1. Introduction

We live in an Era where severe weather events are becoming common, and precise and high-resolution forecasting is not just a scientific challenge but a societal need. Coarse-resolution data sets play the most essential role in predicting the future; however, scale discrepancy limits the coarse-resolution datasets from being directly used for impact assessments and decision-making (Wang et al., 2021), nevertheless, in this age where accurate predictions are imperative rather than optional. The need for finer-resolution data sets is ever-increasing. When going through the Intergovernmental Panel on Climate Change (IPCC) assessment reports, the fifth report highlights the importance of regional-scale climate change (IPCC, 2013). To get over it, we use techniques to infer high-resolution data from these low-resolution coarse data, i.e. GCM-simulated predictors, to obtain locally downscaled values. The method we use to tackle this issue is called downscaling.

* 1. What is Downscaling

Downscaling just as it says is the process of reduction in size, scale, or extent (Merriam-Webster); As such, it is the method of decreasing the pixel size of remotely sensed images (Atkinson, 2012) as a means to translate info from global and other large climate models to the scale of local/regional models. Then, why not just use regional models? The issue with regional models is that the computational cost is high; not only that, even if you have that computational power, the regional model is very boundary condition dependent as it derives from a GCM (Global Climate Model. Furthermore, if the GCM is biased, this model inherits and amplifies those biases, as it works in a “garbage in, garbage out” manner (Rummukainen, 2010). Also, the initiation process would need localized observations, which are hard to come by, and running it requires expertise from different fields. These are but a few of the many problems faced while running an RCM (Regional Climate Model). Downscaling holds a strong hand in the production of regional-scale predictions and circulation models GCMs (e.g., Francisco Et al., 2020). The process of Downscaling is commonly done through 2 major methods: Statistical Downscaling and Dynamic Downscaling.

* 1. Dynamic downscaling

Dynamical downscaling relies on a regional climate or numerical weather model to provide high-resolution climate factors by simulating the physical processes of the coupled land-atmosphere system (Rummukainen, 2010). A high-resolution grid is nested on the coordinates required for the high-resolution global model to infer high-resolution data from a low-resolution global model. Even so, it has its limitations, such as requiring a large amount of processing power to produce a reasonable output while being prone to model errors and—most importantly—not being able to be applied to another domain and deliver results with the same impact. There is also the fact that most of the nested models are optimised for the northern hemisphere, even when many countries in the southern hemisphere encounter extreme weather conditions (B.C Hewitson Et al.,1997), a topic of discussion for another day. Then there is the even more significant fact that it is still challenging for the current downscaling approaches to adequately represent spatial and temporal variability of the local-scale climate and capture local-, small-scale features such as local extreme events (Hertig et al., 2019; Maraun et al., 2019).

* 1. Statistical Downscaling

Statistical Downscaling is the basis for developing an empirical relationship between local high-resolution and low-resolution data either through linear regression or analog techniques and, in some cases, even both (Gutiérrez et al., 2018). This gets more accurate as the quality and quantity of these data increases. This relationship or empirical formula is derived from comparing the local model data to observed and high-resolution data using mathematical or statistical equations. Although the Statistical downscaling method is more computationally efficient than nested models, it also has drawbacks, such as a lack of universality; if the relation is derived for an area, it cannot be used for another area, as the forcing functions may differ. Thus, finding a method to obtain accuracy and universality is necessary. Thus, we have come here to this topic of study – the application of Deep Learning as a method of downscaling.

* 1. Machine Learning & Machine Learning for Downscaling

The first applications of neural networks date back to the late 1990s (Wilby et al., 1998); it is only in recent years that the growth of Artificial Intelligence and Machine Learning (AI/ML) brought a new wave of change to the research world, including meteorological research. This was majorly due to the advent of more powerful computers and efficient programs (Price et al., 2023). Compared to the more equation-driven computer-intensive Numerical Weather prediction methods, the Machine Learning methods offer a more data-driven alternative, where the relationship between the different meteorological datasets, both spatially and temporally, is considered, which uses this to give a prediction, this is primarily due to the ability of DL models which allows to extract high-level feature representations hierarchically due to its (deep) layered structure (Schmidhuber, 2015). This method has been shown to have significant potential in downscaling. As a result, AI/ML is an excellent tool for improving the accuracy of weather predictions. Various machine learning techniques address the limitations of the dynamic approach in downscaling and compare their efficiency during the monsoon season of 2024. Specifically, Deep Learning is a subset of Machine Learning that specialises in extracting data by emphasising successive layers of neural networks (Chollet, 2021). Convolutional Neural Networks (CNNs, LeCun et al., 1989) are a type of neural network that relies on the convolution operation. The convolution operation involves sliding a weight tensor that operates on the data, usually starting from the top left, so only a few input grid points contribute to a given output, and the weights are reused since they are applied to multiple locations in the input. This is a transformation that takes advantage of the gridded nature of the data. Moreover, new efficient computational frameworks like keras/TensorFlow, regularization of data by using dropouts, and new learning methods have been the big thing from the recent popularization of DL techniques, allowing convolutional neural networks to learn efficiently from big data and avoid overfitting freely.

In this paper, we will evaluate the accuracy of NCUM operational forecast. Then, use statistical Bias correction methods by use historical observational data to correct systematic errors in model outputs. By adjusting the outputs to take into account the bias, we can bring the simulations closer to observational data. Later, we use AI/ML in downscaling by training a deep learning CCN to generate high-resolution predictions from global model output (Global NCMRWF Unified Model, NCUM-G) and compare them to the dynamical downscaling output (Regional NCMRWF Unified Model, NCUM-R). While AI-based downscaling has shown promise, its performance against traditional approaches remains an open question. As such, we will delve into the effectiveness of DL-based downscaling compared to dynamical and statistical methods and how computationally efficient it is while looking through how DL-based downscaling techniques deal with the inherent bias of the GCMs.

In the second section of our project, we will discuss the different aspects of the climate models referred to here in greater detail. Later, in section 3, we will verify the downscaling performance of the Dynamic core, statistical, and deep learning-based methods. Finally, a comparison study of these three methods will be conducted in Section 4.

2. Modelling

2.1. Introduction to NCUM

The NCUM or NCMRWF unified model is a numerical weather prediction model operated by Partnerships' Unified Model (UM) since 2012. It works using a Unified Model (UM) developed by the UK Met Office. NCUM is used for both global (NCUM-G) and regional (NCUM-R) weather prediction.

2.1.1. NCUM-G

From the latest model verification report, ‘NCUM Global Model Verification: Pre-monsoon (MAM) 2024,’ we can see that the global model runs with a resolution of 25 km. Moreover, the latest iteration of the model, NCUM-G V7, has a horizontal grid resolution of 12 km with 70 vertical levels in the atmosphere reaching 80 km. This model is adapted from Unified Model version 11.2 (UM11.2), part of the latest “Operational Global Suite” (PS43) of UK Met Office runs based on an advanced “ENDGame” (Even Newer Dynamics for General atmospheric modelling of the environment) dynamical core. This is adept at solving compressible non-hydrostatic equations of motion with semi-Lagrangian advection and semi-implicit time stepping. The model runs in 5-minute time steps, providing a 10-day forecast. Similar to every NWP (Numerical Weather Prediction) model, this model also needs Quality control that has the least amount of error, and it is provided with a Hybrid 4D Var data assimilation process, where a hybrid of climatology and forecast ensemble is used; where the sub-grid process undergoes physical parameterisation (GA7.2).

2.1.1. NCUM-R

Among them, NCUM-R is a high-resolution domain for the Indian region and is configured with the UM seamless prediction for various applications. The model, which ran on Singapore version 2 (SINGV2) from 2016, now uses the UK Met Office’s “Regional Atmosphere and Land version 3 (RAL3) to run from October of 2022. This collective of the high-resolution models includes 3 models, each with different resolutions: the 4km - All India Model – which is of most interest to us, a 1.5km - Delhi and neighbourhood model and finally a 330m - Delhi city model. NCUM-R has a rotated (slightly tilted) latitude-longitude horizontal grid (to make the ‘Equator’ on top of the area of interest) with Arakawa-C staggering and a terrain-following hybrid vertical coordinate with Charney-Philips staggering. The 4km domain model, which is run all year round, covers the area of 62⁰E-106 ⁰E; 6 ⁰S 41 ⁰N respectively with a time step of two minutes. The horizontal and vertical extent of the model are 1200x1200 grid points horizontally and 90 hybrid levels vertically, with a top at 40km. With initial conditions set at 00UTC and 12UTC. These high-resolution initial conditions are set with the “4D-Var” data assimilation system, which is used in NCUM-G. Even then, NCUM-R updates its lateral boundary conditions from operational global model forecasts (NCUM-G) every hour. It is to be noted that even when the different parametrization exists in the model, convection is explicit, and the subgrid scale deep convection parametrization is absent.

1. Evaluation of NCUM-R

Dynamical downscaling improves regional-scale weather and climate predictions by refining coarse-resolution global model outputs. However, the effectiveness of the dynamical downscaling process is heavily dependent on the Global climate models (GCMs) and parametrisation schemes used. In this study, we will compare the effectiveness of this approach to see how much this method improves upon the parent model and how accurately it represents observed climate patterns. To assess the performance of the NCUM-R model**,** we compare its simulated rainfall against observations and the coarse NCUM-G dataset. The verification focuses on mean rainfall, bias, and variance for the JJAS monsoon seasons from 2021 to 2024. Additionally, case studies of major extreme rainfall events in the same period are also included.

We analyse the biases in NCUM-R to determine whether it improves upon NCUM-G in capturing spatial and temporal rainfall patterns. Extreme rainfall events are also examined to evaluate the model’s ability to capture heavy precipitation events using intensity and frequency distributions.

So, for the verification of the dynamical downscaling method, we take data from 3 major data sets

1. NCUM-R (Regional Model) – Daily surface rainfall (APCP\_Surface) output from the high-resolution regional model.
2. NCUM-G (Global Model) – Daily surface rainfall (APCP\_Surface) output from the lower-resolution global model.
3. IMD-MSG (Observed Dataset) – The observed rainfall(rf) dataset is used as a reference.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Resolution** | **Temporal Coverage** | **Variable Used** |
| NCUM-G | ~12 km | Multiple forecast lead times (Day 1, 3, 5) | Daily precipitation |
| NCUM-R | ~4 km | Multiple forecast lead times (Day 1, 2, 3) | Daily precipitation |
| IMD-MSG | ~10 km | Observed rainfall data | Daily precipitation |

3.1. Primary Analysis and Results

For this analysis, we have taken the daily rainfall 2021–2024 for the JJAS (June–September) monsoon period in the spatial range of 65⁰E-106 ⁰E, 6 ⁰S 41 ⁰N. This approximately matches the domain size of the NCUM-R. (The western boundary 65⁰E is taken into account to compensate for the errors in the IMD-MSG observation dataset, which ensures consistent comparison across all datasets.) The dataset is masked using India's shapefile so that only data over the country is considered before further analysis. This is done for all the 3 datasets before doing any further analysis. We analyse the Mean Rainfall for the period, Variance, Biases, and some forecast verification of the 2 models. Along with some model comparisons on some extreme weather events in the period.

* + 1. Mean rainfall

The spatial distribution of mean rainfall is plotted for each dataset to highlight discrepancies and improvements in NCUM-R over NCUM-G. To evaluate how well NCUM-R captures regional rainfall patterns compared to the global NCUM-G.

For this, first, all the datasets need to be gridded to the same resolution of the IMD-MSG observation dataset. This is done by the ‘nearest-neighbour interpolation’ method, where each point in the new grid is assigned the value of the nearest data point in the original grid. It does not perform any averaging or weighting of surrounding points. It is a simple and fast method but may not be as smooth as other interpolation methods like linear or cubic interpolation (xarray Documentation).

After this step, we calculate the mean rainfall over the selected period. Mathematically, the mean rainfall (i,j) for grid point (i,j) is calculated by;

Where Is the rainfall at grid point (i,j) at time t, and T is the total number of time steps. Using that, we find the mean precipitation for the observation dataset and Day 1, Day 2, and Day 3 of NCUM-G and NCUM-R, respectively. For ease of comparison, we make a figure of both separately but with consistent colour scales.

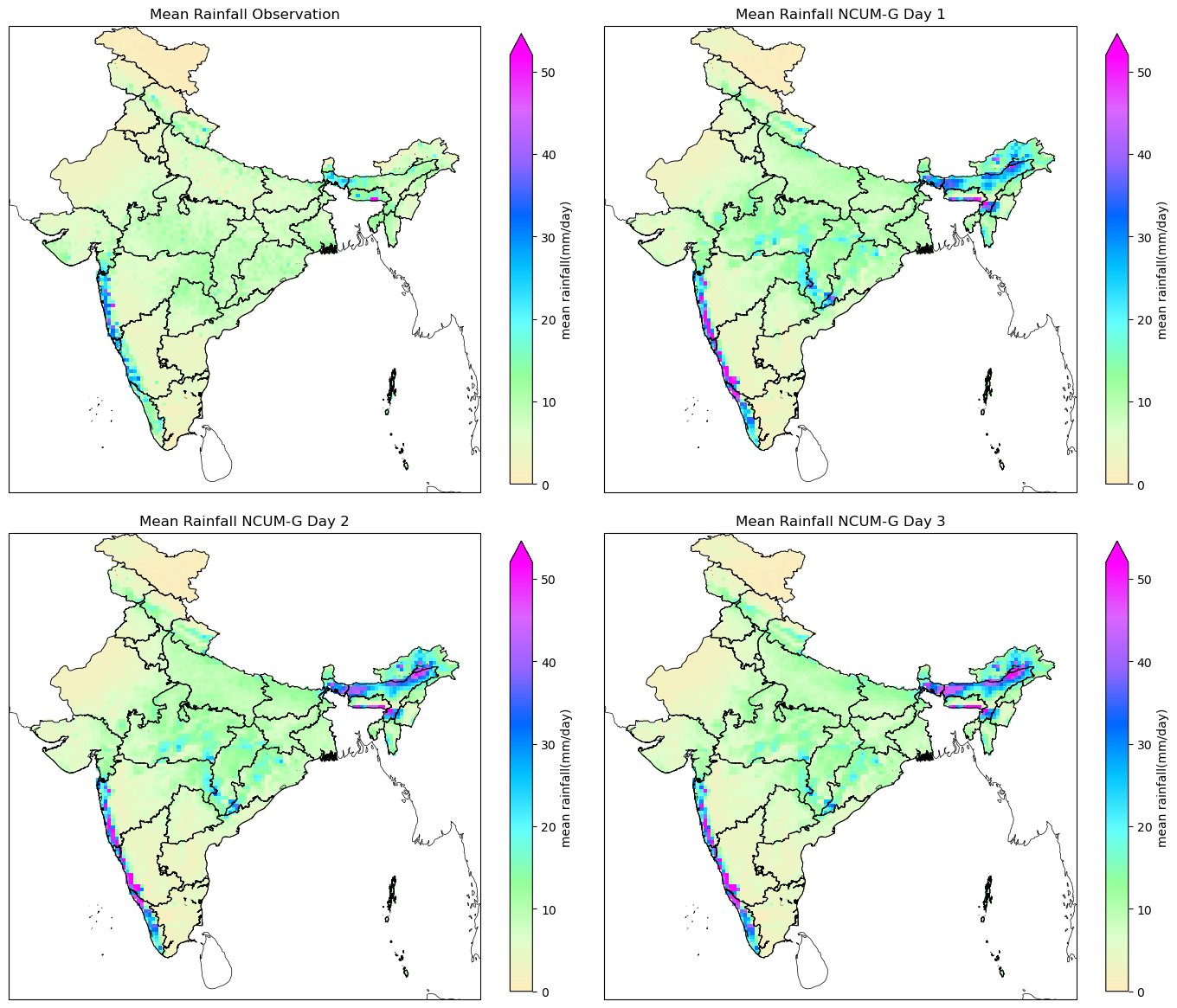


Figure A : Mean daily rainfall (mm/day) over India from observations and NCUM-G forecasts. (Top-left) Mean observed rainfall (IMD-MSG). (Top-right) NCUM-G Day 1 forecast. (Bottom-left) NCUM-G Day 2 forecast. (Bottom-right) NCUM-G Day 3 forecast. The colour scale represents rainfall intensity, ranging from 0 mm/day (beige) to over 50 mm/day (purple)

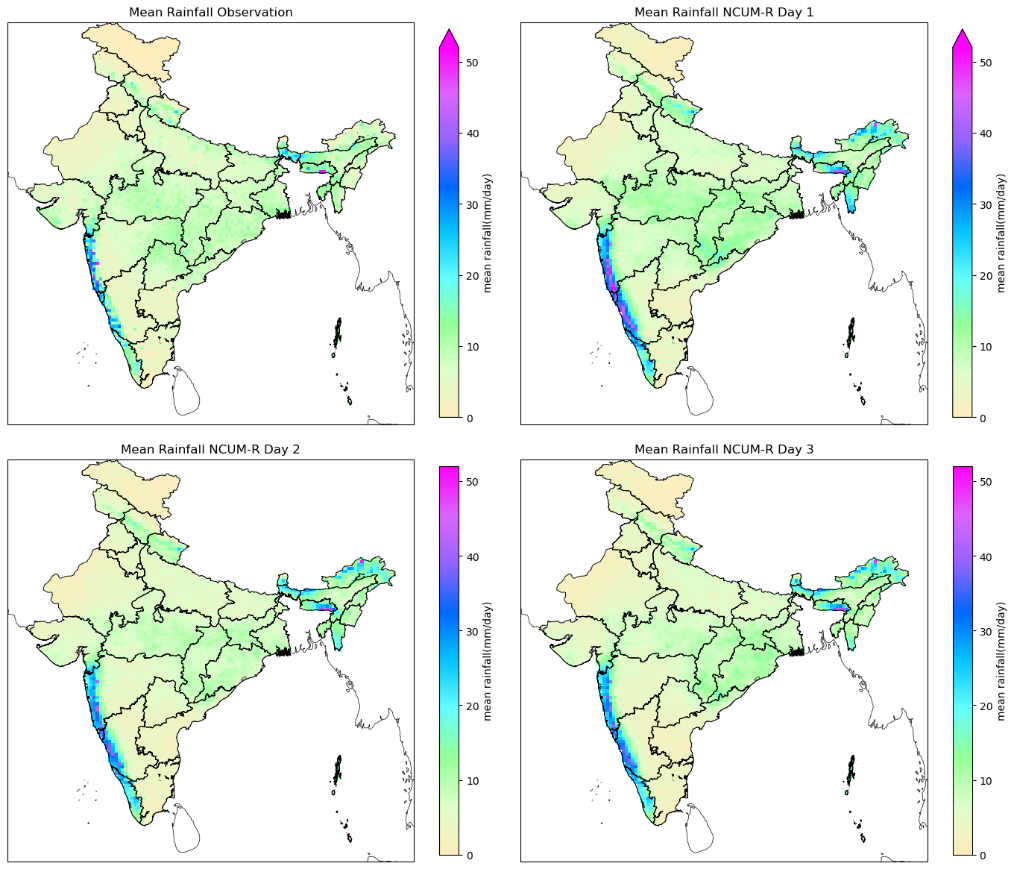


Figure B : Mean daily rainfall (mm/day) over India from observations and NCUM-R forecasts. (Top-left) Mean observed rainfall (IMD-MSG). (Top-right) NCUM-R Day 1 forecast. (Bottom-left) NCUM-R Day 2 forecast. (Bottom-right) NCUM-R Day 3 forecast. The colour scale represents rainfall intensity, ranging from 0 mm/day (beige) to over 50 mm/day (purple)

Result

Looking at figures K and L, we can see that the figure K.A and L. a Both depict the observation mean rainfall. In it, high rainfall is concentrated in the Western Ghats (Kerala, Karnataka, Goa, Maharashtra) and Northeast India (Assam, Meghalaya). This serves as a benchmark for evaluating both models.

NCUM-G Model mean rainfall for Day 1, Day 2 and Day 3 are seen in the figures K.B to K.D, respectively. Looking at it, the Western Ghat rainfall has more localized high-intensity rainfall and more scattered patches of heavy rainfall compared to observations, and some overestimation can also be seen on the west coast. For the Northeast India Rainfall, the rainfall is captured well, but it shows a misaligned peak rainfall and severe overestimation. It also shows rainfall over Parts of Eastern India (Odisha, West Bengal). Day 1 forecast is closest to observations but with overestimated intensities**.** Days 2 & 3: Rainfall decreases, but the model smoothens with lead time.

When looking at the figure L.B to L.D, which shows the NCUM-R model statistics for the Western Ghats rainfall, a more continuous and uniform pattern along the coast is seen. NCUM-R captures the widespread nature of coastal rainfall better than NCUM-G. However, it overestimates rainfall intensity in regions with high orography. And as for Northeast India Rainfall, it aligns slightly better with observations, but there are some displacement issues for peak rainfall in some areas. The Day 1 forecast is more realistic compared to that of NCUM-G. As for Days 2 and 3, the rainfall prediction decreases compared to Day 1.

* + 1. Variance

Similar to the Mean rainfall distributions, the variance distribution also uses the daily rainfall data re-gridded to match the resolution of the observation Dataset. Then, the variance (σ2) of forecasted and observed values is calculated by the equation:

σ2

Where:

* represents individual forecast or observed values,
* Is the mean of the dataset,
* N is the total number of data points.

The variance plot is generated to visualise the dispersion of forecasted values relative to observations. Variance is generated for both NCUM-R and NCUM-G for their Day-1, Day-2, and Day-3 forecasts. These dispersion plots help in understanding the spread of model outputs by showing whether the models overestimate or underestimate variability to the observed data in various circumstances during the period of the study.

Fluctuations in variance values correspond to weather events in the forecast, where high variance shows high rainfall events. The difference between the model’s variance and the variance of observation indicates whether the model over-predicts or under-predicts, depending on the sign of the difference.

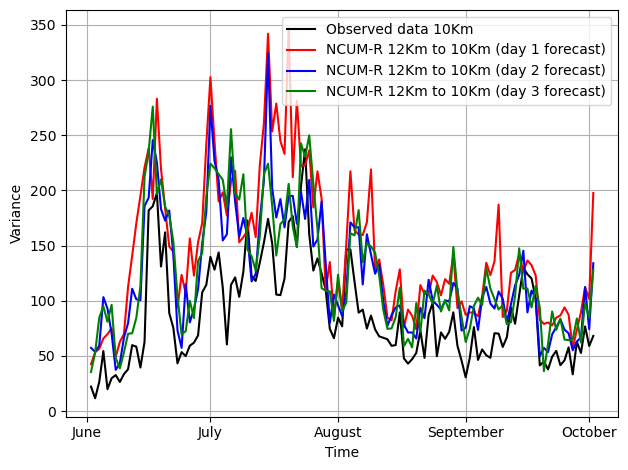
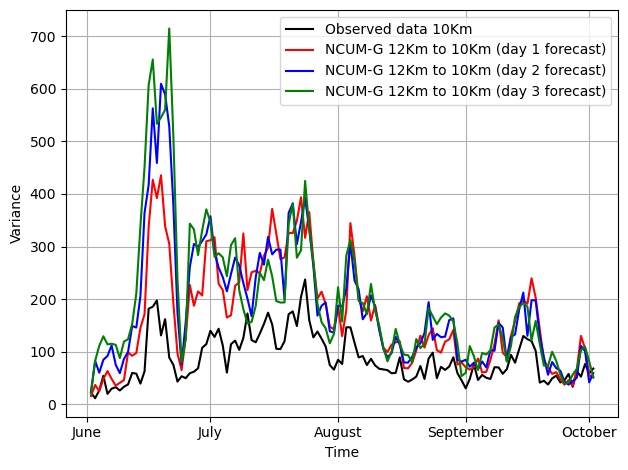


Figure and :Temporal variance of daily rainfall over a JJAS months. The black line represents observed data, while the red, blue, and green lines represent 1. NCUM-G, 2. NCUM-R forecasts for Day 1, Day 2, and Day 3, respectively.

Results

In the NCUM-G model (Figure A), we can see that, although the model captures the general trend of observed values, it shows higher variance values during the peak monsoon months of June and July, reducing thereafter. This means the model captures the extreme events but over-amplifies variability, causing it to predict too much precipitation compared to observation in high-impact weather scenarios. The variance is seen to be greatest for the Day 1 forecast (red line) and reduces with lead time, indicating that predictions smooth out with lead time. Increased variability suggests growing uncertainty with longer lead times. This may be due to the excessive sensitivity of the model to initial conditions taken or the underlying model physics.

Now, if we take a look at the NCUM-R model, as shown in figure B, we can see that the model variance follows the trend of observations variance very well. The values are seen to be closer to the observation compared to the NCUM-G variance. This means that NCUM-R offers a better representation of local-scale processes compared to NCUM-G. Even then, the Day 1 forecast (red line) slightly overshoots on days of high rainfall. Similarly, Day 2 and Day 3 forecasts (green and blue line) also tend to stabilize variance better compared to NCUM-G. It also appears to underestimate variance in some cases, especially in the late monsoon season (months of September–October). This may be due to a damping effect in long-range forecasts of the NCUM-R model.

From all this, we can infer that the NCUM-G model by itself is not enough to meet our needs because the global model generally shows a higher variance than observations, meaning that it overreacts to changes in rainfall, leading to potentially exaggerated forecasts. As seen, while NCUM-R provides more controlled variance, it still tends to both underestimate variance in some cases and overestimate it in others, which makes it imperfect.

* + 1. Bias

The Bias quantifies the difference between forecasted and observed rainfall. Studying bias will help identify systematic overestimation/underestimation trends in the models. The same dataset used before is also used for this study. The bias is calculated by the equation:

Bias = Forecast Rainfall − Observed Rainfall

Where the Forecast Rainfall (F) refers to the Rainfall predicted by NCUM-G or NCUM-R. And Observed Rainfall(O), which is our IMD gridded rainfall data. And this is done for all Day 1, Day 2, Day 3 forecasts of both the models. A positive bias means that the model overpredicts rainfall, and a negative bias indicates that the model underpredicts.

Since our dataset is four-dimensional with Rainfall, Latitude, Longitude and Time. We do Bias analysis in 2 ways: 1. Spatial Bias (spatial) and 2. Time Series-Bias (temporal).

* + - 1. Spatial Rainfall Bias

For each grid point (i,j):

Bias(i,j) = F(i,j) − O(i,j)

Then, the bias is separately found for NCUM-G and NCUM-R for multiple forecast lead times (Day 1, 2, 3). The bias is averaged over time to create mean bias maps. The spatial bias is visualized using plots red for underestimation and blue for overestimation, as seen in Figures C and D.

Results

Spatial Bias Pattern

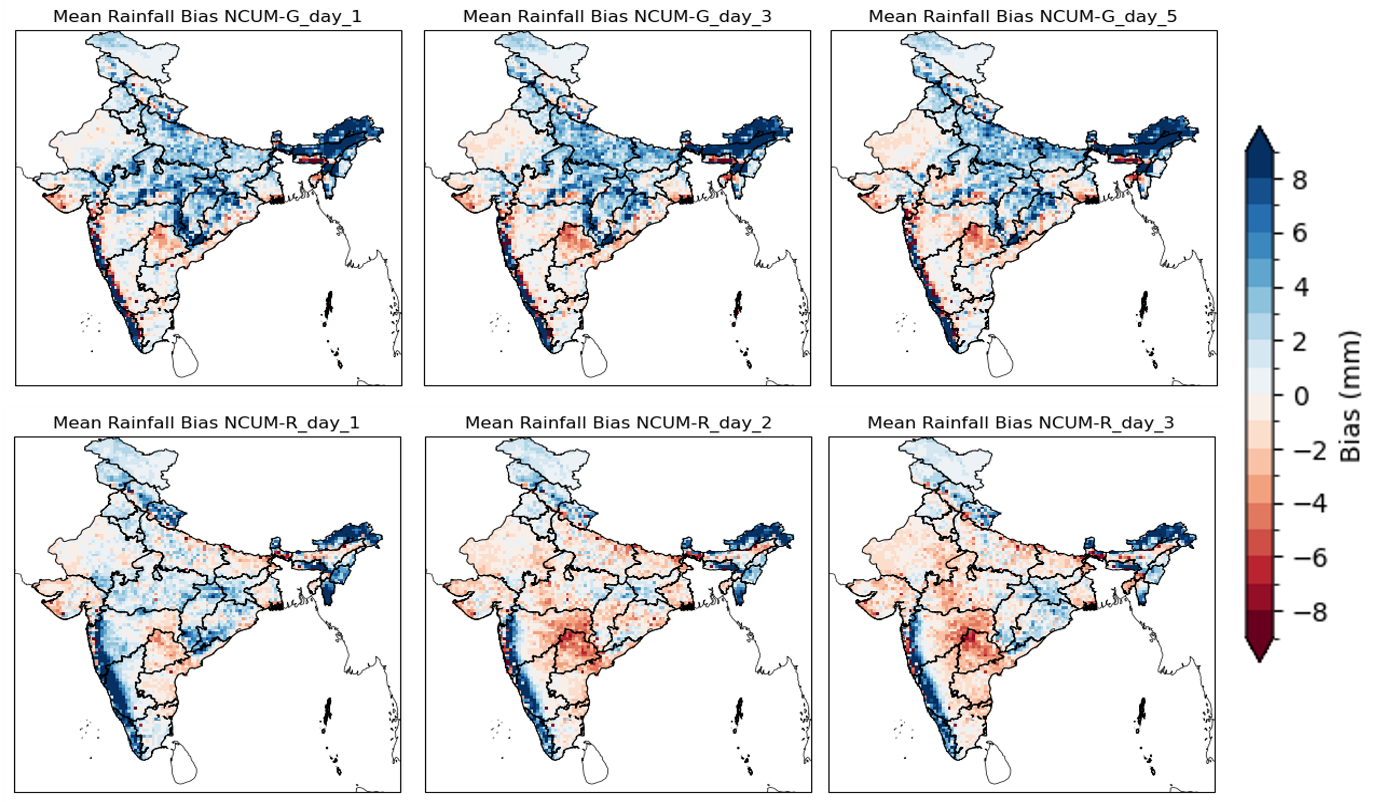


Figure :Mean daily rainfall bias (mm) over India for NCUM-G and NCUM-R forecasts compared to observations. (Top row) NCUM-G biases for Day 1, Day 3, and Day 5 forecasts. (Bottom row) NCUM-R biases for Day 1, Day 2, and Day 3 forecasts. Positive biases (blue) indicate overestimation, while negative biases (red) indicate underestimation.

Figure F shows the spatial bias maps of NCUM-G for forecasts of lead times Day 1, Day 2, and Day 3. We can see that both overestimation and underestimation are seen spread out on the whole of India region. Especially in central India and the interior peninsular regions, underprediction (red regions) is dominant. It is also to be noted that the bias intensity slightly reduces from Day 1 to Day 3, which indicates error dampening over longer lead times.

Similarly, from Figure G, we can see the spatial Bias maps of NCUM-R for forecasts of lead times Day 1, Day 2, and Day 3. We can infer that NCUM-R has strong overprediction tendencies. In the northeastern region and coastal areas (especially along the western coast), the overprediction is particularly high, even higher than in NCUM-G. Underprediction, although mild, is more widespread in the central regions of India. But the key thing to note is that, unlike the NCUM-G, the bias structure remains almost the same in all lead times, which means that it only has somewhat of a systematic model bias. This model seems to have stronger positive biases over high-rainfall regions (Western Ghats, Northeast) but underestimates rainfall over dry regions, this may very well be an effect due to the model not capturing the variation by the virtue of orography.

From all this, the key takeaway is the fact that NCUM-G bias shows less variability across different regions compared to NCUM-R; at the same time, it shows NCUM-G bias changes slightly over time, whereas NCUM-R shows a more consistent bias.NCUM-G has a strong positive bias, especially in the northeastern and western ghats parts of India, and this bias has been increasing noticeably in the years. Compared to that, the NCUM-R has a more balanced bias. But it also shows some underestimation in different parts of India.

* + 1. Forecast verification

Forecast verification is about checking how well weather forecasts match observed data. It assesses how accurate and how reliable the model skill is. By using verification metrics, we can quantify forecast performance and identify model strengths and weaknesses. WE will verify both NCUM\_G and NCUM\_R over different lead times. The key metrics used are Probability of Detection (POD), False Alarm Ratio (FAR), Bias, and Equitable Threat Score (ETS). We will also evaluate them over 1 to 3-day lead times to see how forecast skill changes as lead time increases. To do this, we will take the dataset and regird it concerning the observation dataset as usual for proper comparison, and we will set a threshold of 0-30 to get the general trend.

The following metrics are used to evaluate the forecast performance:

Probability of Detection (POD):

POD measures the fraction of observed events that were correctly forecasted. A POD value of 1 indicates that all observed events were correctly forecasted, while a value of 0 indicates none were detected. Mathematically, it can be described by,

POD =

False Alarm Ratio (FAR):

FAR measures the fraction of forecasted events that did not occur. A FAR value of 0 indicates no false alarms, while a value of 1 indicates that all forecasts were false alarms.

FAR =

Bias:

Bias indicates whether the model tends to over-predict or under-predict events. A Bias value of 1 indicates perfect balance, while values greater than 1 indicate over-prediction, and values less than 1 indicate under-prediction. Mathematically, it is represented by:

Bias =

Equitable Threat Score (ETS):

ETS measures the accuracy of the forecast relative to random chance, accounting for hits, false alarms, and misses. ETS ranges from -1/3 to 1, where 1 indicates a perfect forecast and 0 indicates no skill. Mathematically, it is represented by:

ETS =

Where Expected hits = (Hits + Misses) x (Hits + False Alarms) / (Hits + Misses + False alarms + Correct Rejection)

Analysis

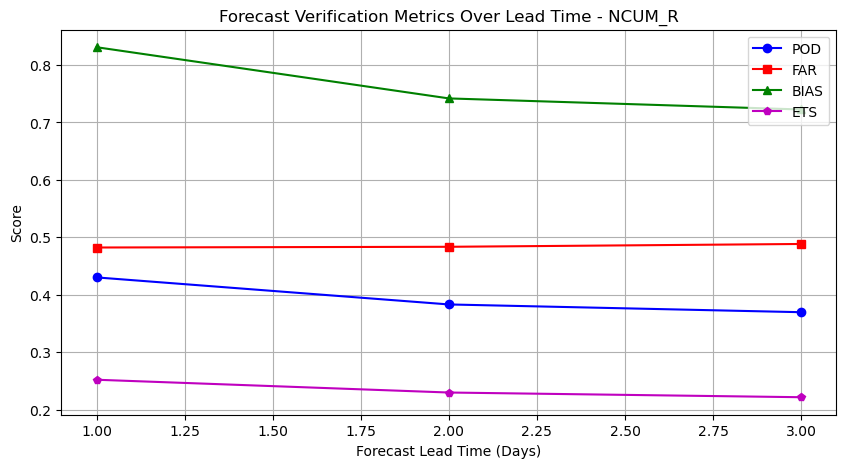
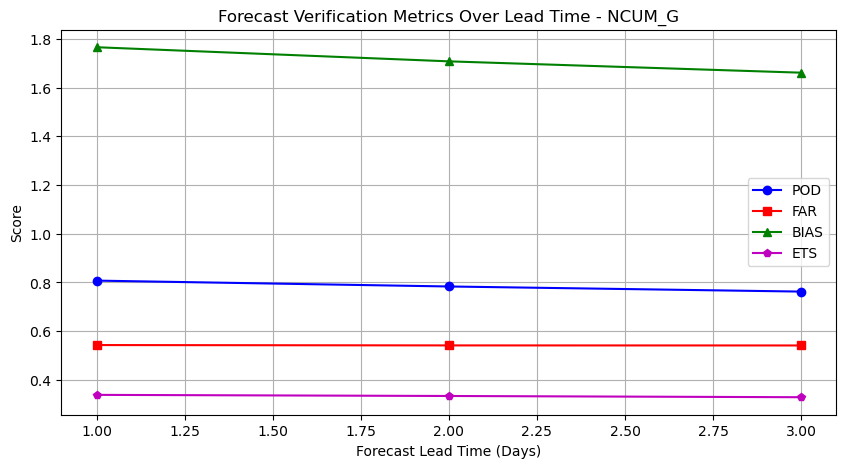


Figure : Forecast verification metrics (POD, FAR, BIAS, and ETS) over different lead times for NCUM-G. The performance remains relatively stable, with minor variations in BIAS and POD as lead time increases. Figure Y: Forecast verification metrics (POD, FAR, BIAS, and ETS) over different lead times for NCUM-R. The skill scores show a gradual decline with increasing lead time, particularly for POD and ETS, indicating a reduction in forecast accuracy.

NCUM\_G Plot Analysis:

POD is the measure of actual events that are correctly forecasted. As we see in Figure J, in the case of NCUM\_G, POD starts high (around 0.8) and decreases slightly as the lead time increases, this means the ability to detect the events of rainfall is decaying over time, that is Day 1 predictions are the most reliable compared to Day 2 and Day 3 forecasts.

FAR is the fraction of forecasted events that did not occur. FAR stays the same with increasing lead with a value little above 0.5, suggesting that the model tends to over-predict events.

Bias indicates whether the model tends to over-predict (Bias > 1) or under-predict (Bias < 1) events. The Bias is quite high, around 1.8 for Day 1 forecasts, and it reduces with lead time. This means that NCUM-G tends to overpredict, and it smoothens out with lead time much like our previous analysis.

ETS measures the accuracy of the forecast relative to random chance. ETS stays relatively stable with lead time, indicating that the forecast skill remains the same, about 0.3, as the lead time increases.

NCUM\_R Plot Analysis:

POD in NCUM\_R starts lower (around 0.4) and decreases slightly over lead time, then remains the same value after Day 2, detection rate of NCUM\_G is better than NCUM-R.

Although POD is lower than NCUM-G, the value of FAR in NCUM\_R is less than 0.5, which means that although it struggles hard to detect the events, the accuracy of the forecast is a bit more than NCUM-G. Similar to NCUM\_G, but it starts from a lower baseline, suggesting fewer false alarms initially.

Although bias in NCUM\_R starts around 0.8, it decreases with an increase in lead time. This means that the NCUM-R tends to give predictions closer to the actual Precipitation, although it shows a bit of underestimation.

ETS in NCUM\_R decreases a bit with lead time and then remains constant. The value is somewhat similar to NCUM\_G, but starts from a lower value, indicating lower initial forecast skill compared to NCUM\_G.

* + 1. QQ plot of Rainfall

A quantile–quantile plot or Q-Q plot indicates bias in is made to understand the reliability of the NCUM-R plot compared to the NCUM-G model in predicting daily surface rainfall. The data set used is the same dataset we have been using. This plot is made by comparing the empirical quantiles of the observed rainfall and model predictions of both NCUM-R and NCUM-G. As the data is spatial data, the first step is to flatten the data to a long 1-dimensional data. After this, Sorting is done where both observed and predicted values are sorted in ascending order. These are done for Day 1, Day 2, and Day 3 of the Model and observation dataset. Then, quantile matching is done; it is the process of plotting corresponding quantiles of observed and predicted distributions against each other. We also add a theoretical 1:1 line (red dashed) to show the perfect agreement for reference. Separate Q-Q plots are generated for NCUM-G and NCUM-R forecasts at lead times of Day 1, Day 2, and Day 3, in python using libraries like NumPy and matplotlib.

The figure N and figure O correspond to NCUM-G and NCUM-R Day 1, Day 2, Day 3 forecasts. Comparing these plots to each other will reveal which model performs better. While highlighting the biases the model has, it will shed light on how and where to correct those biases.

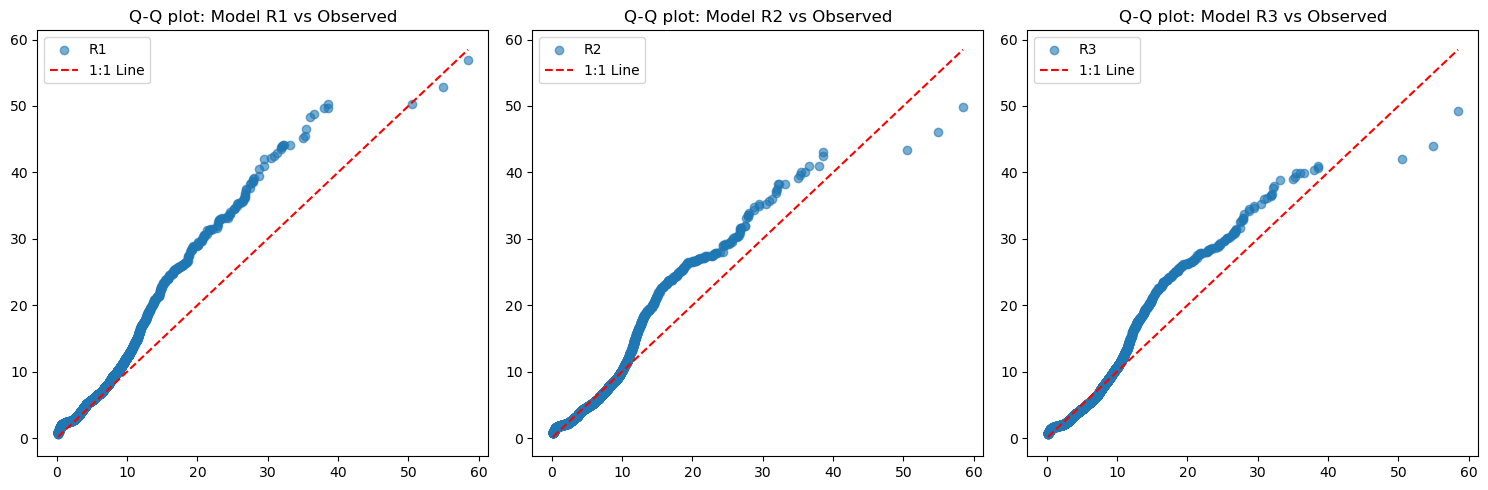
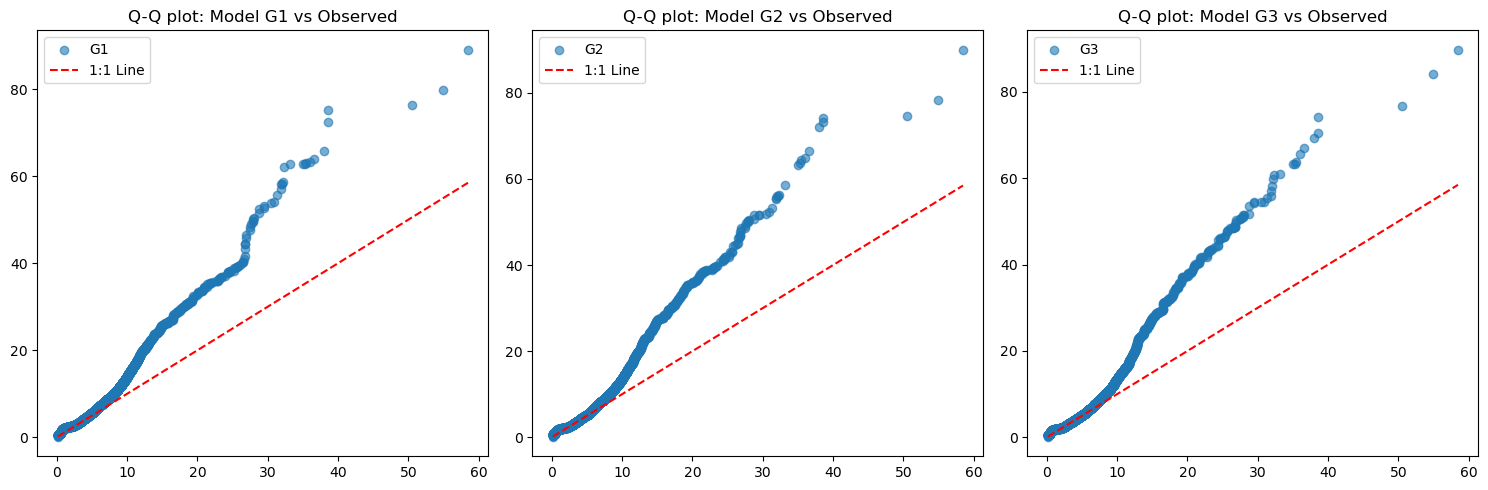


Figure : Quantile-Quantile (Q-Q) plots comparing the distribution of rainfall from NCUM-G (G1, G2, G3) with observed data. Deviations from the 1:1 line indicate biases in the model, particularly for higher rainfall values. Figure : Quantile-Quantile (Q-Q) plots comparing the distribution of rainfall from NCUM-R (R1, R2, R3) with observed data. The models show better agreement with observations at lower values but exhibit biases at higher extremes.

For NCUM-G, although at lower values (near zero), the predictions are relatively close, and it shows a significant deviation from the 1:1 line, particularly in the upper quantiles. It is seen that as the values increase, the deviation from the 1:1 line becomes stronger; this shows the systematic underestimation of higher values by the model. As lead time increases, the model doesn’t have any apparent changes, suggesting that the model is consistent with lead times.

Compared with that of NCUM-R, it shows a much closer alignment to the 1:1 line. And stay close to it for bigger quantiles compared to that of NCUM-G. It is also to be noted that, while there are deviations in the upper quantiles, these discrepancies are noticeably smaller in comparison to the NCUM-G model. Which implies the NCUM-R model is less biased and is better at capturing a bigger range of observed values.

NCUM-G has a more pronounced bias, especially for larger values. These values deviate significantly upwards in higher quantiles, meaning that predicted values tend to be much larger than observed values, confirming overprediction. The curvature of the Q-Q plot is much predominant on NCUM-G; this proves the earlier statement.

NCUM-R is a superior model compared to NCUM-G in terms of distributional fit. The poor fit of NCUM-G suggests that the model is not well-calibrated for the observed data, possibly due to incorrect assumptions, model limitations, or inadequate parameter tuning. This may be due to a myriad of different reasons like lack of flexibility in model structure, improper training or tuning of parameters, Bias in training data, or Transformation issues.

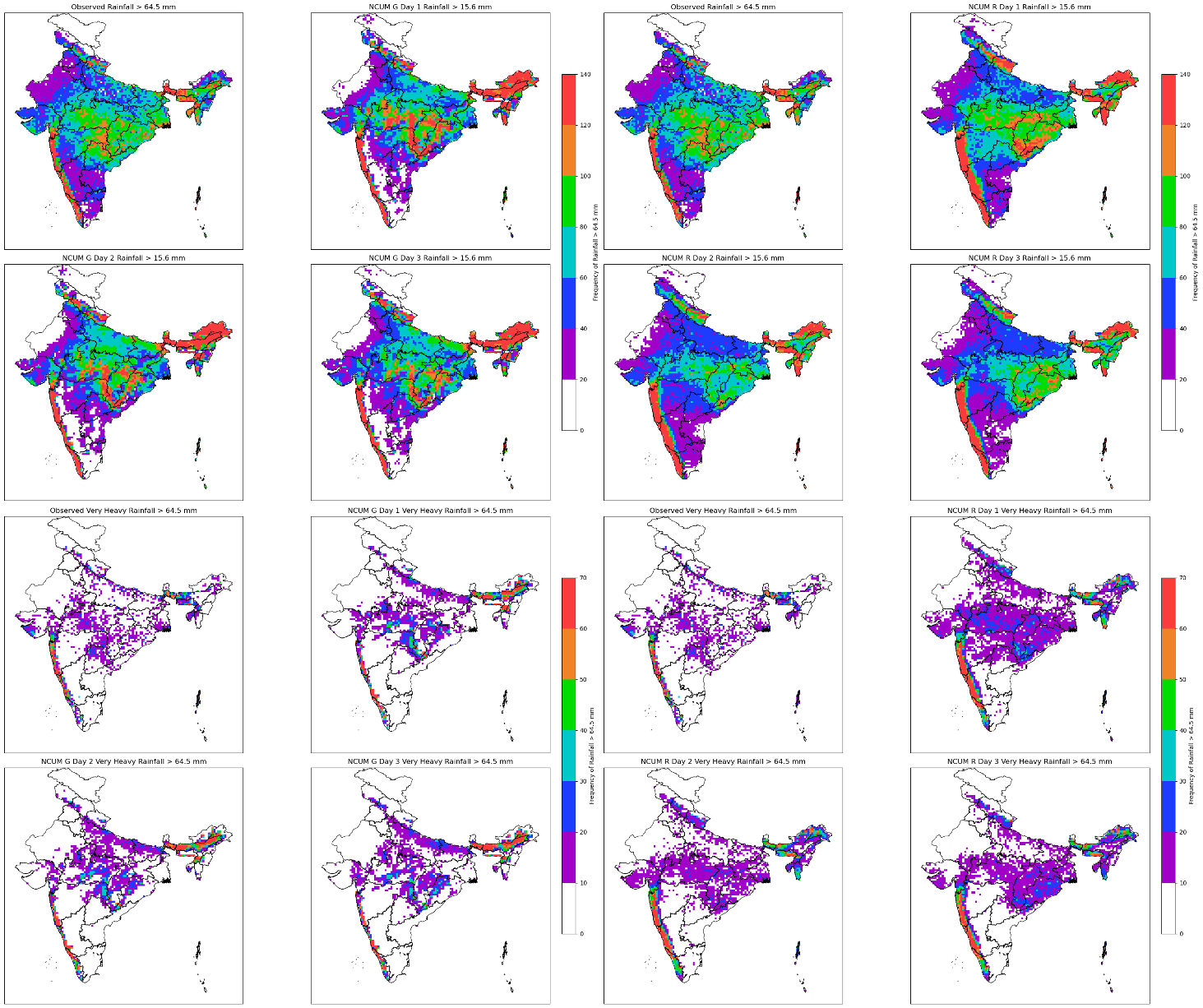
Refining the NCUM-G model by introducing more flexible parameterization, adjusting training strategies, or adding transformations to better capture extreme values. , investigating residual errors, could confirm and identify certain systematic patterns that exist in NCUM-G. The made bias was corrected to make it an ideal data for modern-day forecasts.

* + 1. Rainfall Frequency

So, we have understood the importance of doing bias correction on the model data to give a better data product, which will yield a better forecast. Now, the next question that remains is the source of these biases. So, for that, we have to analyse the Frequency plot of Intense rainfall. These are done primarily for two thresholds: 1. moderate rainy days (“15.6 mm/day”) and 2. heavy rainy days (“64.5 mm/day”).

In figures M and N, the model is the Frequency of moderate rainfall, i.e. events exceeding 15.6 mm across India. The analysis is conducted over three forecast lead times to assess how well each model captures the rainfall events. The colour scale represents the frequency of rainfall occurrences, ranging from low (blue) to high (red).

The first panel of both the plots shows the observed dataset, in it the region which receives the most frequent moderate rainfall is Northeastern India (Assam, Arunachal Pradesh, Nagaland, and Meghalaya), Western coastal regions (Kerala, Karnataka, Konkan region in Maharashtra, and Goa), Central India (Madhya Pradesh, Chhattisgarh, and Odisha), Northern states like Jammu & Kashmir, Himachal Pradesh, and Uttarakhand and the Northwest India (Rajasthan, Gujarat, and parts of Punjab and Haryana) which have lower rainfall frequencies are identified as Drier regions.

**Figure :** Spatial distribution of observed and modelled rainfall over India for different lead times. The top two rows compare total rainfall from **NCUM-G** and **NCUM-R** models against observations during moderate rainfall (≥ 15.6mm), while the bottom two rows highlight areas experiencing very heavy rainfall (≥ 64.5 mm). Differences in spatial patterns indicate model biases in predicting extreme rainfall events.

When looking at the NCUM-G model, it is seen to capture the general rainfall pattern well, especially in the Northeast and Western Ghats. But it is seen to overestimate rainfall in parts of North India (Punjab, Haryana, western UP) and underpredict in Central India, particularly in Odisha and parts of Maharashtra. With an increase in lead time, the model degrades while the overestimations of wet regions persist, the model gets drier, and the drier biases regions get more apparent. This inconsistency shows the decay of forecast skill with lead time.

Comparing it to the NCUM-R model, the NCUM-R has more accurate spatial alignment and matching frequency with observed rainfall. The Western Ghats and Northeast India are well captured, and it has fewer overpredictions compared to NCUM-G. But even then, there are some underestimations still present in Odisha and parts of Maharashtra. The lower overprediction reflects thelow false alarm rate of NCUM-R. It is also noted that as the lead time of forecast increases, the model doesn’t decay much; instead, it smoothens out. NCUM-R is a better spatial match with observations. With lead time, it retains reasonable accuracy and shows less overprediction. In the Western Ghats, Northeast India, an accurate representation with realistic intensity and distribution is displayed.

NCUM-R or dynamical downscaling in general is a sure-fire way to a give a consistent way to refine global model outputs, but it has its challenges. Systematic biases creep in due to model parameterizations, boundary conditions, and the limitations of the driving model itself. This leads to noticeable discrepancies, especially for parameters like rainfall, which is highly sensitive to model physics, terrain influences and parametrisation schemes. If left uncorrected, they can reduce the reliability of high-resolution predictions.

This is where statistical bias correction comes in. Instead of relying on model physics alone, statistical methods use historical observational data to correct systematic errors in model outputs. By adjusting the outputs to take into account the bias, we can bring the simulations closer to reality, making them more useful for modern - day applications. In the next section, let’s explore these techniques in detail - how they work, their strengths, and how they improve the reliability of downscaled data.