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

Khouider, B., Sabeerali, C. T., Ajayamohan, R. S., Praveen, V., Majda, A. J., Pai, D. S., … Rajeevan, M.

2020

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

BOUALEM KHOUIDER

Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada

C. T. SABEERALI, R. S. AJAYAMOHAN, AND V. PRAVEEN

Center for Prototype Climate Modelling, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

ANDREW J. MAJDA

Department of Mathematics and Center for Atmosphere and Ocean Sciences, Courant Institute of Mathematical Sciences, and Center for Prototype Climate Modelling, New York University, Abu Dhabi, United Arab Emirates

D. S. PAI

Climate Services Division, India Meteorological Department, Pune, India

M. RAJEEVAN

Ministry of Earth Sciences, New Delhi, India

(Manuscript received 2 July 2019, in final form 2 February 2020)

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 adequately distributed datasets. Objective interpolation methods such as inverse distance weighting (IDW) are the most widely used and have been employed to produce some of the most popular gridded daily rainfall datasets (e.g., India Meteorological Department gridded daily rainfall). Unfortunately, the skill of these techniques becomes very limited to nonexistent in areas located far away from existing recording stations. This is problematic as many areas of the world lack 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 that assigns a binned rainfall value, from a given number of bins to each lattice site or grid cell, with a certain probability according to the rainfall amounts observed in neighboring sites and a background climatological rain rate distribution, drawn from the available data. Grid cells 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 compared against two widely used gridded daily rainfall datasets over the Indian landmass for data from the summer monsoon seasons (June–September) for 1951–70.

1. Introduction

Rainfall is one of the essential meteorological parameters on which the lives and the well beings of many living organisms and mainly humans depend. The spatial and temporal variability of rainfall is directly linked to the socioeconomic development of people in tropical

Denotes content that is immediately available upon publication as open access.

Corresponding author: B. Khouider, khouider@uvic.ca

DOI: 10.1175/JHM-D-19-0143.1

continents. 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 (from hourly to monthly) is imperative to assessing water resources in mountains, arid regions, and river basins. Many modeling groups try to understand the characteristics of precipitation using general circulation models. The underlying models need to be verified using the observed gridded datasets to improve their performance and prediction skills. The observed daily precipitation is also required to monitor and forecast the subseasonal variability such as monsoon intraseasonal oscillations (MISO) and Madden–Julian oscillations (Madden and Julian 1971; Wheeler and Weickmann 2001; Wheeler and Hendon 2004; Zhang 2005; Yasunari 1980; Sikka and Gadgil 1980; Lawrence and Webster 2002; Wang et al. 2006; Lau and Waliser 2012; Suhas et al. 2013; Sabeerali et al. 2017).

Despite the progress in estimating the precipitation from satellite, the rain gauge observations have 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 and are the most accurate measurement of precipitation for a specific location, whereas the satellite estimates and model predictions of precipitation are indirect in nature and over the land; it is still difficult to estimate accurate precipitation using satellite. Hence, the satellite-estimated precipitation needs to be verified or calibrated using ground-based observations (Xie and Arkin 1995). The rainfall obtained from rain gauges are used as a ''ground truth'' for evaluating both satellite and radar-derived precipitation products as well as in improving satellite retrieval algorithms (Collier 1986; Prakash et al. 2019).

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 modeling, 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 substantially 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 datasets 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. They are typically based on variants of Shepard's weighted interpolation method that are both distance and direction aware. Shepard's method is discussed in section 2c. Moreover, an analysis of the monthly anomalies is performed to maintain the basic climatology gradient in regions of sparse data. The daily gridded precipitation product under the Asian Precipitation–Highly Resolved Observational Data Integration Toward Evaluation of Water Resources (APHRODITE) project (Yatagai et al. 2012), covering the whole of 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 a regular grid. Despite the abundance of gridded products, the pertaining analyses do not provide estimates of the precipitation variability and the impact of human-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) has reported that precipitation analysis is not sensitive to the algorithms used in regions with a dense network of rain gauge stations. In contrast, the bias is likely to exist over the regions with 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 (Bastin et al. 1984; Rudolf et al. 1994; Xie and Arkin 1995; Xie et al. 1996; Chen et al. 2008; Hofstra et al. 2010; Gervais et al. 2014; Prakash et al. 2019; Herrera et al. 2019) 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 (1963). Gandin's method assumes that the rainfall rate at a given grid point is the weighted 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 (Bussières and Hogg 1989). This method, thus called the optimal interpolation (OI) technique, uses remote information, namely, the rainfall variability, in addition to the localized station values.

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 (Bussières and Hogg 1989). Bussières and Hogg (1989) found an optimal radius of influence, for the four techniques they assessed, of about 40 km, for their particular network of pseudogauge data, but they choose to set it to about 110 km for all methods to avoid missing data points on their prescribed grid of $0.05° \times 0.05°$ resolution.

To construct the best possible gridded rainfall products, comparative studies of many different OA techniques are routinely conducted. For instance, Bussières and Hogg (1989) compared the empirical or functional OA algorithms of Barnes (1973), Shepard (1968), and Cressman (1959), and the OI method of Gandin (1963) using an unevenly distributed network of pseudorainfall 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 is 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.

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 compared 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 with local interactions of Khouider (2014) (see also Khouider et al. 2010) for organized tropical convection. It is based on the concept of a 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 domain-mean climatological information, namely, the rainfall rate distribution, to stochastically propagate the station gauge values to neighboring points on the given regular grid. In this sense, the proposed method is closer to the statistical method of Gandin (1963), but instead of minimizing the expected errors, it actually samples an estimated probability density at each grid cell 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 daily rainfall data are based on a dense network of 6955 stations, here, the new SLM algorithm employs only 1380 stations on purpose. To have a meaningful comparison, we also use Shepard's OA algorithm on the same 1380 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, and the new SLM algorithm as well as an overview of Shepard's method. The five daily rainfall data products, including the high-resolution IMD datasets, the APHRODITE datasets, 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 analyzed and compared to each other in section 3. In particular, we first provide an assessment of the SLM 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 rainfall climatologies and daily rainfall estimates, using statistical metrics such the root-mean-square error (RMSE), the absolute relative error, and the cross-correlation maps

of high-resolution IMD datasets versus each one of the four remaining products. The section is concluded with the analysis of the interannual and daily rainfall variabilities. A summarizing discussion is given in section 4, and a few concluding and outlook remarks are given in section 5.

2. Data and algorithm

a. The Indian rain gauge station data

The Indian subcontinent possesses one of the oldest networks of rain gauge data in the world. 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 1901–70. 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 IMD scientists to quality control the wide network of rain gauge station data in India and to generate gridded datasets that represent the rainfall characteristics in a realistic manner (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-h accumulated rainfall ending 0300 UTC. For pedagogical reasons, daily 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 daily 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 Pai et al. (2014), which is based on a much denser network of 6955 stations. This data product will be referred to as IMD6955. It is to be noted that there are no precipitation datasets that are true. Here, we adopt IMD6955 as the standard dataset that we aim to recover with the new stochastic method when the thinner number of 1380 stations is used as input.

As already stated, the specific question asked here is whether the SLM scheme can improve the precipitation estimate over grids with poor rain gauge coverage. Not all stations have recorded good quality rainfall data every day. The 1380 stations used in this study have a minimum of 70% data availability during the analysis period 1951–70. However, the data density is not uniform over the Indian subcontinent. While the gauge network over southern India and northwest and central India are dense, it is scattered over the northeast and eastern coastal region (Fig. 1a). Note that in this data source, there are no stations reported with precipitation over Jammu and Kashmir.

The probability of occurrence of daily rain rate events using all existing stations, in India from 1951 to 1970, is shown in Fig. 2b together with a power-law fit. The fit f with coefficients $a = -1.482$ and $b = 0.376$ on probability of occurrence x has the form $f = ax^b$. The daily rainfall distribution over the Indian subcontinent seems to follow the fitted power law reasonably well except at the high rain rates tail. The maximum probability of daily rain rate occurs in the range of 0–100 mm day^{-1}, and then the probability decreases rapidly with the intensity of rainfall (Fig. 2b). This climatological rain rate distribution is used as an external potential for the SLM. It is worthwhile noting that regional climatologies can be used instead for specific climatic zones of India, such as north India versus south or central India, for instance. However, such partitioning leads to coarsely resolved distributions in regions of low station density, in the rain rate spectrum space, with many empty bins. One can, of course, use a power-law fit, such as the one shown in Fig. 3b. Still, we refrain from doing this here and use the actual all-India climatology at once to avoid the eventually large errors associated with the extrapolation of a discontinuous variable such as rain rate. Besides, even for the all-India dataset, the power law has a hard time fitting the tail of the distribution, as can be seen in Fig. 3b. The use of heavy-tailed distribution or a mixture model may be necessary (Foss et al. 2013). Moreover, the SLM algorithm uses this background climatology as prior knowledge whose weight in the inferred distribution is controlled by a specific model parameter, which in effect sets the strength of the "connection" between nearest-neighbor sites and may thus rely less on the background climatology. Thus, getting the most accurate background distribution is not of paramount importance, as it turns out.

b. The stochastic model on a triangular lattice for daily 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.

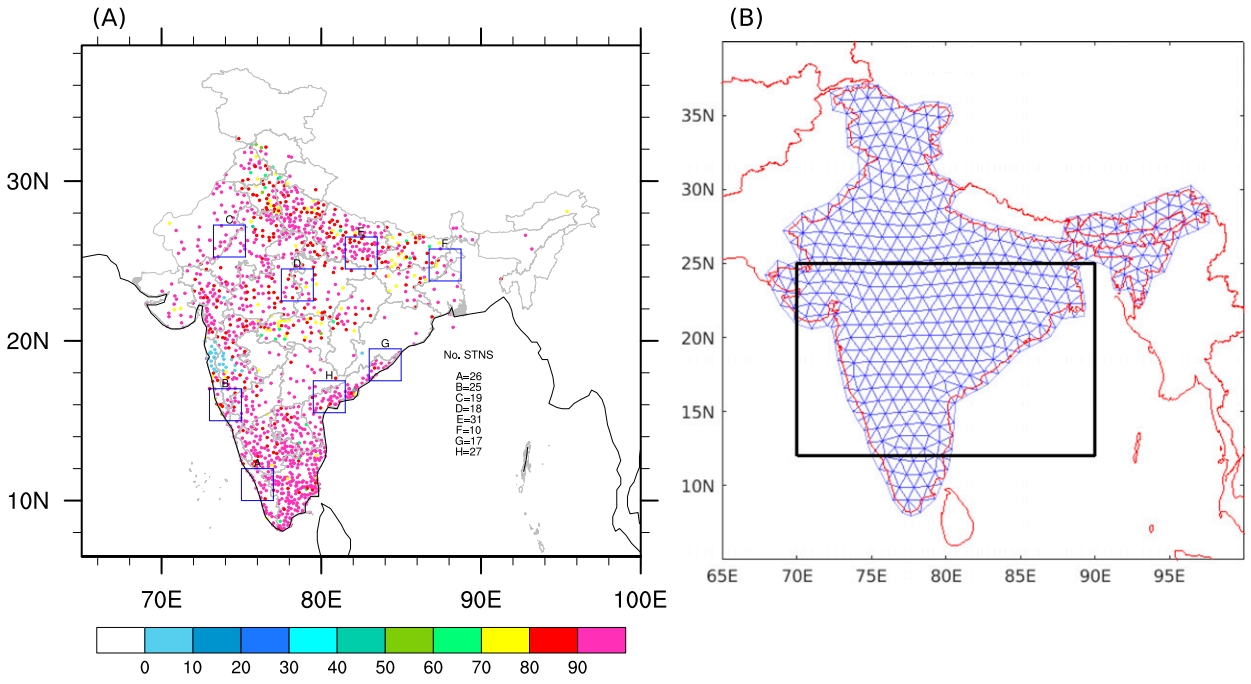

FIG. 1. (a) Location of the 1380 rain gauge stations used by the SLM interpolation scheme to produce the CPCM1380 datasets and by Shepard's scheme to produce the IMD1380 and IMD1380-relaxedR datasets. Colors indicate the percentage of days with rainfall data. Eight validation points, labeled A–H (listed at the bottom right), are marked by the blue squares each representing a 2° square box surrounding the corresponding validation point and the associated number of gauge stations within each box, which is withdrawn when performing the validation tests in section 3a. (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 $M = 11\,921$ triangles. The black box in (b) represent the central India domain used to average the rainfall.

The Indian subcontinent is divided into M triangular mesh elements, as shown in Fig. 1b. The triangular mesh is created using the Delaunay triangulation algorithm in MATLAB. In our analysis, we consider $M = 11\,921$, which is approximately equivalent to a 0.25° spatial resolution.

At any given time, t, spanning the period of interest, a given triangle I, $I = 1, 2, \ldots, M$, on the triangulation lattice may or may not contain station data. Station data will be present at the 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 a case, the average of all these station values is computed and assigned, as an observation value, and the corresponding triangle or cell j is considered as an observation cell and marked by an asterisk below. All other cells must be filled in by the OA-SLM procedure.

To illustrate, Fig. 2a shows a day to day variation of the number of cells containing stations with recorded rainfall data from 1951 to 2004. The data density is satisfactory and more or less uniform till 1995. Out of a total of 11 921 triangular cells over the Indian subcontinent, on average, around 1200 cells with rainfall is available. However, during 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 $I = 1, 2, \ldots, M$, 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 $M = 11\,921$ to the number of cells with recorded data in Fig. 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. The binary values of $\mathscr{M}_t$ should not be confused with the actual recorded rainfall data, and $\mathscr{M}_t = 0$ does not mean that the triangle does not experience any precipitation. It just means that we have no observation of it. It is the job of the SLM interpolation method to fill up those gaps.

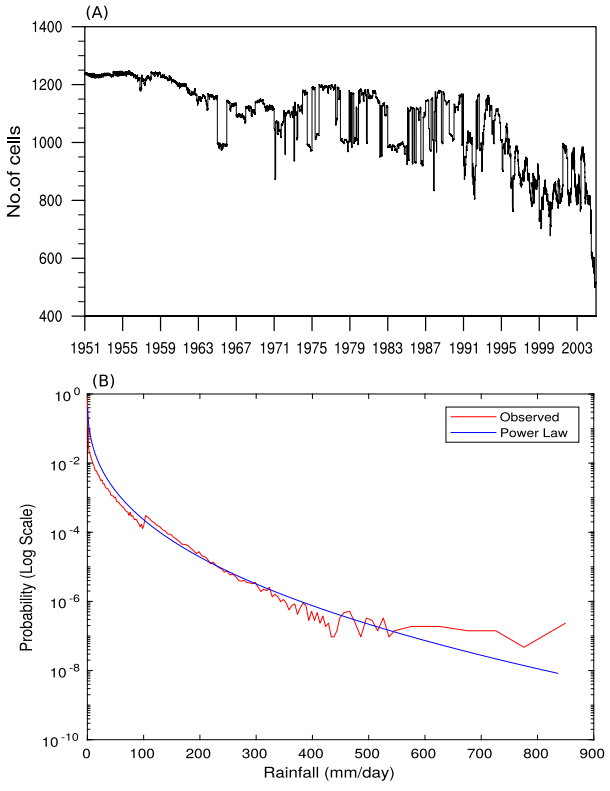

FIG. 2. (a) The number of triangular cells per day containing rain gauge rainfall data. (b) Probability of occurrence of rain rates in each range of rain intensity in the rain gauge datasets produced by the 1380 stations from 1951 to 2004.

The new SLM, introduced here, is based on the concept of multitype 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 (here, rainfall bin index) to another depending on prescribed probabilistic rules, based on the states of the nearest neighbors. In the present study, the station daily rainfall data are binned into N rain rates, corresponding to the N states of the SLM. To better accommodate the distribution of the recorded daily rainfall, we adopt a piecewise-uniform binning strategy. Various bin configurations have been tested, with a total number of bins ranging from $N = 51$ to $N = 137$. Our results indicate that the finer the bin sizes are, the more accurate the interpolated rainfall is. However, the finer bins are associated with a larger number of bins, 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 in the appendix for the sake of streamlining.

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

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

The bin resolution is thus set to be higher in the parts of the spectrum 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 rain rate bin number or none, if there is no rainfall. This partitioning of the rain rate spectrum into a finite number of bins has the advantage of eventually flattening the rain rate distribution as this is consistent with our algorithm described below, based on uniform random draws that do not discriminate between the bins. This is particularly the case if the partitioning is coarse. An excessively large bin number will make the model about to be described inefficiently. Thus, a middle ground needs to be found, and our strategy of using the background distribution ρ to use a fine resolution in places of large gradient seems to work nicely.

To be precise, we consider the order parameter

$$\sigma_I^t = j, \quad j = 0, 1, \ldots, N-1 \tag{3}$$

on a given lattice site $I, I = 1, 2, \ldots, M$, and at any given time t, according to whether there is a rain event within the bin $j, j = 0, 1, 2, \ldots, N-1$, in that cell at that

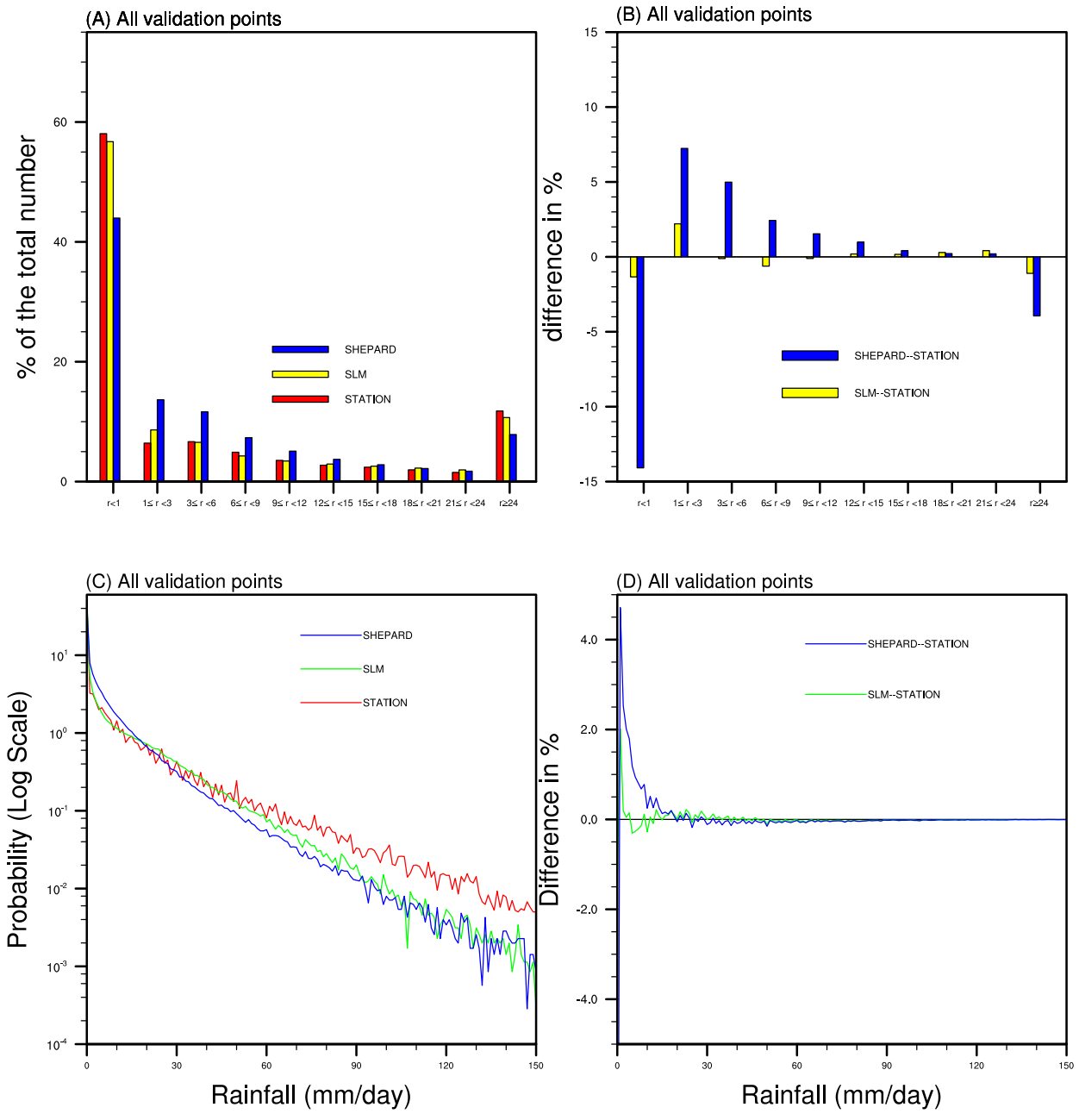

FIG. 3. (a) Probability density function (PDF; %) of the aggregated daily rain rates from all station locations contained in all eight validation point boxes shown in Fig. 1a. The PDFs of daily precipitation rates corresponding to the station (red bars), SLM interpolation technique (yellow bars), and Shepard interpolation technique (blue bars) are shown. (b) The difference in PDFs of daily precipitation rates corresponding to the SLM (yellow bars) and Shepard (blue bars) interpolation techniques from the PDF of daily station precipitation rates. (c),(d) As in (a) and (b), but for the full PDF of rainfall and its differences. See text for details. The x axes indicate different rain rate (mm day^{-1}) bins.

time t. Let R_j be the rain rate associated with bin j, $j = 0, 1, 2, \ldots, N-1$. 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. While at lattice site I^* with available measurement, the value of $\sigma^t_{I^*}$ is well defined; it is assumed to be random at all other sites $I \neq I^*$. It is the job of the SLM to provide and sample the distribution of these random values.

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 in each panel in Fig. A1.

Rainfall (mm day^{-1})	Bin size (mm day^{-1})	No. 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

To account for correlations between nearest-neighbor rain events, the bin particles are set to interact and ''exchange information'' in a seamless fashion, depending only on the distance between particle sites and the information itself, namely, particle type. This is somewhat the idea of particle interacting systems in a heat bath. The heat bath acts as an infinite reservoir of energy allowing the particles to constantly bounce around and interact with each other at the microscopic level. Particle interacting systems in a heat bath, follow the Gibbs canonical distribution,

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

as their equilibrium measure (Liggett 1999; Thompson 1971; Katsoulakis et al. 2003b), where H is the Hamiltonian energy which includes the energy from a local interaction potential, allowing nearest neighbor sites to excite each other as well as an external energy source, β is the inverse temperature of the heat bath, 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

$$H(\sigma) = -\frac{1}{2}\sum_{I}\sum_{I'} J(\sigma_I, \sigma_{I'}) + \sum_{I} h(\sigma_I), \tag{5}$$

where J is the internal interaction potential between neighboring 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 (MCMC) sampling method based on Arrhenius dynamics (Thompson 1971; Katsoulakis et al. 2003a), where for any fixed physical time t, we introduce a pseudotime s and view the order parameter, which we re-denote by σ^s, as a Markov process that makes random transitions at random lattice sites, over a long enough period of pseudotime s, until it reaches a statistical equilibrium, whose distribution is the Gibbs measure conditional on the climatology and the instantaneous station data. In other words, σ^s obeys an iterative process with a varying step size in the pseudotime s while the physical time t remains fixed. In this fashion, each realization σ^s is viewed as a random sample of the random variable σ^t.

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

$$\begin{aligned}\Delta^I_+\tilde{H}(\sigma) &= J_0\Big\{\max_{I'}[|R(\sigma_I+1)-R(\sigma_{I'})|] - \max_{I'}[|R(\sigma_I)\\ &\quad - R(\sigma_{I'})|]\Big\} + h(\sigma_I+1) - h(\sigma_I),\\ \Delta^I_-\tilde{H}(\sigma) &= J_0\Big\{\max_{I'}[|R(\sigma_I-1)-R(\sigma_{I'})|] - \max_{I'}[|R(\sigma_I)\\ &\quad - R(\sigma_{I'})|]\Big\} - h(\sigma_I+1) + h(\sigma_I),\end{aligned} \tag{6}$$

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. In (5), $J_0 > 0$ represents the strength of local interactions and is considered as an interpolation parameter and $R(x)$ is the rain rate R_x associated with bin x, $0 \le x \le N$ and the maximum is taken over all neighboring cells I' of I. Our tests indicate that the optimal J_0 value depends on the number of bins N and $J_0 = 1.05$ seems to be close to being optimal 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 actual equilibrium distribution, the jump rates of the Markov process σ^s, from a given configuration σ to its two closest ''neighbors'' in the configuration space, are given by

$$C_+^{I,j}(\sigma) = [1 - \mathscr{M}_t(I)]e^{(-1/2)\Delta_+^I H(\sigma)} + \frac{\mathscr{M}_t(I)}{\tau}\{\max[e^{-\alpha(\sigma_I - \sigma_I^\star)}, 1.0] - 1.0\},$$
$$C_-^{I,j}(\sigma) = [1 - \mathscr{M}_t(I)]e^{(-1/2)\Delta_-^I H(\sigma)} + \frac{\mathscr{M}_t(I)}{\tau}\{\max[e^{-\alpha(\sigma_I - \sigma_I^\star)}, 1.0] - 1.0\}. \quad (7)$$

Here, α and τ are positive parameters that are specified in Table 2 together with the other model parameters while $\mathscr{M}_t$ is the binary mask function in (1) and $0 \le \sigma_I^* \le N-1$ is a fixed bin index (does not vary with s) corresponding to the observed rainfall data at the given cell I, if available, that is, $\sigma_I^* = \sigma_{I*}^t$ if $\mathscr{M}_t(I) = 1$. The external potential $h(\sigma)$ satisfies the relation $\rho_j = e^{-h(j)/2}$, when $\sigma_I = j, j = 0, 1, 2, \ldots, N$, so that when interactions between nearest neighboring cells are ignored, $J_0 \equiv 0$, independently on the cell index I, the transition rates reduce to the background values,

$$\tilde{C}_+^j = \frac{1}{\tau}\frac{\rho_{j+1}}{\rho_j}, \quad j = 0, 1, 2, \ldots, N-1,$$
$$\tilde{C}_-^j = \frac{1}{\tau}\frac{\rho_{j-1}}{\rho_j}, \quad j = 1, 2, \ldots, N, \quad (8)$$

accordingly, so the stochastic process is in detailed balance with the bare background distribution ρ_j in (2),

$$\tilde{C}_+^j \rho_j = \tilde{C}_-^{j+1} \rho_{j+1}, \quad (9)$$

(i.e., the probability of transition from bin j to $j+1$ is equal to the probability of transition from bin $j+1$ to bin j) and thus admits ρ_j as the default equilibrium distribution of the stochastic process. Moreover, the binary factors $[1 - \mathscr{M}_t(I)]$ and $\mathscr{M}_t(I)$ force the transition rates toward $(1/\tau)\{\max[e^{\alpha(\sigma_I - \sigma_I^*)}, 1.0] - 1.0\}$ at the lattice cells with available observations, that is, when $\mathscr{M}_t(I) = 1$, and to exponential energy differences $e^{(-1/2)\Delta_+^I H(\sigma)}$, otherwise. In this fashion, the stochastic process is forced to rapidly relax (at the pseudotime scale τ) to the observed values σ_I^*, when available.

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

$$\text{Prob}\{\sigma_I^{s+\Delta s} = \sigma_I^s + 1\} = C_+^I(\sigma^s)\Delta s + o(\Delta s),$$
$$\text{Prob}\{\sigma_I^{s+\Delta s} = \sigma_I^s - 1\} = C_-^I(\sigma^s)\Delta s + o(\Delta s),$$
$$\text{Prob}\{\sigma_I^{s+\Delta s} = \sigma_I^s\} = 1 - [C_+^I(\sigma^s) + C_-^I(\sigma^s)]\Delta s + o(\Delta s), \quad (10)$$

for small pseudotime increment Δs, $\Delta t/\tau \ll 1$, of the pseudotime s, used to iterate the process to equilibrium.

The definition of the transition rates in (7) and (8) 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^s converges to $G(\sigma)$ in the long run (Khouider 2014). Therefore, according to the MCMC theory, the time series of the process σ^s 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 (7) on the mask function $\mathscr{M}$ is such that the convergence of the process to the observed values σ_I^* occurs on an exponentially fast time scale, at all lattice sites with station data, independently on the background climatology distribution and on the state of the neighboring sites; σ^s becomes quickly (almost) deterministic at those locations. The rate of this convergence is set by the parameter α, which bears a large value of $\alpha = 4$. The station values are then used to update the values of its neighboring cells, which then transmit the information to their own neighbors and so on. The process goes back and forth until statistical convergence. Our tests indicate that fixing the values to $\sigma_I^s = \sigma_I^*$ at the cells with observation data leads to the same results but 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

$$S_R = \sum_I [C_+^I(\sigma) + C_-^I(\sigma)]. \quad (11)$$

Also, to avoid the occurrence of unphysical values of σ, we enforce the following "boundary conditions," so that the lattice value cannot transition out of the physical range of bin index values:

$$C_-^I(\sigma) = 0, \quad \text{if} \quad \sigma_I = 0 \quad \text{and}$$
$$C_+^I(\sigma) = 0, \quad \text{if} \quad \sigma_I = N, \quad (12)$$

at each lattice cell $I = 1, 2, \ldots, M$. Here C_+^I, C_-^I are transition rates at a given lattice site I, from bin $\sigma_I = k$

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 h
J_0	Strength of local interaction potential	1.05
N	Number of bins	137
M	Number of lattice cells	11 921
T_0	Pseudo iteration time	24 h

to bin $\sigma_I = k + 1$ and from bin $\sigma_I = k$ to $\sigma_I = k - 1$, respectively, while σ is the configuration array of all bin values, corresponding to all lattice sites as a whole and σ_I is precisely the bin value at site I.

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

1) Read the station data at the given physical time (day of the year between 1951 and 1970 here) and set $T = T_0$.
2) Compute the up and down transition rates C_+^I and C_-^I using (7) at every cell I, $I = 1, 2, \ldots, M$ and compute the total rate S_R using (11).
3) Draw a uniform random number U between 0 and 1 and set $s = -[1/(S_R)\log(U)]$.
4) If $s \le T$, make a single transition at a random site I in the following way.
   (i) Renumber the rates $C_+^I(\sigma)$ and $C_-^I(\sigma)$ from 1 to $2M$, say, $C_1 = C_+^1$, $C_2 = C_-^1$, $C_3 = C_+^2$, $C_4 = C_-^2$, $\ldots$, $C_{2M-1} = C_+^M$, $C_{2M} = C_-^M$. Compute the probabilities $P_k = C_k/S_R$ and their cumulative sums $S_k = \sum_{l=1}^{k} P_l$, $k = 1, 2, \ldots, 2M$.
   (ii) Draw a second random number U^1, uniformly between 0 and 1 and independent of U, and find the first k_0 such that $S_{k0} \ge U^1$ and perform the transition associated with C_{k_0}:

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

   (iii) 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 neighbors need to be recalculated every time step 1 is called again. Also, it is worth noting that the variable s defines a pseudotime step for the Markov chain that changes randomly at each step according to step 3 of the algorithm above.

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 summarize, the new SLM technique relies on three sources of information to transform the sparsely and irregularly distributed daily rain gauge data into a gridded rainfall data product, in a given period. First, the domain of interest (here India) is partitioned into nonoverlapping triangles and the triangles containing existing rain gauge data are identified and marked by a binary mask $\mathscr{M}_t(I)$ in (1), for each time t (here day) of the given period, where I is the triangle index. The average rain rate from all station values recorded within each triangle, at the given day, are then calculated and assigned a marked bin value σ_I^*. Second, the binned rain rates from nearest neighbors, in the triangular lattice, mutually exchange information through the interaction potential in such a way that the rain rate/bin values, σ_I, assigned to the lattice as a whole are distributed (or nearly so) according to the Gibbs measure in (4) while constrained by the readily available station data, $\sigma_I = \sigma_I^*$ at all marked triangles. This assignment can be understood as an energy minimization process under the constraint of the available rain gauge data. Third, in addition to the information exchanged between neighboring triangle cells, the minimized energy takes also into account the climatological rain rate distribution ρ_j in (2) which is approximated by a power law in Fig. 3. When the nearest-neighbor interaction is ignored, the Gibbs measure reduces to the climatological distribution ρ_j.

The last 10% of the converged MCMC chains, associated with each triangle and physical time (day), are used to obtain a set of interpolated rain rate values at all the lattice triangles. More discussion on the convergence of the MCMC chains is given in the appendix.

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 the regular latitude–longitude grid (0.25° × 0.25°) using the bilinear interpolation. It is worth noting that the bilinear interpolation is not the most suitable method for rainfall data as it assumes a certain degree of smoothness and a fractional coverage conversion between the triangular and rectangular grids will be more accurate. We used the bilinear interpolation technique as an easy and quick method to convert from triangular mesh to regular latitude–longitude grid because the available rainfall products are in the regular

latitude–longitude grid and therefore, it is easy to analyze and compare. The fractional coverage conversion requires element-by-element remapping and will be computationally cumbersome. However, we have checked the probability density function (PDF) of rain rates against station data PDF, and it is found that the PDF shape is fairly preserved, which means that the bilinear interpolation is not as bad as it seems at least for this validation. Alternatively, we could avoid this bilinear interpolation by introducing a rectangular mesh instead of triangular mesh in order to perform SLM, but the advantage of triangular mesh is that it can accommodate the geographical border of Indian subcontinent quite well. Future studies over the global domain can be addressed using the rectangular mesh instead of a triangular mesh.

Furthermore, given that the triangular and rectangular grids have the same resolutions of 0.25°, it is expected that the error induced 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. The convergence of the MCMC time series and sensitivity to parameters of the SLM scheme are discussed in the appendix. This newly gridded dataset is named as the CPCM1380 data product, in reference to the Center for Prototype Climate Model (CPCM) at New York University Abu Dhabi, where this research was conducted and to the 1380 rain gauge stations used.

c. *Shepard weighted interpolation method and its relaxation*

As already mentioned, the SLM interpolation technique is assessed in comparison to the high-resolution (0.25° × 0.25°) daily 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 fewer 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 comparison with the SLM method on the whole Indian domain.

In Shepard's method, the interpolated values at a grid node are computed from a weighted sum of the neighborhood observations. 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 are used directly and no interpolation is required, otherwise, the rainfall rate at P_i is given by

$$R_i : f(P_i) = \frac{\sum_s W_i^s Z_s}{\sum_s W_i^s}, \tag{13}$$

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 that 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 overrepresentation 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 whether $0 < d_i^s \le D_x/3$ or $D_x/3 < d_i^s \le D_x$ and station data not used if $d_i^s > D_x$. The interested reader is referred to Rajeevan et al. (2006) and Pai et al. (2014) for details.

Following the previous studies (Rajeevan et al. 2006; Pai et al. 2014), we considered a limited number of neighboring points (minimum 1 and maximum 4) within a search distance (radius of influence D_x) of 1.5° around the grid node where we want to compute the interpolated values. We termed this product as the IMD1380 station product. We note in particular that because of 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 already mentioned low station density regions. To gauge the performance of the SLM technique on the whole Indian domain, we decided to push Shepard's method beyond its limits and have relaxed the radius of influence restriction and reproduced a full coverage gridded daily rainfall data for the Indian continent based on the same 1380 stations. We termed these data as the IMD1380-relaxedR product. 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 Bussières 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 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 all-India domain.

3. Results

In the following analysis, we compared various statistical metrics of CPCM, IMD, and APHRODITE gridded datasets for the sake of consistency. Since all grid datasets are going to be method dependent, there is no perfect

reference gridded rain gauge data. Nonetheless, we assume that the high-resolution IMD6955 is the default-standard dataset and compares the other products on the Indian continental scale. However, before doing so, we also provide local comparisons between the new SLM product and Shepard's method based on how they can infer the rainfall distribution at a given station when that station is withdrawn from the pool of existing measurements. Alternatively, one can use a strategy of thinning a region of highly dense stations and compared the recovered and observed data over the whole regions as opposed to removing all stations in a relatively small region as done below. The gradual thinning of station data as in Chen et al. (2008), for example, is a good alternative to prove convergence in the limit of full station coverage, but it is beyond the scope of the current study as we are more interested in the case of coarse station distribution.

In addition to the traditional RMSE and correlation estimates, deviations between the various data products are estimated according to the following equation, which is namely, the accumulated relative error (ARE). 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

$$N_{12} = \sum_x \sum_t \frac{2|R^1(x,t) - R^2(x,t)|}{R^1(x,t) + R^2(x,t)}. \tag{14}$$

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 performed, and the interpolated datasets or products, CPCM1380 and IMD1380-relaxedR, respectively, on the 0.25° × 0.25° grid are constructed for the 20 years (1951–70), using the procedures outlined above. Here, we report the results of the comparative tests of these products against each other and the high-resolution IMD6955 and APHRODITE products. Notice that because rainfall is very rare to nonexistent 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 rain rate distribution skill

First, we assess how well the new SLM and the Shepard technique reproduce the observed local daily rain rate intensity PDFs. Following Chen et al. (2008), we have selected eight validation points over the Indian landmass, and the daily precipitation from all the stations in a 2° square around each validation point is withdrawn from the datasets. These squares correspond to boxed regions shown in Fig. 1a. With two boxes (A and B) along the west coast and two along the east coast (G and H) of the southern peninsula, and four boxes (C, D, E, and F) distributed along the east–west extent 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 neighborhoods. The validation point locations are representative of the complexity of the Indian rain gauge datasets 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 daily precipitation rate intensity is computed by aggregating the values of precipitation from all withdrawn station locations contained in all eight validation point boxes, leading to an aggregated PDF of all validation points and for each algorithm. The estimated PDFs of each algorithm are compared to the corresponding PDF of the withdrawn station observed precipitation to assess the accuracy of the two algorithms in reproducing the precipitation intensity distribution. The aggregated PDF of daily precipitation rate intensity at all station locations over the eight 2° square boxes are given in Fig. 3a. As can be surmised from Fig. 3, the PDF estimates are given in terms of rainfall events falling into the 10 bins

$$\begin{aligned} &R < 1, \quad 1 \le R < 3, \quad 3 \le R < 6, \quad 6 \le R < 9, \\ &9 \le R < 12, \quad 12 \le R < 15, \quad 15 \le R < 18, \\ &18 \le R < 21, \quad 21 \le R < 24, \quad 24 \le R \end{aligned} \tag{15}$$

where R is the rainfall rate (mm day^{-1}).

In general, the PDF of daily precipitation rate intensity is dominated by weak to no rain events ($R < 1$ mm day^{-1}). The PDF of station precipitation rate is largely dominated by the no rain events, which has a frequency of occurrence 58%, while the probability of heavy rainfall ($R \ge 24$) is 12.5%. The Shepard method underestimates the frequency of no-rain events and heavy rain events (blue bars), whereas it overestimates the frequency of occurrence of light precipitation ($1 \le R \le 18$) events (Fig. 3a). These results are clearer from the differences of PDF of each method from the PDF of corresponding station data (Fig. 3b). 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 (Figs. 3a,b). The frequency of no-rain event in the SLM method is 57%, which is comparable to the station

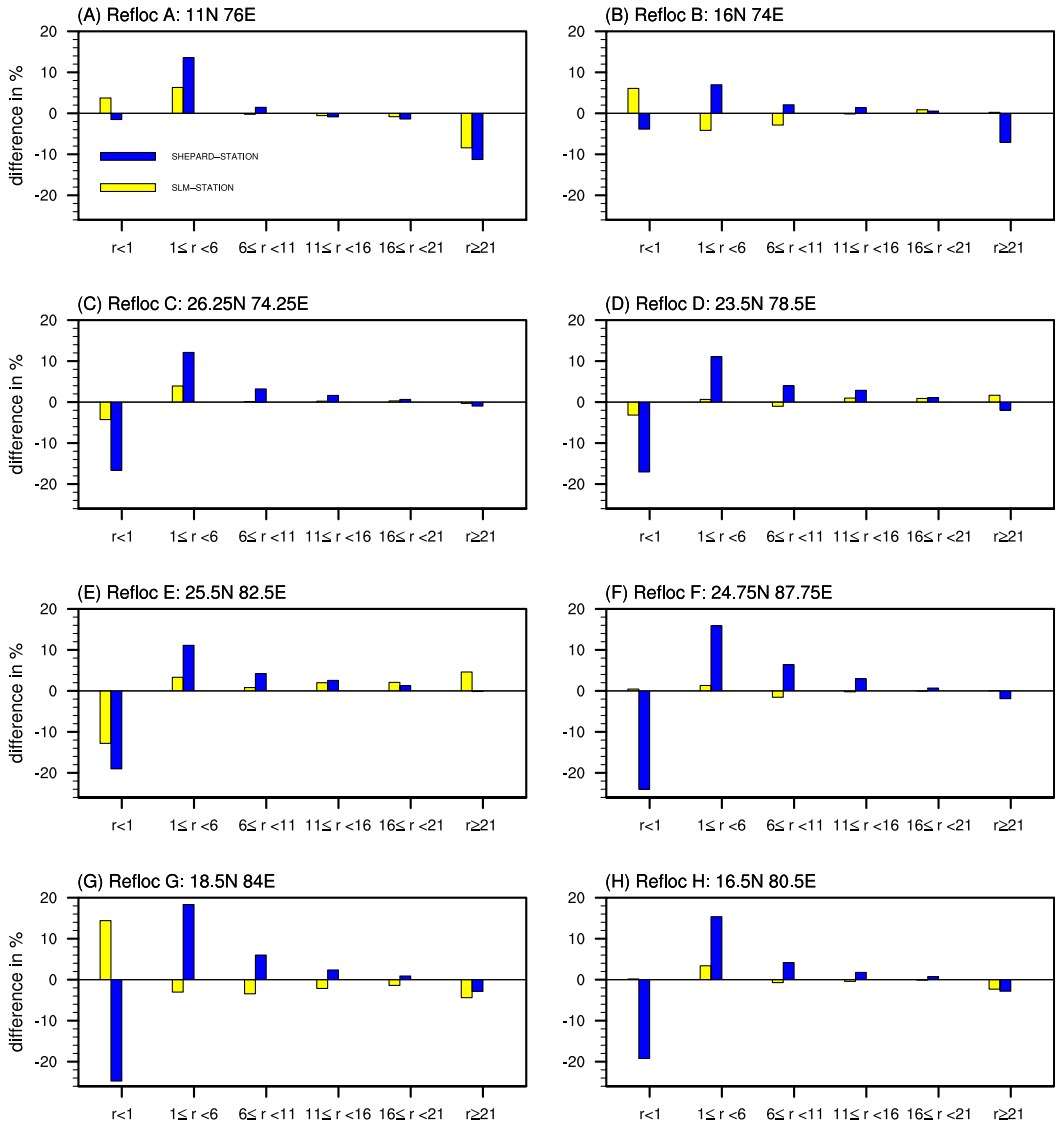

FIG. 4. (a)–(h) Difference in PDF (%) of daily precipitation rates corresponding to the SLM (yellow bars) and Shepard (blue bars) interpolation techniques from the PDF of daily station precipitation rates are shown at eight validation points (A–H). See text for details. The x axes indicate different rain rate (mm day^{-1}) bins.

precipitation (Fig. 3a). Similarly, the frequency of occurrence of light rainfall events ($1 \leq R \leq 18$) and moderate or heavy rainfall ($R \geq 18$) in the SLM method is also comparable to the station precipitation (Figs. 3a,b). Figure 3 may seem to indicate that Shepard's is as good as the SLM in estimating moderate rain events within the range $18 \leq R < 24$ but looking back at the local panels (Fig. 4) this is clearly due to cancellations of errors some of which is also inevitably true for the SLM results, though to a much lesser extent. In Figs. 3c and 3d, we extend this comparison to the full PDF of rain rates and the corresponding differences between the PDF of each method and the PDF of the station data. This confirms that the SLM method outperforms the Shepard method in representing the PDF of rain events, especially at weak to moderate rainfall events of up to 60 mm day^{-1} (Figs. 3c,d). At very high rainfall events, the log scale in Fig. 3c suggests that the two methods perform similarly, but as we can see from the actual PDF differences in Fig. 3d, in both cases, the mismatch is negligible.

The localized PDF of daily precipitation rate intensity for each validation point, and each algorithm is also computed by aggregating the values of precipitation of all withdrawn station locations in each box around each validation point (figure not shown). It is found that the frequency of occurrence of low- to no-rain events varies

strongly between the validation points. In terms of the station data, it goes from as high as 80% at the northwest validation point C to less than 40% at the southwest point B located at the northern tip of the Western Ghats mountain range (figure not shown). We have shown the differences of PDF of each algorithm from the PDF of corresponding station precipitation data for each validation points (Fig. 4). According to Fig. 4, except for the two validation points A and B, the no rain events are better represented in the SLM algorithm (yellow bars) compared to Shepard's method (blue bars). These two validation points are located over the windward side of the 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 orography. The number of stations reporting the precipitation is also large on these locations (number of 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 (Fig. 4). The moderate and heavy precipitation events ($R > 16$ mm day $^{-1}$) are also well represented by the SLM method except for two validation points (Figs. 4e,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 (Fig. 4g), the moderate and heavy precipitation events are underestimated by the SLM. 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 point G is located on the eastern coast of India, where the monsoon depression/low pressure systems normally first hit land. However, around this validation point, the number of stations reporting precipitation data is comparatively less (17 stations).

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

It is worthwhile noting that refining the bin distribution in (15) to bins of 3 mm day^{-1} yields quantitatively and qualitatively similar results (results not shown).

b. *Seasonal mean and daily rainfall direct comparisons*

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

Consistent with 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 CPCM1380 (Fig. 5c) datasets. Even with the significantly reduced number of stations, CPCM1380 (Fig. 5c) is in good agreement with the high-resolution IMD6955 and APHRODITE gridded rainfall products all over the Indian continent while IMD1380 in Fig. 5d misses large areas, namely the northern and northeastern parts of India, because of the lack of station coverage. The IMD1380-relaxedR climatology, on the other hand, shows significant biases, especially over the northern/northwestern part of the Indian subcontinent, whereas it is close to IMD6955 in the northeast region.

The seasonal rainfall averaged over the Indian subcontinent of all the four products are contrasted in Table 3. The seasonal precipitation of APHRODITE is the smallest among all precipitation products. Seasonal rainfall of CPCM1380 and IMD6955 are almost identical, whereas that of IMD1380-relaxedR overestimates it by about almost 60 mm day^{-1}, compared to IMD6955. This suggests again that gridded rainfall data are method dependent and that the station density is less important if one is interested only in the climatological regional mean values.

The daily precipitation maps for the three different active days of precipitation (12 July 1956, 9 July 1958, and 25 August 1965) in three gridded daily precipitation products are compared in Fig. 6. In all these three active days, the precipitation is mainly concentrated over the Western Ghats and central/eastern India. The precipitation is well organized in these regions. This snapshot map shows that all three gridded rainfall products reasonably capture this pattern of precipitation over the Western Ghats and central/eastern India. However, APHRODITE precipitation variability is relatively smooth (Fig. 6), and its magnitude is underestimated when compared to IMD6955, and CPCM1380 gridded rainfall products. However, the precipitation variability in CPCM1380 and IMD6955 are comparable, and more importantly, both these datasets show greater spatial details than APHRODITE.

When taking into account the fact that the SLM method used to produce the CPCM1380 datasets is based on rainfall binning with bin sizes of 2 mm day^{-1}

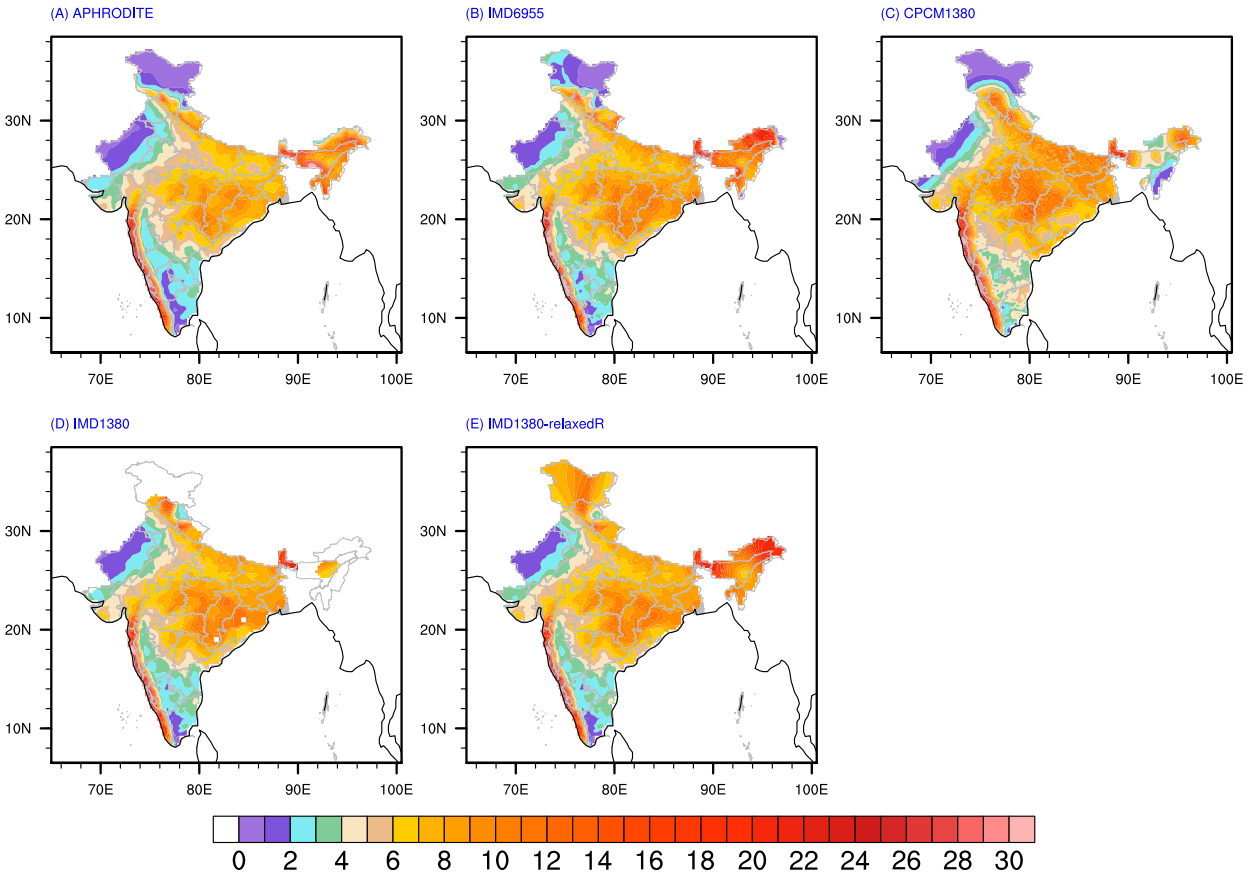

FIG. 5. JJAS rainfall climatology (mm day^{-1}) of the Indian subcontinent for the period 1951–70 obtained from the five datasets. (a) APHRODITE, (b) IMD6955, (c) CPCM1380, (d) IMD1380, and (f) IMD1380-relaxedR. See the text for details.

and larger, according to Table 1, errors in the range from 1 to even 3 mm days^{-1} are expected and are deemed acceptable since they are within the bin size. As shown in Table A1, increasing 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 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 (see section 5 for more information).

c. *Statistical metrics*

We show in Fig. 7 the maps of the 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 datasets. 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 the Western Ghats. Generally, the RMSE is a minimum over the low elevation plains, such as central India. However, compared to APHRODITE and IMD1380 datasets, the CPCM1380 datasets show 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 Fig. 7b. This is expected from the CPCM1380 product because of the combination of the stochasticity of the SLM method and the coarseness of the bin size used to implement it.

TABLE 3. Seasonal mean rainfall in different rainfall products (mm).

Rainfall product	Seasonal mean (mm)
IMD6955 stations	864
APHRODITE	756
CPCM1380 stations	863
IMD1380-relaxedR	920

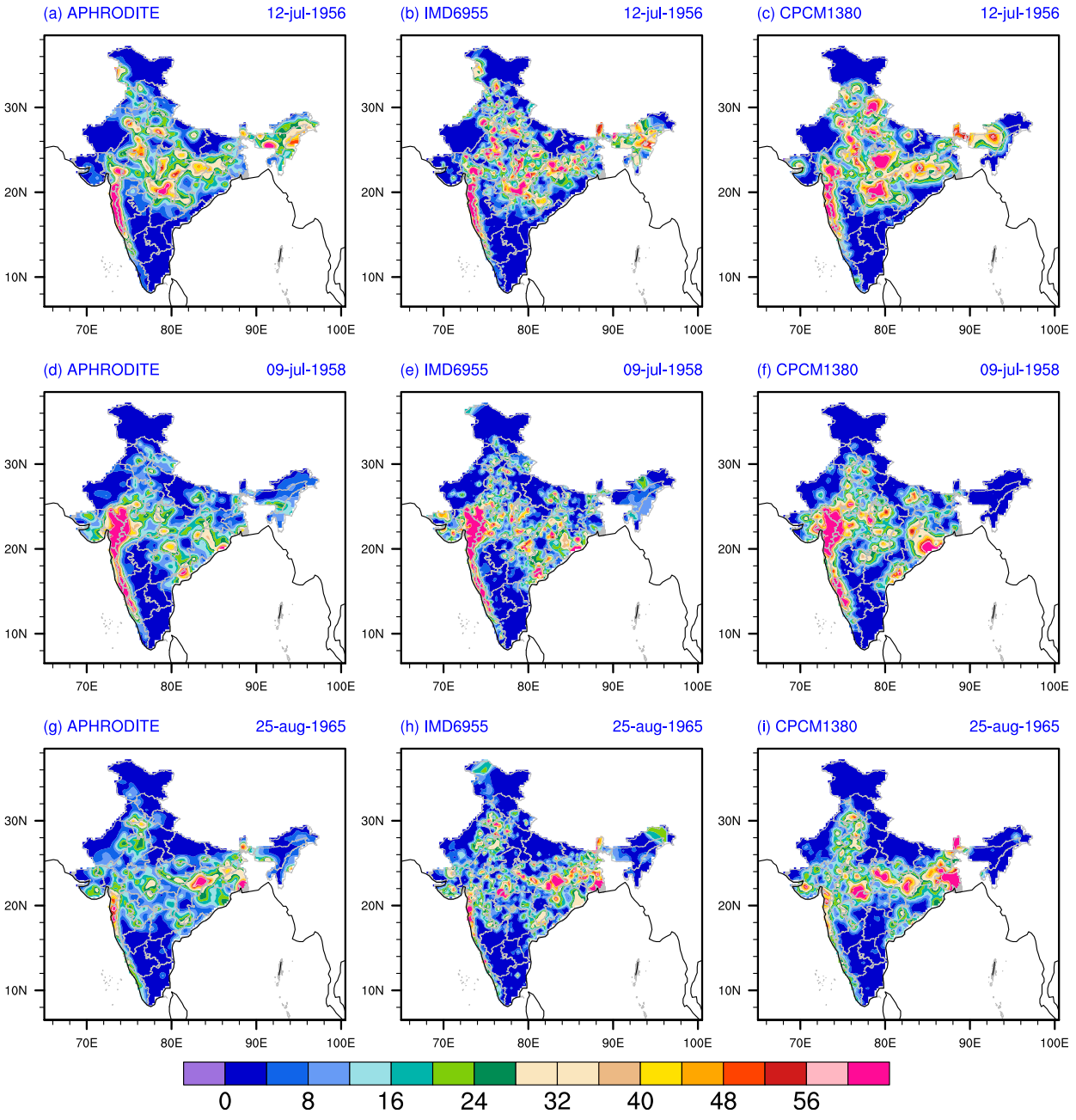

FIG. 6. Daily rainfall (mm day^{-1}) in the three different gridded products for the three active monsoon days (a)–(c) 12 Jul 1956, (d)–(f) 9 Jul 1958, and (g)–(i) 25 Aug 1965. Shown are (left) APHRODITE, (center) IMD6955, and (right) CPCM1380.

Nonetheless, the RMSE displayed by the CPCM1380 datasets remains comparable to those displayed by the APHRODITE and the IMD1380 datasets. As expected, large errors are associated with the IMD1380-relaxedR datasets over the regions of low station data coverage in the northern and northeastern tips of India.

In Table 4, we reported the absolute relative error (N_{12} between the IMD6955 data and the other precipitation products using the equation in (14), and the RMSE. From Table 4, 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, the daily error estimated from (14) is slightly less in CPCM1380 than it is in APHRODITE, however, the RMSE of seasonal mean ISMR is larger in CPCM1380 than

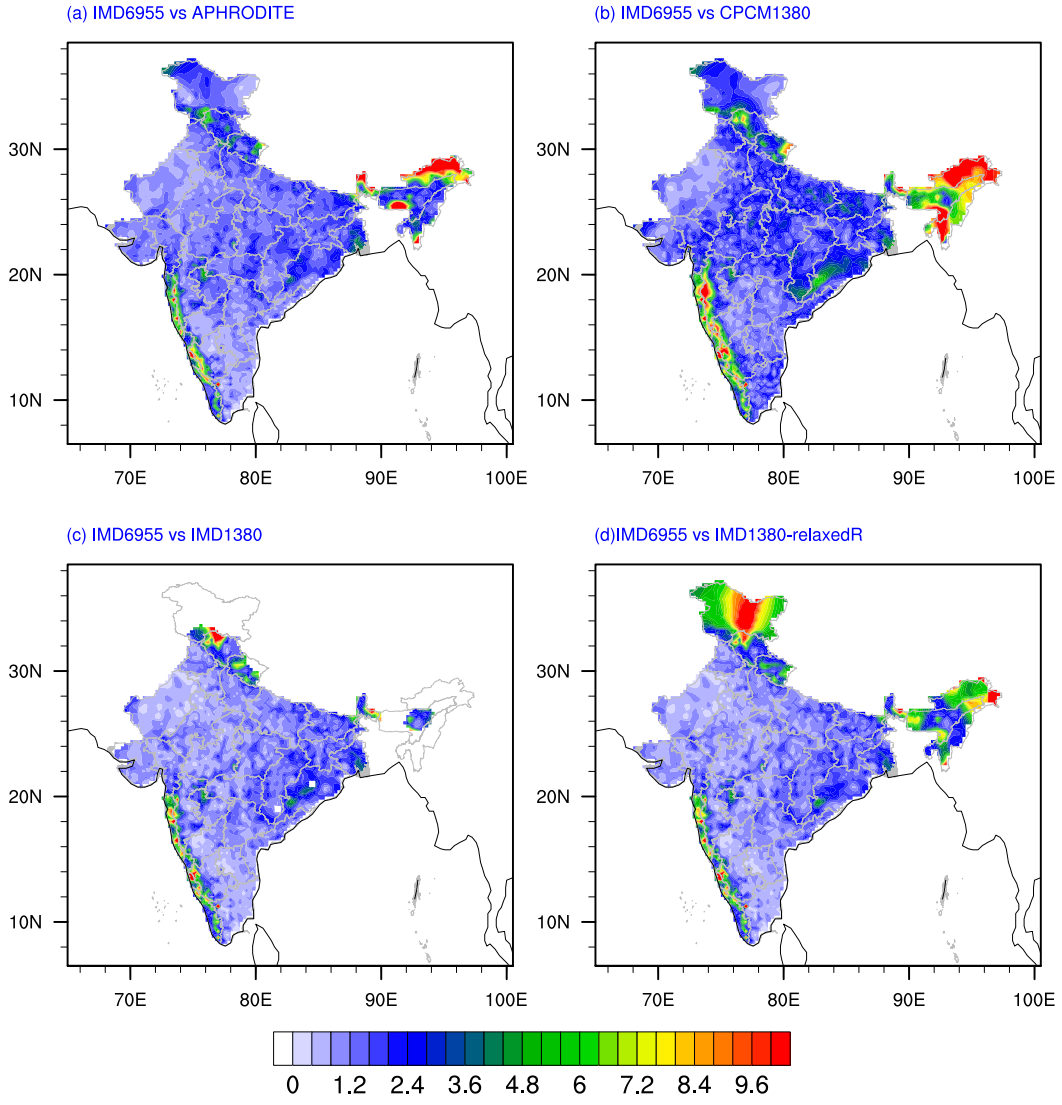

FIG. 7. RMSE (mm day^{-1}) (a) between IMD6955 and APHRODITE, (b) between IMD6955 and CPCM1380, (c) between IMD6955 and IMD1380, and (d) between IMD6955 and IMD1380-relaxedR, for all JJAS seasons from 1950 to 1970.

it is in APHRODITE. It is also true that the absolute relative error between APHRODITE and CPCM1380 is larger than it is between APHRODITE and IMD6955.

Figure 8 represents the seasonal correlation between IMD high-resolution analysis (IMD6955) and 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 (e.g., the northeast, Jammu, and Kashmir regions). Note however that the caveat, in all these analyses, is of course in the fact that we assumed IMD6955 as the standard for convenience while as already stated the OA products are always going to be method dependent and Shepard's algorithms show large bias in representing daily rainfall intensity with respect to station data in regions with sparse data (Fig. 3).

d. Interannual rainfall variability

The interannual variation of India summer monsoon (JJAS) rainfall (ISMR; precipitation averaged over the Indian subcontinent) is plotted in Fig. 9 for the five data products. The ISMR time series of IMD6955, CPCM1380, and IMD1380 datasets nearly match each other in terms

TABLE 4. Absolute relative error (14) and RMSE of seasonal mean Indian summer monsoon rainfall between various data products, as indicated.

Rainfall products	Error estimated from (1)	RMSE of seasonal mean rainfall (mm day^{-1})
IMD6955 vs IMD 1380 stations relaxedR (all India)	0.76	2.25
IMD6955 vs IMD 1380 stations relaxedR (inside Rinf)	0.69	1.60
IMD6955 vs IMD 1380 stations relaxedR (Outside Rinf)	1.14	6.30
IMD6955 vs CPCM1380 stations (all India)	0.87	2.77
IMD6955 vs CPCM1380 stations (inside Rinf)	0.85	2.33
IMD6955 vs CPCM1380 stations (outside Rinf)	0.99	5.96
IMD6955 stations vs APHRODITE (all India)	0.88	2.03
APHRODITE vs CPCM1380 stations (all India)	1.02	2.58

of magnitude and phase. However, the magnitude of the ISMR time series derived from the APHRODITE is underestimated in all years compared to both IMD6955 and CPCM1380 gridded rainfall, which is consistent with the results in Table 3. 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

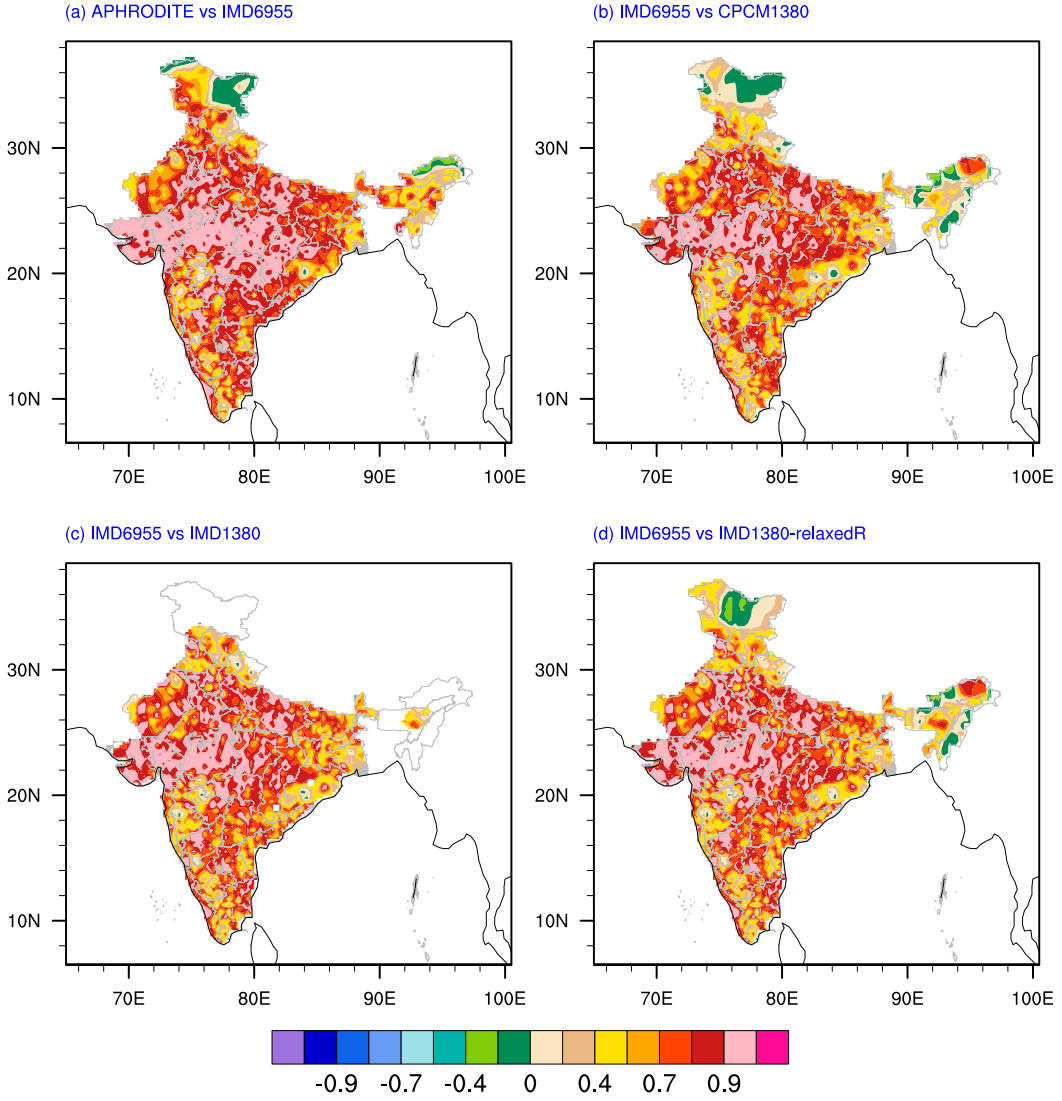

FIG. 8. (a) Grid point correlation of JJAS mean rainfall between IMD6955 and (a) APHRODITE, (b) CPCM1380, (c) IMD1380, and (d) IMD1380-relaxedR for the period 1951–70.

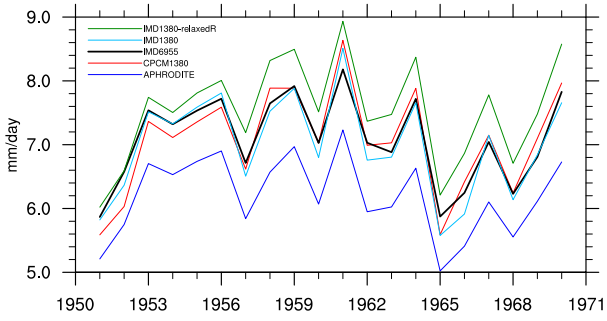

FIG. 9. Interannual variation of all India summer monsoon rainfall (mm day^{-1}; averaged over Indian landmass and averaged over JJAS season): IMD6955 (black), APHRODITE (blue), CPCM1380 (red), IMD1380 (sky blue), and IMD1380-relaxedR (green).

the IMD1380-relaxedR have relatively higher magnitudes compared to the other ISMR time series.

The daily variation of rainfall anomaly averaged over central India (black box in Fig. 1b) for three monsoon season (1951, 1960, 1970) are given in Fig. 10. In CPCM1380 analysis, the daily variation of central India rainfall anomalies is in line with other rainfall products. It is clear that the CPCM1380 daily rainfall product is quite good in capturing the signs of rainfall anomaly over central India in agreement with the other precipitation products, such as IMD6955 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 exceeds 0.95 in all these three monsoon seasons.

4. Discussion

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 technique is applied to the Indian Meteorological Department rain gauge datasets, which started since 1901. While the Indian station network totals 6955 stations, we chose to use a selection of 1380 stations dispersed unevenly over the Indian subcontinent to implement and test the SLM technique.

Existing studies (Bussières and Hogg 1989; Chen et al. 2008) found that the statistical optimal interpolation (SOI) method of Gandin (1963) 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 neighboring stations. Arguably, it is because the SOI method minimizes the expected error over all the existing stations and as such it uses remote 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 Bussières and Hogg (1989) and Chen et al. (2008), the SOI results are very closely followed by those obtained by the inverse distance weighted method of Shepard (1968).

The existing IMD6955 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 1380 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), which are in a way both used as the standard or target to achieve or beat. To have meaningful comparison, we decided to lift the radius of influence restriction on Shepard's method to produce a data product based on the smaller set of 1380 stations 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 1971), when the particles are actually rain rate bin indices at the corresponding grid points forming the lattice, conditional to the existing station data at the local station sites and the associated domain-mean climatology. As such it draws information from the neighboring locations, the available rain gauge data, and the rain rate climatology distribution to build statistical estimates at remote grid points. In this sense, the SLM method has this remote-information-gathering feature in common with the SOI method of Gandin (1963).

5. Conclusions and outlook

After selecting a reference 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 the daily rain rate event PDFs obtained by the SLM and Shepard's methods at selected, widely separated, areas of the Indian landmass, consisting of 2° × 2° square boxes within each all existing station data have been removed and corresponding rainfall values are inferred from the remaining stations. The associated PDFs are compared to the preexisting station data within each one of the boxes and in terms of the aggregated data from all the boxes (Figs. 4 and 3). This test revealed that the SLM method is superior to Shepard's method in terms of the daily rain rate event PDF accuracy. Shepard's method tends to underestimate the no rain and very light rain events

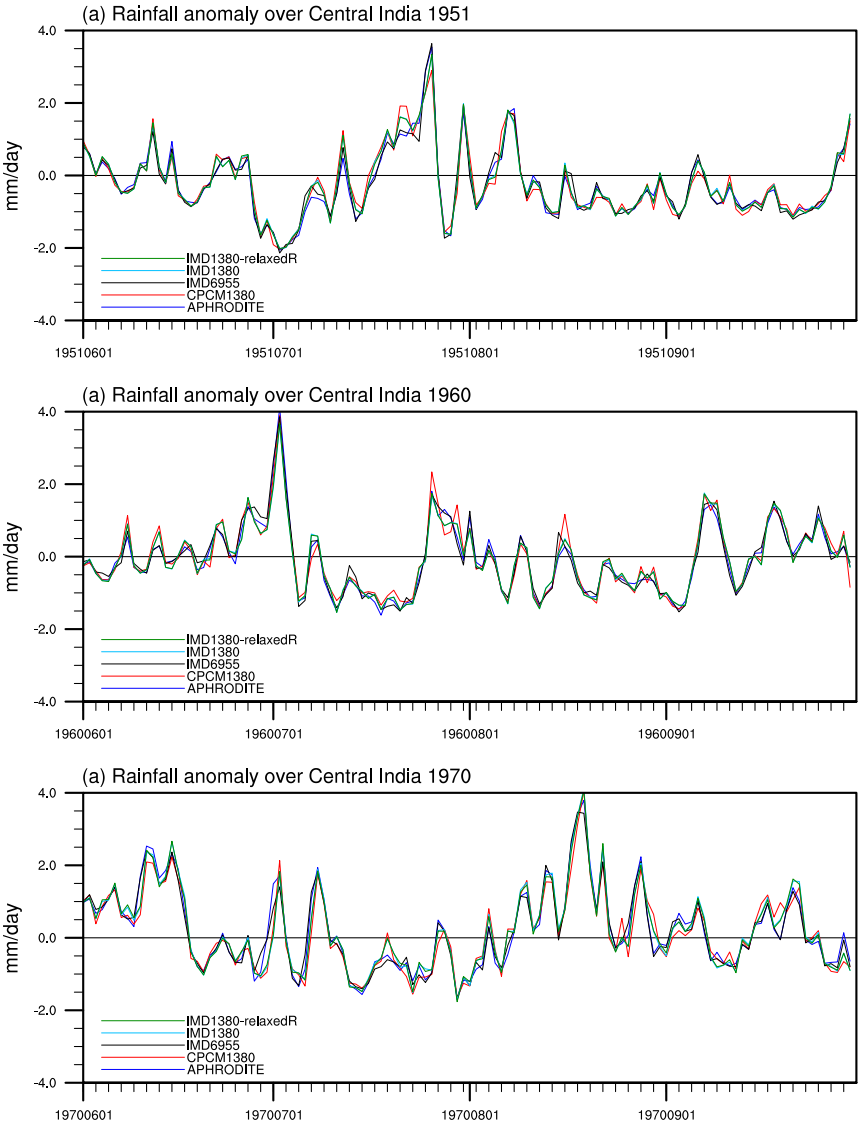

FIG. 10. Daily variation of the rainfall anomaly (mm day^{-1}) over central India (12°–25°N, 70°–90°E; black box in Fig. 1b) for the (a) 1951, (b) 1960, and (c) 1970 JJAS seasons.

of less than 1 mm day^{-1}, to underestimate the high rain events, greater or equal to 21 mm day^{-1}, and to overestimate light to moderate rain events between 2 and 21 mm day^{-1}.

Moreover, the statistical analysis of RMSE, ARE, and cross correlations with respect to the standard IMD6955 dataset as well as to APHRODITE revealed that the SLM is capable of producing a dataset that can outperform conventional methods in sparse station regions. The interannual and daily spatial means comparisons, on the other hand, showed that the SLM-based product is more in line with the IMD6955 and IMD1830 products than is the APHRODITE product, which especially shows a systematic underestimation bias in terms of the annual rainfall while the IMD1830-relaxedR has an overestimation bias. It is interesting to note that the smallest errors are associated with IMD6955 versus IMD1380 inside the radius of influence while on the Indian subcontinent level, IMD1380 and CPCM1380 exhibit comparable errors. The same is true for the all-India errors of APHRODITE with respect to both IMD6955 and CPCM1380.

As demonstrated by the sensitivity tests in Table A1, besides the exhibited acceptable accuracy of the CPCM1380 dataset, generated over all-India, including low station density regions, there is a 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 may be 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 uncertainty 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 to understand the true meaning of 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 such a 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.

The strength of the new SLM method resides in its capability to combine the climatology on the continental scale and the nearest available station data to yield grid rain gauge data in widely distribution observation stations. However, care must be taken as the usefulness of this method in station sparse regions may be limited to regions of the globe that are characterized by large scale weather systems such as it happens during the monsoon season over India. The method may fail in regions with localized climatology—controlled by sea breeze and/or topography, for example—in the absence of nearby measurement stations as will do any other OA method. Nonetheless, the method is fairly flexible and can accommodate other types of measurement by introducing artificial stations to obtain acceptable values in such places, for example.

Finally, error bars can be produced by the SLM algorithm to account for sampling error. One can readily estimate an uncertainty due to the MCMC chain fluctuations themselves, but these are very restricted, as can be surmised from Fig. A1. The true error bars due to the sampling error of the Gibbs measure can be estimated by making ensemble runs where each ensemble member is drawn from using a different seed of the random number generator. Such error bars will be produced and assigned to the CPCM1830 product in the future and will be reported elsewhere by the authors.

Acknowledgments. 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 MM/SERP/NYU/2014/SSC-01/002). This research was initiated during a visit of BK to NYUAD during spring 2017. All data used herein are listed in the references. The high-resolution gridded rainfall data of India Meteorological Department (IMD6955) are now freely available to download from the IMD Pune website (http://www.imdpune.gov.in/Clim_Pred_LRF_New/Grided_Data_Download.html). The quality controlled daily station rainfall data of 1380 stations over the Indian subcontinent are also collected from the NDC, India Meteorological Department, Pune, India (http://www.imdpune.gov.in/ndc_new/Request.html). The daily gridded APHRODITE data used in this study are retrieved from https://climatedataguide.ucar.edu/climate-data/aphrodite-asian-precipitation-highly-resolved-observational-data-integration-towards. The authors thank the editor and two anonymous reviewers for their time and valuable comments.

APPENDIX

Convergence of the MCMC Time Series and Sensitivity to Parameters of the SLM Scheme

As already mentioned, the MCMC algorithm consists of running an ergodic Markov process to equilibrium, whose equilibrium distribution is the one, one wishes to sample, and use the converged pseudotime 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 pseudotime accordingly. The results from this exercise led us to choose a conservative iteration time $T_0 = 24$ h. For the sake of illustration, we plot in Fig. A1 the MCMC pseudotime series corresponding to the lattice point with latitude–longitude coordinates 28°N, 80.75°E, and the day 19 July 1951, for six different bin sizes. As we can see from Fig. A1, after a transient period of up to ~3 h (reached roughly in 10 000 pseudotime steps with an average step size $s = 1.08$ s), the chains enter a statistical steady state where they fluctuate up and down within their stochastic variability range. As can be surmised from Fig. A1, 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 and the variability range is shorter for the larger number of bins (137). 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 10% of each chain as the

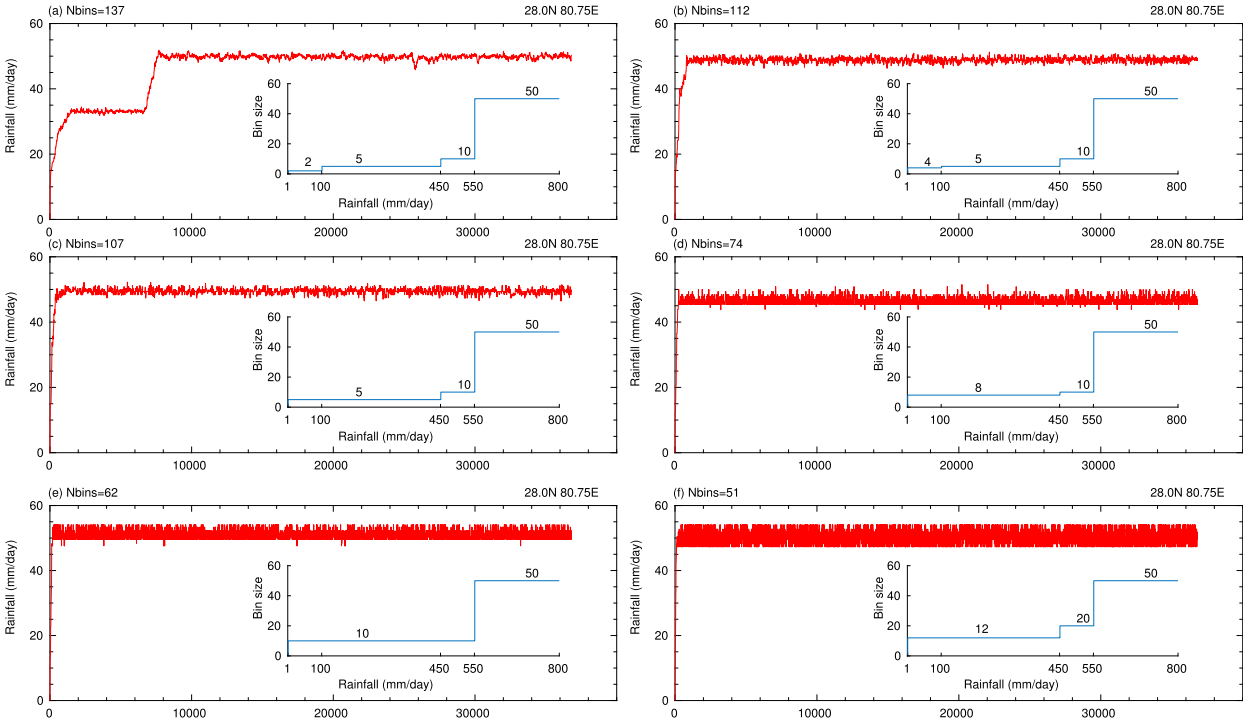

FIG. A1. (a)–(f) Convergence at latitude–longitude point 28°N, 80.75°E for the day 19 Jul 1951. The y axis represents rainfall (mm day^{-1}), and the x axis is the iteration count of the MCMC simulation over the pseudotime, from 0 to $T_0 = 24$ h. The broken blue curves in the middle of each panel represent 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 bins of size δ, where δ is indicated right on top of that segment. For the case of bin number $N = 51$ in (f), for example, the rainfall rate segment between 0 and 450 mm day^{-1} is roughly divided into 37 bins of the same size $\delta = 12$.

interpolated rainfall value at the corresponding lattice cell. To take full advantage of the stochastic nature of the scheme, the associated variances can also be recorded to provide some measure of uncertainty in the interpolated data. This will be done in the future.

To ensure the convergence of MCMC runs in our SLM scheme, we have also shown the 50th percentile of the last 100, 50, and 10 iterations for three different active days of precipitation (12 July 1956, 9 July 1958, and 25 August 1965) in Fig. A2. It is found that the 50th percentile of precipitation looks similar for the last 100, 50, and 10 iterations in all the three active days of monsoon, which again confirms that the MCMC runs in our SLM scheme satisfactorily converged.

Preliminary tests indicated that the scheme is most sensitive to the values of J_0 and the number of bins N. In Table A1, we report the RMSE between the interpolated and regridded rainfall data based on the SLM scheme, CPCM1380, and the high-resolution IMD6955 datasets for various values of J_0 and bin number N, integrated over the totality of the structured grid for the 1951 monsoon (JJAS) season. As we can see from this table, for a fixed J_0 the RMSE typically increases with decreasing bin number while its variation with respect to J_0 is more subtle. For a fixed bin number, 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 A1, $J_0 = 1.05$ and $N = 137$ seem to be an optimal choice in terms of minimizing the RMSE in comparison to the high-resolution IMD6955 datasets. 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 A1. 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 A1, which suggests that the scheme is robust and can 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}.

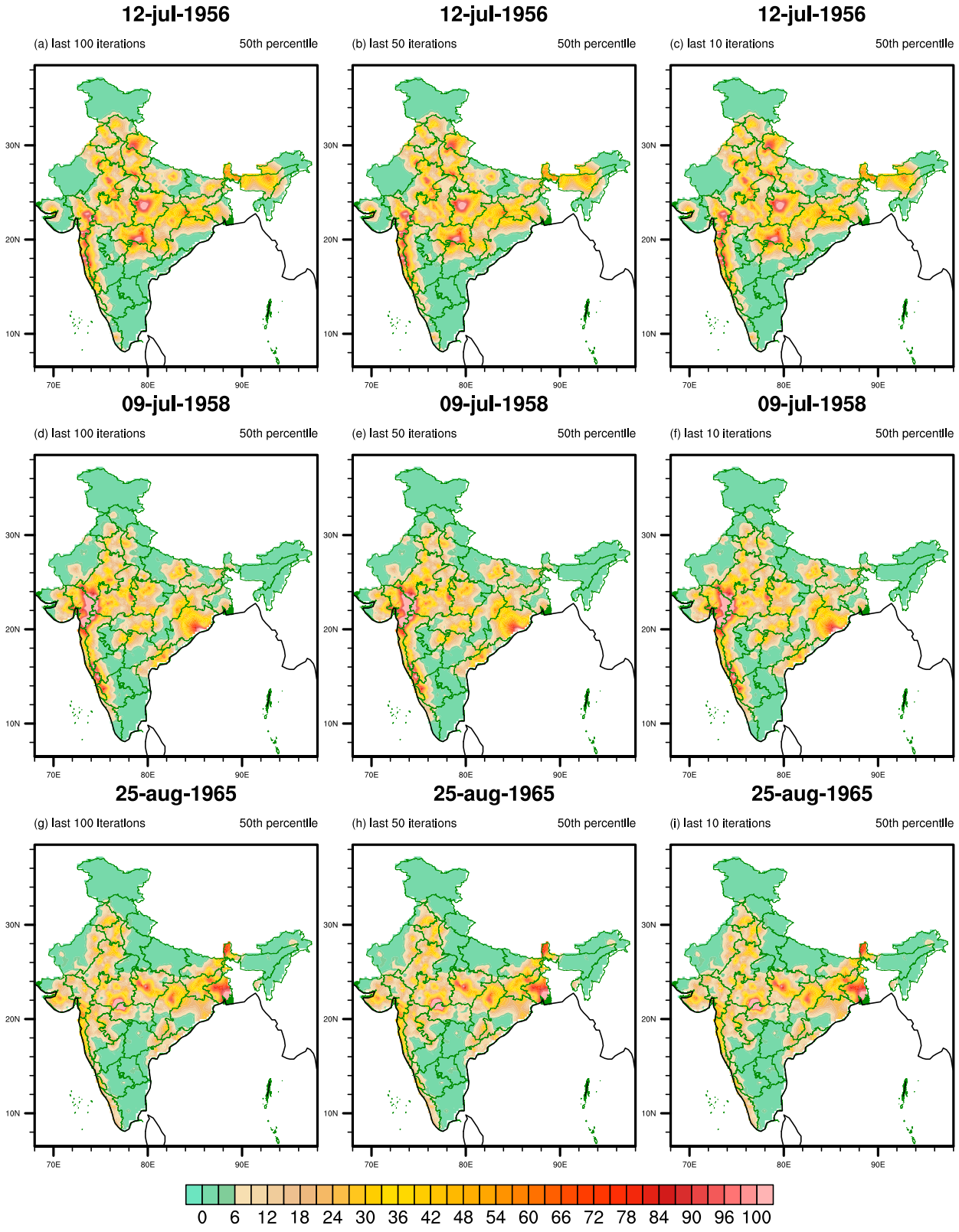

FIG. A2. The 50th percentile (mm day^{-1}) of last 100, 50, and 10 iterations for three active monsoon days (a)–(c) 12 Jul 1956, (d)–(f) 9 Jul 1958, and (g)–(i) 25 Aug 1965. Shown are the 50th percentile of the last (left) 100, (center) 50, and (right) 10 iterations.

TABLE A1. 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. The bold numbers indicate the optimal parameter regime.

	Bin No. N					
J_0 (day mm^{-1})	**137**	112	107	74	62	51
0.9	1.16	1.19	1.21	1.23	1.32	1.49
0.95	1.12	1.16	1.17	1.19	1.30	1.47
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	1.46
1.1	1.09	1.12	1.13	1.13	1.27	1.46
1.2	1.11	1.13	1.12	1.12	1.26	1.45
1.4	1.23	1.21	1.19	1.13	1.25	1.44
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	1.73	1.63	1.42	1.13	1.24	1.43
2.4	—	1.68	1.44	1.14	1.24	1.44
2.5	—	1.70	1.47	1.14	1.25	1.44

REFERENCES

Adler, R. F., and Coauthors, 2003: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.*, **4**, 1147–1167, https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

Barnes, S. L., 1973: Mesoscale objective map analysis using weighted time-series observations. NOAA Tech. Memo. ERL NSSL-62, 60 pp., http://docs.lib.noaa.gov/noaa_documents/OAR/NSSL/NOAA_TM_ERL_NSSL/ERL_NSSL_62.pdf.

Bastin, G., B. Lorent, C. Duqué, and M. Gevers, 1984: Optimal estimation of the average areal rainfall and optimal selection of rain gauge locations. *Water Resour. Res.*, **20**, 463–470, https://doi.org/10.1029/WR020i004p00463.

Bussières, N., and W. Hogg, 1989: The objective analysis of daily rainfall by distance weighting schemes on a mesoscale grid. *Atmos.-Ocean*, **27**, 521–541, https://doi.org/10.1080/07055900.1989.9649350.

Chen, M., P. Xie, J. E. Janowiak, and P. A. Arkin, 2002: Global land precipitation: A 50-yr monthly analysis based on gauge observations. *J. Hydrometeor.*, **3**, 249–266, https://doi.org/10.1175/1525-7541(2002)003<0249:GLPAYM>2.0.CO;2.

——, W. Shi, P. Xie, V. B. Silva, V. E. Kousky, R. W. Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.*, **113**, D04110, https://doi.org/10.1029/2007JD009132.

Collier, C., 1986: Accuracy of rainfall estimates by radar, Part I: Calibration by telemetering raingauges. *J. Hydrol.*, **83**, 207–223, https://doi.org/10.1016/0022-1694(86)90152-6.

Cressman, G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev.*, **87**, 367–374, https://doi.org/10.1175/1520-0493(1959)087<0367:AOOAS>2.0.CO;2.

Foss, S., D. Korshunov, and S. Zachary, 2013: *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer, 157 pp.

Gandin, L. S., 1963: *Objective Analysis of Meteorological Fields* (in Russian). Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), 242 pp.

Gervais, M., L. B. Tremblay, J. R. Gyakum, and E. Atallah, 2014: Representing extremes in a daily gridded precipitation analysis over the United States: Impacts of station density, resolution, and gridding methods. *J. Climate*, **27**, 5201–5218, https://doi.org/10.1175/JCLI-D-13-00319.1.

Gruber, A., X. Su, M. Kanamitsu, and J. Schemm, 2000: The comparison of two merged rain gauge–satellite precipitation datasets. *Bull. Amer. Meteor. Soc.*, **81**, 2631–2644, https://doi.org/10.1175/1520-0477(2000)081<2631:TCOTMR>2.3.CO;2.

Hartmann, D. L., and M. L. Michelsen, 1989: Intraseasonal periodicities in Indian rainfall. *J. Atmos. Sci.*, **46**, 2838–2862, https://doi.org/10.1175/1520-0469(1989)046<2838:IPIIR>2.0.CO;2.

Herrera, S., S. Kotlarski, P. M. Soares, R. M. Cardoso, A. Jaczewski, J. M. Gutiérrez, and D. Maraun, 2019: Uncertainty in gridded precipitation products: Influence of station density, interpolation method and grid resolution. *Int. J. Climatol.*, **39**, 3717–3729, https://doi.org/10.1002/joc.5878.

Hofstra, N., M. New, and C. McSweeney, 2010: The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data. *Climate Dyn.*, **35**, 841–858, https://doi.org/10.1007/s00382-009-0698-1.

Huffman, G. J., and Coauthors, 1997: The Global Precipitation Climatology Project (GPCP) combined precipitation dataset. *Bull. Amer. Meteor. Soc.*, **78**, 5–20, https://doi.org/10.1175/1520-0477(1997)078<0005:TGPCPG>2.0.CO;2.

Katsoulakis, M. A., A. J. Majda, and D. G. Vlachos, 2003a: Coarse-grained stochastic processes and Monte Carlo simulations in lattice systems. *J. Comput. Phys.*, **186**, 250–278, https://doi.org/10.1016/S0021-9991(03)00051-2.

——, ——, and ——, 2003b: Coarse-grained stochastic processes for microscopic lattice systems. *Proc. Natl. Acad. Sci. USA*, **100**, 782–787, https://doi.org/10.1073/pnas.242741499.

Khouider, B., 2014: A coarse grained stochastic multi-type particle interacting model for tropical convection: Nearest neighbour interactions. *Commun. Math. Sci.*, **12**, 1379–1407, https://doi.org/10.4310/CMS.2014.v12.n8.a1.

——, J. Biello, and A. J. Majda, 2010: A stochastic multicloud model for tropical convection. *Commun. Math. Sci.*, **8**, 187–216, https://doi.org/10.4310/CMS.2010.v8.n1.a10.

Krishnamurthy, V., and J. Shukla, 2000: Intraseasonal and interannual variability of rainfall over India. *J. Climate*, **13**, 4366–4377, https://doi.org/10.1175/1520-0442(2000)013<0001:IAIVOR>2.0.CO;2.

——, and ——, 2007: Intraseasonal and seasonally persisting patterns of Indian monsoon rainfall. *J. Climate*, **20**, 3–20, https://doi.org/10.1175/JCLI3981.1.

——, and ——, 2008: Seasonal persistence and propagation of intraseasonal patterns over the Indian monsoon region. *Climate Dyn.*, **30**, 353–369, https://doi.org/10.1007/s00382-007-0300-7.

Lau, W. K.-M., and D. E. Waliser, 2012: *Intraseasonal Variability in the Atmosphere-Ocean Climate System*. Springer, 613 pp.

Lawrence, D. M., and P. J. Webster, 2002: The boreal summer intraseasonal oscillation: Relationship between northward and eastward movement of convection. *J. Atmos. Sci.*, **59**, 1593–1606, https://doi.org/10.1175/1520-0469(2002)059<1593:TBSIOR>2.0.CO;2.

Liggett, T., 1999: *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*. Grundlehren der Mathematischen Wissenschaften, Vol. 324, Springer, 335 pp.

Madden, R. A., and P. R. Julian, 1971: Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.*, **28**, 702–708, https://doi.org/10.1175/1520-0469(1971)028<0702:DOADOI>2.0.CO;2.

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*, **65** (1), 1–18.

Parthasarathy, B., and D. Mooley, 1978: Some features of a long homogeneous series of Indian summer monsoon rainfall. *Mon. Wea. Rev.*, **106**, 771–781, https://doi.org/10.1175/1520-0493(1978)106<0771:SFOALH>2.0.CO;2.

Prakash, S., A. Seshadri, J. Srinivasan, and D. Pai, 2019: A new parameter to assess impact of rain gauge density on uncertainty in the estimate of monthly rainfall over India. *J. Hydrometeor.*, **20**, 821–832, https://doi.org/10.1175/JHM-D-18-0161.1.

Rajeevan, M., and J. Bhate, 2009: A high resolution daily gridded rainfall dataset (1971–2005) for mesoscale meteorological studies. *Curr. Sci.*, **96**, 558–562.

——, ——, J. D. Kale, and B. Lal, 2006: High resolution daily gridded rainfall data for the Indian region: Analysis of break and active. *Curr. Sci.*, **91**, 296–306.

——, ——, and A. K. Jaswal, 2008: Analysis of variability and trends of extreme rainfall events over India using 104 years of gridded daily rainfall data. *Geophys. Res. Lett.*, **35**, L18707, https://doi.org/10.1029/2008GL035143.

Rudolf, B., H. Hauschild, W. Rueth, and U. Schneider, 1994: Terrestrial precipitation analysis: Operational method and required density of point measurements. *Global Precipitations and Climate Change*. M. Desbois and F. Désalmand, Eds., NATO ASI Series (Series I: Global Environmental Change), Vol. 26, Springer, 173–186.

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 Dyn.*, **49**, 3031–3050, https://doi.org/10.1007/s00382-016-3491-y.

Shepard, D., 1968: A two-dimensional interpolation function for irregularly-spaced data. *Proc. of the 1968 23rd ACM National Conf.*, ACM, New York, NY, 517–524, https://doi.org/10.1145/800186.810616.

Sikka, D., and S. Gadgil, 1980: On the maximum cloud zone and the itcz over Indian, longitudes during the southwest monsoon. *Mon. Wea. Rev.*, **108**, 1840–1853, https://doi.org/10.1175/1520-0493(1980)108<1840:OTMCZA>2.0.CO;2.

Suhas, E., J. Neena, and B. Goswami, 2013: An Indian Monsoon Intraseasonal Oscillations (MISO) index for real time monitoring and forecast verification. *Climate Dyn.*, **40**, 2605–2616, https://doi.org/10.1007/s00382-012-1462-5.

Thompson, C. J., 1971: *Mathematical Statistical Mechanics*. Macmillan, 278 pp.

Walker, G. T., 1910: On the meteorological evidence for supposed changes of climate in India. *Indian Meteor. Mem.*, **21** (Part I), 1–21.

Wang, B., P. Webster, K. Kikuchi, T. Yasunari, and Y. Qi, 2006: Boreal summer quasi-monthly oscillation in the global tropics. *Climate Dyn.*, **27**, 661–675, https://doi.org/10.1007/s00382-006-0163-3.

Wheeler, M. C., and K. M. Weickmann, 2001: Real-time monitoring and prediction of modes of coherent synoptic to intraseasonal tropical variability. *Mon. Wea. Rev.*, **129**, 2677–2694, https://doi.org/10.1175/1520-0493(2001)129<2677:RTMAPO>2.0.CO;2.

——, and H. H. Hendon, 2004: An all-season real-time multivariate mjo index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932, https://doi.org/10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

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. *Amer. Cartogr.*, **12**, 5–16, https://doi.org/10.1559/152304085783914686.

Xie, P., and P. A. Arkin, 1995: An intercomparison of gauge observations and satellite estimates of monthly precipitation. *J. Appl. Meteor.*, **34**, 1143–1160, https://doi.org/10.1175/1520-0450(1995)034<1143:AIOGOA>2.0.CO;2.

——, and ——, 1996: Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J. Climate*, **9**, 840–858, https://doi.org/10.1175/1520-0442(1996)009<0840:AOGMPU>2.0.CO;2.

——, and ——, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.*, **78**, 2539–2558, https://doi.org/10.1175/1520-0477(1997)078<2539:GPAYMA>2.0.CO;2.

——, B. Rudolf, U. Schneider, and P. A. Arkin, 1996: Gauge-based monthly analysis of global land precipitation from 1971 to 1994. *J. Geophys. Res.*, **101**, 19 023–19 034, https://doi.org/10.1029/96JD01553.

Yasunari, T., 1980: A quasi-stationary appearance of 30 to 40 day period in the cloudiness fluctuations during the summer monsoon over India. *J. Meteor. Soc. Japan Ser. II*, **58**, 225–229, https://doi.org/10.2151/JMSJ1965.58.3_225.

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, https://doi.org/10.2151/SOLA.2005-050.

——, 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. *Bull. Amer. Meteor. Soc.*, **93**, 1401–1415, https://doi.org/10.1175/BAMS-D-11-00122.1.

Zhang, C., 2005: Madden-Julian Oscillation. *Rev. Geophys.*, **43**, RG2003, https://doi.org/10.1029/2004RG000158.