To be submitted to BAMS

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

Boualem Khouider

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

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

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

Andrew J. Majda

Department of Mathematics & Center for Atmosphere and Ocean Sciences,

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

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

D. S. Pai

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

M. Rajeevan

Ministry of Earth Sciences, New Delhi, India.

*

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 in order to obtain an adequately distributed dataset. Objective interpolation methods such as the inverse-distance weighting (IDW) are the most widely used and have been employed to produce some of the most used gridded rainfall datasets (e.g. Climate Prediction Center Merged Analysis of Precipitation, Global Precipitation Climatology Project, and the India Meteorological Department). Unfortunately, the skill of these techniques becomes very limited to non existent in areas located far away from existing recording stations. This is problematic as many areas of the world are lacking an adequate rain gauge coverage throughout the recording history. Here, we introduce a new probabilistic interpolation method in an attempt to address this issue. The new algorithm employs a multitype particle interacting stochastic lattice model which assigns a binned rainfall value, from an arbitrary number of bins, to each lattice site or grid cell, with a certain probability according to the rainfall amou nts assigned to neigghbouring sites and a background climatological rainfalml distribution, drawn from the available data. Grid points containing recording stations are not affected and are being used as "boundary" input conditions by the stochastic model. The new stochastic model is successfully tested and validated against two standard gridded rainfall datasets, over the Indian land mass.

1. Methods

a. Pre-defining a grid array

We defined a grid array over the Indian subcontinent containing M number of triangular mesh (In our analysis, we consider M=11921 which is approximately equal to 0.25 resolution).

Let I be the index of arbitrary grid cell. I=1,2,3,4,.....M

First, at each time t read the observed station rainfall data and then it is assigned to each triangular cell. If there is more than one station in a triangular cell, then it is averaged out.

At each time t define a mask M_I such that

M_I=1 if there is station data

M_I=0 if there is no station data

These stations rainfall events over the Indian subcontinent, binned into N rain rates. Here the binning is as follows.

b. Binning

Rainfall below 1 mm/day is categorized into one bin and rainfall above 800 mm/day is categorized into another bin. Rainfall between 1 mm/day and 450 mm/day is binned into different bins with a bin size of 5 mm/day. Similarly, rainfall between 450 mm/day to 550 mm/day is binned with a bin size of 10 mm/day and rainfall between 550 mm/day to 800 mm/day is binned using the bin size of 50 mm/day. So the total number of bin is, N=107. Then the probability of occurrence at each bin is computed.

Let ρ_j be the probability of occurrence of rain rates, where $j = 0, 1, 2, ...N-1$. We constraint ρ_j to the observed rainfall over the Indian subcontinent.

c. Lattic Method

One can think of lattice as containing particle. Various numbers of particles are contained at different site. At any given time t, each lattice site is either occupied a rainfall at a particular bin or there is no rainfall at all. We assume σ_I^t be an order parameter that takes values 0,1,.....N-1 on a lattice site (I $\in \mathscr{L}$) according to whether there is a rain events of R_j, j=0,1,2,.....N-1 occurring at site I and at time t

Now compute the Hamiltonian masked energy differences $\triangle_+^I \tilde{H}(\sigma)$ and $\triangle_-^I \tilde{H}(\sigma)$ at each lattice site including where rainfall data is available according to its nearest neighbors.

We have

$$\triangle_+^I \tilde{H}(\sigma) = J_0[max(|\sigma_I + 1 - \sigma_J|) - max(|\sigma_I - \sigma_J|)] + h$$

$$\triangle_-^I \tilde{H}(\sigma) = J_0[max(|\sigma_I - 1 - \sigma_J|) - max(|\sigma_I - \sigma_J|)] - h$$

Where σ_I represent the order parameter at lattice site I and σ_J represents the same at its neighboring sites. Here $J_0 > 0$, which represent the strength of local interactions. Here we chose $J_0 = 8$ an ideal choice for the better result. We also tried with different values of J_0, For examples $J_0 = 3$ and $J_0 = 10$. Increasing the J_0 values will give less weight to the climatological equilibrium distribution. Here h is ignored for the time being and we set it to 0.

After computing the masked energy difference $\triangle_+^I \tilde{H}(\sigma)$ and $\triangle_-^I \tilde{H}(\sigma)$ at each grid cell, we compute rates $R_+^I(\sigma), R_-^I(\sigma)$ at each cell

$$R_+^{I,j}(\sigma) = (1 - M_I)\tilde{R}_+^I e^{(-1/2)\triangle_+ H(\sigma)} + M_I \Gamma(max(e^{-\beta(\sigma_I - \sigma_I^\star)}, 1.0) - 1.0)/\tau$$

$$R_-^{I,j}(\sigma) = (1 - M_I)\tilde{R}_+^I e^{(-1/2)\triangle_- H(\sigma)} + M_I \Gamma(max(e^{\beta(\sigma_I - \sigma_I^\star)}, 1.0) - 1.0)/\tau$$

Here β is a positive parameter, and we set it to 4.0. M_I represent the mask of available data stations and it may vary with time. $\sigma_I^\star$ = represents the observed station rainfall data at each lattice site I.

Here the background rates $\tilde{R}_+^j$ and $\tilde{R}_-^j$ can be computed from the $\rho_j \rightarrow$, probability of occurrence of rain rate using the following equations.

$\tilde{R}_+^j = (1/\tau)((\rho_j + 1)/\rho_j), j = 0,1,2,...N-1$

$\tilde{R}_-^j = (1/\tau)$, j=1,2,....N

where τ are arbitrary time scale and we set it to 5 hours in our analysis. We define ρ_j as the equilibrium distribution when the Hamiltonian energy difference is zero i.e when local interactions are ignored.

Now compute the total rate from all grid cells as $S_R = \Sigma(R_+^I(\sigma) + R_-^I(\sigma))$

To limit $R_-^I(\sigma) = 0$ if $\sigma_I = 0$

$R_+^I(\sigma) = 0$ if $\sigma_I = N$

Now sample a uniform random number U between 0 and 1 and let $s = -(1/(S_R))\log(U)$

The order parameter σ_I^t can jumb up by one unit or jumb down by one unit with some probabilities depending on whether its neighbors have more or less particle. A jump at a site I occurs according to Markov jump process.

$Prob\{\sigma_I^{t+\triangle t} = \sigma_I^t + 1\} = R_+^I(\sigma^t)\triangle t + o(\triangle t)$

$Prob\{\sigma_I^{t+\triangle t} = \sigma_I^t - 1\} = R_-^I(\sigma^t)\triangle t + o(\triangle t)$

To limit $0 \leq \sigma_I^t \leq N$:

$R_+^I(\sigma) = 0$ if $\sigma_I = N$

$R_-^I(\sigma) = 0$ if $\sigma_I = 0$

If $s \leq \triangle t$, (i.e during the time interval $(t, t+\triangle t)$, $\triangle t \rightarrow$ time remaining until next observations,), make a single transition at site I in the following way.

Number the rates $R_+^I(\sigma)$ and $R_-^I(\sigma)$ from 1 to 2M. Where M is the number of cells. e. g. $R_1, R_2.....R_{2M}$. Then define the probabilities $P_j = R_j/S_R$ and its cumulative sums $S_j = \sum_{k=1}^{j} P_k$, j=1,2,....2M

Pick a random number U^1 between 0 and 1 which is independent of U and find the first $S_{j0} \geq U^1$ and perform the transition associated with R_{j_0} such that

$\rightarrow$ if $R_{j_0} = R_+^{I_0}(\sigma)$, then $\sigma_{I_0} = \sigma_{I_0} + 1$ and $\rightarrow$ if $R_{j_0} = R_-^{I_0}(\sigma)$, then $\sigma_{I_0} = \sigma_{I_0} - 1$

d. Parameters used

Strength of local interactions. We tried with different values of J0, For examples, $J0 = 8.0$, $J0 = 10.0$ and $J0 = 3.0$. We found that $J0 = 8.0$ is an ideal choice. By increasing the $J0$ values, it will give less weight to the climatology, which means local interactions are significant.

Positive parameter $\beta = 4.0$

Time scale $\tau = 5hrs$

$\Gamma = 1.0$

$h = 0.0$

2. Unstructured cell to regular grid

Finally, we converted the sigma values back to the rainfall values by taking the middle value of bin (For example, if bin explain the rainfall between 5 mm/day and 10 mm/day we take it as 7.5 mm/day). Then the triangular cell output (unstructured) is converted to regular lat-lon grid (0.25°X 0.25°) using the bilinear interpolation.

For identifying our product from the other existing product we call it as CPCM rainfall product and its statistics are compared against 0.25°X 0.25°IMD rainfall (Pai et al. 2014) and APHRODITE rainfall (Yatagai et al. 2012) period 1951-1970.

3. Shepard interpolation method for high resolution rainfall

In order to compare the efficiency of our new Lattice model interpolation technique over the conventional shepard method, we also reproduced the high resolution (0.25x0.25) rainfall product (similar to (Pai et al. 2014)) using the inverse distance weighted shepard interpolation method but here we used less number of station data (1380 stations) as opposed to 6955 stations used in (Pai et al. 2014).

Using the Shepard method, the interpolated values at a grid node is computed from a weighted sum of the neighborhood observations. The interpolated values are more influenced by the nearby observational points and less influenced by the distant points. Following the previous study of (Pai et al. 2014), here we considered limited number of neighboring points (minimum 1 and maximum 4) within a search distance of 1.5 degree around the grid node where we want to compute the interpolated values.

For example, for the grid point P, the distance based weighting factors are defined as follows.

if the data point distance $d_i = 0$, we used the station data directly

if $0 < d_i < D_x/3$: weighting based on $1/d_i$

$D_{x/3} < d_i < D_x$: weighting based on $27((D_i/D_x) - 1)^2/4D_x$

$d_i > D_x$, station data not used. Here the D_x is the search radius.

In addition to the distance based weighting factor we also used a direction factor, which represented the shadowing of the influence of a data point from P by a nearer on in the same direction.

The directional weighting factor for the data point D_i near to the grid node P is defined as

$t_i = \sum S_j[1 - cos(D_iPD_j)]/\sum S_j$, where D_j represent data points near to grid point P except data point D_i

Therefore the new weighting function is

$W_i = (S_i)^2 X(1+t_i)$

Finally the interpolated value at grid node P is defined as $f(P) = \sum W_i Z_i / \sum W_i$ if $d_i \neq 0$

$f(P) = Z_i$ if $d_i = 0$

Since we used less number of stations (1380 stations) for the high resolution gridded datasets as opposed to 6955 stations used in (Pai et al. 2014), there are possibilities that no data points are found within the search distance D_x. Therefore, lot of missing points are noted in the final gridded product as opposed to (Pai et al. 2014). This limitations are covered in our Lattice based product.

Error estimate between each data products are computed from the following equation. Error between rainfall product 1 and 2 is calculated as

$$N_{12} = \sum_{x_i} \sum_{t_i} (2|R^1(x_i,t_i) - R^2(x_i,t_i)|)/(R^1(x_i,t_i) + R^2(x_i,t_i)) \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Here R^1 and R^2 represent rainfall of data product 1 and 2

4. Results

a. Parameter J0=8.0

Fig 1 left panel shows the spread of cells with rainfall ($\sigma_I^\star$) over the Indian subcontinent for the day 01-July-1951 and the right panel show the corresponding model output for the same day (σ). The model give the same value of observed rainfall over the cells where the rainfall data is available, and elsewhere the model give the rainfall value with some probabilities depending on whether its neighbors have more or less rainfall. From Fig 1 left panel it is quite clear that there is a large data gap in the northeast India, Jammu and Kashmir, and eastern coast of India. In spite of this, the model reasonably give the rainfall over this region.

Location of 1380 rain gauge stations are given in Figure 2. Colors indicate percentage of days with rainfall reported from the station (period considered 1951-2004). It is found that most of the stations have minimum 70% data availability during the analysis period 1951-1970.

Fig 3 shows day to day variation of number cells with rainfall. Out of total 11921 triangular cells over the Indian subcontinent, on an average around 1200 cells with rainfall is available. However, during the recent time there is a drop in the number of cells with rainfall.

In Figure 4 JJAS mean climatology of CPCM product (present rainfall product) is compared against the existing high resolution rainfall products (APHRODITE and IMD 0.25deg high resolution product). The heavy precipitation over the windward side of Western Ghats, copious rainfall over the central India are well captured in our new product. It is found that by using only less number of stations (around 1380), our product reasonably matches with other gridded rainfall products with a pattern correlation of 0.85 with both IMD high resolution datasets and APHRODITE dataset. Note that IMD used 6955 stations to produce 0.25°gridded rainfall product whereas here we used only maximum 1380 stations. Obviously the rainfall over the northeast India is slightly underestimated and it may be due to absence of sufficient number of stations over this region (only two stations are available in the present dataset) and we believe that this problem can be solved by updating the station data from IMD. However, the pattern correlation between CPCM product and TRMM3b42 is 0.80. Comparatively less correlation here may be due to different period used in TRMM3B42 to compute climatology (1998-2014) whereas in all other products we used a common period 1951-1970. By using the less number of stations the shepard interpolation technique is not able to reproduce the rainfall over the northwestern India and northeastern India Figure 4e. In the daily time series, number of missing points are shown in other regions also (see Figure 5).

The seasonal mean rainfall in all the three products are given in Table 1. It is found that seasonal mean is comparable in all products except APHRODITE where it is underestimated. Seasonal mean of our product is 872 mm which is very close to the IMD product where is 864 mm.

The standard deviation of seasonal mean rainfall shows an overall agreement between each data products (Figure 6). In APHRODITE the standard deviation over the central India and northeast India is weak compared to both IMD and CPCM product. The standard deviation of IMD 0.25deg product and our product (CPCM) matches very well.

The mean rainfall bias between each rainfall product is presented in Figure 7. In most of the Indian landmass the CPCM product estimate more precipitation than both APHRODITE and IMD. However, in northeast region, CPCM product estimate less rainfall compared to other two rainfall products. In the northeast region the number of stations are very less (only two), it may be reason for this less rainfall in CPCM product. It is also found that IMD product estimate more rainfall over most of the Indian landmass compared to APHRODITE. However, it underestimate orographic rainfall over the Western Ghats compared to APHRODITE.

The RMS error also show large difference between the CPCM product and other rainfall products over northeast India and western Ghats whereas this differences are quite negligible in other regions of Indian landmass (Figure 8). It is clear that there is always an uncertainty in the rainfall over the northeast India which is quite clear from the RMS difference of APHRODITE and IMD rainfall (Figure 8c)

The spatial pattern of temporal correlation between CPCM rainfall product and other products are generally high (more than 0.8). Over the central India the correlation between CPCM and APHRODITE exceed 0.9. However, in northeast India and Jammu and kashmir region the correlation value is less and sometime goes to negative correlation.

The interannual variation of all India summer monsoon rainfall (Rainfall averaged over the Indian subcontinent) is compared in Figure 10a. The inter-annual variability of IMD high resolution product and our present rainfall product matches very well both in terms of magnitude and phase (Figure 10a). This is reflected in the high correlation value (0.95) between these two gridded product. The magnitude of the rainfall time series derived from the APHRODITE is underestimated in all years compared to both IMD and CPCM gridded rainfall. However, in most years the time series in APHRODITE are in phase with the time series derived from other products (CPCM and IMD)) and therefore,the correlation between the APHRODITE and CPCM is 0.92.

Similarity, the interannual variability of central India summer monsoon rainfall is compared in Figure 10b. Compared to both IMD and APHRODITE, the rainfall time series of CPCM product have slightly higher values over the Central India in almost all years. However, it is is phase with the rainfall time series of both gridded IMD and APHRODITE product. The correlation value of rainfall time series of IMD with that of CPCM is 0.98. Similarly, the correlation value of rainfall time series of APRODITE with that of CPCM is also 0.98.

The daily variation of rainfall anomaly over the central India for five monsoon season (1951,1955,1960,1965,1970) are given in Figure 12. It is clear that the present rainfall product is quite good in capturing the signs of rainfall over the central India in agreement with other products for examples IMD and APHRODITE. In all the five monsoon season analyzed, the active and break phases identified are in good agreement with other products.

Acknowledgements

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

References

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

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

LIST OF TABLES

Rainfall product	Seasonal Mean
IMD 6955 statns	864 mm
APHRODITE	756 mm
CPCM 1380 statns	872 mm
TRMM3B42	887 mm
IMD 1380 statns	854 mm
IMD 1380 statns (RelaxedR)	920 mm

TABLE 1. Seasonal Mean Rainfall in different rainfall products.

Rainfall products	Error
IMD 6955 statns vs IMD 1380 statns(inside Rinf)	0.69
IMD 6955 vs IMD 1380 statns relaxedR (Global)	0.76
IMD 6955 vs IMD 1380 statns relaxedR (inside Rinf))	0.69
IMD 6955 vs IMD 1380 statns relaxedR (Outside Rinf))	1.14
IMD 6955 statns vs CPCM 1380 statns (Global)	0.83
IMD 6955 statns vs CPCM 1380 statns (inside Rinf)	0.80
IMD 6955 statns vs CPCM 1380 statns (outside Rinf)	0.96
Aphrodite vs IMD 6955 statns (Global)	0.88
CPCM 1380 statns vs Aphrodite (Global)	0.97

TABLE 2. Error estimate between each rainfall products calculated via eq (1).

LIST OF FIGURES

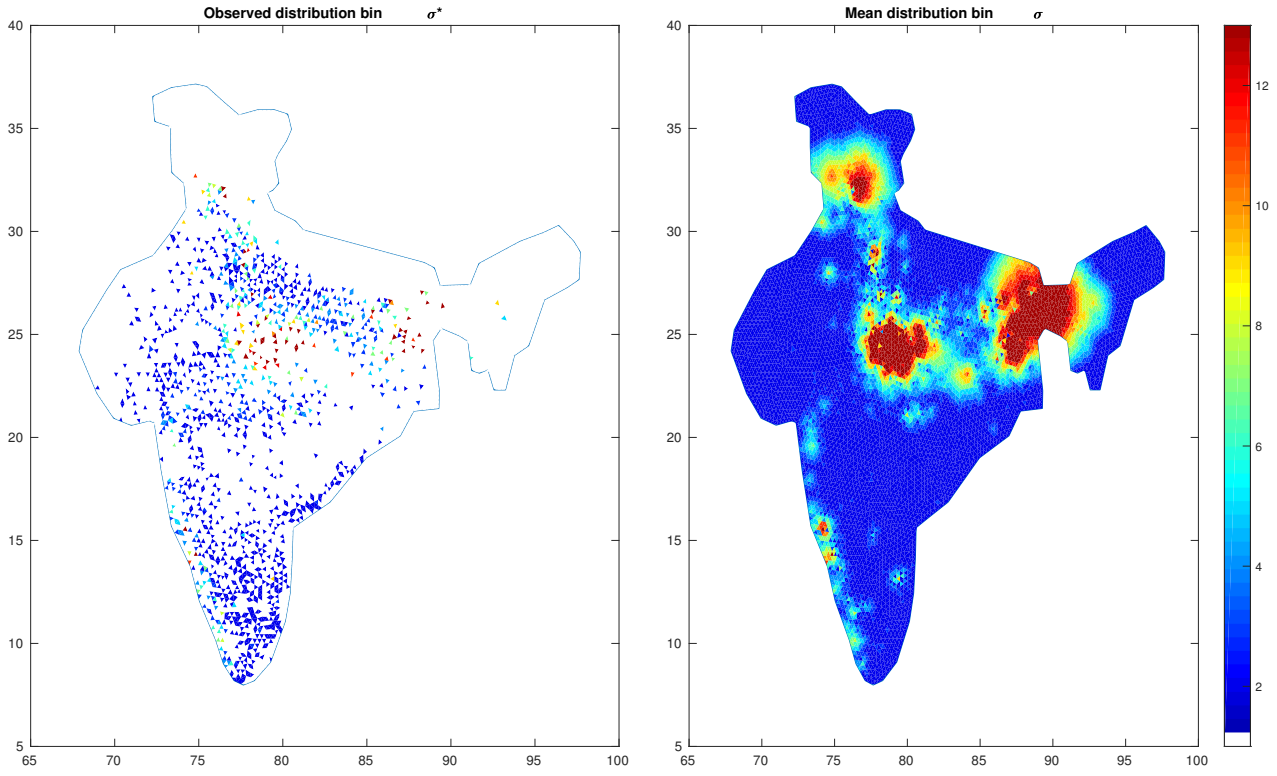


FIG. 1. Left panel: Observed sigmastar values at each stations for 02-July-1951. Right panel: Model value of sigma for the same day

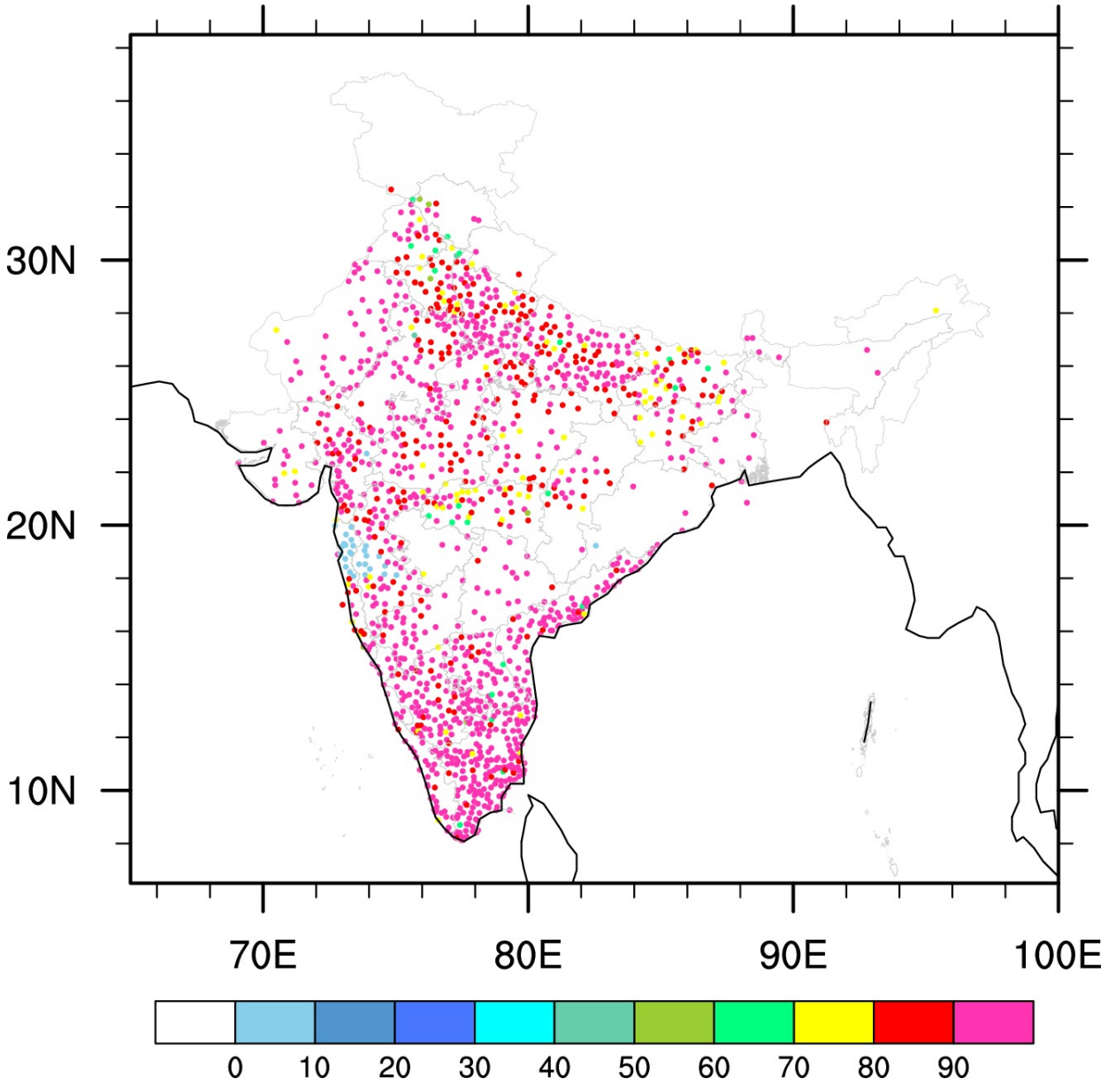


FIG. 2. Location of 1380 rain gauge stations. Colors indicate percentage of days with rainfall data.

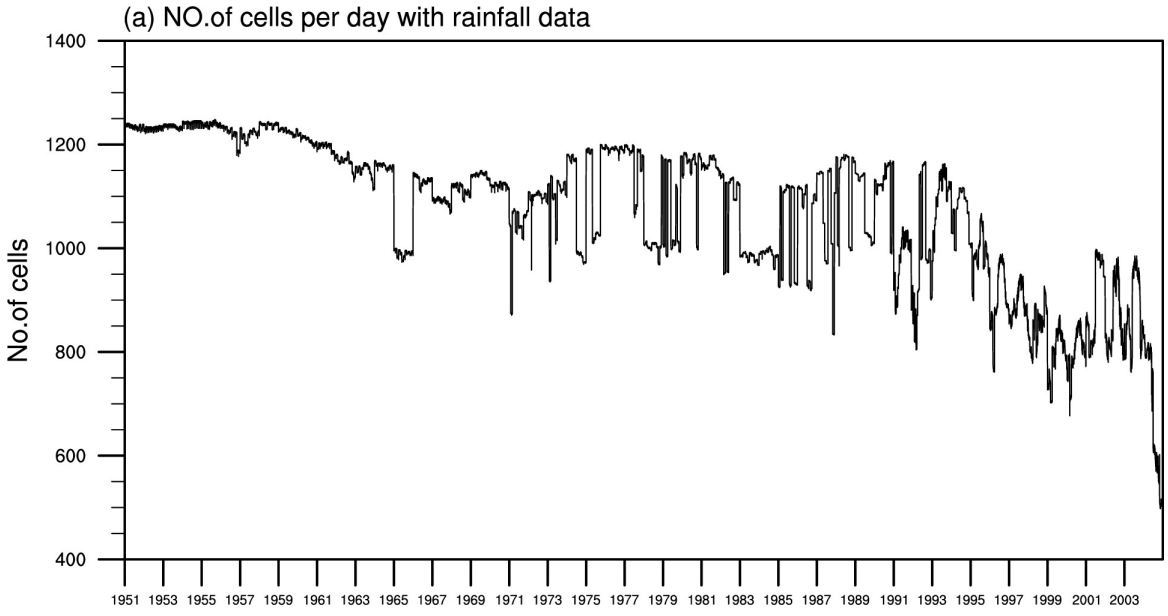


FIG. 3. Number of cells per day with rainfall. Total number of cells over the the Indian subcontinent is 11921 which is roughly equal to 0.25 deg spatial resolution.

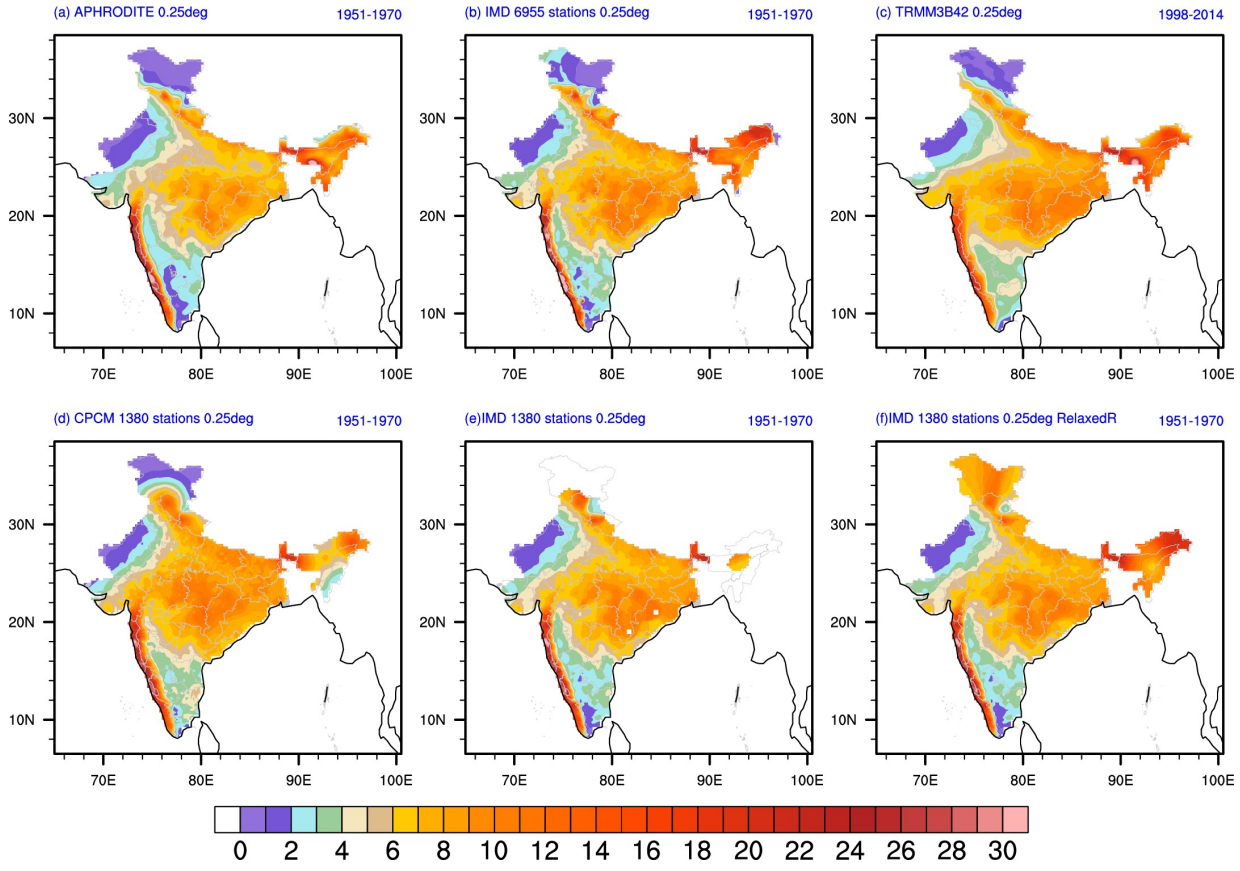


FIG. 4. Comparison of JJAS rainfall climatology (a) APHRODITE product (b) IMD 6955 station product (c) TRMM product (d) Our product (CPCM; 1380 stations) (e) IMD 1380 station product (f) IMD 1380 relaxed radius of influence station product. Units mm/day. JJAS climatology of TRMM is computed for the period 1998-2014. For all other products the climatology is computed for the period 1951-1970.

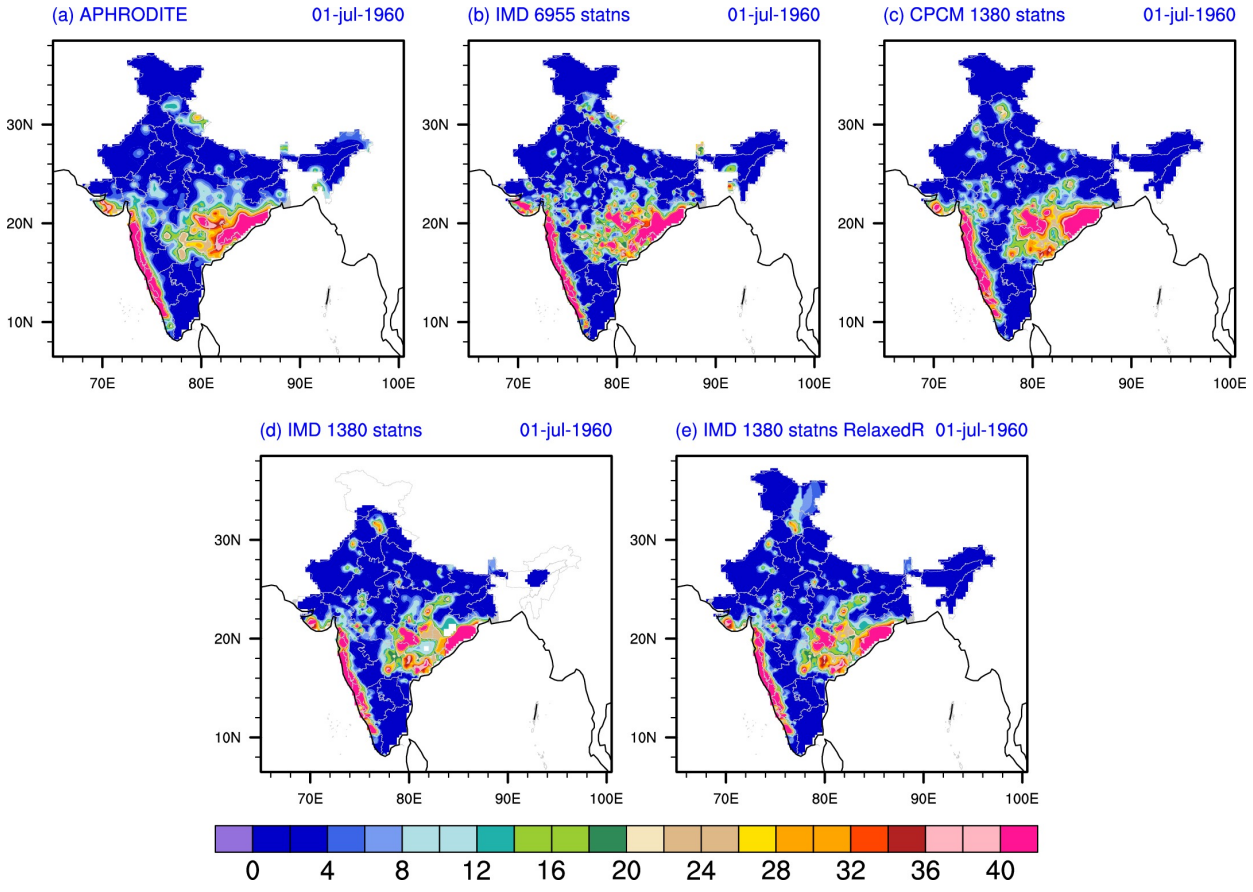


FIG. 5. Comparison of 01-July-1960 (a typical case) rainfall in different gridded products (a) APHRODITE product (b) IMD 6955 station product (c) CPCM 1380 station product (d) IMD 1380 station product (e) IMD 1380 relaxed radius of influence station product. Units mm/day.

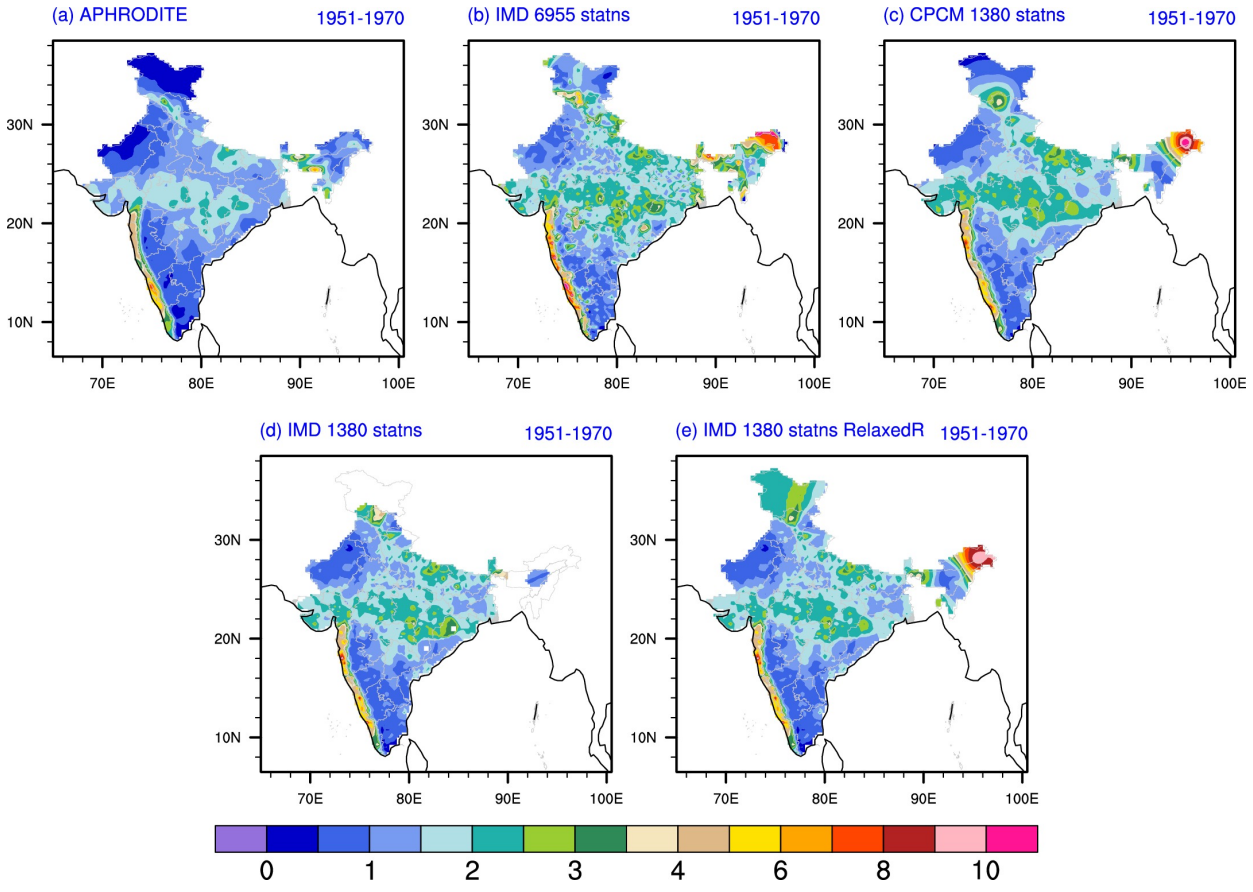
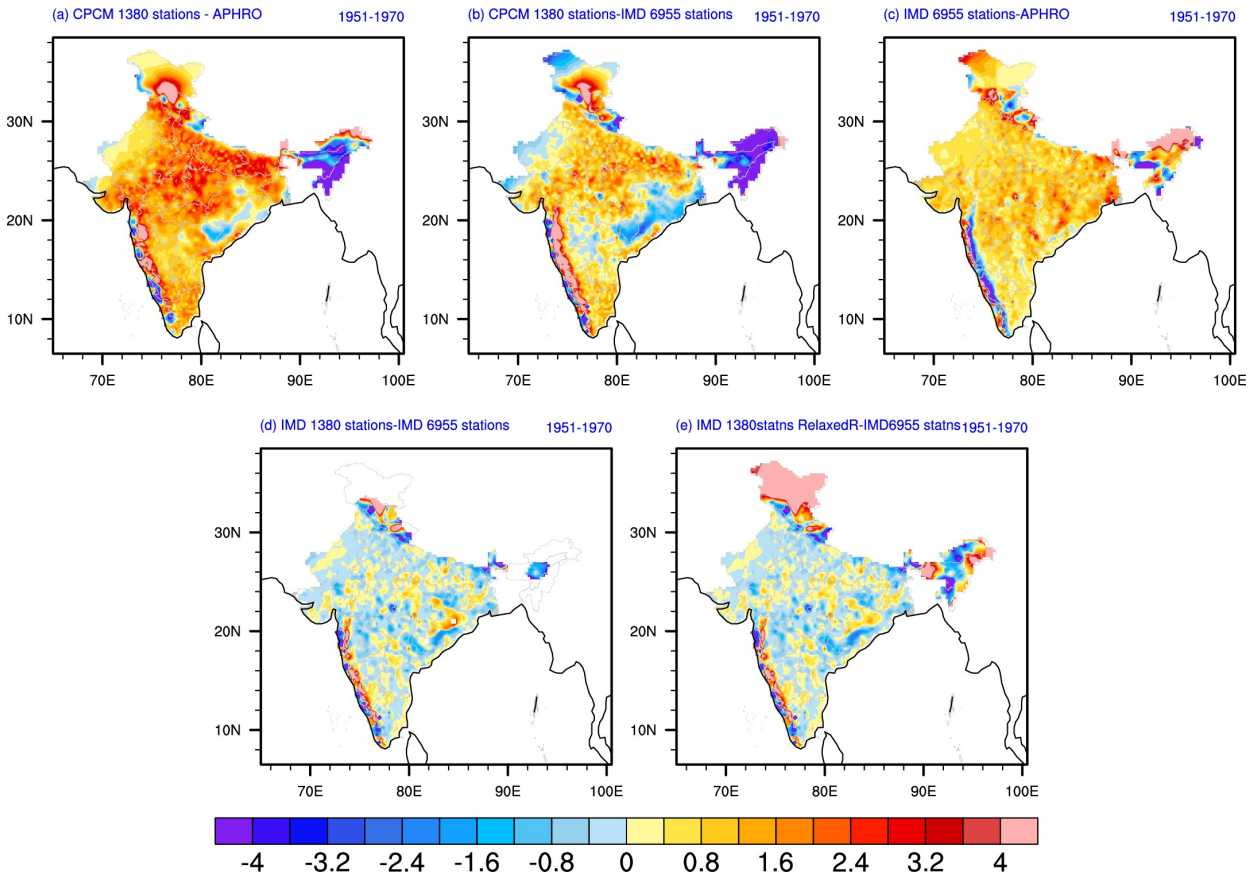


FIG. 6. Comparison of JJAS rainfall standard deviation (a) APHRODITE product (b) IMD 6955 station product (c) CPCM1380 product (d) IMD 1380 station product (e) IMD 1380 relaxed radius of influence station product. Units mm/day. Standard deviation is computed over the JJAS mean rainfall for the period 1951-1970.



309 FIG. 7. JJAS mean rainfall difference between (a) CPCM and APHRODITE product (CPCM-APHRODITE)
310 (b) CPCM and IMD 6955 station product (CPCM-IMD) (c) IMD 6955 station product and APHRODITE (IMD-
311 APHRODITE) (d) IMD 1380 station product and IMD 6955 station product (IMD 1380 station product-IMD
312 6955 station product) (e) IMD 1380 relaxed radius of influence station product and IMD 6955 station product
313 (IMD 1380 relaxed radius of influence station product-IMD 6955 station product) . Units mm/day. Difference
314 is computed over the JJAS mean rainfall for the period 1951-1970.

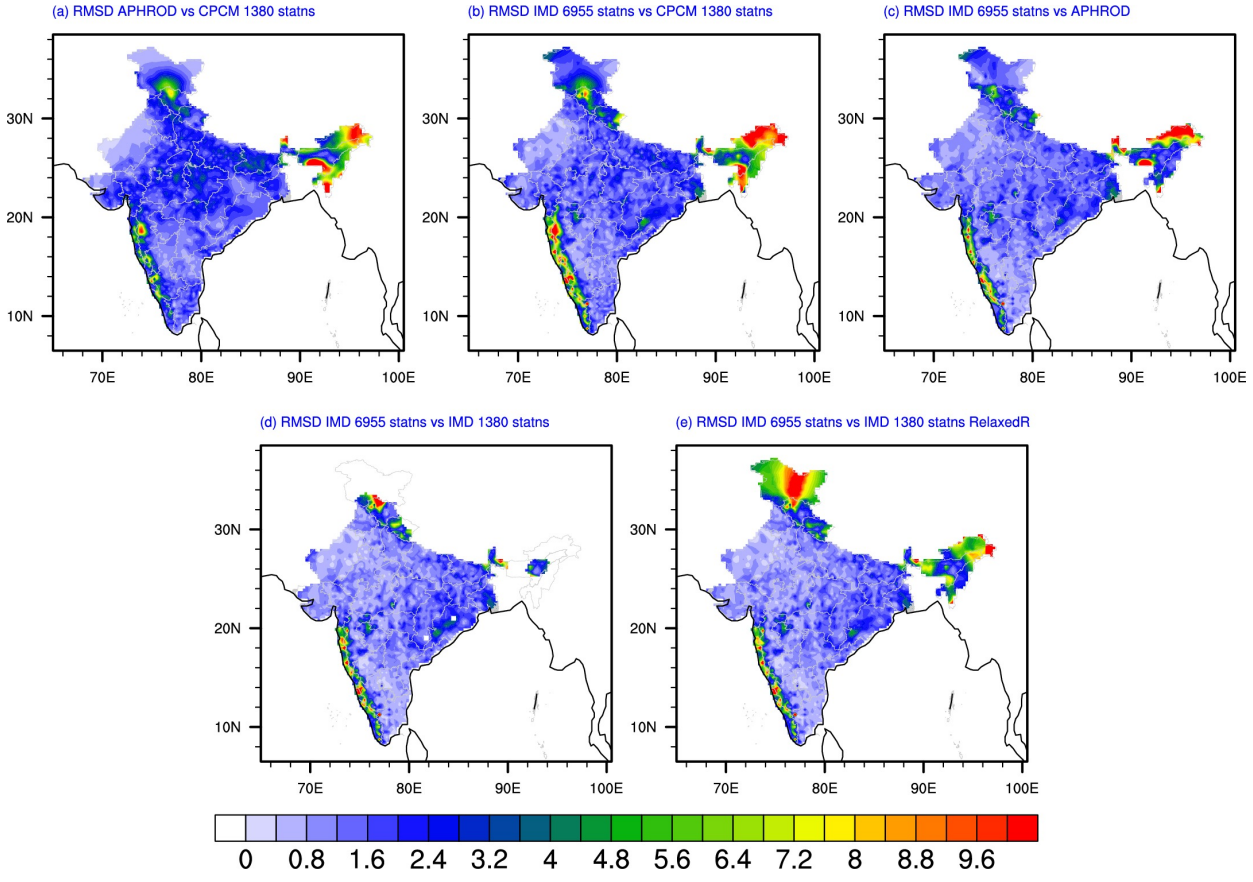


FIG. 8. (a) RMS deviation between APHRODITE product and our product (b) RMS deviation between IMD 6955 station product and our product (c) RMS deviation between IMD 6955 station product and Aphrodite (d) RMS deviation between IMD 6955 station product and IMD 1380 station product (e) RMS deviation between IMD 6955 station product and IMD 1380 relaxed radius of influence station product.

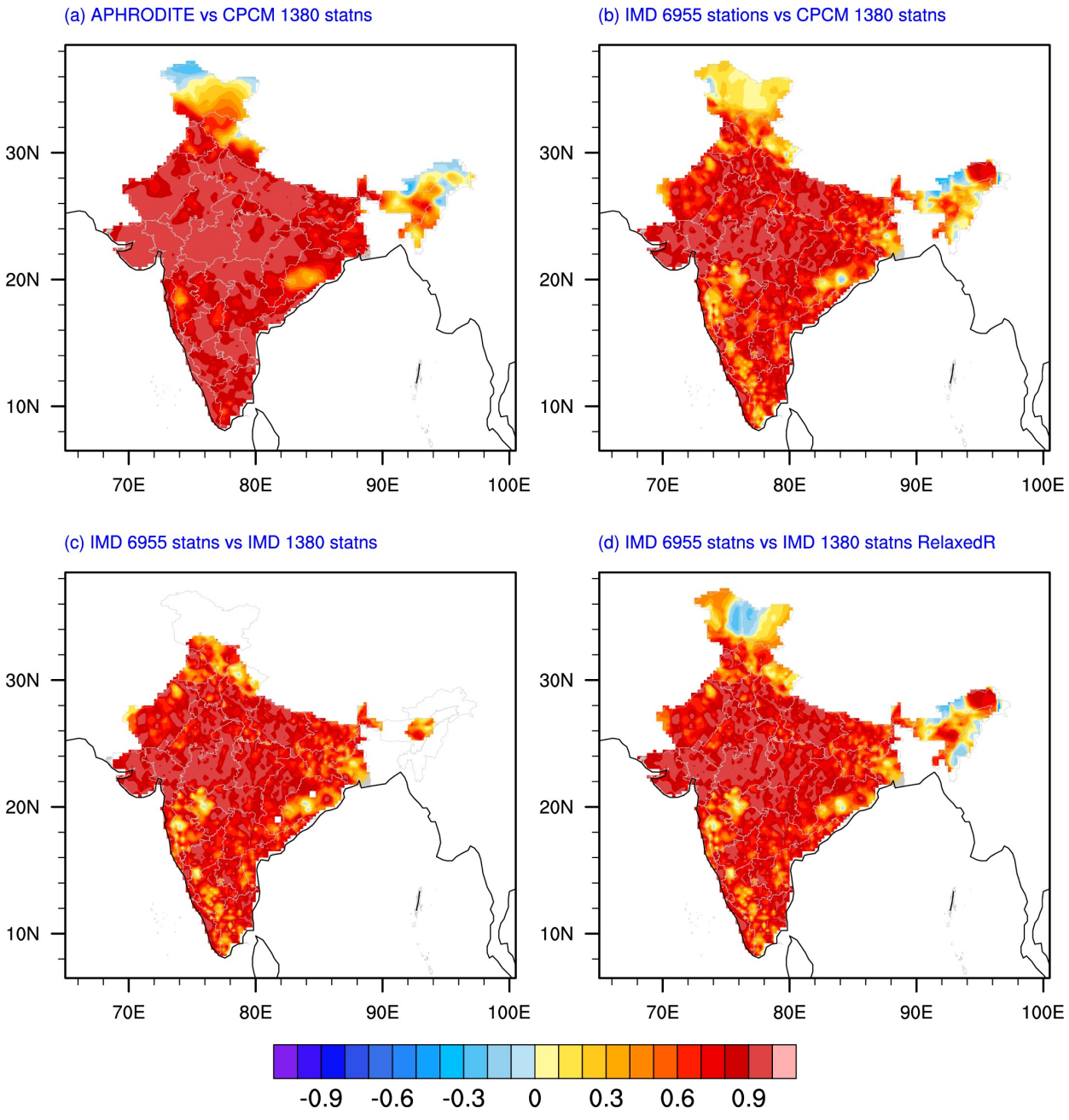


FIG. 9. (a) Grid point correlation of JJAS mean rainfall (a) APHRODITE rainfall vs CPCM1380 rainfall (b) IMD 6955 station product rainfall vs CPCM1380 rainfall (c) IMD 6955 station product rainfall vs IMD 1380 station product rainfall (d)IMD 6955 station product rainfall vs IMD 1380 relaxed radius of influence station product. Period of analysis: 1951-1970.

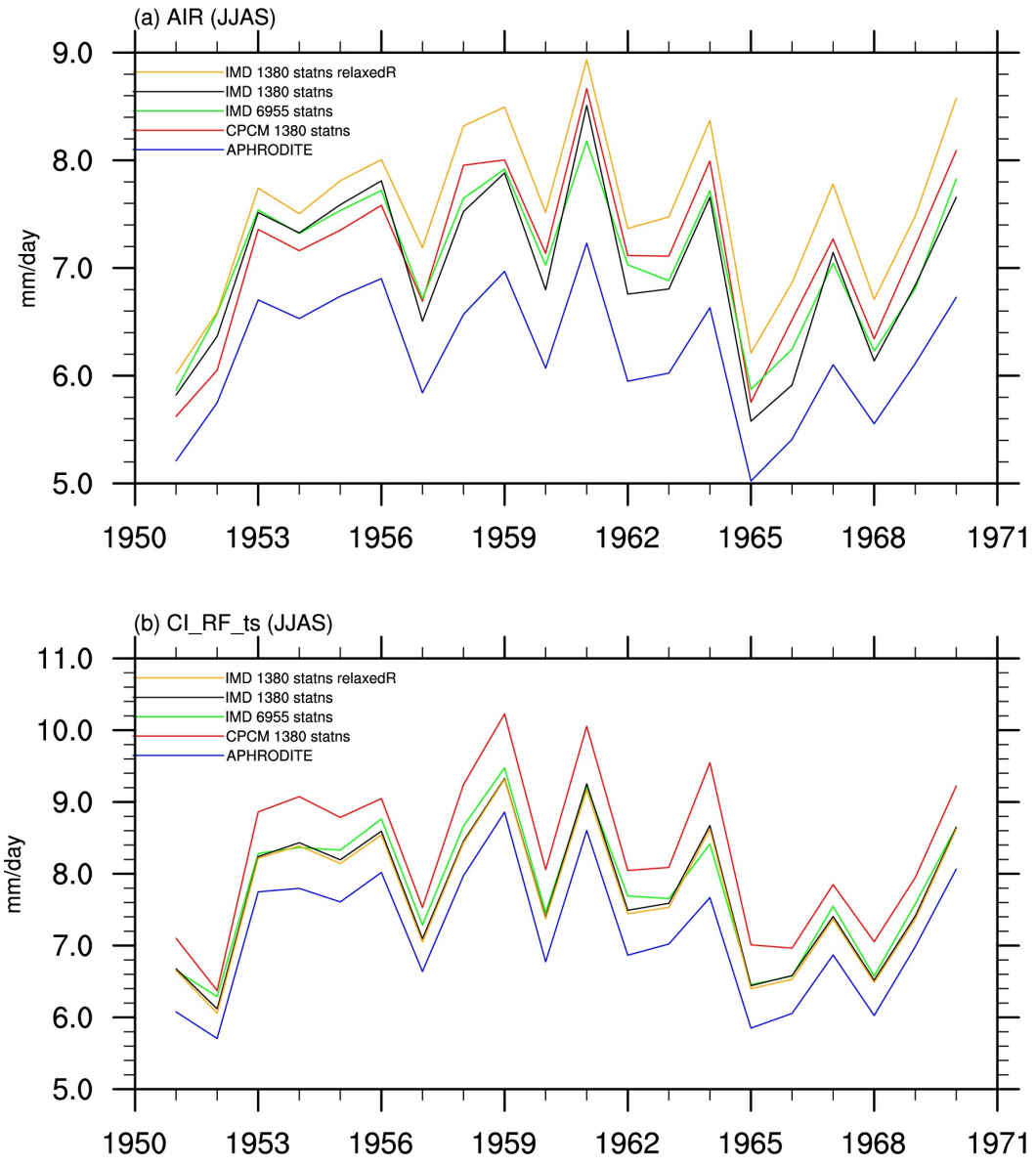


FIG. 10. (a) Interannual variation of All India summer monsoon rainfall (averaged over Indian landmass and averaged over JJAS season). IMD 6955 stations (Green) , APHRODITE (blue), CPCM 1380 station product (Red), IMD 1380 station product (black) IMD 1380 station RelaxedR (Orange) (b) Same as (a) but rainfall is averaged over Central India (12N-22N, 70E-90E).

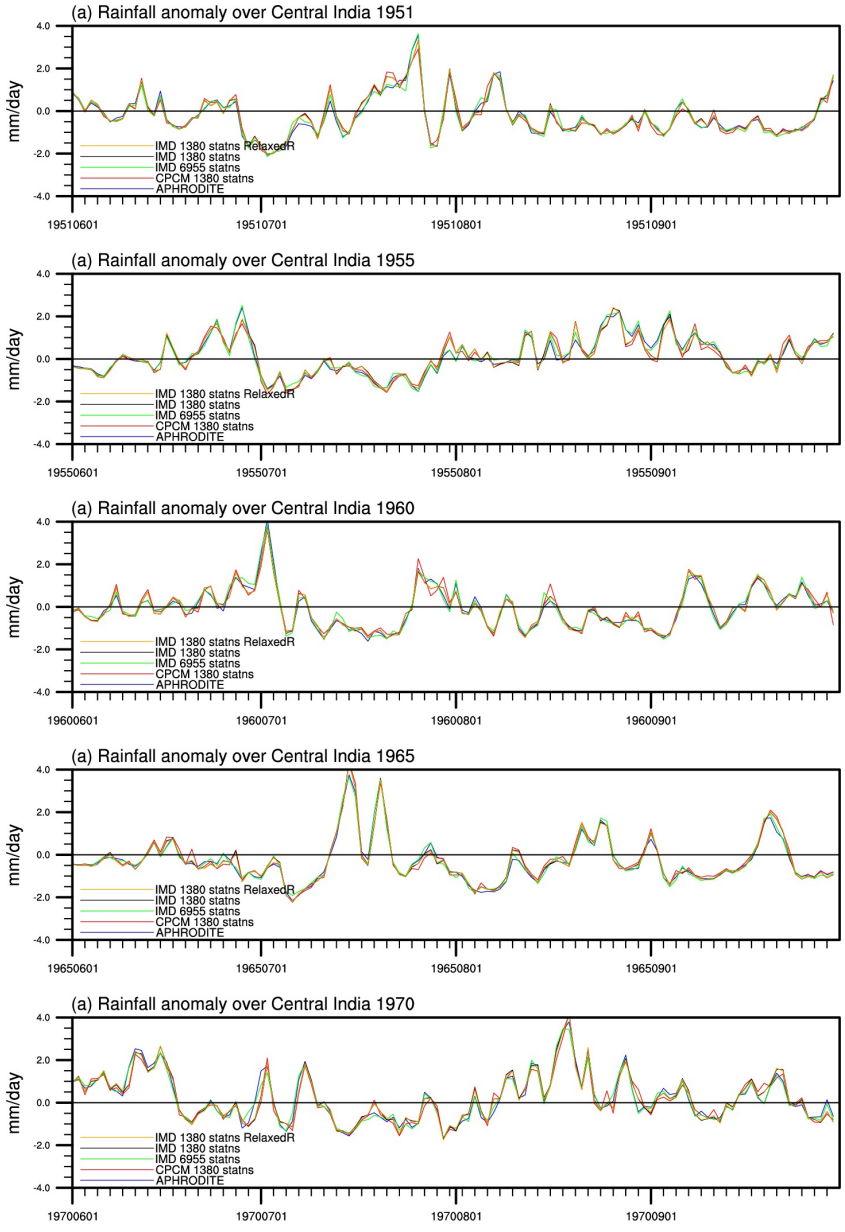
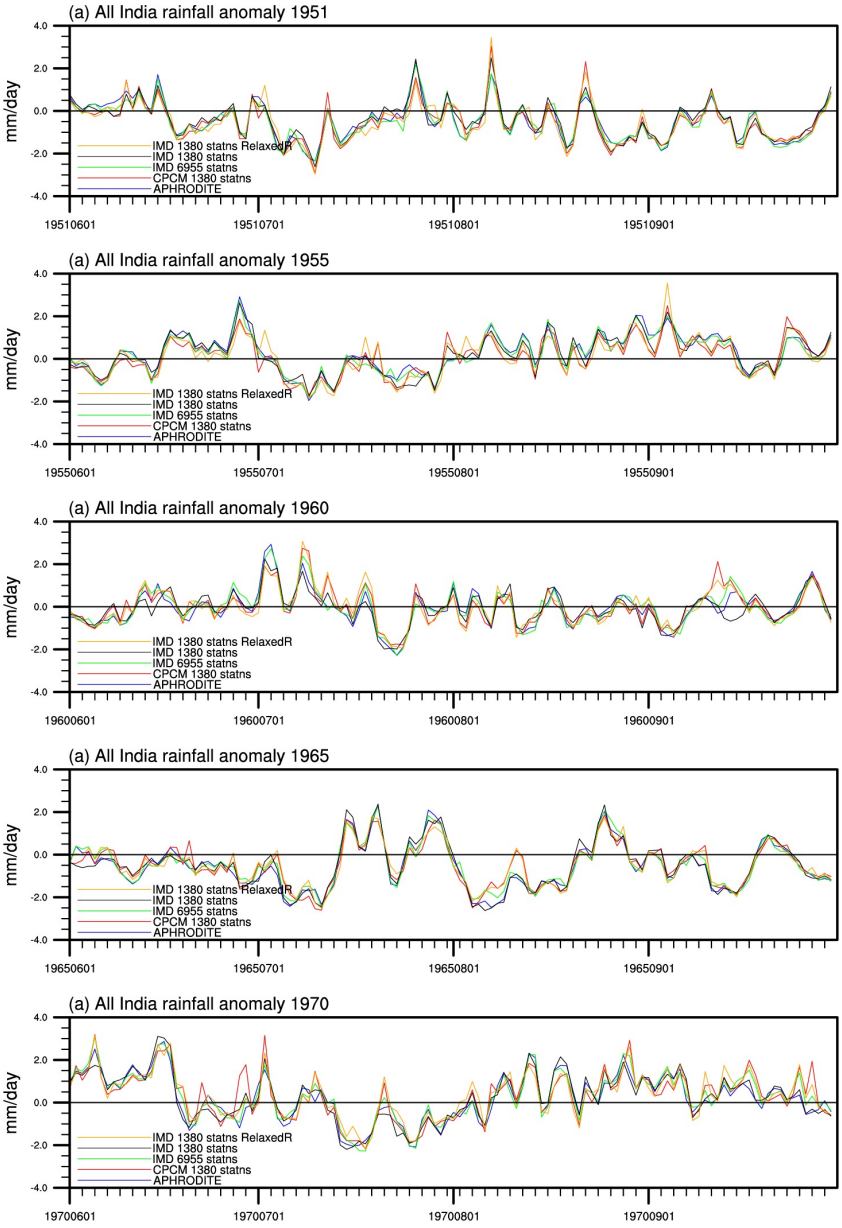


FIG. 11. Daily variation of rainfall anomaly over the Central India (a) for 1951 JJAS period (b) 1955 JJAS period (c) 1960 JJAS period (d) 1965 JJAS period (e) 1970 JJAS period.



329 FIG. 12. Daily variation of All India rainfall anomaly (a) for 1951 JJAS period (b) 1955 JJAS period (c) 1960
330 JJAS period (d) 1965 JJAS period (e) 1970 JJAS period.