed by the CPCM1380 dataset remains comparable to those displayed by the APHRODITE and the IMD1830 datasets. As expected large errors are associated with the IMD1830-relaxed dataset over the regions of low station data coverage, in the Northern and Northeastern tips of India.

In Table 5, we reported the absolute relative error (N_{12}) between the IMD6955 data and the other precipitation products using the equation in (9), and the RMSE. From Table 5, it is clear that outside the radius of influence the error is larger for the IMD1380-relaxedR dataset than it is for the CPCM1380 product, implying once again that our lattice model method outperforms Shepard method in data sparse regions. Over the entire Indian subcontinent (global) the daily error estimated from Equation 9 is slightly less in CPCM1830 than it is APHRODITE, however, the RMSE of seasonal mean ISMR is larger in CPCM than it is in APHRODITE. It is also true that the absolute relative error between APHRODITE and CPCM1830 is larger than it is between APHRODITE and IMD6955. Note however that the caveat, here, is of course in the fact that we assumed IMD6955 as the truth for convenience while as already stated the OA products are going to always be method dependent.

Figure 10 represents the seasonal correlation between IMD high resolution analysis (IMD 6955) against the rest of the precipitation datasets. All the precipitation products exhibit close agreement with IMD high resolution analysis especially over Central India and Northwest India. In general, correlations higher than 0.9 are observed over the central and northwestern parts of India. Meanwhile all the precipitation products show poor correlations with IMD6955 over areas with a sparse station network (for example, the Northeast, Jammu and Kashmir regions).

d. Interannual daily rainfall variability

The inter-annual variation of all India summer monsoon (JJAS) rainfall (ISMR; precipitation averaged over the Indian subcontinent) is plotted in Figure 11 for the five data products. The ISMR time series of IMD6955, CPCM1380 and IMD1380 datasets nearly match each other in terms of magnitude and phase. However, consistent with the analysis in Figure 6a the magnitude of the ISMR time series derived from the APHRODITE is underestimated in all years compared to both IMD 6955 and CPCM gridded rainfall, which is consistent with the results in Figure 5 and Table 4. However, in most years the ISMR time series derived from APHRODITE are in phase with the time series derived from other rainfall products. On the other hand, the ISMR time series derived from the IMD1380-relaxedR have relatively higher magnitudes compared to the other ISMR time series.

The daily variation of rainfall anomaly averaged over Central India for three monsoon season (1951,1960,1970) are given in Figure 12. In CPCM analysis, the daily variation of Central India rainfall anomalies are in line with other rainfall product. It is clear that the CPCM1830 rainfall product is quite good in capturing the signs of rainfall anomaly over Central India in agreement with the other precipitation products, such as IMD 6955 and APHRODITE. In all the three monsoon seasons, shown here, the easily identifiable active and break phases of the monsoon, associated with the five data products are in good agreement. The correlation between the IMD6955 rainfall time series and other datasets exceed 0.95 in all these three monsoon season.

The corresponding daily variation of rainfall anomalies averaged over the entire Indian subcontinent, shown in Figure 13, display some large differences in the magnitude of rainfall anomalies among different rainfall products. However, all the rainfall products follow a nearly identical daily variation; In most of the days the magnitude of daily rainfall anomalies are slightly larger in the CPCM1380 product compared to the IMD6955 product however the APHRODITE time series shows a much smoother variability and underestimates the magnitudes at times.

4. Conclusion

Rain gauge datasets are often used to compensate and validate satellite precipitation data which in turn is used for climatological and hydrological studies and to validate earth system models. However, they are also important in their own right as they constitute accurate and reliable data sources for local studies and long term weather and climate projections, especially, over land areas where they are routinely recorded by climate and weather centers around the world (Xie and Arkin 1996). Satellite observations appeared only during the last decades while rain gauge data collection dates back to the late century. However, the measurement stations are unevenly distributed across the continents and in time and many areas are only sparsely covered if at all. Several OA techniques have been devised and used to convert (interpolate) these unevenly distributed rainfall data into a regular grid to ease their usage for theoretical, forecasting, and modelling purposes alike. Unfortunately, all the existing OA techniques have limitations in areas with sparse gauge station coverage and the gridded data is method dependent over such areas (Xie and Arkin 1996).

We proposed a new stochastic OA method for rain gauge data based on the theory of stochastic particle interacting systems on a lattice (Liggett 1999; Khouider 2014), here abbreviated SLM for stochastic lattice model. The SLM thechnique is applied to the Indian Meteorological Department rain gauge dataset which started since 1901. While the Indian station network totals 6955 stations, we advertently used a selection of 1830 stations dispersed unevenly over the Indian subcontinent to implement and test the SLM technique.

Existing studies (Bussieres and Hogg 1989; Chen et al. 2008) found that the statistical optimal interpolation (SOI) method of Gandin (1965) is superior to the so-called empirical or function methods that aim to approximate the rainfall at a given grid point using a weighted average of the neighbouring stations. Arguably, it is because the SOI method minimizes at once the expected error at the existing stations and as such its uses global information as well as local information. However, this method is also restricted to a radius of influence region from the station network and according to the results shown in both Bussieres and Hogg (1989) and Chen et al. (2008), the SOI results are very closely flowed by those obtained by the inverse distance weighted method of Shepard (1968).

The existing IMD 6955 station data has been recently quality controlled and gridded using Shepard's technique (Rajeevan et al. 2006; Pai et al. 2014). We thus also run Shepard's algorithm on the same 1830 stations and assessed the new SLM scheme (CPCM1380 product) against Shepard's scheme (IMD1380) in the light of two existing high resolution data products over the Indian subcontinent, namely the IMD6955 and APHRODITE (Yatagai et al. 2012). To have a fair comparison, we decided to lift the radius of influence restriction on Shepard's method to produce a data product that equally covers all of India (IMD1380-relaxedR).

In a nutshell, the SLM method attempts to sample the Gibbs grand canonical measure of a large lattice particle interacting system, as in statistical mechanics (Thompson 1972), when the particles are actually rainfall bins at the corresponding grid points forming the lattice, conditional to the existing station data at the local station sites and the associated global climatology. In this sense the SLM method has this "globality" feature in common with the SOI method of Gandin (1965).

After selecting a default set of parameters that minimizes the RMSE of the 1380 station interpolated rainfall data, with respect to the high resolution IMD6955 data product, as Chen et al. (2008) did, we first compared in Figure 4 the rainfall event PDFs obtained by the SLM and Shepard's methods at select, widely separated, areas of the Indian land mass, consisting of $2^{\circ} \times 2^{\circ}$square boxes within each all existing station data has been removed and corresponding rainfall values are inferred from the remaining stations. The associated PDFs are compared to the pre-existing station data within each one of the boxes and in terms of the aggregated data from all the boxes. This test revealed that the SLM method is superior to Shepard's method in terms of the rainfall event PDF accuracy. Shepard's method tends to underestimate the no and very light rain events of less than 1 mm day^{-1}, underestimate the high rain events, greater or equal to 21 mm day^{-1}, and overestimate light to moderate rain events between 2 and 21 mm day^{-1}.

The mean seasonal climatologies of the five datasets are compared in Figures 5 and Table 4 while the associated mean biases, with respect to the IMD6955 dataset of the other four products, are reported in Figure 6. Except for the IMD1380-relatedR, which appears to be at odds with the rest in the low station density areas, these results indicated that the five products are more or less consistent with each other in many respects. However, APHRODITE seems to underestimate everywhere the seasonal rainfall associated with the IMD6955 whereas CPCM1380 appears to overestimate it in Central India and on the shadows of the Western Ghats and underestimate it in Northern India and the east coastal regions. Remarkably, the bias errors are within the controlled bin size and the globally accumulated mean monsoon rainfall of CPCM1380 and IMD6955 nearly match (Table 4) while the other products showed significant discrepancies though it is very small (10 mm) for IMD1380.

In terms of interannual variability, CPCM1380 seems to be the only product to capture the high standard deviation of IMD6955 over Central India, though it seems to exaggerate it in the low station density regions (Figure 7). CPCM1380 also appears to be the one to better capture the high rainfall events over the Western Ghats and near the east coast of Central India happening on the typical monsoon day on 1-Jul-1960.

The RMSE and ARE of IMD6955 with respect to the other four products were also considered (Figure 9 and Table 5). Again all the products seem to agree with each other in the bulk part except for IMD1380 that misses large areas and IMD1380-relaxedR which is faulty in those areas. It is interesting to note that the smallest errors are associated with IMD6955 v.s IMD1380 inside the radius of influence while both globally and outside the radius of influence IMD1380 and CPCM1380 exhibit comparable errors. The same is true for the global errors of APHRODITE both with respect to IMD6955 and CPCM1380. Moreover, all the product showed strong correlation with respect to the IMD6955 product but the places of low station coverage (Figure 10).

The Interannual and daily spatial means in Figures 11-13 are also consistent between all products both in terms of phasing and amplitudes although APHRODITE shows a systematic underestimation of the interannual rainfall while IMD1380-relaxedR overestimates it. Also, APHRODITE appears to be the smoothest in terms of daily precipitation consistent with the observed low standard deviation in Figure 7.

As demonstrated by the sensitivity tests in Table 3, besides the demonstrated acceptable accuracy of the CPCM1380 dataset, generated globally all over India, including low station density regions, there is promise that the accuracy can be improved specifically by increasing the number of bins N. However, the sweet spot in the underlying parameters specifically J_0 may not be the same as for the bin size $N = 137$, thus some retuning maybe required if the bin size has to be increased. Importantly, given the stochastic nature of the SLM algorithm one can easily infer and assign some degree of uncertainly to each interpolated value by simply estimating the standard deviation for each Markov chain of the MCMC runs. Consistently this uncertainty appears to decrease with the bin size. However, this remains to be thoroughly tested in order to understand the true meaning of the this uncertainty in comparison to available station data. This will be the subject of a future study.

Given the success of the SLM method on a such reduced number of stations, it is natural to expect that a dataset produced by this method using all of the existing 6955 stations will be a better product than the existing IMD6955 product. The same method can be applied to other regions of the world with a contiguous climate.

Acknowledgements

The research of B.K. is supported partly by a discovery grant from the Natural Sciences and Engineering Research Council of Canada. The Center for prototype Climate Modeling is fully supported by the Abu Dhabi Government through New York University Abu Dhabi Research Institute grant. This research is supported by the Monsoon Mission project of the Earth System Science Organization, Ministry of Earth Sciences (MoES), Government of India (Grant No.MM/SERP/NYU/2014/SSC-01/002). This research was initiated during a visit of BK to NYUAD during spring 2017.

References

Adler, R. F., and Coauthors, 2003: The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *Journal of hydrometeorology*, **4 (6)**, 1147–1167.

Barnes, S. L., 1973: *Mesoscale objective map analysis using weighted timeseries observations*. [NTIS COM7310781], NOAA Tech. Memo. ERL NSSL62, National Severe Storms Laboratory, Norman, OK 73069, 60 pp.

Bussieres, N., and W. Hogg, 1989: The objective analysis of daily rainfall by distance weighting schemes on a mesoscale grid. *Atmosphere-Ocean*, **27 (3)**, 521–541.

Chen, M., W. Shi, P. Xie, V. B. Silva, V. E. Kousky, R. Wayne Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *Journal of Geophysical Research: Atmospheres*, **113 (D4)**.

Chen, M., P. Xie, J. E. Janowiak, and P. A. Arkin, 2002: Global land precipitation: A 50-yr monthly analysis based on gauge observations. *Journal of Hydrometeorology*, **3 (3)**, 249–266.

Cressman, G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev*, **87 (10)**, 367–374.

Gandin, L. S., 1965: Objective analysis of meteorological fields, translated from Russian by R. Hardin. *Israel Program for Sci. Transl, Jerusalem*, 242pp.

Gruber, A., X. Su, M. Kanamitsu, and J. Schemm, 2000: The comparison of two merged rain gauge–satellite precipitation datasets. *Bulletin of the American Meteorological Society*, **81 (11)**, 2631–2644.

Hartmann, D. L., and M. L. Michelsen, 1989: Intraseasonal periodicities in indian rainfall. *Journal of the Atmospheric Sciences*, **46 (18)**, 2838–2862.

Huffman, G. J., and Coauthors, 1997: The global precipitation climatology project (GPCP) combined precipitation dataset. *Bulletin of the American Meteorological Society*, **78 (1)**, 5–20.

Khouider, B., 2014: A coarse grained stochastic multi-type particle interacting model for tropical convection: nearest neighbour interactions. *Comm. Math. Sci., in press*.

Khouider, B., J. Biello, and A. J. Majda, 2010: A stochastic multicloud model for tropical convection. *Commun. Math. Sci.*, **8 (1)**, 187–216.

Krishnamurthy, V., and J. Shukla, 2000: Intraseasonal and interannual variability of rainfall over india. *Journal of Climate*, **13 (24)**, 4366–4377.

Krishnamurthy, V., and J. Shukla, 2007: Intraseasonal and seasonally persisting patterns of indian monsoon rainfall. *Journal of climate*, **20 (1)**, 3–20.

Krishnamurthy, V., and J. Shukla, 2008: Seasonal persistence and propagation of intraseasonal patterns over the indian monsoon region. *Climate Dynamics*, **30 (4)**, 353–369.

Liggett, T., 1999: *Stochastic interacting systems: contact, voter and exclusion processes*. Grundlehren der Mathematischen Wissenschaften, vol. 324, Springer, Berlin.

Pai, D., L. Sridhar, M. Rajeevan, O. Sreejith, N. Satbhai, and B. Mukhopadhyay, 2014: Development of a new high spatial resolution (0.25× 0.25) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *Mausam*, 1–18.

Parthasarathy, B., and D. Mooley, 1978: Some features of a long homogeneous series of indian summer monsoon rainfall. *Monthly weather review*, **106 (6)**, 771–781.

Rajeevan, M., and J. Bhate, 2008: *A high resolution daily gridded rainfall data set (1971-2005) for mesoscale meteorological studies*. Pune, India Meteorological Department.

Rajeevan, M., and J. Bhate, 2009: A high resolution daily gridded rainfall dataset (1971–2005) for mesoscale meteorological studies. *Current Science*, 558–562.

Rajeevan, M., J. Bhate, and A. K. Jaswal, 2008: Analysis of variability and trends of extreme rainfall events over india using 104 years of gridded daily rainfall data. *Geophysical Research Letters*, **35 (18)**.

Rajeevan, M., J. Bhate, J. D. Kale, and B. Lal, 2005: *Development of a high resolution daily gridded rainfall data for the Indian region*. Pune, India Meteorological Department.

Rajeevan, M., J. Bhate, J. D. Kale, and B. Lal, 2006: High resolution daily gridded rainfall data for the indian region: Analysis of break and active. *Current Science*, **91 (3)**, 296–306.

Sabeerali, C. T., R. S. Ajayamohan, D. Giannakis, and A. J. Majda, 2017: Extraction and prediction of indices for monsoon intraseasonal oscillations: an approach based on nonlinear laplacian spectral analysis. *Climate Dynamics*, **49 (9-10)**, 3031–3050.

Shepard, D., 1968: A two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the 1968 23rd ACM national conference*, ACM, 517–524.

Thompson, C., 1972: *Mathematical Statistical Mechanics*. Princeton Univ. Press, Princeton, NJ, USA.

Walker, S. G. T., 1910: *On the meteorological evidence for supposed changes of climate in India*, Vol. 21. Indian Meteor. MemoGeneral Government Branch Press, 1-21 pp.

Willmott, C. J., C. M. Rowe, and W. D. Philpot, 1985: Small-scale climate maps: A sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. *The American Cartographer*, **12 (1)**, 5–16.

Xie, P., and P. A. Arkin, 1995: An intercomparison of gauge observations and satellite estimates of monthly precipitation. *Journal of Applied Meteorology*, **34 (5)**, 1143–1160.

Xie, P., and P. A. Arkin, 1996: Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *Journal of climate*, **9 (4)**, 840–858.

Xie, P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bulletin of the American Meteorological Society*, **78 (11)**, 2539–2558.

Xie, P., B. Rudolf, U. Schneider, and P. A. Arkin, 1996: Gauge-based monthly analysis of global land precipitation from 1971 to 1994. *Journal of Geophysical Research: Atmospheres*, **101 (D14)**, 19 023–19 034.

Yatagai, A., K. Kamiguchi, O. Arakawa, A. Hamada, N. Yasutomi, and A. Kitoh, 2012: APHRODITE: Constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges. *Bulletin of the American Meteorological Society*, **93 (9)**, 1401–1415.

Yatagai, A., P. Xie, and A. Kitoh, 2005: Utilization of a new gauge-based daily precipitation dataset over monsoon asia for validation of the daily precipitation climatology simulated by the mri/jma 20-km-mesh agcm. *Sola*, **1**, 193–196.

LIST OF TABLES

TABLE 1. Example of a bin configuration corresponding to the case $N = 137$ bins adapted as the default in this study. The configurations associated with all the binning cases considered can be surmised from the broken blue curves on each panel in Figure 3.

Rainfall (mm/day)	Bin Size (mm/day)	Number of Bins
< 1	1	1
1-100	2	50
100-450	5	70
450-550	10	10
550-800	50	5
> 800	∞	1
Total		137

TABLE 2. Parameters values of the SLM interpolation scheme.

Parameter	Description	Value
α	Sets strength of transition rate to station data cell	4.0
τ	Transition time scale	5 hours
J_0	Strength of local interaction potential	1.05
N	Number of bins	137
M	Number of lattice cells	11921
T_0	Pseudo iteration time	24 hours

TABLE 3. RMSE between the CPCM1380 and IMD6955 products for different J_0 values (left column) and bin number, N, (top row) based on data from the 1951 JJAS season.

—-Bin-Number, N, J_0 (day mm^{-1})—-	**137**	112	107	74	62	51
0.8	1.27	-	-	-	-	-
0.9	1.16	1.19	1.21	1.23	1.32	-
0.95	-	1.16	1.17	1.19	1.30	-
1.0	1.10	1.14	1.15	1.16	1.29	1.47
1.05	**1.09**	1.11	1.12	1.15	1.28	-
1.1	1.09	1.12	1.13	1.13	1.27	-
1.2	1.11	-	-	-	-	-
1.4	1.23	-	-	-	-	-
1.5	1.28	1.26	1.22	1.13	1.24	1.44
2.0	1.60	1.55	1.39	1.13	1.24	1.44
2.2	-	-	-	-	-	1.43
2.4	-	-	-	1.14	1.24	1.44
2.5	-	-	-	-	1.25	-
2.6	-	-	-	-	-	1.43

TABLE 4. Seasonal Mean Rainfall in different rainfall products (Unit:mm).

Rainfall product	Seasonal Mean (mm)
IMD 6955 stations	864
APHRODITE	756
CPCM 1380 stations	863
IMD 1380 stations	854
IMD 1380 (RelaxedR)	920

TABLE 5. Absolute relative error (9) and RMSE of seasonal mean Indian summer monsoon rainfall between various data products, as indicated.

Rainfall products	Error estimated from eqn (1)	RMSE of Seasonal Mean rainfall (Unit:mm/day)
IMD 6955 vs IMD 1380 stations relaxedR (Global)	0.76	2.25
IMD 6955 vs IMD 1380 stations relaxedR (inside Rinf)	0.69	1.60
IMD 6955 vs IMD 1380 stations relaxedR (Outside Rinf)	1.14	6.30
IMD 6955 vs CPCM 1380 stations (Global)	0.87	2.77
IMD 6955 vs CPCM 1380 stations (inside Rinf)	0.85	2.33
IMD 6955 vs CPCM 1380 stations (outside Rinf)	0.99	5.96
IMD 6955 stations vs APHRODITE (Global)	0.88	2.03
APHRODITE vs CPCM 1380 stations (Global)	1.02	2.58

LIST OF FIGURES

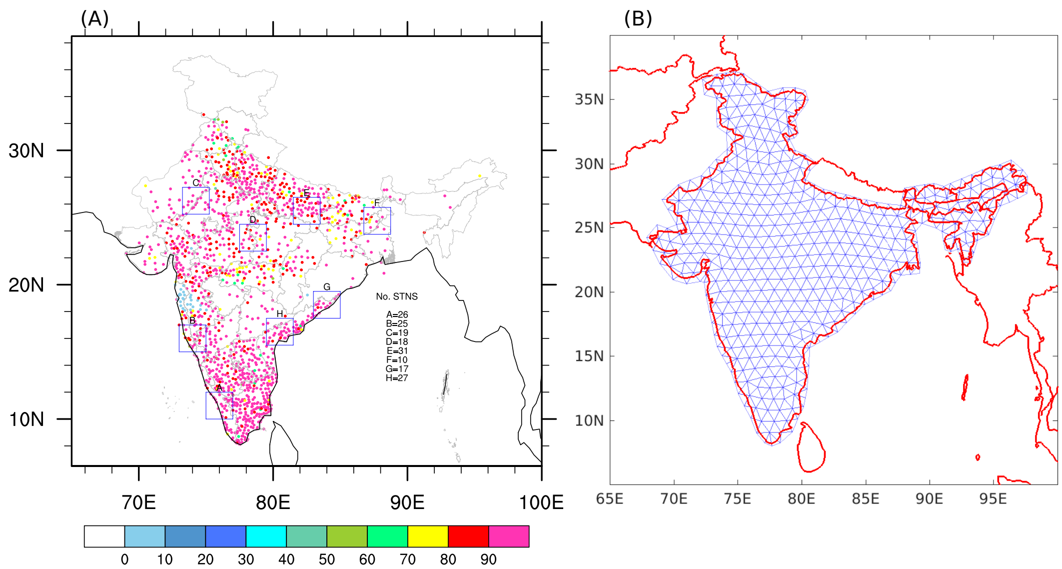


FIG. 1. A: Location of the 1380 rain gauge stations used by the SLM interpolation scheme to produce the CPCM1380 data set and by Shepard's scheme to produce the IMD1380 and IMD1380-relaxedR datasets. Colours indicate percentage of days with rainfall data. Eight validation points, labeled A-H, are marked by the blue squares each representing a two degree square box surrounding the corresponding validation point and the associated number of gauge stations within each box, that are withdrawn when performing the validation tests in Section 3a, are listed at the bottom right part of the panel. B: triangular lattice on which the SLM takes discrete values with $M = 802$ triangles, yielding roughly a 1° resolution. Notice that in the actual application, we used $M = 11921$ triangles.

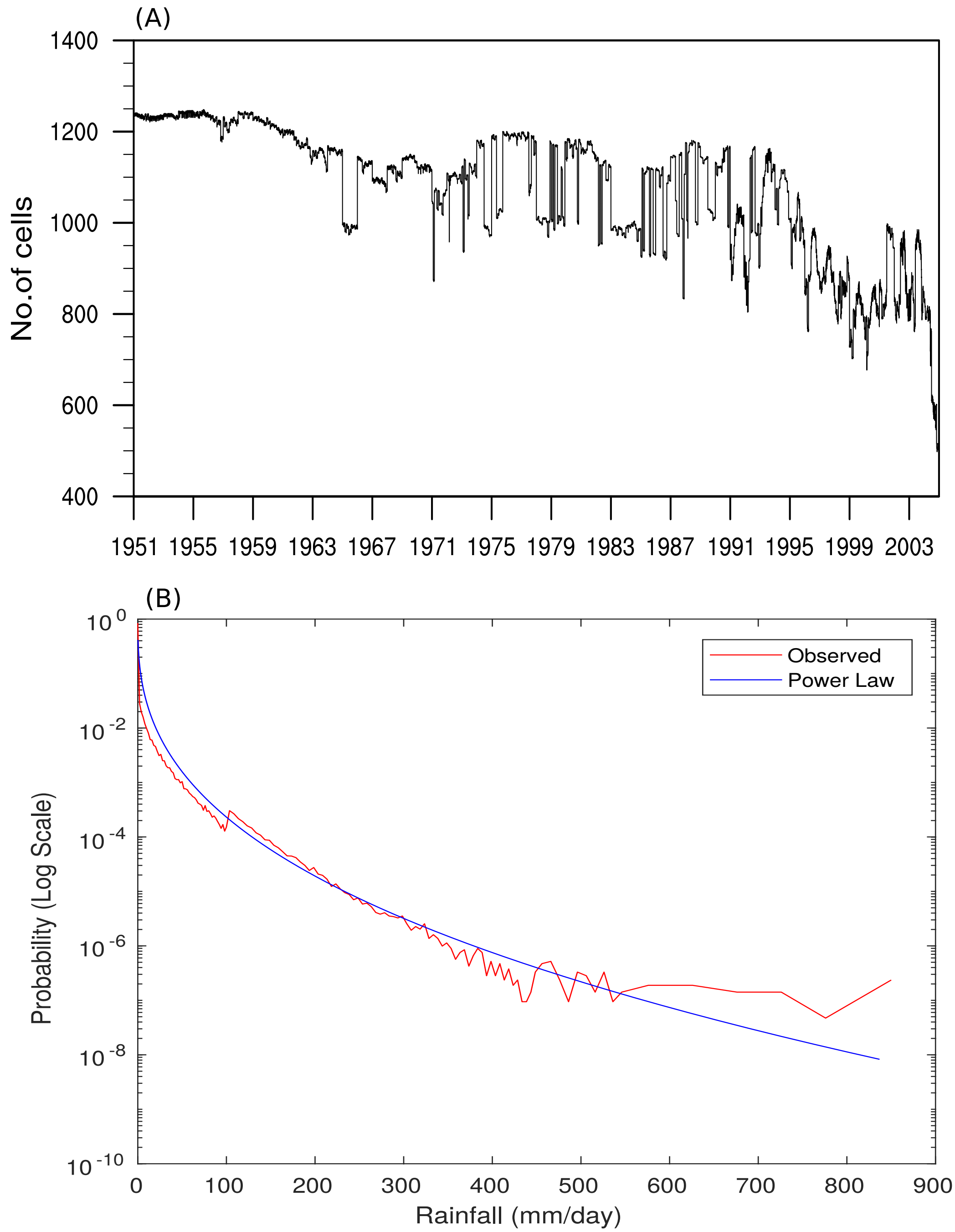


FIG. 2. A: Number of triangular cells per day containing rain gauge rainfall data. B: Probability of occurrence of rainfall in each range of rain intensity in the rain gauge dataset produced by the 1380 stations from 1951 to 1970.

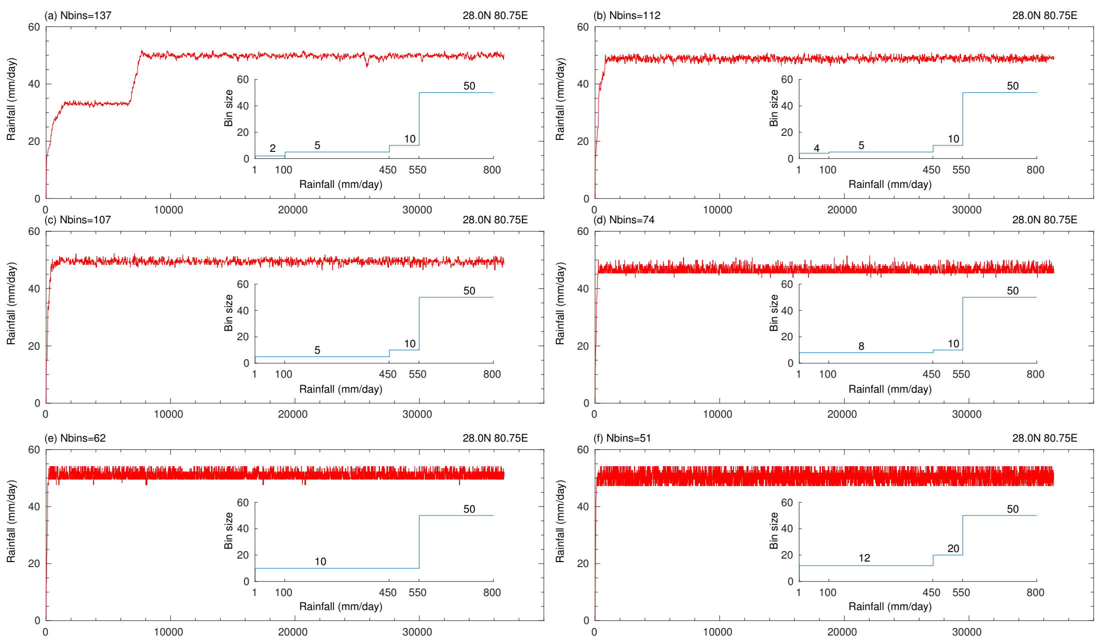


FIG. 3. Convergence at lat-lon point 28N 80.75E for the day 19-Jul-1951. y-axis represents rainfall in mm day^{-1} and x-axis is the iteration count of the MCMC simulation over the pseudo-time, from 0 to $T_0 = 24$ hours. The broken blue curves in the middle of each panel represent the bin configurations for each corresponding bin number case. Each horizontal segment of the broken curve represents an interval of rainfall rates that is uniformly divided into n bins where n is the number indicated right on top of that segment. For the case of Bin Number $N = 51$ in panel (f), for example, the rainfall rate segment between 0 and 450 mm day^{-1} is divided into 12 bins of the same size.

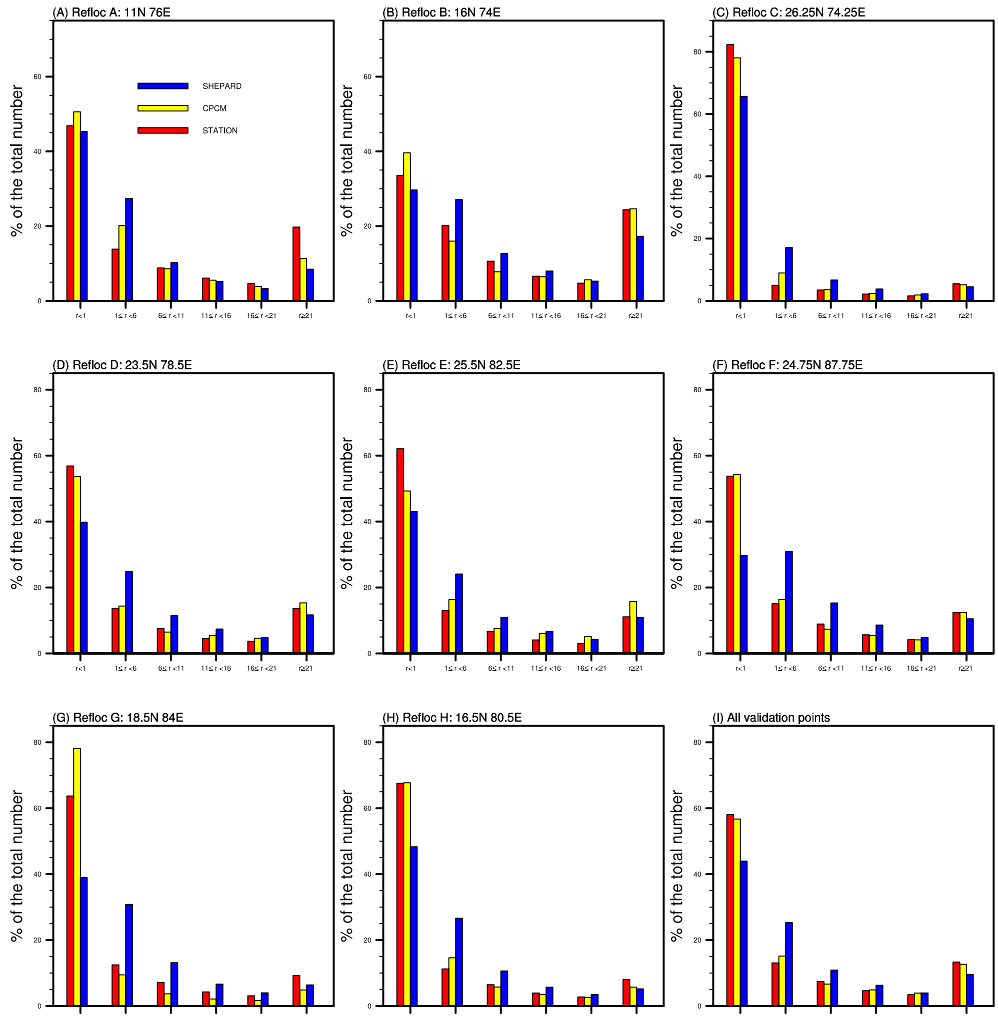
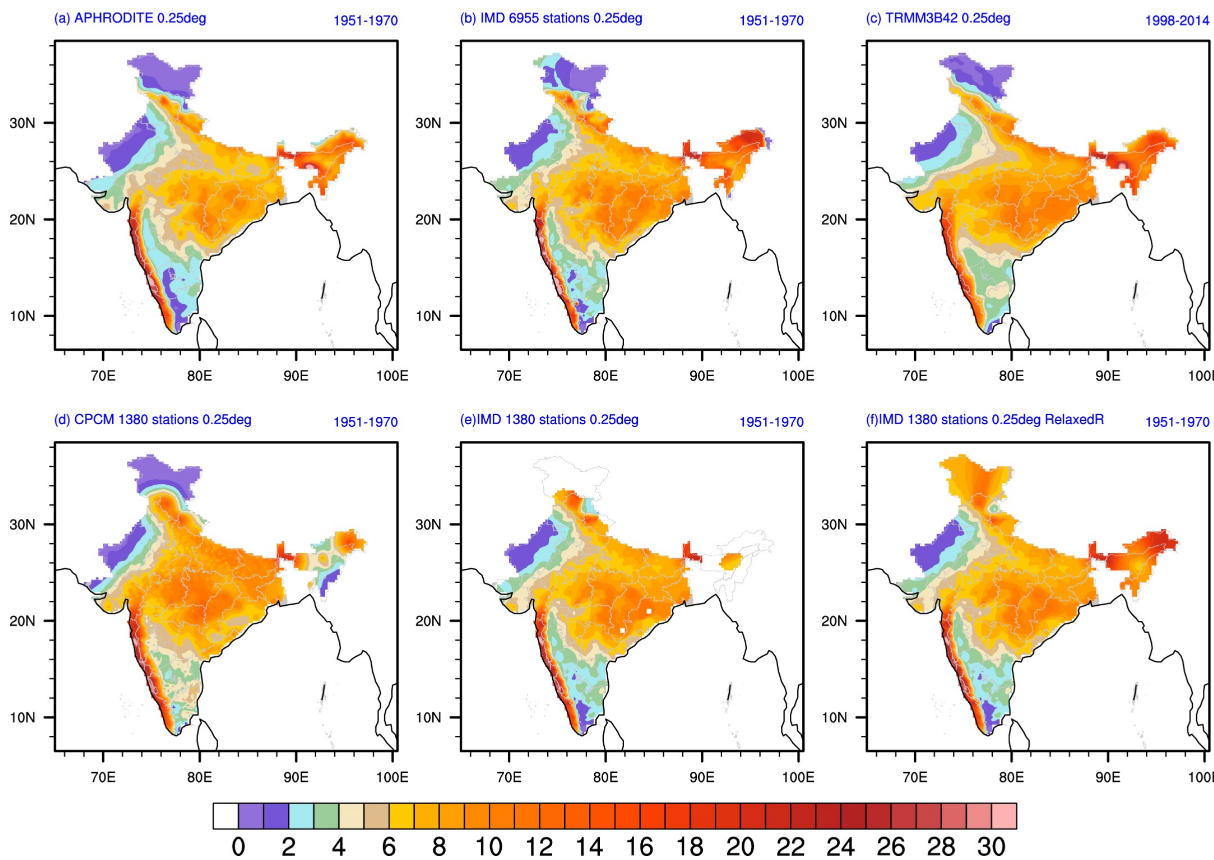
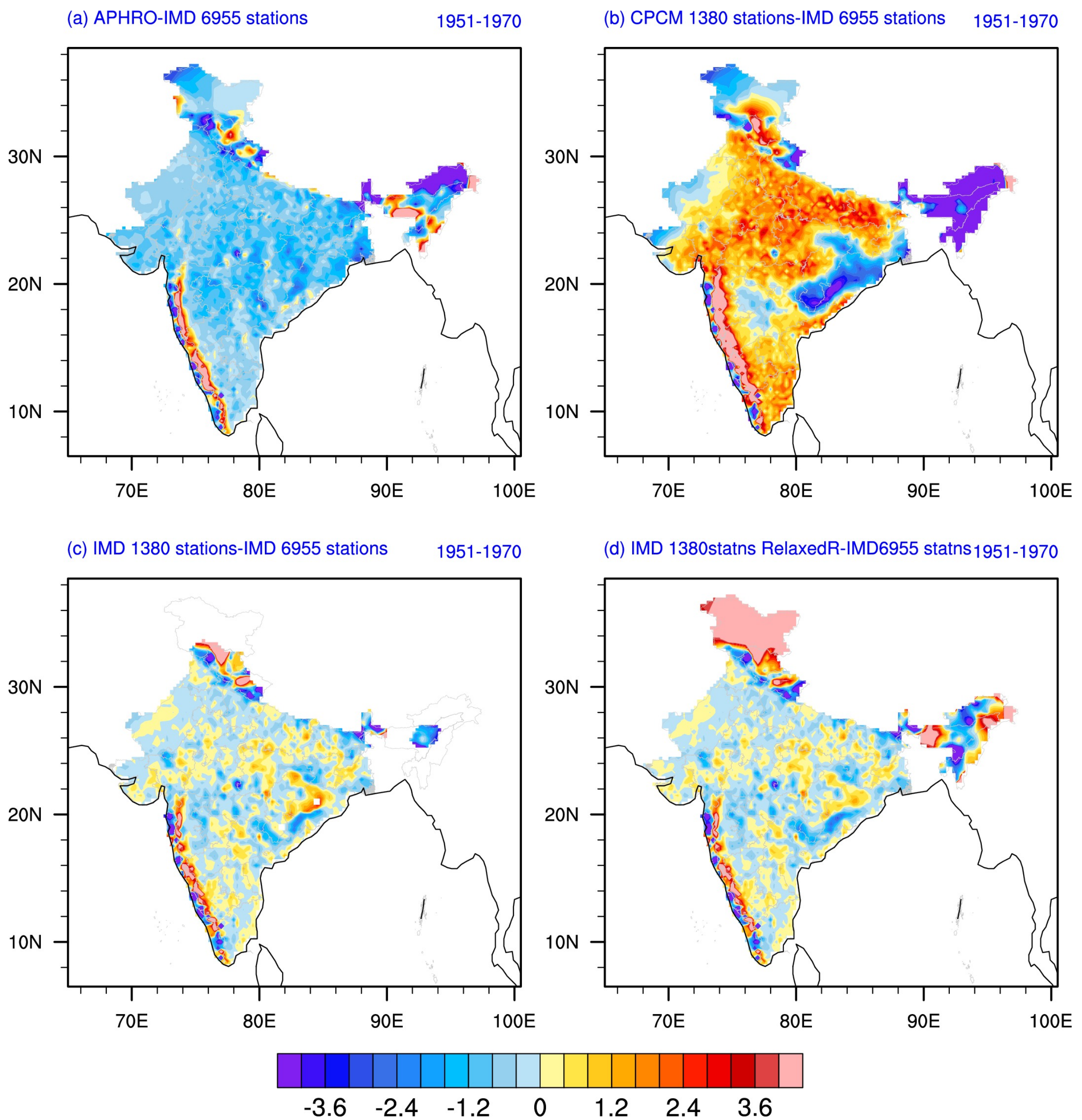


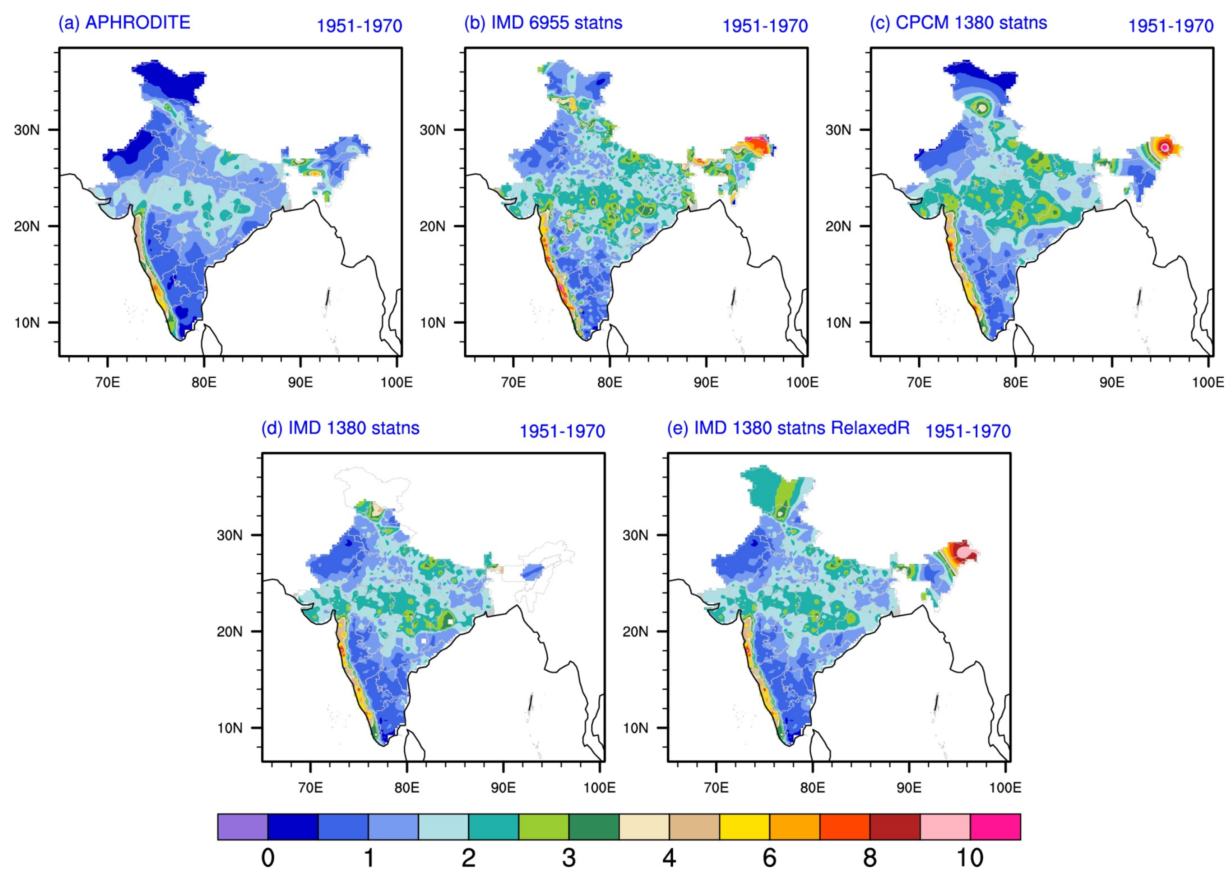
FIG. 4. Probability density function (PDF; units %) of box averaged daily precipitation corresponding to the SLM (yellow bars) and Shepard (blue bars) interpolation techniques over the eight validation point boxes indicated on the corresponding panels (a-h) as well as the PDF of the aggregated rainfall from all stations contained in all validation point boxes (i). See text for details. x axis indicates midpoint rainfall (mm day^{-1}) in each bin. The red bars represent the PDFs of the rain gauge data.



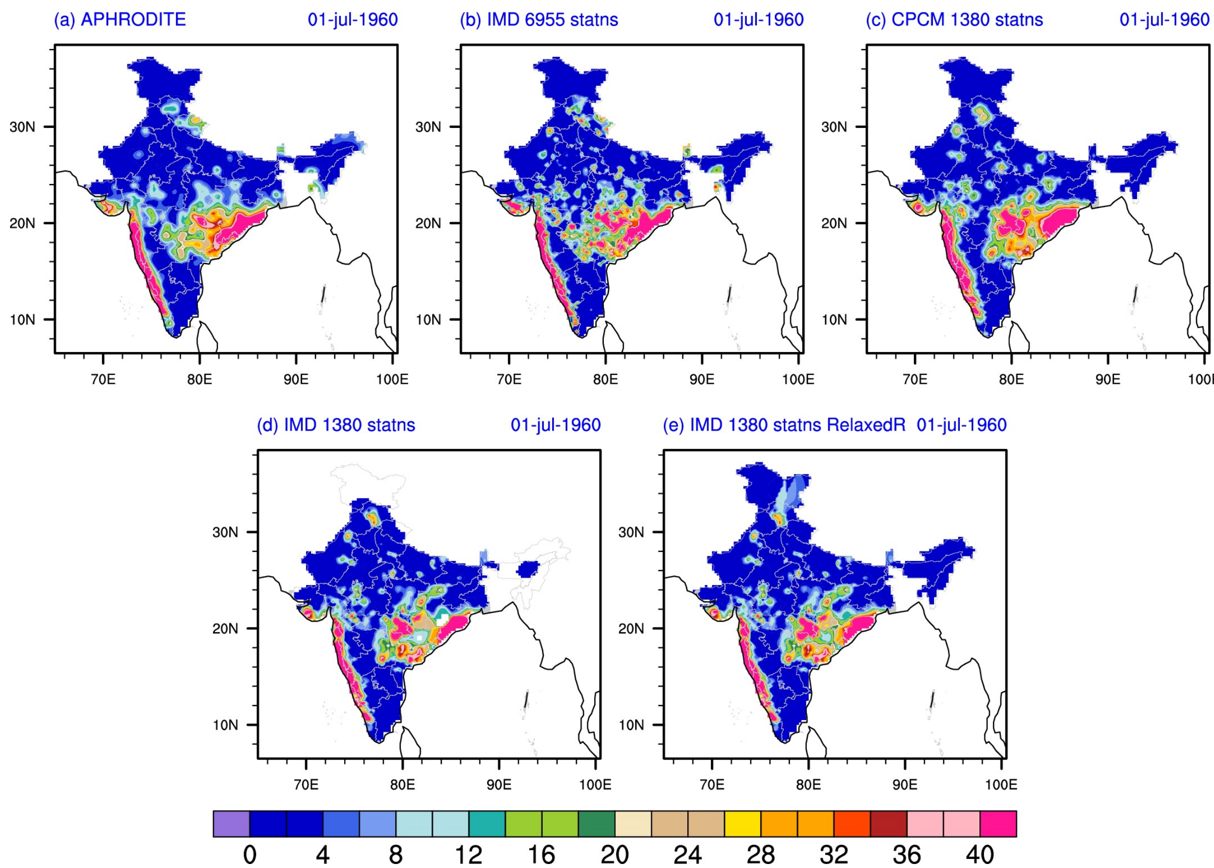
934 FIG. 5. JJAS rainfall climatology of the Indian subcontinent for the period 1951-1970 obtained from the five
935 datasets. (a) APHRODITE, (b) IMD6955, (c) CPCM1380, (d) IMD1380, and (f) IMD1380-relaxedR. See text
936 for details. Units are in mm day^{-1}.



937 FIG. 6. JJAS mean rainfall difference between (a) APHRODITE minus IMD6955 (b) CPCM minus IMD6955
938 (c) IMD1380 minus IMD6955 (d) IMD1380-relaxedR minus IMD 6955. Units are in mm day^{-1}. Differences
939 are between the JJAS mean rainfall averaged over all seasons from 1951 to 1970.



940 FIG. 7. Standard deviation of JJAS mean rainfall (interannual variability) in (a) APHRODITE, (b) IMD6955, (c) CPCM1380, (d) IMD1380, and (e) IMD1380-relaxedR data products. Units mm/day. Standard deviation is for JJAS mean rainfall of all seasons from 1951 to 1970.



943 FIG. 8. Daily rainfall in the five different gridded products for the typically monsoon day of 01-July-1960: (a)
944 APHRODITE, (b) IMD6955, (c) CPCM1380, (d) IMD1380, (e) IMD1380-relaxedR. Units are in mm day^{-1}.

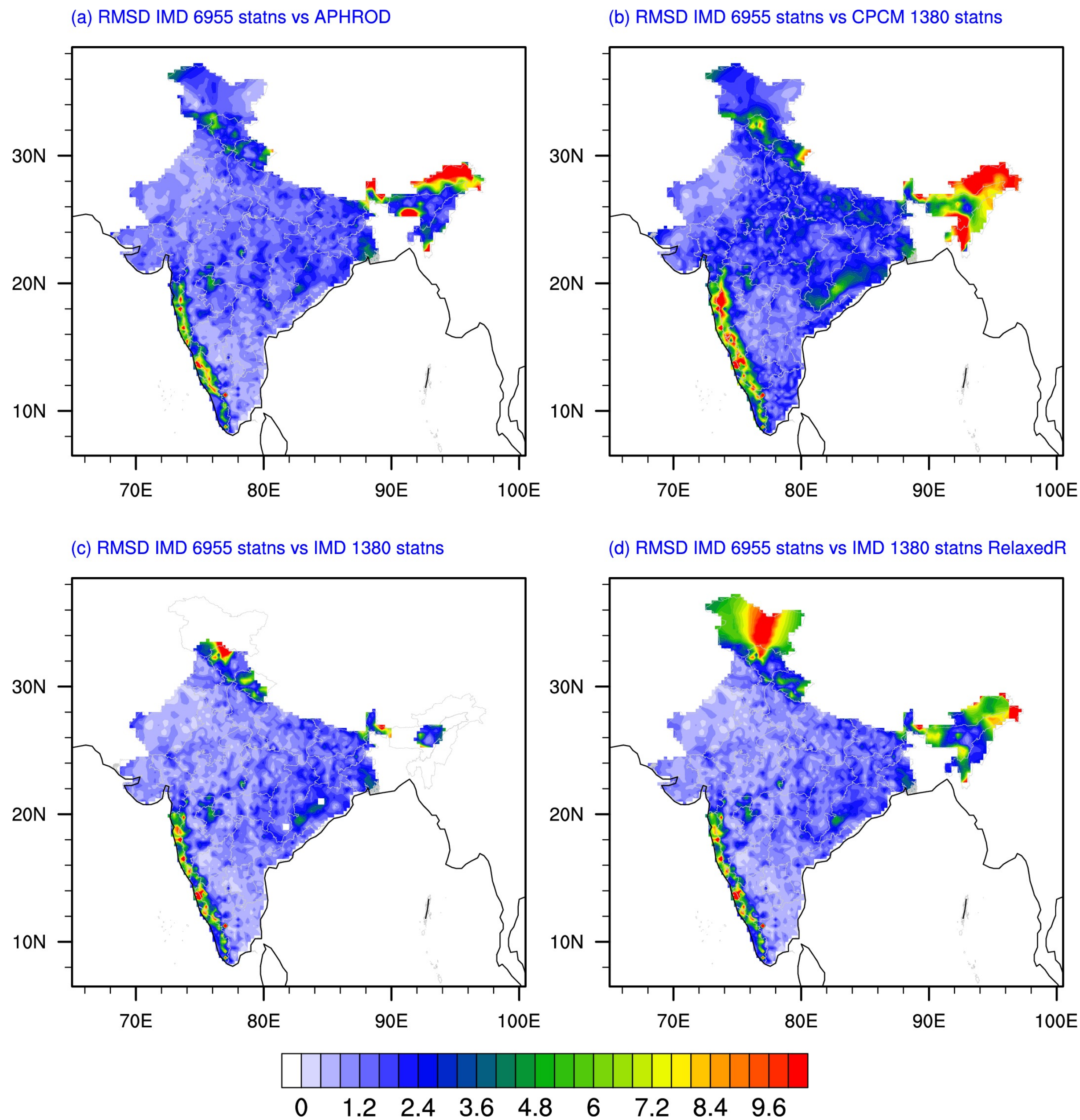
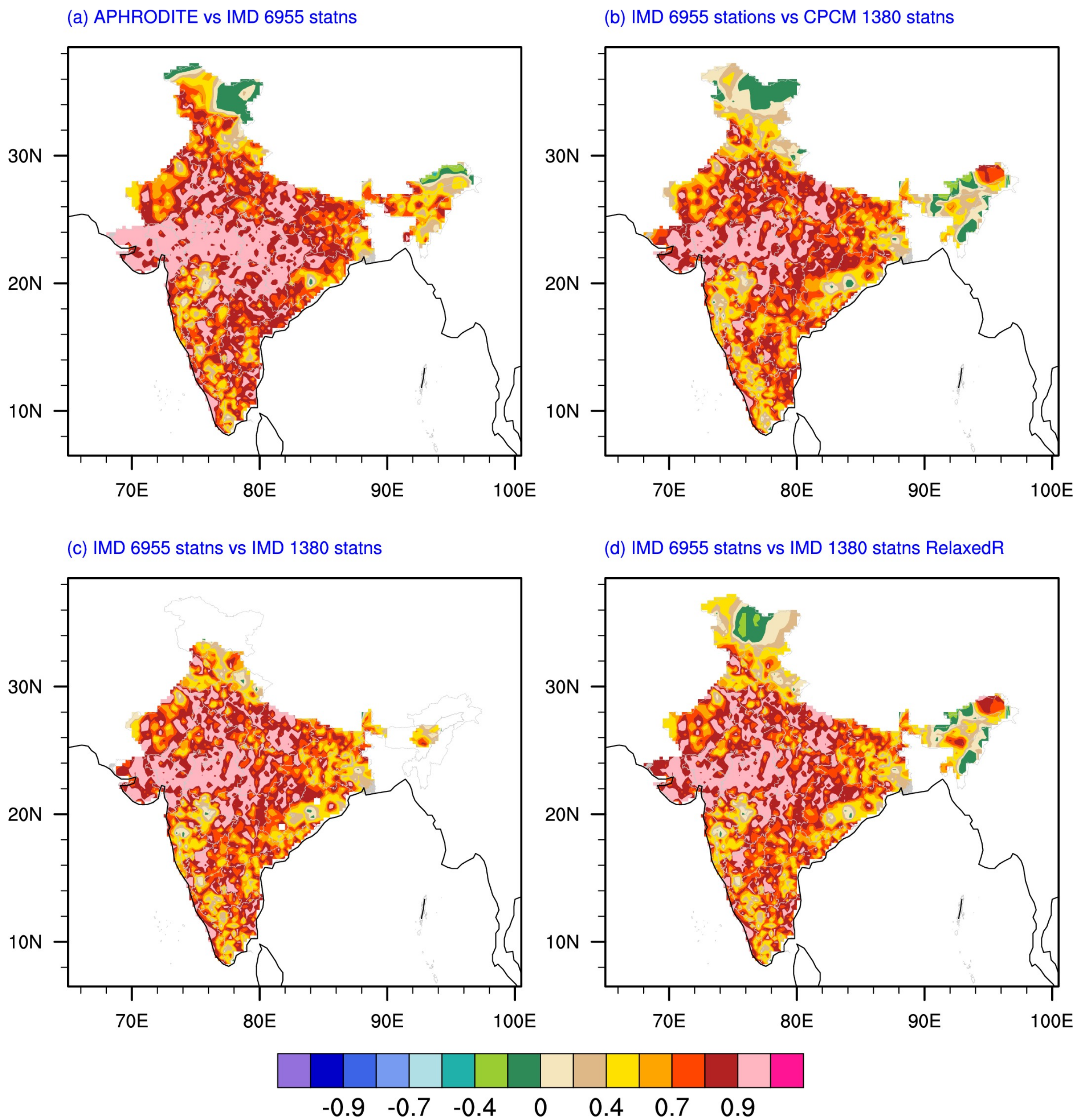


FIG. 9. RMSE between (a) IMD6955 and APHRODITE, (b) IMD6955 and CPCM1380, (c) IMD6955 and IMD1380, (d) IMD6955 and IMD1380-relaxedR. Units are mm day^{-1}.



947 FIG. 10. (a) Grid point correlation of JJAS mean rainfall between IMD6955 and (a) APHRODITE, (b)
948 CPCM1380 (c) IMD1380, and (d) IMD1380-relaxedR for the period 1951-1970.

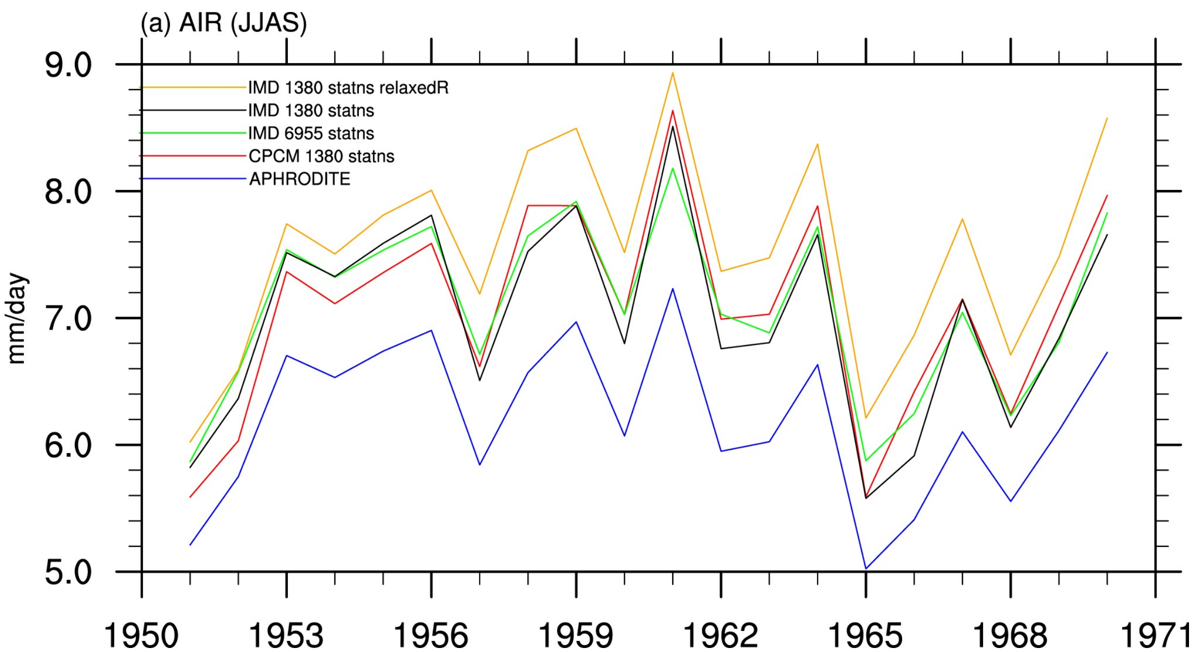


FIG. 11. (a) Interannual variation of all India summer monsoon rainfall (averaged over Indian landmass and averaged over JJAS season): IMD 6955 (green), APHRODITE (blue), CPCM1380 (red), IMD1380 (black), and IMD1380-relaxedR (orange). Units mm day^{-1}.

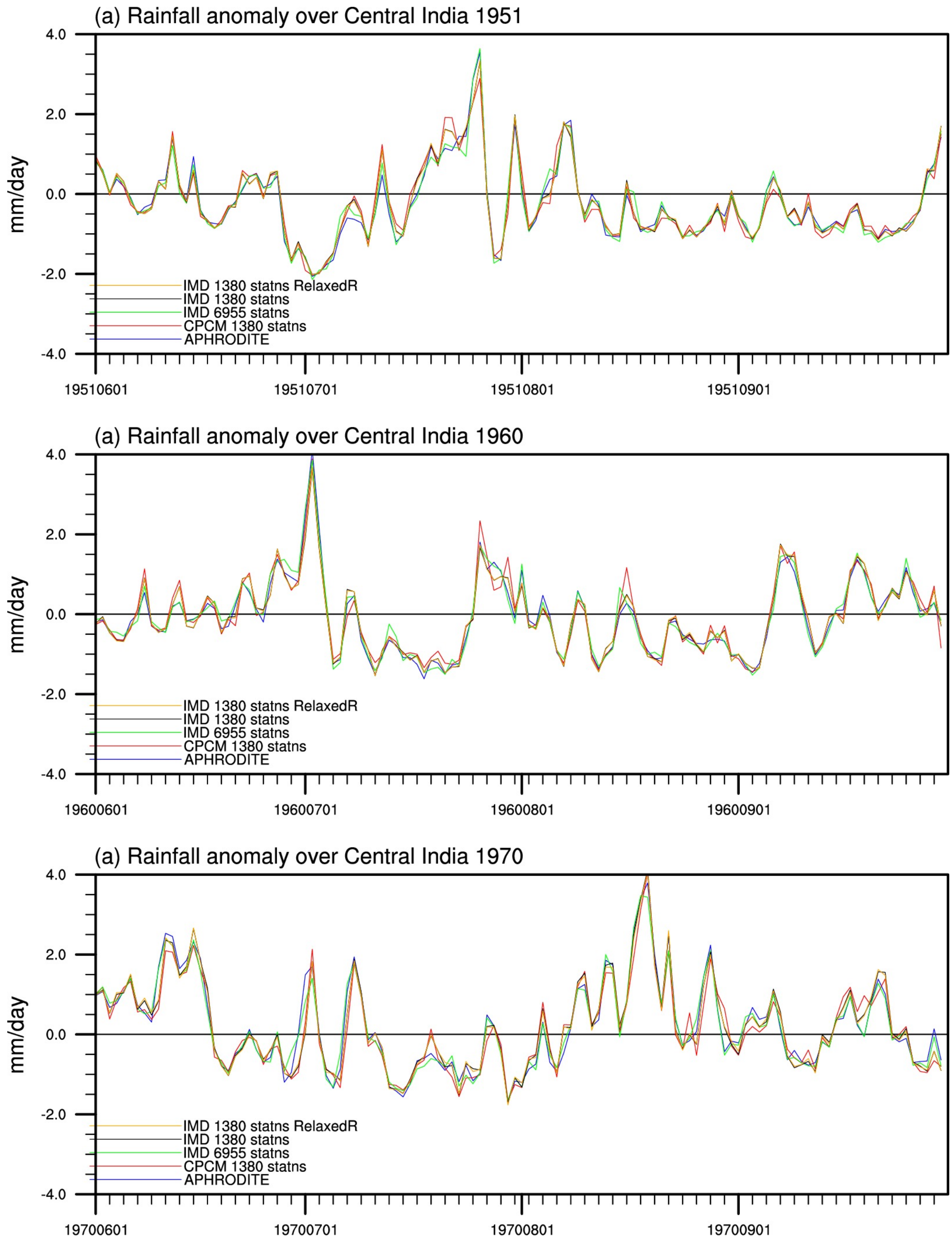


FIG. 12. Daily variation of rainfall anomaly over Central India (coordinates of averaging region) for the (a) 1951, (b) 1960, and (c) 1970 JJAS seasons. Units: mm day^{-1}

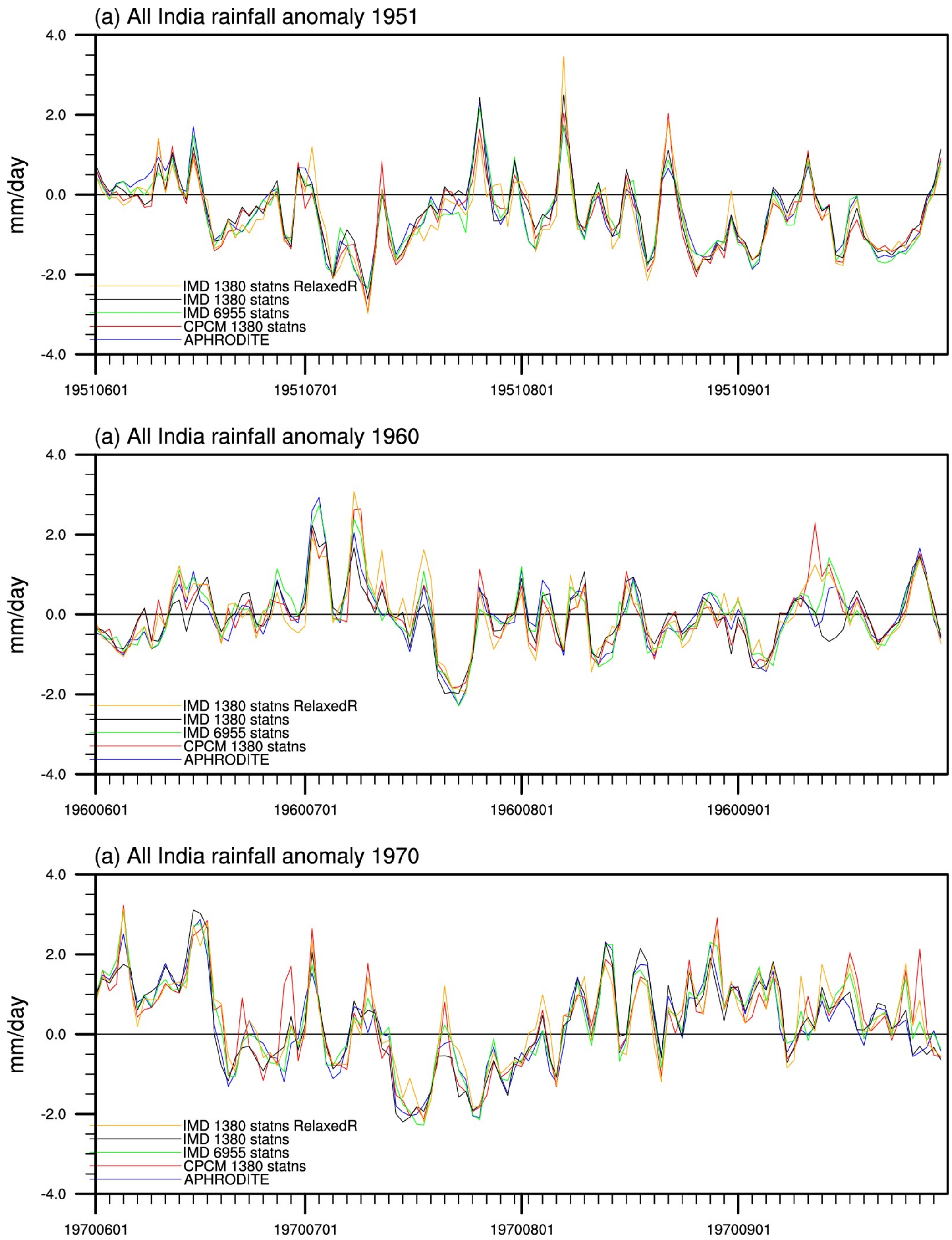


FIG. 13. Same as Figure 12 but the spatial average is over all India.