Determination of secondary peaks in daily precipitation over the Central South Indian peninsula by time series analysis and tracking their causal rain systems (April to August)

Anindya Guria

February 2, 2023

Abstract

The presence of statistically significant maxims other than the monsoon peak in the daily precipitation over central south India (and its deviation from the national distribution) has been claimed and justified. The periodicity and auto-correlation of such a time series have been analyzed. A novel speed-adaptive tracking technique has been used over GPM IMERG half-hourly precipitation products to gather preliminary insight into the kinetics of systems that cause this anomaly, especially their frequency and origin. Results from this have been used to predict, with admittedly limited capacity, the possible reason behind the observed trend. We discussed further ways to follow up this study by using the general circulation of the atmosphere in the form of reanalysis data.

1 Introduction

This study has been motivated by a search for spatial and temporal periodicity and a general trend in the nature of off-monsoon rainfall in Bangalore and the adjacent central southern Indian peninsula[1]. This effort is partly based on Dr. Sukhatme's prior experience regarding the abnormal yearly rainfall pattern in Bangalore. There hasn't been much prior study regarding this, evident from the lack of literature available.

I have used IR and microwave combined rainfall products of TRMM [8] and GPM 3IMERGHH half-hourly data [9] for the time series analysis and precipitation tracking, respectively. These data sets provide wide area coverage with high spatial and temporal resolution compared to ground station-based data (For e.g., IMD data for this region has too many missing values).

[1]I will henceforth refer to this region as"Bangalore" comprising of a $2^\circ \times 2^\circ$ square ($76.5E$ to $78.5E$ and $12.5N$ to $14.5N$).

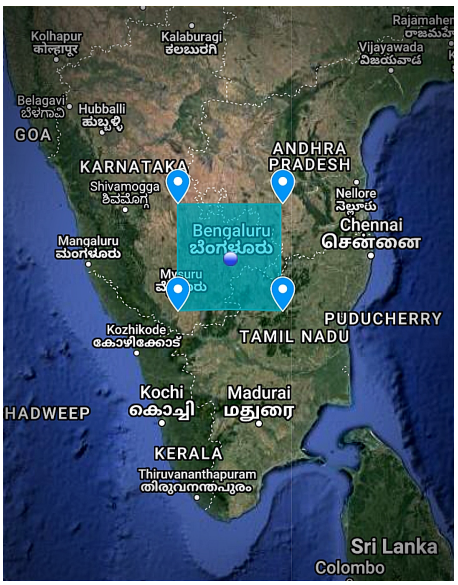


Figure 1: The highlighted area centered around Bangalore has been studied primarily

2 Time series of the rainfall data

The TRMM daily precipitation product over the period of 1998-2020 was used to obtain the daily rain over the Bangalore region as defined.

The TRMM daily netCDF4 files were read to form a 4-D grid ${d_{Yt}}^{xy}$. In this notation, Y is the year, t refers to the t^{th} day in a year, x is the longitude index, and y is the latitude index [2]
This was further summed over spatially and normalized, per $1^\circ \times 1^\circ$

$$\tilde{d}_{Yt} = \frac{1}{x_n y_m} \sum_{x=1}^{x_n} \sum_{y=1}^{y_m} {d_{Yt}}^{xy}$$

In the TRMM data set, the spatial resolution of the data was $0.1^\circ \times 0.1^\circ$. So $x_n = 20$ and $y_m = 20$.

[2]This notation will be used throughout.

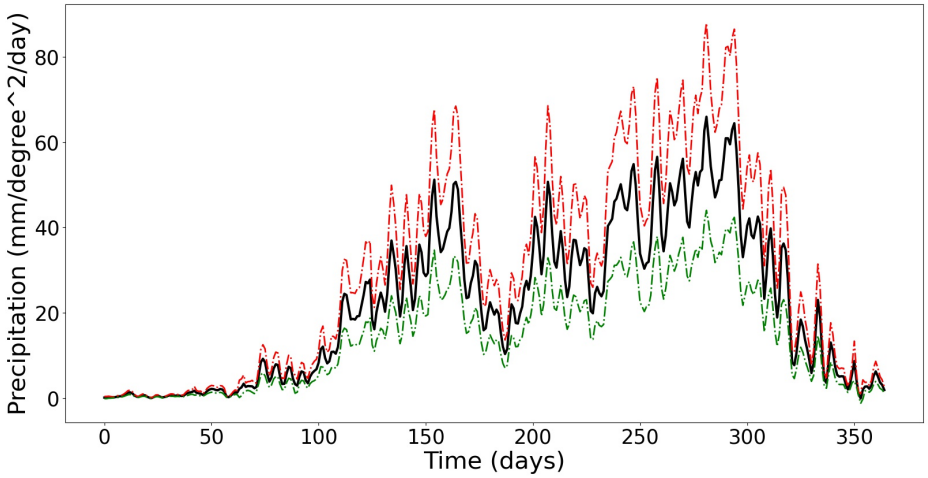


Figure 2: **Black curve:** the mean precipitation (μ_t), **Red curve:** the upper 90% and **Green curve:** lower 90% confidence interval (U_t and L_t respectively)

2.1 Noise removal

It was necessary to remove the high-frequency noise (fluctuations with a time period of fewer than three days) since we are interested only in the long-term trend. I implemented a Butterworth low pass filter at the 10^{th} order, with a cut of frequency $f_cut = 0.33 day^{-1}$.

$$\hat{d}_{Yt} = Butterworth_{\text{ order=10, } f_{cut}=0.33} \left(\tilde{d}_{Yt}\right)$$

For comparison a figure showing $\hat{d}_{2020,t}$ and $d_{2020,t}$ has been shown.

2.2 Hypothesis testing

I averaged over all the years taken for the data and upper and lower 90% confidence intervals. I assumed that for each day, the rainfall values were a sample from Gaussian random distribution (under the central limit theorem).

$$\mu_t = \langle \tilde{d}_{Y,t} \rangle$$

$$U_t = \mu_t + \frac{t_{0.05,22}}{\sqrt{2020-1998}} s_t$$

$$L_t = \mu_t - \frac{t_{0.05,22}}{\sqrt{2020-1998}} s_t$$

$$s_t = \frac{1}{2020-1998-1} \sum_{Y=1998}^{2020} \left(\tilde{d}_{Y,t} - \mu_t\right)^2$$

Here μ_t and s_t are the means and sample standard deviation of t^{th} day. U_t and L_t are the 90% upper and lower confidence intervals based on t-distribution.

μ_t,L_t and U_t are plotted. A quick visual inspection shows the presence of 3 distinct heavy rainfall periods. The t values at which the maxima occur are 154 (end May), 200 (mid-end July), and the highest peak at 286 (mid-September, as expected during monsoon).
To demonstrate that this observed pattern is different from the overall national rainfall trend, the corresponding mean precipitation of mainland India has been plotted along with the "Bangalore" average. Both have been re-scaled by appropriate factors to compensate for the different spatial areas.

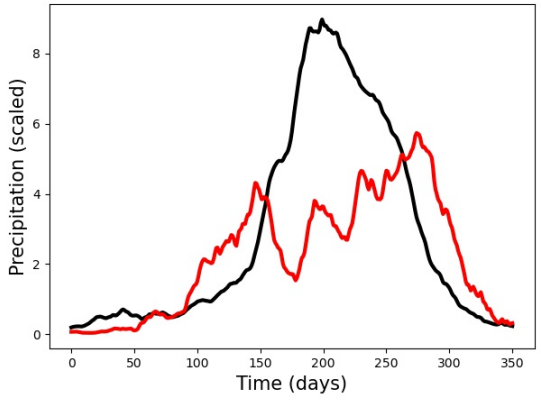


Figure 3: **Red curve:** is μ_t with a moving average of 15 days. **Black curve:** precipitation of mainland India (15 days moving average)

A 15 day moving average was added to both of

them so that the short-term fluctuations get suppressed. The three smaller peaks in Fig.2 are distinctly different from the national curve.
Based on all the data and plots, it is being claimed that there exist yearly peaks in Bangalore rainfall that doesn't quite overlap with the national pattern.

3 Fourier analysis and auto-correlation study of the time series

Once the secondary peaks are confirmed, it is of our interest to examine the dynamic properties of the events that lead to these deviations. In this section, the periodicity of the rain events and the auto-correlation of the events on subsequent events were the primary focus.

3.1 Power spectrum and absence of a regular periodicity

The first possibility was checked simply by a Fast Fourier Transform (FFT) of the data. The entire daily data set ${d_{Y_t}}^{xy}$ was used for this purpose, without the low-pass filter, so that the spectrum we obtain is representative of all the possible frequencies. I used Welch's [12] periodogram with the "hann" [5] windowing function. Two linear Welch periodograms have been shown; Figure 2 with 5-year segments and Figure 3 with 100 days segments.

This was done so that both long and short-term periodicities are examined. None of the curves show any significant peaks of interest. The maxima at the low-frequency end correspond to the harmonics of the segment size (e.g., in the 100 days segmentation case, the maxima occurred at 100 days, 50 days, 25 days periods).

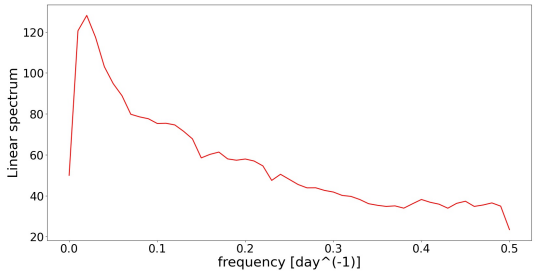


Figure 4: Welch periodogram of the entire 22-year time series with "hann" windowing and **100 days segment**

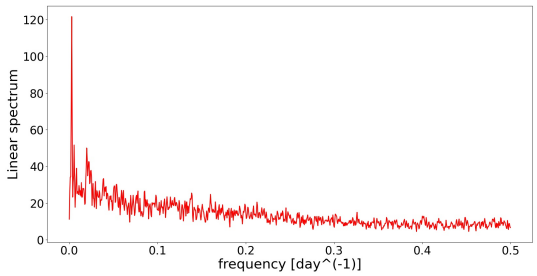


Figure 5: Welch periodogram of the entire 22-year time series with "hann" windowing and **5-year segment**

3.2 Auto-correlation function(ACF)

The impact of one rainfall event to the subsequent time series would be best revealed by studying the autocorrelation with different lags. A 95% confidence interval was constructed using Bartlett's formula in statsmodel library [10].

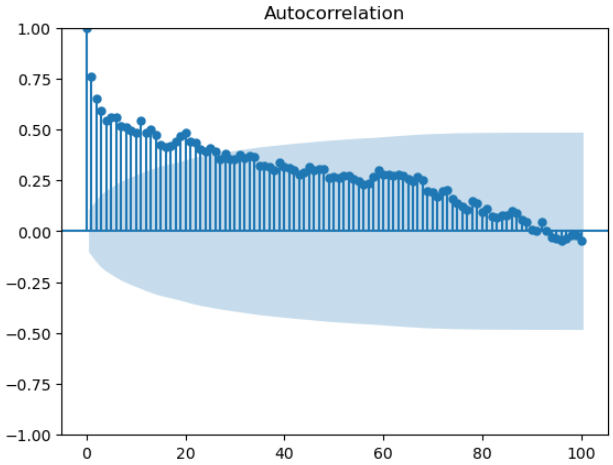


Figure 6: The x-axis is the lag (in days). The blue region depicts the boundary for the 95% confidence interval.

We found that after around 26 days, there is no more significant correlation between the time series. This gives us a rough estimate of the time scale at which the dynamic processes is likely to affect this distribution.

4 Kinetics of the causal rain system and their categorization

A natural question to consider is: *Where do these rain systems originate? Are there any seasonal patterns in the location of their sources?* To determine this, the GPM Half Hourly data was used [9]. Since the focus is on the secondary peaks, the data from April to August [3] every year (2001 to 2021). Each individual rain system was tracked and categorized. Details of the tracking algorithm will be discussed in the subsequent section. The reader may wish to skip to the results section for final outcome.

4.1 The tracking algorithm

There are multiple well-established rain cell tracking algorithms in the literature, for example, some using Particle Image Velocimetry (PIV) [4] on radar reflectivity data. There have been attempts involving estimating probability density mixtures to classify different cells [1] Apart from complex neural network-based approaches e.g. [2], there are also many simple techniques like Iterative Rain cell Tracking (IRT) [7]. Almost all of these methods divide the problem into two sub-parts: (a) rain cell identification and (b) finding spatial correlations between identified cells in subsequent time steps. Part (b) is often done by finding advection due to steering flow, the overlap of rain cells, or a certain combination of both.
In our context, we have data with a resolution much coarser than radar imaging (typically $1km \times 1km$) and a temporal resolution of 30 min compared to 5 min in the case of radar imaging. We also need to consider that since we are interested in classifying the origin of each rain cell, which may be several thousand kilometers away, a sufficiently large tracking domain is required. Significant inspiration has been taken from the prior work of [13] and [11] in tracking rain systems in satellite precipitation products like TRMM 3B42 (the data set used in our time series analysis). Although the FiT (Forward in Time) algorithm they used hasn't been used here, we have adopted a back-tracing approach analogous to theirs. The tracking algorithm we need must also be robust enough to follow the same cell for a long duration throughout its development and subsequent dissipation, as has been exemplified in [6].

[3] As seen in the time series analysis

4.1.1 Speed Adaptive Peak Search (SAPS)

To suit our purposes, a novel adaptation of the IRT and FiT was implemented. We have used the location of the maxima of each rain cell as the point marker that we track reverse in time. The zone where we search for the maxima is determined based on the latest observed speed of the cell's propagation. A faster speed would imply greater variance in the location of the cell in the previous time step, thus requiring a larger area to search the spatial peak. It should be noted that SAPS doesn't take into account the size of each cell. In later studies, it is possible to do so in the way [11] did in FiT.

The details of our algorithm have been elaborated on below:

1. A tracking domain of 62.75E to 96.35E and 1.45S to 28.65N was used. The spatiotemporal data set for this domain is denoted by ${\phi_t}^{xy}$, where t is in steps of 30 minutes, x and y are latitude and longitudes.

2. The timestamps of all the peaks in the raw $\tilde{d}_{Yt}$ data[4] was listed for the time period of April 1st to August 31st. We will call this set E_t.

3. For each of the events e in E_t, we wish to document where it originated. We first find the spatial maxima of ${\phi_{t=e}}^{xy}$, say $M_0 = (x_{m,0}, y_{m,0}, t = e)$, within the Bangalore domain. This is how we identify the dominant rain cells causing the spike in precipitation.

4. Now, a modified advective algorithm has been used to back-trace the origin of this cell. Centering at point M_0, we search for the nearest spatial peak in ${\phi_{e-\delta t}}^{xy}$. Assuming that we found this maximum M_1 at a distance r_1 from M_0. We repeat the same process to search maxima in ${\phi_{e-2\delta t}}^{xy}$ by centering at M_1 with an initial search radius of ar_1. (a chosen to be 1.5 for the final run)
This is based on the assumption that the speed of the rain cells is expected to remain similar within contiguous time intervals. Limiting the search radius for the maxima aids in the computation speed. In particular, in cases where the maxima were not found in this initial search zone, the zone is extended till $r = 2ar_1$.

5. This process iteratively continues and the sequence $\{M_i\}$ is saved for each events.

   A crucial concern in designing the tracking system is the choice of parameters or conditions

[4] The unfiltered precipitation curve on Bangalore

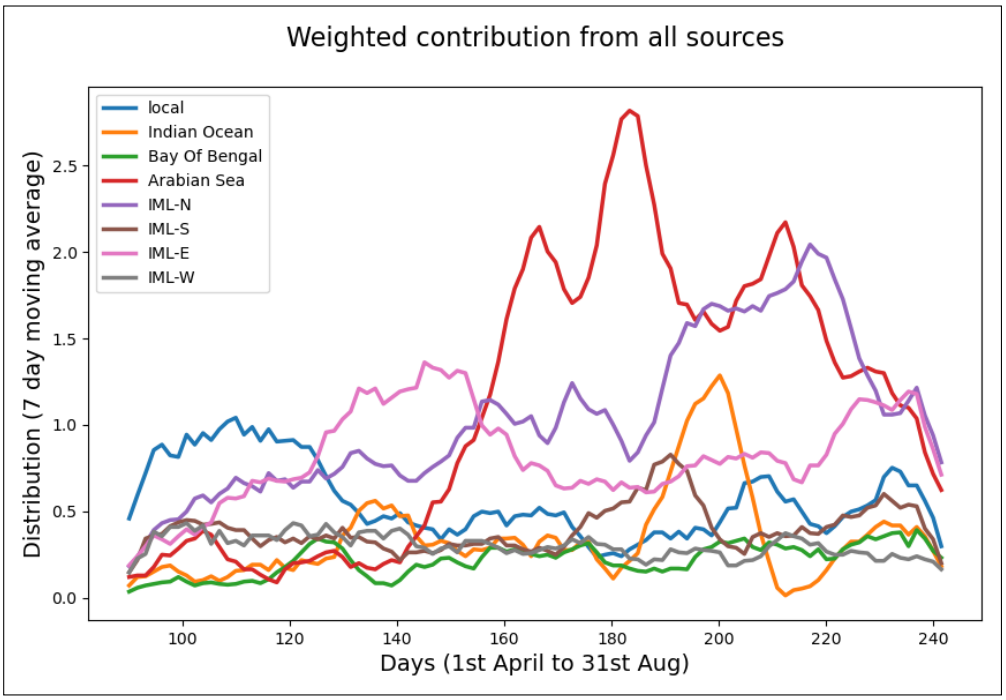


Figure 7: The frequency distribution of events from each category has been plotted. The X-axis is the index of the days in a year. Y-axis is the average number of events of any particular source on that day per year

for terminating the process. The endpoint is what we will document as the source of the rain system.

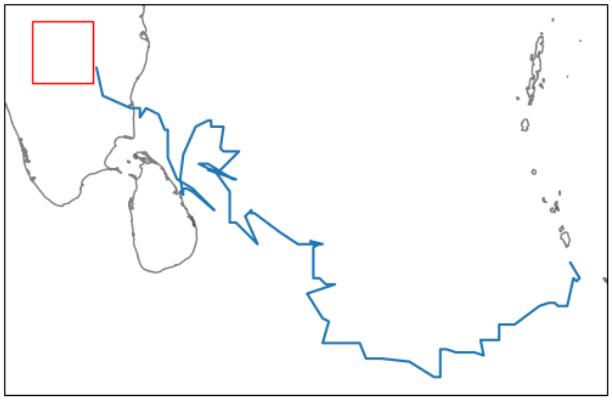


Figure 9: This is the trajectory of one of the rain cells that originated in the Bay of Bengal generated by our algorithm.

We have used the following set of conditions to terminate the process:

- The magnitude of ${\phi_{e-n\delta t}}^{xy}$ at M_n is less than $\langle {\phi_{e-n\delta t}}^{xy} \rangle_{x,y} + 0.5\sigma$. This is to set a lenient bound to distinguish between a true nascent rain cell and random noise (with a confidence of $\approx 70\%$, assuming a Gaussian white noise).
- No peak found even after extending search radius to $r = 2ar_{n-1}$. The last entry in the M_i list, i.e., M_{n-1}, is the origin of the cell.
- The search went beyond the boundary of the tracking domain. In this case, too, the last known position is assumed to be the origin.

4.1.2 Categorisation

We sorted all the events into eight categories based on the location of their origin: Bay of Bengal, Indian Ocean, Arabian Sea, India Mainland(IML) North, East, West, South, and finally, locally generated within "Bangalore." The distinction was made by verifying if the origin points were present within a polygonal approximation of the peninsular coastline (turfy and geojson libraries were used).

Category	No. of events 2001-2021
Local	1851
Indian Ocean	1126
Bay Of Bengal	713
Arabian Sea	3575
IML North	3384
IML South	1294
IML East	2674
IML West	980

Table 1: Data for the number of events in each defined category.

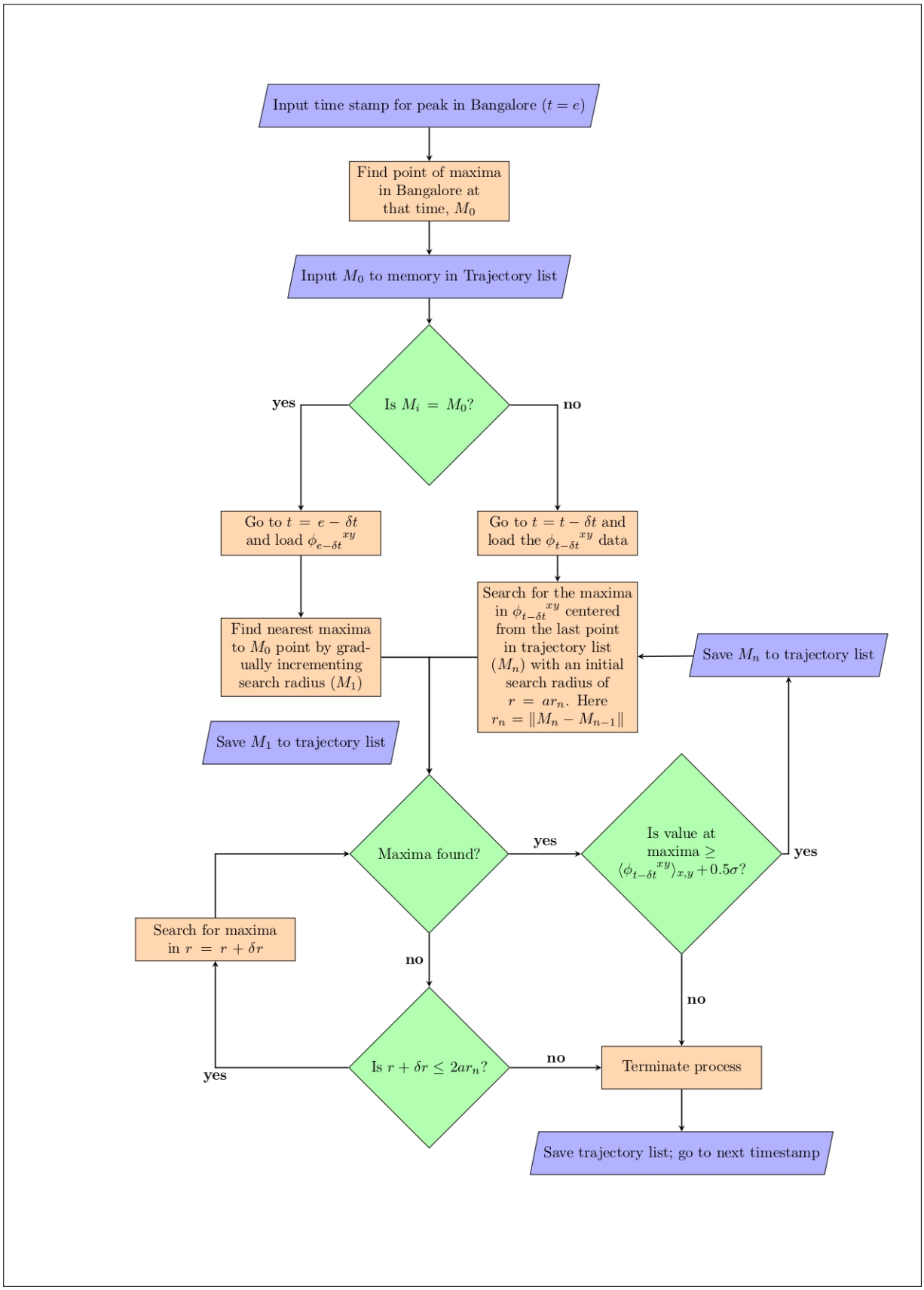


Figure 8: ***Flowchart of the SAPS algorithm***

4.2 Results and discussion

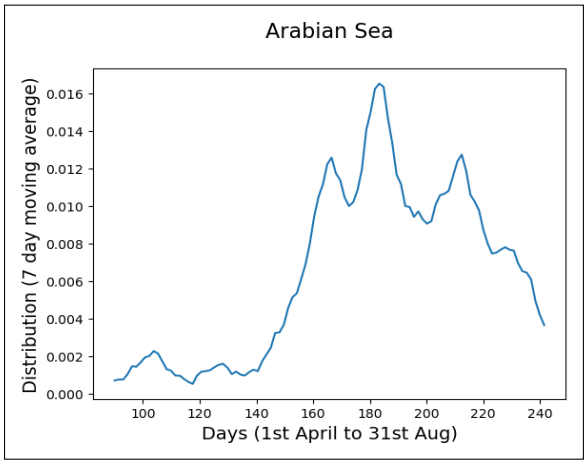
A huge variety of analyses can be done from the tracking data that has been generated. In this article, we will focus on two significant ones which seem to provide a hint towards the origin of the secondary off-monsoon peaks in precipitation. Figure 7 shows the frequency distribution as a function of time for each category and highlights the relative dominance of each source.

In the next section, we will look at the major contributors of special interest and observe their trends.

4.2.1 Frequency distribution of key sources

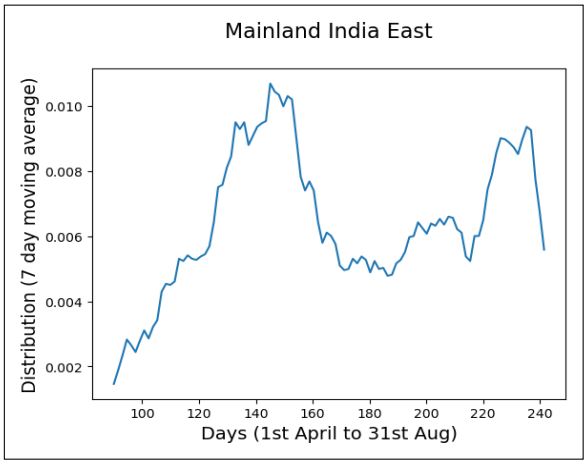
To understand the time dependence of the activity in each category, a frequency distribution is produced with 100 bins in the entire date range for each category. To clarify the trends, a 7 day uniformly weighted moving average has been applied to the data. Being the major contributors, Arabian Sea, IML-East, Indian Ocean, local, and IML-North calls for more detailed observation.

Arabian sea Fig.4.2.1 is evidently the largest source of most rain cells. It starts rising from around mid-May. An oscillating trend can be seen when it is active (June-August). It might be of further interest to look into ERA reanalysis data to verify the source of these oscillations.

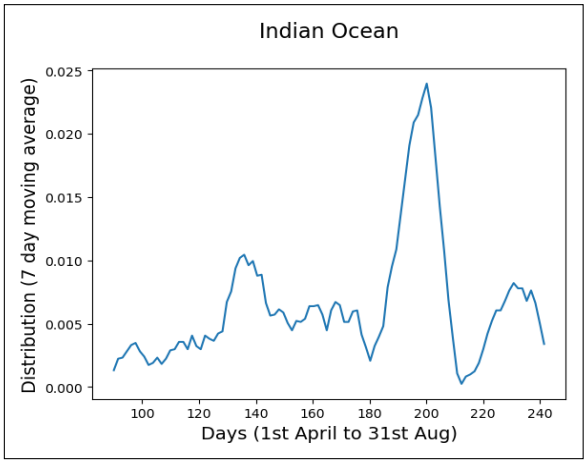


The region to the east of the "Bangalore" region (***IML-E***) Fig.4.2.1 seems to be the strongest source of rain cells in the off-monsoon period. The activity picks up from the beginning of April and peaks in May end (150). It decreases till June-end and stabilizes at a constant rate; eventually, it starts rising (much more gradually than the first peak) with the onset of monsoon. Comparing with other sources in Fig. 7, it is expected that this is the largest contributor to the "off-monsoon" rain.

In the half-hourly animation prepared, we observed several systems being produced near the East coast and traveling for a short distance into "Bangalore". This hypothesis potentially also explains the distance peak in the 1° range (Fig.4.2.3)

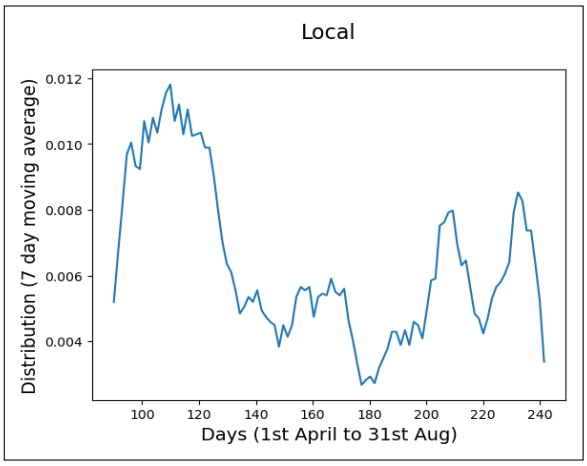


The ***Indian Ocean*** Fig.4.2.1 has been defined as south of the 8.5N latitude. A distinct peak is visible around day 190 in the month of July. There is some relatively minor activity spread across May to June (day 130), within the expected period of the first peak.

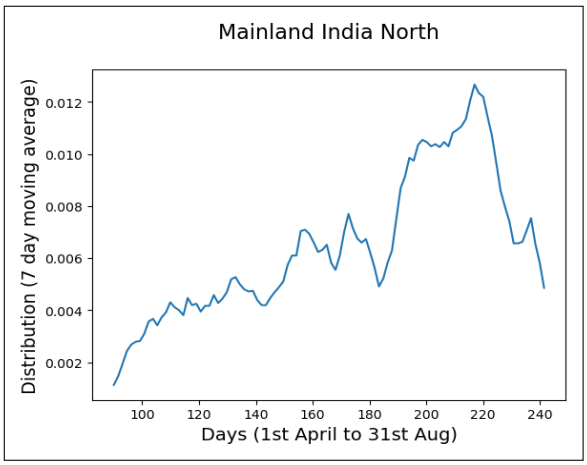


The ***locally generated*** Fig.4.2.1 events within our defined Bangalore region seem to be the most active source during April. There are some locally produced systems in the monsoon period. Although, Fig.8 shows that it is relatively insignificant. They form the bulk of short-distance events, as seen in Fig.4.2.3.

The ***Northern mainland*** Fig.4.2.1 of "Bangalore" is another dominant source of rain cells in our observation period. The frequency of events steadily increases, with a sudden drop near June end. The large count could be biased because of the massive area in this category.

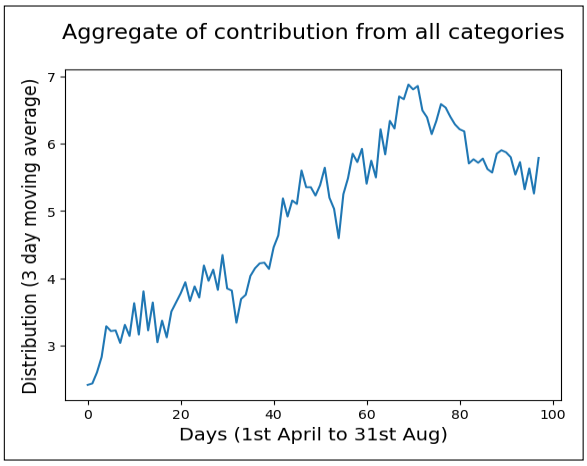


The rest of the categories, as shown in Fig.7, are relatively insignificant contributors to the total rain-cell count. This is of interest to note that the Bay of Bengal seems to be relatively inactive all throughout, although the density of rain cells from IML-E is high. It is suspected that many IML-E sources are driven by moisture transported from the Bay of Bengal, which would explain the observed data. This claim needs to be verified with reanalysis data to ensure that it is not due to some fallacy in our categorization system.



4.2.2 Comparison with TRMM time series trend

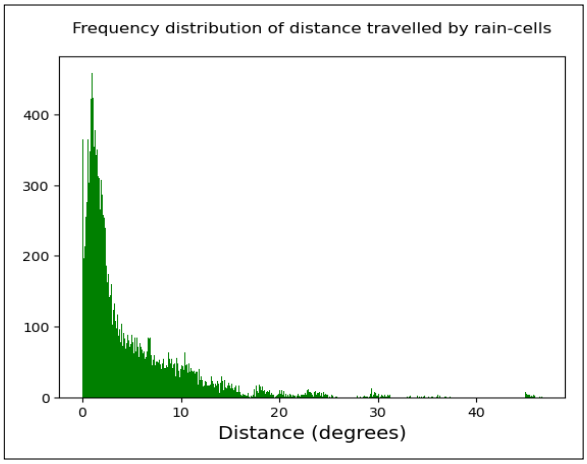
It should be verified if our tracking algorithm using the 30 minute GPM data can reproduce the yearly time series we got by analyzing the TRMM data. All the individual contribution from the categories has been superposed together to find the curve shown below.



Qualitatively, the two minor peaks before the major monsoon peak are visible. We should not expect a perfect overlap as the tracking algorithm didn't consider the size of each cell (which would influence the actual precipitation value) but rather the number of events. There is a fundamental limitation to using only microwave-IR satellite data for tracking precipitation events; we don't have a direct idea about the pressure and humidity distribution. This limitation can be overcome by involving reanalysis data like ERA. Nonetheless, the qualitative agreement bestows confidence in the consistency of our methods used.

4.2.3 Distribution of distance traversed

This is another piece of statistics that would be useful in searching for the underlying process behind our observations. An idea regarding this would provide a geographical scale of the events.



The distances are calculated by finding the Euclidean distance between the starting and the ending point of each trajectory. This does not capture the actual path length of the rain cell. We see a peak at 0.93 degrees.

5 Summary and conclusion

It has been strongly indicated that the rainfall pattern in the central south Indian peninsula or "Bangalore" region deviates from the rest of India. There are two minor peaks before the Monsoon peak. We have done the necessary statistical analysis to support this hypothesis.

A new tracking algorithm has been developed with our specific purpose in mind. The data generated by this algorithm has provided us with several clues behind this anomaly in question. Specifically, we have seen that mostly the locally generated and mainland India to the east of "Bangalore" contribute to the "off-monsoon" peaks. The frequency distribution of the rain events has revealed a peak in 0.93°. This dominance of rain systems generated at a short distance, along with the local and IML-E time series, correlates with each other. The mechanism behind the oscillating trend of the Arabian sea is yet to be understood. The passiveness of the Bay of Bengal in this is also interesting since a lot of monsoon dynamics are otherwise influenced by it [3].

In terms of further scope, the dynamics of these multiple peaks should be explored using reanalysis data (ERA). Questions like, what causes the large frequency of locally generated events in the first place, can only be answered using reanalysis. The trajectory of the rain cells we produced needs to be correlated with the moisture transport in the general atmospheric circulation simulations.

We hope this article has brought to notice a very peculiar and interesting nuance in the Southern Indian monsoon and laid some preliminary groundwork for further studies into this phenomenon.

References

[1] Veneziano D. and Villani P. "Identification of rain cells from radar and stochastic modeling of space-time rainfall". In: *Meccanica* 31 (1996), pp. 27–42. DOI: https://doi.org/10.1007/BF00444153.

[2] X. Ding, T. Denœux, and F. Helloco. "Tracking Rain Cells in Radar Images using Multilayer Neural Networks". In: *ICANN '93*. Ed. by Stan Gielen and Bert Kappen. London: Springer London, 1993, pp. 962–967. ISBN: 978-1-4471-2063-6.

[3] Sulochana Gadgil. "The Indian Monsoon and Its Variability". In: *Annual Review of Earth and Planetary Sciences* 31.1 (2003), pp. 429–467. DOI: 10.1146/annurev.earth.31.100901.141251. eprint: https://doi.org/10.1146/annurev.earth.31.100901.141251. URL: https://doi.org/10.1146/annurev.earth.31.100901.141251.

[4] Ting He et al. "New Algorithm for Rain Cell Identification and Tracking in Rainfall Event Analysis". In: *Atmosphere* 10 (Sept. 2019). DOI: 10.3390/atmos10090532.

[5] Peter Kahlig. "Some aspects of Julius von Hann's contribution to modern climatology". In: *Interactions Between Global Climate Subsystems: The Legacy of Hann*. Geophysical monograph. Washington, D. C.: American Geophysical Union, 1993, pp. 1–7.

[6] Yukie MORODA et al. "Structure and Evolution of Precipitation Cores in an Isolated Convective Storm Observed by Phased Array Weather Radar". In: *Journal of the Meteorological Society of Japan. Ser. II* 99.3 (2021), pp. 765–784. DOI: 10.2151/jmsj.2021-038.

[7] Christopher Moseley, Peter Berg, and Jan O. Haerter. "Probing the precipitation life cycle by iterative rain cell tracking". In: *Journal of Geophysical Research: Atmospheres* 118.24 (2013), pp. 13, 361–13, 370. DOI: https://doi.org/10.1002/2013JD020868. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2013JD020868. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JD020868.

[8] NASA Goddard Earth Sciences Data and Information Services Center. *TRMM (TMPA) Precipitation L3 1 day 0.25 degree x 0.25 degree V7*. 2018.

[9] Precipitation Processing System (PPS) at NASA GSFC. *GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06*. 2019.

[10] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[11] Gregor Skok et al. "Object-Based Analysis of Satellite-Derived Precipitation Systems over the Low- and Midlatitude Pacific Ocean". In: *Monthly Weather Review* 137.10 (2009), pp. 3196–3218. DOI: 10.1175/2009MWR2900.1. URL: https://journals.ametsoc.org/view/journals/mwre/137/10/2009mwr2900.1.xml.

[12] P Welch. "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms". In: *IEEE Trans. Audio Electroacoust.* 15.2 (June 1967), pp. 70–73.

[13] R. H. White, D. S. Battisti, and G. Skok. "Tracking precipitation events in time and space in gridded observational data". In: *Geophysical Research Letters* 44.16 (2017), pp. 8637–8646. DOI: https://doi.org/10.1002/2017GL074011. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017GL074011. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL074011.