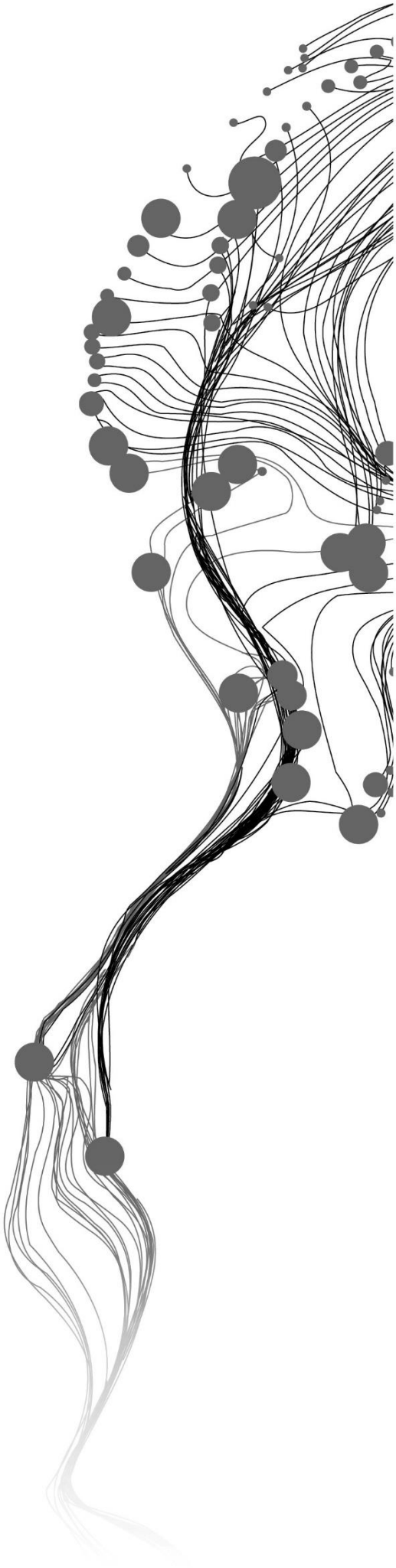
COMPARATIVE ANALYSIS OF RAINFALL SURFACE GENERATION USING DETERMINISTIC, STOCHASTIC AND DEEP LEARNING APPROACHES

SWARAJ SAHA
September 2024

SUPERVISORS:
Mr. Prabhakar Alok Verma
Dr. Ir. Frank Osei



COMPARATIVE ANALYSIS OF RAINFALL SURFACE GENERATION USING DETERMINISTIC, STOCHASTIC AND DEEP LEARNING APPROACHES

SWARAJ SAHA
Enschede, The Netherlands, [September,2024]

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: M-GEO (Geoinformatics)

SUPERVISORS:
Mr. Prabhakar Alok Verma
Dr. Ir. F.B. Osei

THESIS ASSESSMENT BOARD:
Prof. dr. ir. A. Stein (Chair)
Prof. dr. R.D. Garg (Civil Engineering Department, IIT Roorkee)

ABSTRACT

Rainfall surface generation plays a critical role in climate science, providing essential insights into the spatial distribution of precipitation. Accurate rainfall data are pivotal for disaster management, water resource planning, and understanding regional climate dynamics, particularly in regions with complex topography like the Himalayas. Despite advances in spatial interpolation techniques, generating high-resolution rainfall surfaces that accurately reflect regional variability remains a challenge. Traditional deterministic and stochastic methods often struggle with capturing fine-scale variations, especially in areas with sparse observational networks. The advent of deep learning, however, offers new opportunities to enhance the precision and efficiency of rainfall surface generation.

This thesis addresses the problem of generating high-resolution, daily rainfall surfaces for the state of Uttarakhand, India, a region characterized by challenging terrain and significant risk of natural disasters. The research identifies a gap in existing methodologies, which often fail to integrate the computational power of deep learning with traditional spatial interpolation techniques. The objective of this study is to develop a novel spatial interpolation framework that leverages deep learning to generate precise, daily gridded rainfall data and to compare its performance against traditional methods.

The methodology involves a comparative analysis of deterministic, stochastic, and deep learning approaches to rainfall surface generation. The study uses daily rainfall observations from automatic weather stations (AWS) in Uttarakhand, employing geostatistical interpolation methods such as Kriging and Regression Kriging. The deep learning model, based on the Graph Neural Network (GNN) framework and Gaussian Mixture Model (GMM) convolutions, is trained to predict rainfall surfaces from these observations. The model's performance is evaluated using Root Mean Square Error (RMSE) and compared with traditional methods and existing gridded data provided by the Indian Meteorological Department (IMD).

Results indicate that while traditional Kriging methods capture general trends, they often fail in areas with sparse data, leading to inaccurate predictions in unobserved regions. The deep learning model, on the other hand, shows promising results, particularly in capturing complex spatial patterns and providing moderate to high accuracy across the study area. The study also reveals that elevation, while statistically significant, has a limited impact on rainfall predictions in this context, suggesting the need to incorporate additional climatic variables for improved accuracy.

The implications of this research are significant for regional planning and disaster management in Uttarakhand. The enhanced rainfall surfaces can improve early warning systems for floods and landslides, contributing to better risk management and spatial planning. The deep learning model, once further refined and trained on larger datasets, could be deployed in real-time applications, providing continuous updates on rainfall patterns.

Future work will focus on expanding the study area, increasing the number of training samples, and incorporating additional environmental variables to improve model accuracy. Moreover, integrating satellite-based rainfall products with ground observations could offer a more comprehensive approach to rainfall prediction on both regional and global scales.

This research demonstrates the potential of deep learning as a powerful tool in climate science, capable of overcoming the limitations of traditional interpolation methods and providing accurate, high-resolution rainfall data critical for environmental management.

ACKNOWLEDGEMENTS

I would like to express my profound gratitude and appreciation to all individuals who have contributed to the completion of this thesis. Their support and involvement have been instrumental in shaping this work, and I am sincerely grateful for their contributions.

First and foremost, I extend my deepest gratitude to my first supervisor, Mr. Prabhakar Alok Verma. Without his invaluable guidance, this thesis would not have been possible. His unwavering support and positive demeanour served as a catalyst for my motivation. Our regular meetings and discussions were a constant source of inspiration, propelling me to present part of this thesis at a conference and submit a research paper for publication. These achievements exceeded my initial expectations, and I fully credit his mentorship for this success. I am truly grateful for the comfort and openness of our discussions, enabling the productive exchange of ideas. I could not have asked for a better first supervisor.

I would also like to express my gratitude to my second supervisor, Dr. Ir. Frank Osei. His invaluable feedback emphasized the significance of meticulous attention to detail in writing, presenting, and overall implementation. His guidance has helped me strengthen the foundation of this work, ensuring its quality and coherence.

I extend my heartfelt appreciation to my dear friends, Narendra, Srihari, Jay and Rachit for their unwavering kindness, support, and availability. Their companionship and the meaningful conversations we shared have been a source of joy and fulfilment. I treasure these connections and look forward to fostering further dialogue with them in the future.

I cannot express enough gratitude to my parents for their unyielding support and provision throughout this academic journey. Their unwavering belief in me has been a constant source of motivation and strength.

To all individuals mentioned above, as well as anyone else who has contributed directly or indirectly to this thesis, I extend my sincerest thanks for your support. I hope that the content presented herein does justice to your efforts.

Thank you, and I wish you a pleasant reading experience.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

Rainfall surface generation is critical in Climate Science since it provides valuable insights into the spatial patterns of precipitation that impact a range of environmental and socioeconomic variables. Robust disaster mitigation techniques, efficient management of water resources, and a comprehensive understanding of regional climate dynamics depend on precise and comprehensive rainfall data.(F. W. Chen & Liu, 2012) Accurately forecasting and interpreting rainfall patterns is essential for tackling various issues, from controlling natural disasters to ensuring sustainable water supply. With substantial significance for both ecological management and societal development, this predictive capacity is not just an academic endeavour.(Ribeiro et al., 2022) For well-informed spatial planning and infrastructure development, especially in areas susceptible to natural disasters, reliable rainfall records are especially crucial. For instance, the rough terrain and steep elevation gradients of the Himalayas present difficulties for rainfall forecasting. These topographical characteristics increase the danger of flash floods and landslides in such geological environments.(Delbari et al., 2013) Reliable precipitation data becomes essential for creating efficient early warning systems and putting risk reduction plans into action. Increased demand for accurate rainfall data due to the growth of human settlements in these risky locations underscores the necessity of techniques that can deliver fast and accurate data to enhance disaster preparedness and lessen the effects of natural disasters.(Y. C. Chen et al., 2008)

The growing human populations in hilly and mountainous places, such as Uttarakhand, India, emphasise the importance of comprehensive rainfall data.(Delbari et al., 2013) These susceptible locations are more likely to experience landslides and flash floods due to rapid urbanisation and infrastructure development, which can have catastrophic consequences for the surrounding populations, infrastructure, and wildlife. Inadequate rainfall data combined with inadequate spatial planning can result in poorly managed development, making people more vulnerable to these dangers. High-resolution, daily rainfall datasets can give the specific information required for good land use planning, allowing high-risk locations to be identified and appropriate preventive measures implemented.(Sati, n.d.) There are also significant ecological ramifications to rainfall fluctuation. Variations in precipitation patterns can have an impact on ecosystem health and distribution, which can impact biodiversity and the long-term viability of natural environments. For instance, changed rainfall patterns may affect soil moisture, which may then impact the development of plants and the accessibility of nutrients for wildlife. Accurate rainfall data is essential for comprehending and reducing the effects of climate change on these delicate settings, especially in areas like the Himalayas, where distinct and diversified ecosystems are present. By guiding management plans that consider the variability and trends in precipitation, detailed rainfall data helps preserve natural ecosystems and save endangered species.(Goovaerts, n.d.)
Spatial interpolation techniques have grown greatly as computer technology has advanced, allowing for more complex strategies to improve rainfall surface creation.(Rodriguez-Ramirez & Fuentes-Mariles, 2023) In climate science, spatial interpolation—especially using geostatistical techniques—has long been a key technique for generating continuous surfaces from discrete data points. These techniques have shown to be crucial in producing comprehensive and trustworthy datasets that reflect the spatial variability of climatic variables like precipitation(Papacharalampous et al., n.d.). They are essential to environmental research and climate modelling because of their capacity to handle intricate spatial patterns and offer useful estimates in

regions with little observational data.(Meuer et al., n.d.) The incorporation of machine learning, and more especially deep learning, has brought strong tools that further improve these methods as computing technology has advanced. Machine learning algorithms, with their capacity to learn from extensive datasets, enhance the ability to analyze complex relationships and produce high-resolution predictions. (Zhang et al., 2022)Deep learning, building on this foundation, leverages advanced computational methods to model intricate patterns and interactions within climate data, offering new potential to capture detailed spatial dynamics that traditional methods might overlook.

1.2. Research Identification

An enhanced framework that can generate high-resolution rainfall surface data that properly reflects regional spatial subtleties while maintaining efficiency and rapid model execution is urgently needed, especially with the recent developments in deep learning algorithms and computing capabilities. These techniques complement and expand upon the capabilities of traditional geostatistical methods by utilising deep learning to improve rainfall estimation precision and drastically cut down on computation time. This research aims to create a unique spatial interpolation framework that uses deep learning to generate precise daily gridded rainfall data and assess its performance compared to traditional approaches. The research findings have significant potential to guide the Uttarakhand government in India, aiding in creating an all-encompassing regional rainfall inventory. In places prone to landslides and floods, where accurate and timely rainfall data is crucial for effective risk management and planning, this inventory is essential for bolstering catastrophe resilience frameworks and optimising spatial design.

1.3. Research Objectives

This project seeks to generate gridded rainfall datasets from daily observations recorded by various Automatic Weather Stations (AWS) within the study area, employing geostatistical interpolation methods. Additionally, the research aims to explore the potential applications of deep learning algorithms for producing gridded rainfall data. The effectiveness of these deep learning techniques will be compared with traditional geostatistical methods and with the gridded data provided by the Indian Meteorological Department (IMD), to assess their relative performance and accuracy.

Sub-Objectives:

1. Implementation of Stochastic Interpolation Techniques for Predicted Rainfall Surface Generation.
2. To develop a novel way of Rainfall Prediction model using Deep Learning (DL)
3. To analyze how changes in Elevation can influence Rainfall Prediction.
4. To evaluate grid quality with the existing IMD grids.

1.4. Research Questions

1. **Referring to Sub-objective 1:**

a. Which model of Interpolation going to be used?

b. How well this model is performing in the prediction?

2. Referring to Sub-objective 2:

a. How to implement Deep Learning for Interpolation?

b. How effectively can Rainfall be predicted using DL?

3. Referring to Sub-objective 3:

a. Is it possible to establish a relation between the elevation and Rainfall, in rainfall prediction analysis? Can this relation influence the predictive performance of the Rainfall Surface Generation?

4. Referring to Sub-objective 4:

a. How the newly developed grids are performing alongside with the IMD, for rainfall prediction?

b. Does the application of deep learning provide any significance?

2. LITERATURE REVIEW

Rainfall surfaces, mainly those formed daily, are critical for evaluating and managing environmental phenomena. These surfaces are created using observational data from many places and show the geographical distribution of rainfall throughout an area. These surfaces offer comprehensive insights into patterns of precipitation, which are essential for climate professionals working in hydrology, meteorology, and disaster management.(Mitra et al., 2009) Daily rainfall surfaces, which record the immediate consequences of meteorological events, provide real-time data necessary for precise forecasting and prompt decision-making, especially in situations where knowing short-term fluctuations is crucial to reducing the effects of floods or droughts. (Navalgund et al., 2018)

The limited distribution of observational networks makes it difficult to estimate climatic variables, particularly rainfall, in unsampled places. Precise forecasting is essential for applications like catastrophe risk reduction, water resource management, and agricultural planning in areas with limited data.(Martorell et al., 2009) In this context, the importance of regional studies is highlighted since they offer localised insights on climate behaviour, which are essential for creating policies that are particular to environmental issues. Rainfall studies contribute vital data to local climate research, helping to anticipate future scenarios and influence regional decision-making processes.(Cecinati et al., 2018)

Addressing the challenge of estimating climate variables in unsampled locations, spatial interpolation techniques generate continuous surfaces from discrete data points. Stochastic, deep learning and deterministic methods improve prediction accuracy and dependability, especially in areas with scant observational data. An extensive analysis of these approaches will be provided in the upcoming sections of this literature review, with a particular emphasis on their use in regional rainfall studies.

2.1. Review on Spatial Interpolation Techniques

Spatial interpolation methods, whether deterministic or stochastic, are critical for estimating values at unsampled locations. While deterministic methods, such spline interpolation and Inverse Distance Weighting (IDW), offer computational efficiency, they frequently fail to capture spatial variability and uncertainty since they rely their estimations on the spatial connections among known data points.(Rodriguez-Ramirez & Fuentes-Mariles, 2023) These drawbacks are overcome by stochastic techniques, such as variogram analysis and kriging, which use probabilistic models to take uncertainty and spatial dependency into account.(Zsolt Farkas et al., 2017) For instance, kriging yields the best linear unbiased predictions; variants such as Ordinary Kriging are helpful in forecasting rainfall patterns, as seen by research conducted in Tunisia, where low RMSE and mean error values supported the accuracy of the estimates.(Feki et al., 2012) Furthermore, comparing multivariate geostatistical techniques shows that adding elevation data to rainfall prediction improves accuracy, especially in areas with moderate rainfall-elevation

correlations (e.g., kriging with variable local means and external drift).(Zareifard et al., 2023) It has been demonstrated that integrating elevation data greatly improves forecast accuracy, particularly in regions with prominent topographic relief. This has been the subject of various research. By incorporating measurement uncertainty, advances in kriging, such Kriging for Uncertain Data (KUD), significantly enhance rainfall estimate. Applications show improved forecasts when gauge density is sufficient. (Rodriguez-Ramirez & Fuentes-Mariles, 2023)Moreover, techniques combining entropy and kriging to optimise rainfall network design show data accuracy and network efficiency improvements. The significance of parameter selection has also been emphasised by IDW studies, with the best power and radius values improving interpolation accuracy, especially in central Taiwan. (Mukhopadhaya, n.d.)Collectively, these developments demonstrate the continued relevance and efficiency of both standard and new spatial interpolation approaches for properly forecasting rainfall and handling hydrological data.

2.2. Review on Deep Learning Techniques

Deep learning-based interpolation algorithms outperform classic geostatistical techniques by capturing complicated, non-linear spatial connections and managing heterogeneity in big datasets. These data-driven techniques go beyond the constraints of preset spatial models employed in conventional approaches like Kriging and provide flexibility in resolving issues related to non-stationary and sparse data.(Liang et al., 2022) Deep learning techniques can better incorporate auxiliary information and adapt to geographical situations by learning from the data independently.(*GlobalSIP : 2014 IEEE Global Conference on Signal and Information Processing : 3-5 December 2014, Atlanta, GA, USA*, 2014) For example, Bayesian deep learning has improved spatial interpolation by learning complex relationships between point-sampled data and auxiliary variables like terrain elevation. This has been successfully applied to mapping calcium concentrations in stream sediment across the United Kingdom (UK), yielding a strong probabilistic calibration and a high coefficient of determination ($R^2 = 0.74$)(Johnson et al., n.d.). By integrating numerous algorithms, an ensemble technique with nine quantile-based learners has improved probabilistic predictions, surpassing conventional quantile regression over a range of quantile levels, with gains in predictions ranging from 3.91% to 8.95%.(Bronstein et al., 2021) The possibilities of deep learning techniques are further demonstrated by the development of the Deep Geometric Spatial Interpolation (DGSI) framework, which uses a multilayer perceptron to predict neighbour weights based on distance and orientation. DGSI outperformed traditional and advanced interpolators, with an average RMSE decrease of 48% across different datasets.(Xiang & Demir, n.d.) All these developments demonstrate how deep learning-based approaches may improve spatial interpolation by overcoming the drawbacks of conventional methods and increasing accuracy in various spatial situations.

2.2.1. Graph-based Models

A graph can be used to conceptualise the spatial distribution of rainfall, with rainfall measurement locations acting as nodes and the spatial interactions between them as edges. Each node in this graph-based structure represents a distinct observation of cumulative precipitation or rainfall intensity. These nodes are connected by edges created depending on their physical closeness or connectivity; the edge weights indicate the degree of spatial correlation or node distance. (Liang et al., 2022)

Graph-structured data is handled using specialised neural network topologies called Graph Neural Networks (GNNs). GNNs utilise neural network operations to process input through nodes, edges, and global context to develop better embeddings for these elements.(Scarselli et al., 2009) GNNs update edges and nodes separately at first. Advanced GNN systems, on the other hand, use message transmission mechanisms, which let nodes and edges interact with nearby entities and integrate data in accordance with the graph's topology.(Kirkwood et al., 2022) Although GNNs can capture complex relationships within the graph structure, they may accomplish complicated tasks like node categorisation, edge prediction, and global graph analysis through this iterative process of collecting information throughout the graph.

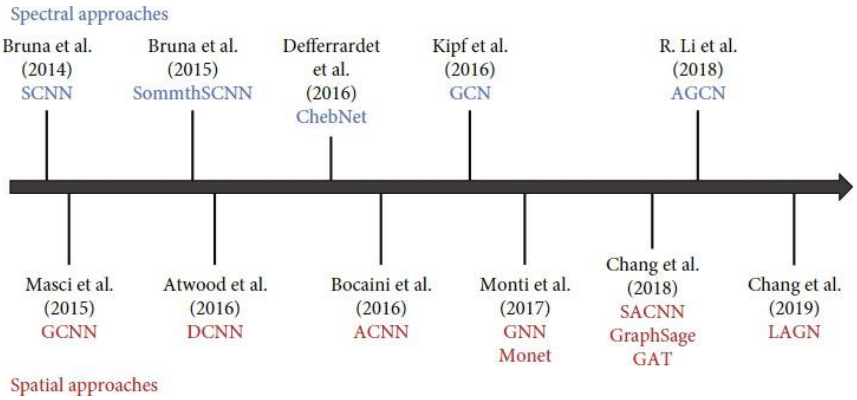


Figure 1 *Current Research Flow of GNN*

There are two types of GNN architectures: spectral and spatial. Using graph Laplacians and eigenvalue decomposition, spectral methods—such as Spectral Networks (SCNN) and Graph Convolutional Networks (GCN)—perform convolutions in the spectrum domain. Nevertheless, they have difficulties in integrating edge characteristics and scaling. Conversely, spatial approaches overcome these drawbacks by directly allowing convolutional procedures to be applied by converting non-Euclidean graph data into Euclidean space. This change improves efficiency and flexibility.(Monti et al., 2016) Notable spatial techniques include Graph Attention Networks (GATs), which incorporate attention mechanisms to dynamically adjust convolution parameters based on node interrelationships, and Graph Sample and Aggregate (Graph SAGE),(Veličkovi´veličkovi´c et al., n.d.) which uses sampling and aggregation techniques to manage unordered neighbour sets. By dynamically allocating significance weights to nodes and their neighbours, GATs enhance the model's efficiency by highlighting the most pertinent data for each node's representation.

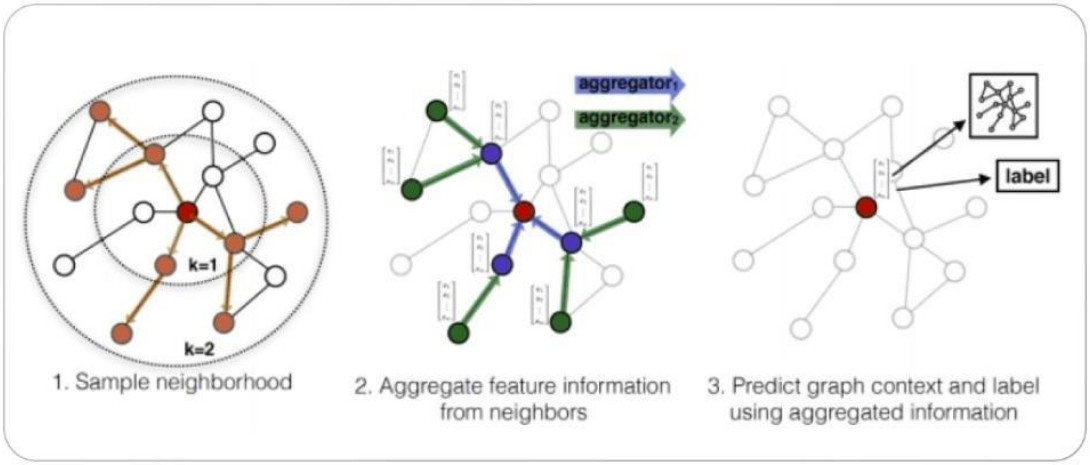


Figure 2 *Framework of Graph SAGE Algorithm(M. Chen et al., 2018)*

2.2.2. Graph Attention Network

Graph Attention Networks (GAT) use attention mechanisms to dynamically assign importance weights to nodes and their neighbours, as opposed to traditional Graph Neural Networks (GNNs) that use fixed weights.(Veličkovi´veličkovi´c et al., n.d.) GAT computes node embeddings by evaluating and attending to the features of neighbouring nodes, enabling the model to emphasize the most relevant neighbours in each node's representation. This adaptive approach improves the model's ability to capture complex relationships within graph-structured data.
In Graph Attention Networks (GAT), the updated feature vector h_v for a node v is computed by leveraging an attention mechanism to dynamically aggregate and weight the feature vectors of its neighbouring nodes $\mathcal{N}_u$. The formulation for this process is given by:

$$h_v = \boldsymbol{\phi}\,(x_u, \oplus_{v\in\mathcal{N}_u}\, \psi(x_u, x_v)) \qquad (1)$$

where: ϕ denotes a transformation function that maps the aggregated information into a new feature space. This function is generally a linear layer followed by a non-linearity, which transforms the combined features into the updated representation for node v. ψ is the attention mechanism that computes the attention score between the feature vector x_u of the central node u and the feature vector x_v of each neighbouring node v. This score denoted as $\psi(x_u, x_v)$, quantifies the relevance or influence of node v on the central node u. Usually, a learnable function—like a single-layer feedforward neural network with a LeakyReLU activation—is used to calculate the attention score. This function produces a scalar weight that represents the significance of the relationship between nodes u and v. $\oplus$ represents the aggregation operation that combines the attention-weighted features from all neighbouring nodes $\mathcal{N}_u$. This operation involves
computing a weighted sum of the feature vectors x_v from each neighbour v, where the weights are determined by the attention scores $\psi(x_u, x_v)$. Through this aggregation stage, the model may incorporate relevantly weighted data from several neighbours.(Veličkovi´veličkovi´c et al., n.d.)

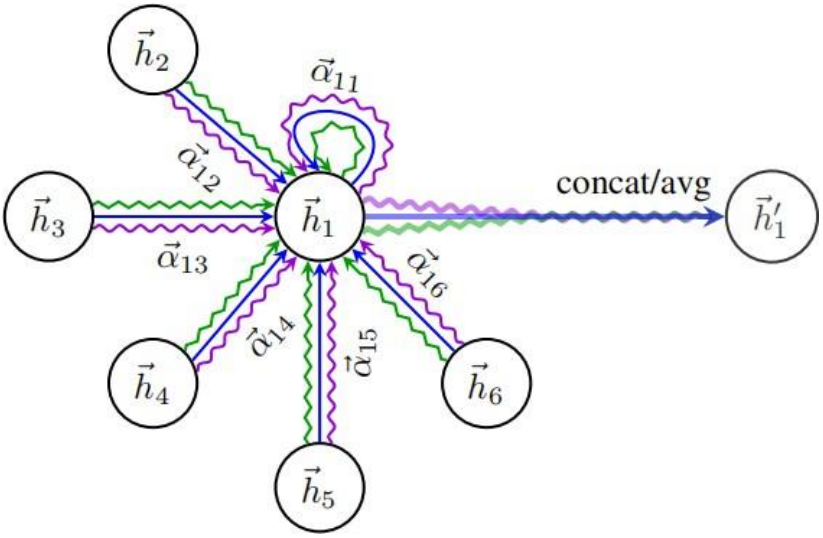


Figure 3 *An example of node 1's multiheaded attention (with K = 3 heads) on its neighbourhood. Independent attention calculations are shown by arrow styles and colours that differ. ~ h01 is obtained by concatenating or averaging the aggregated characteristics from each head.(Veličkovi´veličkovi´c et al., n.d.)*

3. STUDY AREA AND DATASETS

3.1. Study Area

Uttarakhand, a state situated in the northwestern region of India amidst the Central Himalayas, shares its borders with Tibet to the north, Nepal to the east, Himachal Pradesh to the northwest, and Uttar Pradesh to the south. The state boasts a varied landscape, ranging from the majestic Himalayan mountains and glaciers in the north to the plains in the south. In this state, around 46,000 sq. km area is hilly out of a total of 53484 sq. km of the geographical area of the state. The elevation in the state ranges from 300m to 7000m above mean sea level. The state receives an annual average rainfall of 1631mm. Southwest monsoon winds are the primary rain-bearing winds for the state, contributing70% of the annual average, or 1162.7mm, between June and September.(Navalgund et al., 2018)

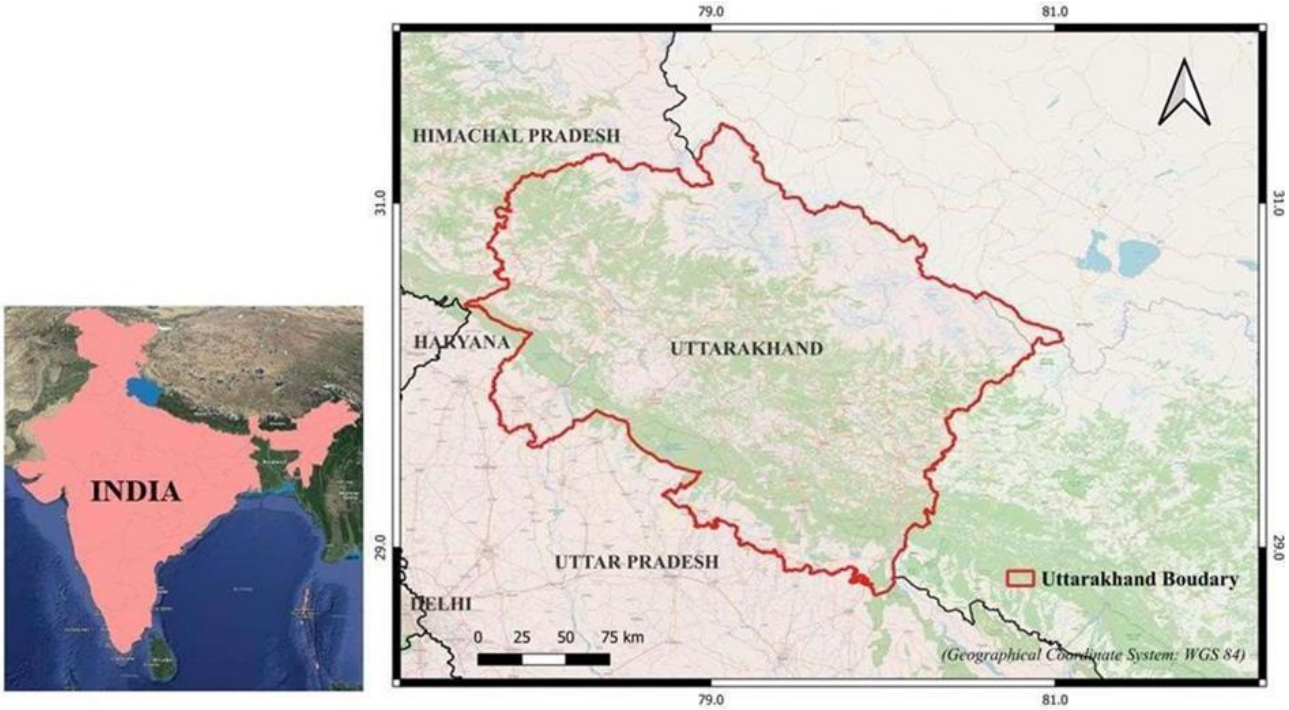


Figure 4 *The Study Area, State of Uttarakhand, India*

3.2. Topography

The state has varied landforms, such as deep valleys, high mountains, glaciers, snow-covered peaks, perennial rivers, streams, creeks, plateaus, and plains. The state also has the second-highest peak in the country, Nanda Devi (7,817 m above mean sea level). The state has different physiographic zones running parallelly from northwest to southeast. The fast-flowing rivers denote the steep slope in the region, which

usually results in landslides and landslips in the rainy season. (Mitra et al., 2009)The following Digital Elevation Model (DEM) is provided by NASA provides SRTM DEM of 30 m spatial resolution.

.

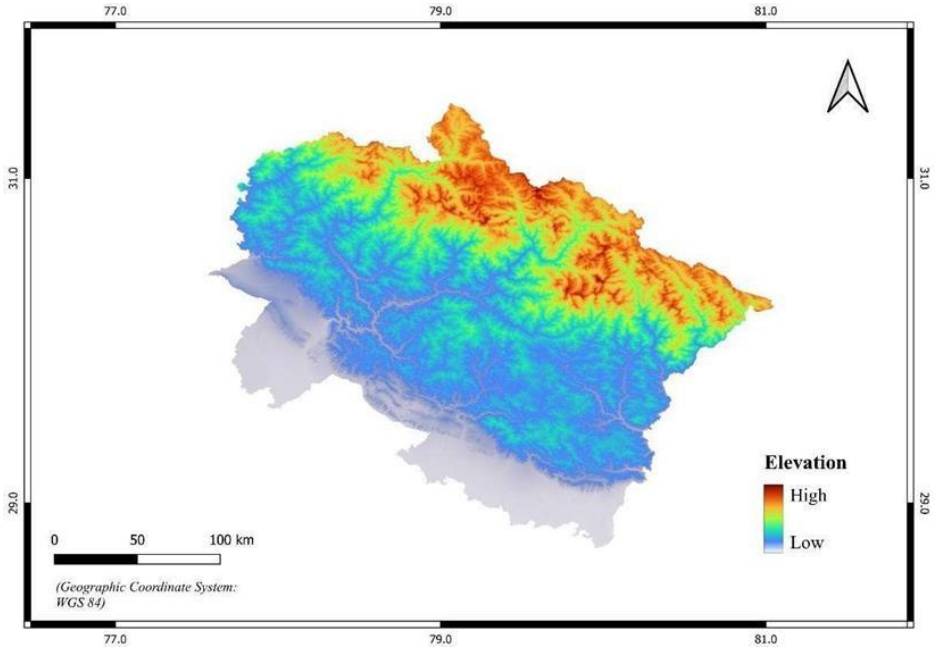


Figure 5 *Digital Elevation Model*

3.3. Datasets

3.3.1. IMD AWS Data

Automatic weather stations (AWS) are automated versions of regular weather stations. Both single-site and networked weather stations are possible. An automated weather station is a variation of a typical weather station that is automated to save staffing costs or to allow readings from far-off locations.(AWS, ARG , AGRO AWS AND ASG (Surface Instrument Division), n.d.) It offers real-time data recording and offline or non-real-time data recording for analysis. Tipping Bucket Rain Gauge, a precise device employed by IMD (Indian Meteorological Department) with a sensitivity of 0.5 mm per tip, is used in AWS to measure rainfall. As of 03 UTC today, the cumulative rainfall reported at 03 UTC today is the amount of rain that fell over the 24 hours that ended at 03 UTC today, starting at 03 UTC yesterday. The rainfall value is reset at 03 UTC and fresh logging and accumulation of the rainfall, if any, takes place as per IMD convention.(AWS and ARG Automatic Weather Stations, n.d.) There are 137 AWS across Uttarakhand situated in different locations.

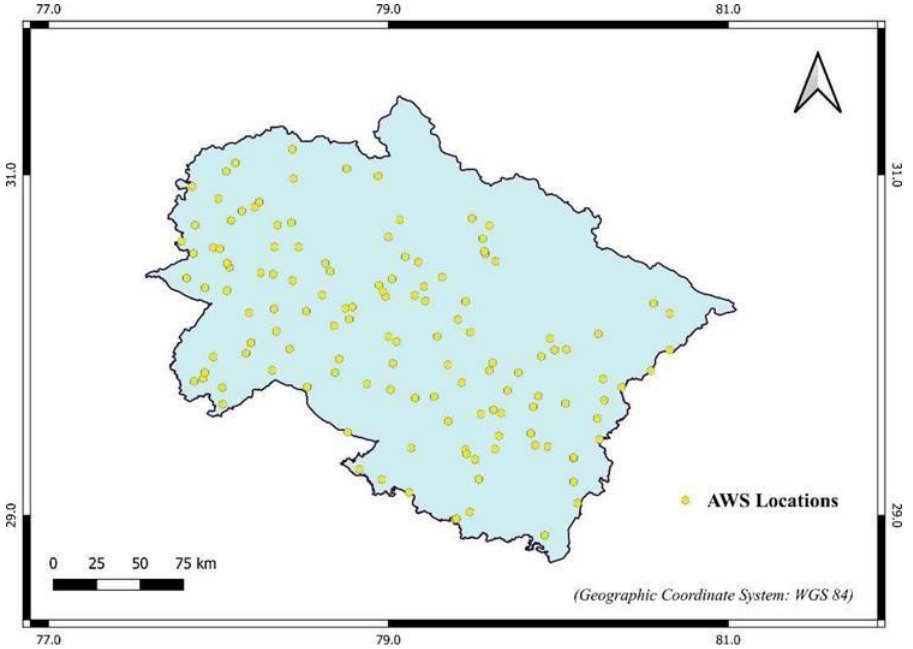


Figure 6 *AWS locations over Uttarakhand*

3.3.2. IMD's Daily Gridded Rainfall Data

The National Data Centre at IMD, Pune, creates and preserves the IMD rainfall gridded using daily rainfall data. These datasets feature a 25-kilometre geographical resolution and a one-day temporal resolution. The daily gridded rainfall dataset for India, known as IMD4, was created with a high geographical resolution of 0.25° × 0.25° and covers the years 1901 to 2010, offering essential insights into the rainfall patterns of the nation. Incorporating daily rainfall records from 6,955 rain gauge stations—the most significant number of stations utilised in such studies—distinguishes this dataset and allows for a more accurate representation of rainfall distribution across different areas of India.(Pai et al., 2014)
These datasets are developed using conventional, deterministic, spatial interpolation techniques, the Inverse Distance Weight (IDW). The formulation of this method:

$$\hat{Z}(x_0) = \frac{\sum_{i=1}^{n} \frac{Z(x_i)}{d(x_0,x_i)^p}}{\sum_{i=1}^{n} \frac{1}{d(x_0,x_i)^p}} \qquad (2)$$

Where,

$\hat{Z}(x_0)$ is the estimated value at location x_0.
$Z(x_i)$ is the known value at location x_i.
$d(x_0, x_i)$ is the distance between x_0 and x_i.
p is the power parameter, which determines the influence of distance.

In this project, these data are treated as a result of deterministic spatial interpolation method.

3.3.3. Platform Used

The computational part of this research work is performed using R and Python Programming languages. Due to several issues of versions of repositories and individual library issues, the R based works could not performed in the RStudio software of the local computing system. So, all the computational works are performed in the Geospatial Computing Platform provided by the ITC, University of Twente, developed by the Centre of Expertise in Big Geodata Science (CRIB) (https://itc.nl/big-geodata). It has distributed and GPU-accelerated capabilities for computation and analysis, out of which, the computation environment with 72 virtual CPUs (vCPUs) based on the Intel x86-64 architecture, 768 GB of RAM, and an NVIDIA RTX A4000 GPU has been used predominantly for this work. For some code checking and exploration purposes, Google Colab, a cloud-based platform provided by Google has been explored. Free version of this offers typically offers an Intel Xeon CPU (2-4 cores) with 12-16 GB RAM, and an NVIDIA Tesla K80, T4, P100, or similar GPU with 12-16 GB VRAM.

4. METHODOLOGY

The following diagram shows the overall methodology of this project, respect to each sub objectives.

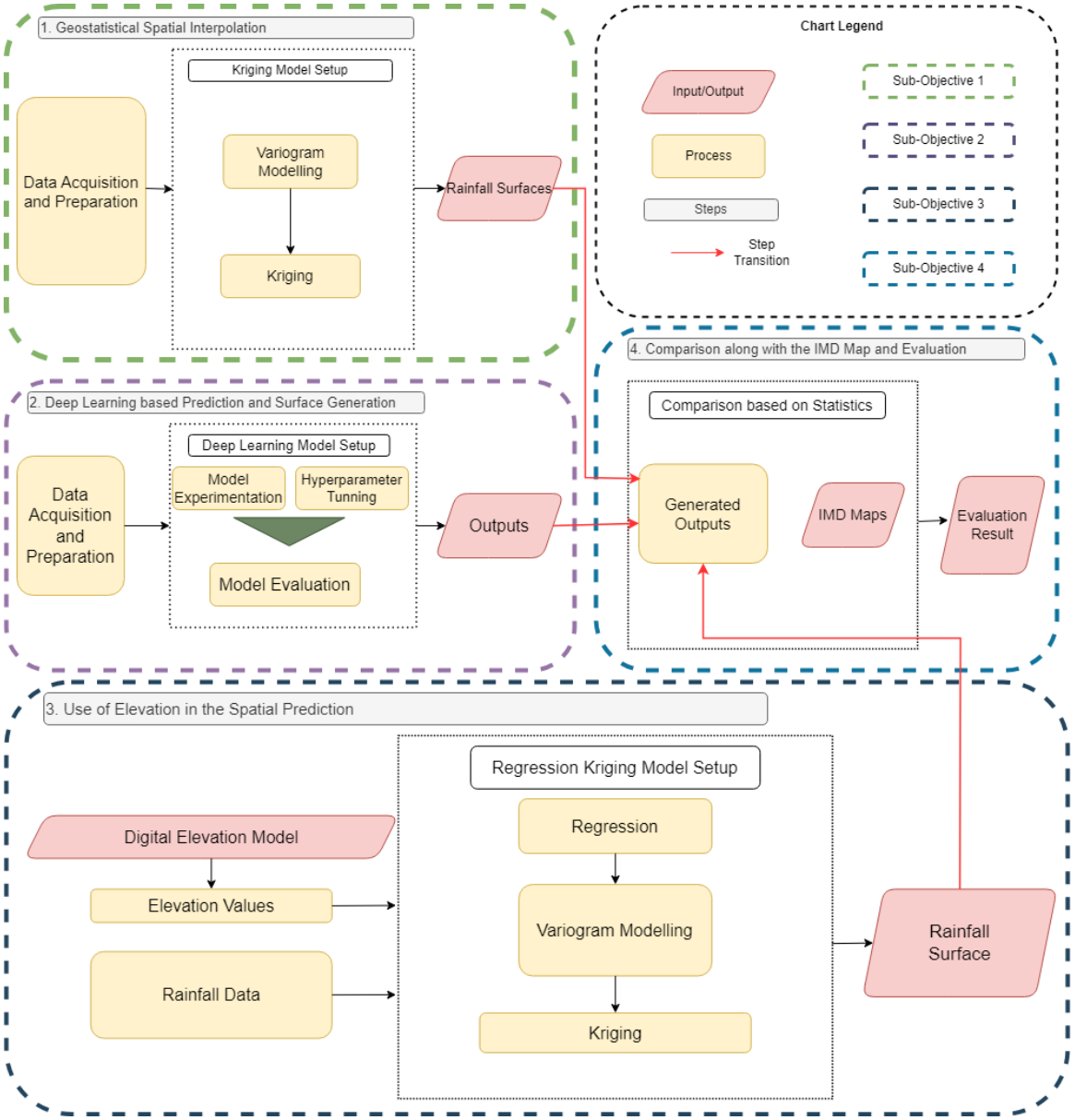


Figure 7 *Overall Framework of Methodology*

4.1. Data Preparation and Processing

The daily observations of rainfall data are provided in a comma-separated value (CSV) format along with the observation sites. However, the elevation values were present in the DEM's raster dataset. Then, those values based on their locations were cross-checked with the locations of the rainfall observations. Eventually, each day, a combined dataset of rainfall observations and corresponding elevation values was made available. Those were converted to CSV, and coarsening modules were used to reduce the accuracy of the measurements.

For have a justified analytical result and following discussion, instead of one or few days, daily data of a whole month was considered. Just to be clarified, this decision was taken to analyze the methodological insights and benefits of each method mentioned in the following passages, not to consider the temporal dependencies or patterns. Detailed discussion about the outcomes is included in Result and Discussion section.

From exploratory analysis, it has been found that the raw data were skewed. To reduce the skewness and stabilize the variance, Box-Cox transformation function was used. The Box-Cox transformation of a variable $Z(x)$ is defined as

$$Z'(x) = \begin{cases} \frac{Z(x)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(Z(x)) & \text{if } \lambda = 0 \end{cases} \qquad (2)$$

The parameter λ is chosen to maximize the likelihood function under the assumption that the transformed data follows a normal distribution. The transformed data is then geocoded by adding transformed coordinate reference system (crs) of UTM 43N zone of WGS 84.

In the Deep Learning module, the input variable, i.e. rainfall values were converted to tensor, a multi-dimensional array like NumPy arrays, optimized for PyTorch's deep learning framework.

4.2. Variogram Modelling

The transformed rainfall values then included for variogram analysis. General expression of variogram function:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i,j} [Z(x_i) - Z(x_j)]^2 \qquad (3)$$

Where:

$\gamma(h)$ is the semi variance at lag h.

$Z(x_i)$ and $Z(x_j)$ are the values of the variable of interest at locations x_i and x_j.

$N(h)$ is the number of pairs at distance h.

4.3. Maximum Likelihood Estimation

The estimation of variogram parameters were achieved through maximum likelihood estimation (MLE), aiming to optimize the likelihood function based on observed spatially correlated data. Starting with a set of observations $Z(x1), Z(x2), \ldots, Z(xn)$ at locations $x_1, x_2, \ldots, x_n$, the variogram model is defined to capture the spatial correlation structure of these observations by modelling the relationship between the variance of the differences of the data at various locations.
model includes a deterministic trend expressed as

$$m(x) = \beta_0 + \beta_1 f_1(x) + \cdots + \beta_p f_p(x) \quad (4)$$

where $f_1(x), \ldots, f_p(x)$ are basis functions, and a stochastic term represented as $Z(x) = m(x) + \epsilon(x)$, where $\epsilon(x)$ is a zero-mean Gaussian random field characterized by a covariance structure dictated by the variogram parameters. The covariance matrix $\Sigma(\theta)$ depends on the variogram parameters θ (nugget, sill, range), and the joint probability density function of the observed data is expressed as

$$L(\theta) = \frac{1}{(2\pi)^{n/2}|\Sigma(\theta)|^{1/2}} \exp\left(-\frac{1}{2}(Z - m)^{\top}\Sigma(\theta)^{-1}(Z - m)\right) \quad (5)$$

To facilitate optimization, the log-likelihood function is computed as,

$$\log L(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}(Z - m)^{\top}\Sigma(\theta)^{-1}(Z - m) \quad (6)$$

By maximizing this log-likelihood function using appropriate optimization techniques, the variogram parameters θ can be estimated.
In R, using the GeoR package, the likfit () function was used to perform MLE on the variogram models.

4.4. Covariance Function

The covariance function with the most considerable log-likelihood function or the one with the least negative log-likelihood function is chosen as the plausible covariance function in the maximum likelihood technique.
The following are plausible variogram models that were employed in this study:

The Spherical function which is given as:

$$C(h) = \sigma^2\left(1 - \frac{3h}{2\phi} + \frac{1}{2}\left(\frac{h}{\phi}\right)^3\right) \text{ for } h \le \phi, \text{ and } C(h) = 0 \text{ for } h > \phi \quad (7)$$

where:

- σ^2 is the sill (variance).
- ϕ is the range parameter.
- h is the lag distance.

The Exponential function which is given as:

$$C(h) = \sigma^2 \exp\left(-\frac{h}{\phi}\right) \quad (8)$$

The Gaussian covariance function is given as:

$$\gamma(h) = c_0 + c\left\{1 - \exp\left(-\frac{h^2}{a^2}\right)\right\} \text{ for } 0 < h$$
$$= 0 \text{ for } h = 0$$

The parameters have the same meaning as in the exponential model. The effective range is given as $r' = \sqrt{3}a$.

4.5. Spatial Prediction

Spatial Prediction was carried out in the 25km grid, using the prediction equation:

$$\hat{Z}(x_0) = \mu(x_0) + \sum^{n}{}_{i=1}\ w_i(x_0)(Z(x_i) - \mu(x_i)) \qquad (9)$$

Where:

- $\hat{Z}(x_0)$ is the estimated value at location x_0,
- $\mu(x_0)$ is the trend or mean value at location x_0,
- $w_i(x_0)$ are the kriging weights for each observation $Z(x_i)$,
- $Z(x_i)$ is the observed value at location x_i,
- n is the number of observed data points.

4.6. Regression Analysis

The transformed elevation values were treated as the explanatory variables, to the dependent variable of transformed rainfall. The Linear Regression Model:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \qquad (10)$$

Where:

Y is the dependent variable, which is the predicted rainfall.
β_0 is the intercept (constant term) of the regression equation.
β_1 is the coefficient of the explanatory variable, representing the change in Y for a one-unit change in X_1.
ϵ is the error term, which accounts for the variability in Y not explained by the linear relationship with X_1.

The residual values of the regression were estimated and used into variogram modelling using Maximum Likelihood Estimations and then Spatial Prediction was performed, using the previously mentioned methodologies.

4.7. Deep Learning based Interpolation

Here, the distribution of the observations in the bounds of Latitude and Longitude is considered as form of a Graph, where each observation is treated as nodes, and links or edges are the proximity between those nodes, to understand the influence of rainfall values of each node can influence each other. Mixture Model Neural Network (MoNet), which was introduced to effectively manage tasks across different domains—specifically image processing, graph analysis, and 3D shape analysis—demonstrating its versatility. While Graph Attention Networks (GAT) effectively compute node embeddings through adaptive attention scores, MoNet introduces a more sophisticated methodology by employing Gaussian Mixture Model (GMM) convolutions.(Zhang et al., 2022) In contrast to GAT's reliance on a single attention mechanism, MoNet leverages a mixture of Gaussian kernels to capture a diverse array of spatial patterns and relationships. The core idea behind MoNet is to adaptively model the influence of neighbouring nodes based on their spatial relationships, using a mixture of Gaussian kernels. This is achieved through a specialized convolution process as described by the following formula:

$$h_v = \frac{1}{|N(u)|} \sum_{v \in N(u)} \sum_{k=1}^{K} w_k(e_{uv}) \odot W x_v$$

Here,

h_v is an updated feature vector for node v that considers its neighbours' effect.

$\frac{1}{|N(u)|}$ Assures that the average of the collected data from nodes within proximity balances the contributions of each neighbour.

$\sum_{v \in N(u)} \sum_{k=1}^{K} w_k(e_{uv})$ Aggregates information from all neighbouring nodes v of node u, with each neighbor's contribution weighted by K Gaussian kernels based on the edge feature e_{uv}.

$w_k(e_{uv}) \odot W x_v$ Applies the Gaussian weights to the feature vector x_v of node v,transforming and combining it with the learned weight matrix W.

4.7.1. Model Architecture

The following is a description of the MoNet-based model's approach for spatial interpolation: To represent the node attributes x_u and target variable y_u, the geographic coordinates (latitude and longitude) and daily cumulative rainfall values are first taken out of the dataset and transformed into PyTorch tensors. After that, a network structure is created by encoding the spatial interactions between these nodes using the k-nearest-neighbors (k-NN) method. The spatial relationships between these nodes are then encoded into a graph structure using a k-nearest neighbours (k-NN) approach, where each node u is connected to its k nearest

neighbours $N(u)$ based on Euclidean distance. The resulting adjacency matrix is converted to an edge index E, representing the graph's connectivity, and is crucial for capturing the neighborhood structure of the data. A PyTorch Geometric Data object is constructed, encapsulating the node features x_u, edge index E, and the target values y_u, which serves as the input to the graph neural network (GNN). To facilitate training and evaluation, the dataset is split into training and validation sets, with binary masks M_{train} and M_{val} created to specify the nodes used for each purpose.

The model is defined as a two-layer GMMNet, where each layer is a GMMConv 2 operation. The GMMConv 2 layers are designed to perform convolution over the graph, where the feature vector of each node h_u is updated by aggregating information from its neighbours $N(u)$. This is done by first computing the pseudo-coordinates e_{uv}, representing the relative spatial positions of neighbouring nodes u and v, i.e., $e_{uv} = x_u - x_v$. The first GMMConv layer applies a learned transformation W_1 to the node features, resulting in transformed features $W_1 x_v$ for each neighbour v. These transformed features are then weighted by Gaussian kernels $w_k(e_{uv})$, where each kernel k captures different spatial patterns based on the edge feature e_{uv}, reflecting the spatial dependency between nodes. The weighted contributions from all kernels are summed and aggregated across all neighbours $N(u)$ for each node, resulting in a new feature representation h_u for each node. The element-wise multiplication $\odot$ of the transformed features and using Gaussian weights, the aggregation process is made robust to the spatial relationships represented in the graph. After adding non-linearity by applying an ELU activation function to the first GMMConv layer's output, a second GMMConv layer refines the node features to get the final interpolated rainfall values.

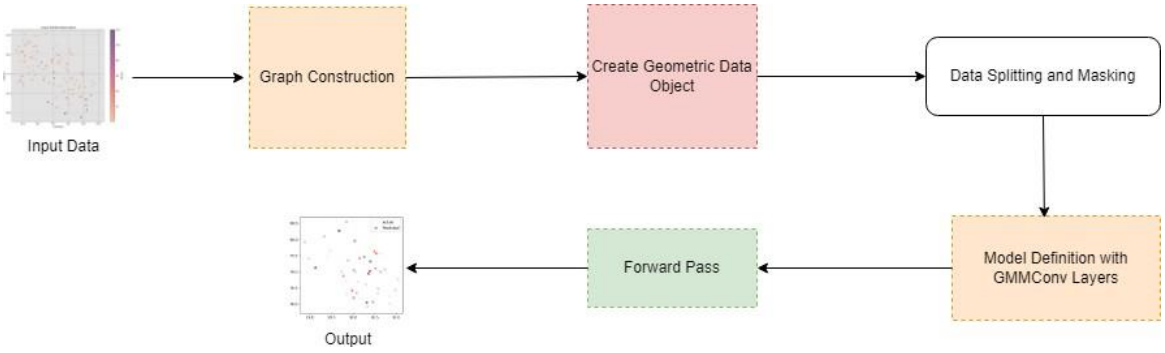


Figure 8 *Architecture of MoNET with GMMConv*

4.7.2. Model Parameters

Followings are the parameters of the MoNET model:

a. K neighbourhood value: This hyperparameter determines how many nodes each node is linked to, controlling the graph's connection. The value of k = 5 was chosen here to initiate the model.

b. Gaussian Kernel Size (K): The total amount of kernels for the GMMConv layers. Each kernel can capture different spatial patterns or connections. In this model, K= 3 is chosen, which allows the model to learn and apply three distinct spatial weighting functions.

c. No of Hidden Units in GMMConv Layers: Determines the feature representation's size following each convolutional layer. The first GMMConv layer transforms the input feature dimension from 2 to 16. This decision establishes a compromise between the necessity for more detailed feature representation and the danger of overfitting. The feature dimension is reduced to 1 by the second GMMConv layer, which corresponds to the target variable (rainfall).
d. Activation Function: Induces the model's non-linearity, which enables it to recognise more intricate patterns in the data. Here, we made use of the Exponential Linear Unit (ELU). It is the method of choice because ELU may reduce the vanishing gradient issue while preserving computing efficiency. It also ensures quick convergence, which accelerates the learning process.
e. Learning Rate: Controls the rate at which the model changes its parameters while being trained. Here, a learning rate of 0.01 is employed. Here, the learning rate for each parameter is adjusted using the Adam Optimizer based on estimations of lower-order moments.
f. Number of Epochs: Determines how long the model is trained. 140 epochs has been used.
g. Weight Decay: It helps keep the model from overfitting the training set and becoming too complicated. 0.0005 weight decay is employed.

4.7.3. Training and Validation

The whole dataset is the size of [106,2], where 106 are unique rainfall observations, with their location values. Data is divided into training and validation sets, to train and evaluate the model. Size of training set is [84,1] and validation sets is [22,1].

4.7.4. Application

The model is applied over the observations of all days of the month July 2023. The Observations are made into graphs, followed by the prediction grid is made, with 0.25-degree spatial resolutions. The prediction is done on the combined grid or graph of observations and the prediction grid.

4.8. Comparison

Statistical comparison of the results of the Deterministic (IDW), Stochastic (Kriging and Regression Kriging) Spatial Interpolation methods, and Deep Learning based interpolation methods, is done using Root Mean Square Error (RMSE) analysis. The general formula to calculate RMSE

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

n is the total number of observations.

y_i is the actual value.

$\hat{y}$ is the predicted value.

Range of RMSE values between 25 to 75 is treated as Moderately Performing, below 25 is Good Performance and above 75 is Low Performance.(Mitra et al., 2009)

4.9. Reproducibity

Separate parts of the methodology are performed in Python and R language consecutively. Alongside with the domain knowledge of statistical analysis, Spatial Interpolation techniques, Machine Learning and Deep Learning algorithms, rigorous studies on the documentation of the used libraries like PyTorch, SciPy, GSTools in Python and Gstats, GeoR, Caret and Raster in R is needed, to reproduce the work. With the availability of advanced cloud computing facilities in free of costs, the methodology can be carried out over any samples. However, for complex and huge data, there is need of High-Performance Computing facilities to run these methods. The codes of the mentioned methods in this section can be found out in this GitHub repository (https://github.com/GrassHopper97/MSc-Thesis).

5. RESULTS

In this chapter, the results from the works of the different parts of the methodology has been presented.

The daily distribution of observed rainfall data over the whole month of July 2023 is illustrated here:

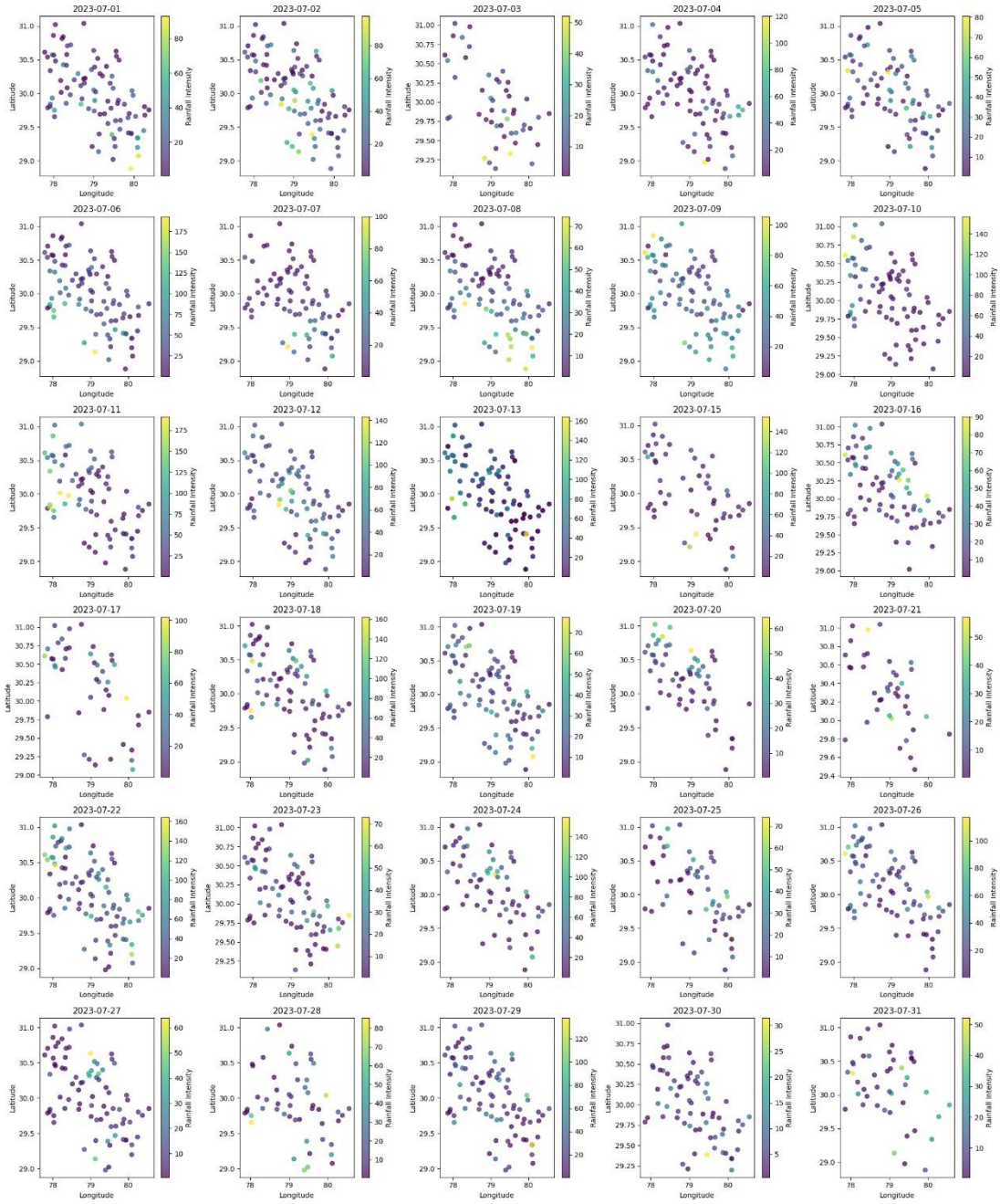


Figure 9 *Daily Distributions of Rainfall over all the days*

5.1. Data Preprocessing

The input daily rainfall data are sorted out by removing any faults and then transformed using the transformation function mentioned in the methodology section. The outcomes of the transformed daily rainfall data are as follows

Figure 10 *Histogram Distribution of Transformed Rainfall*

5.2. Variogram Generation and Maximum Likelihood Analysis

The transformed daily rainfall data is the input for the variogram generation. The parameters of those variograms are estimated using MLE, and then used to fit the empirical variogram models. Estimated parameters are shown here:

Table 1: Variogram Parameters and Covariance models used for Kriging

Date	Nugget	Partal Sill	Practical Range (km)	Covariane Model
01-07-2023	1.04	1.94	23.6	Exponetial
02-07-2023	0.09	4.52	29.13	Exponetial
03-07-2023	0.07	1.59	10.6	Exponetial
04-07-2023	0.037	1.44	28.3	Gaussian
05-07-2023	0.052	5.59	30.22	Exponetial
06-07-2023	4.38	2.81	98.7	Exponetial
07-07-2023	0.46	0.96	78.7	Spherical
08-07-2023	0.49	1.019	82.73	Spherical
09-07-2023	16.3	1.078	86.76	Spherical
10-07-2023	0.647	0.419	103.4	Exponetial
11-07-2023	2.015	1.966	85.5	Exponetial
12-07-2023	3.383	3.513	92.53	Exponetial
13-07-2023	7.751	0.86	27.23	Spherical
15-07-2023	0.261	2.28	29.04	Spherical
16-07-2023	0.794	2.51	37.5	Exponetial
17-07-2023	0.401	7.145	44.89	Exponetial
18-07-2023	0.657	2.94	44.23	Exponetial
19-07-2023	0.664	3.16	46.85	Gaussian
20-07-2023	0.772	1.59	44.44	Exponetial
21-07-2023	0.744	3.13	36.35	Exponetial
22-07-2023	0.806	3.34	38.86	Exponetial
23-07-2023	0.166	2.9	43.37	Exponetial
24-07-2023	0.283	3.3	49.49	Exponetial
25-07-2023	1.144	1.45	32.21	Exponetial
26-07-2023	0.457	2.51	27.25	Exponetial
27-07-2023	1.146	3.11	42.98	Exponetial
28-07-2023	0.476	2.63	38.25	Exponetial
29-07-2023	0.172	2.84	49.27	Gaussian
30-07-2023	0.628	1.36	34.17	Gaussian
31-07-2023	0.595	1.48	47.2	Exponetial

Based on the estimated parameters, the modelled variograms are shown here:

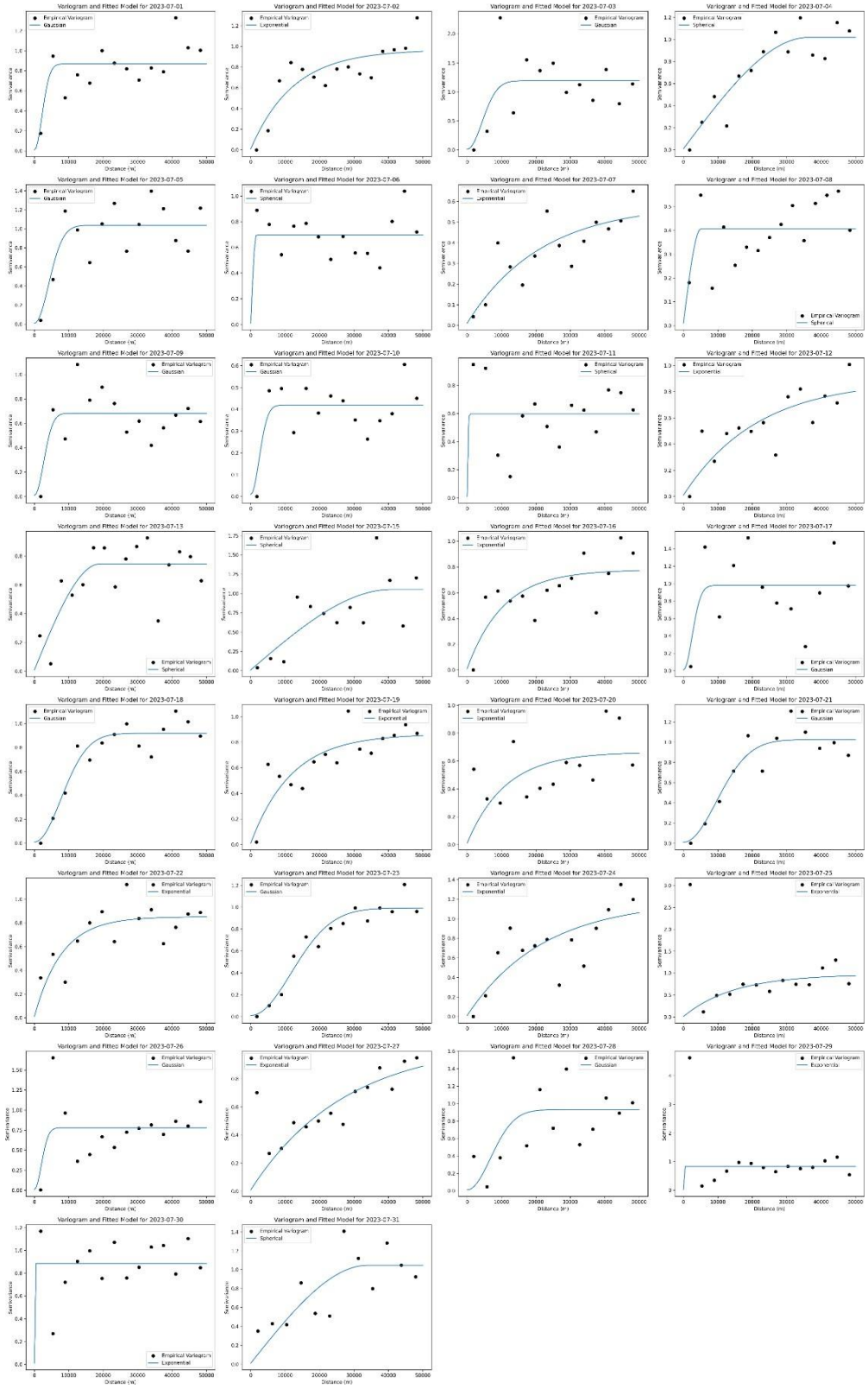


Figure 11 *Fitted Variograms*

5.3. Spatial Prediction

Following Variogram models are used for the spatial prediction into the kriging model. The daily predicted rainfall maps or surfaces are shown:

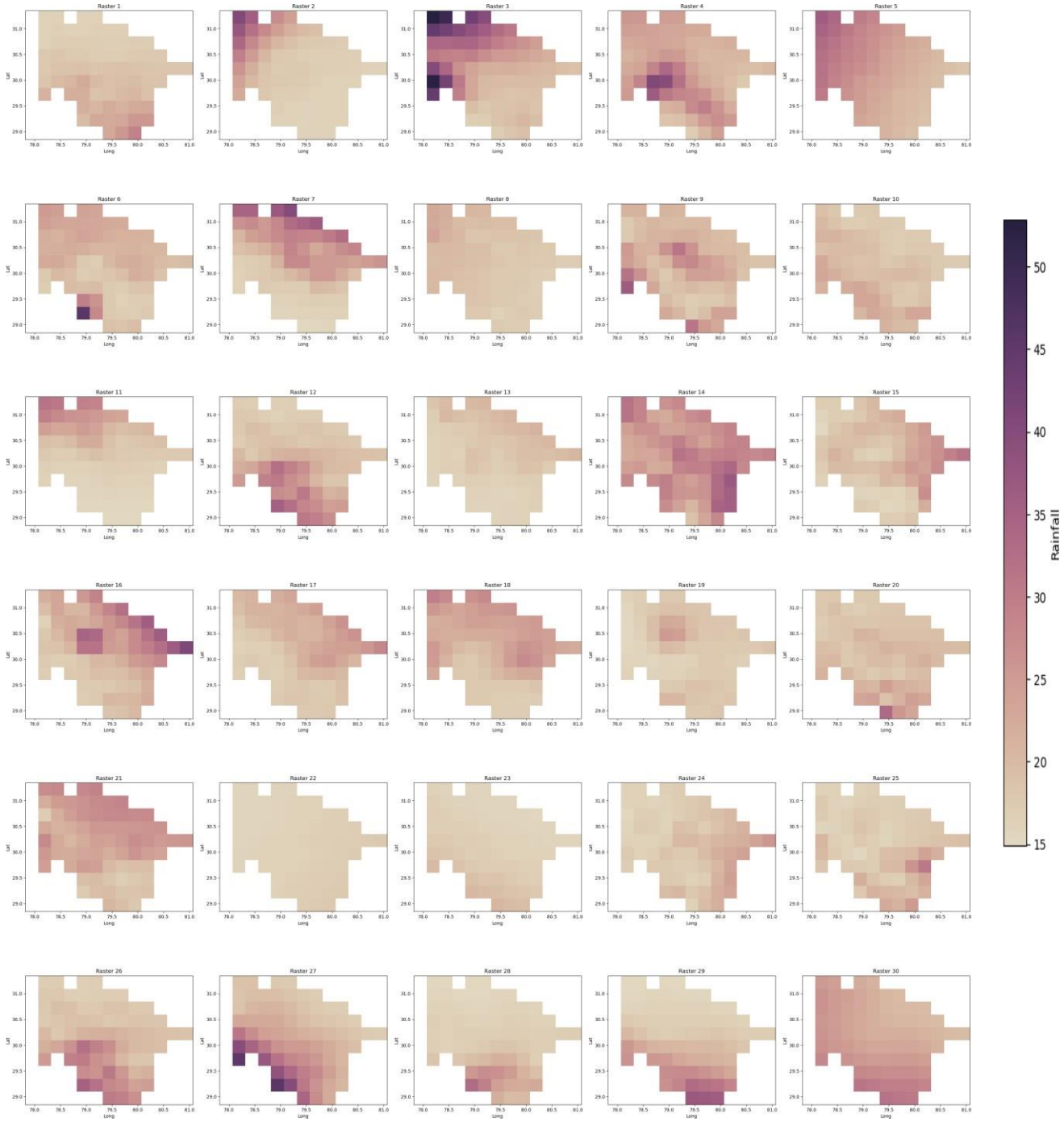


Figure 12 *Spatial Predicted Maps from Kriging*

5.4. Regression Work and Prediction

With the transformed elevation data, the linear regression analysis is performed. The results of the regression coefficients and intercepts are shown here

Table 2: Regression Coefficients and intercepts

Date	Regr Coefficient	Intercepts
01-07-2023	0.254	2.61
02-07-2023	0.137	1.522
03-07-2023	-0.182	2.054
04-07-2023	0.193	2.343
05-07-2023	0.274	2.306
06-07-2023	-0.196	1.81
07-07-2023	0.257	2.622
08-07-2023	0.151	1.899
09-07-2023	-0.198	2.649
10-07-2023	0.266	2.035
11-07-2023	0.166	2.326
12-07-2023	-0.137	2.777
13-07-2023	0.286	1.625
15-07-2023	0.279	2.245
16-07-2023	-0.243	1.25
17-07-2023	0.153	2.107
18-07-2023	0.228	1.585
19-07-2023	-0.139	1.204
20-07-2023	0.287	2.591
21-07-2023	0.165	1.192
22-07-2023	-0.166	1.874
23-07-2023	0.26	2.422
24-07-2023	0.26	1.729
25-07-2023	0.221	1.969
26-07-2023	0.282	1.411
27-07-2023	0.136	2.368
28-07-2023	-0.191	2.759
29-07-2023	0.162	2.262
30-07-2023	-0.243	2.68
31-07-2023	-0.22	2.185

Using the residuals of the regression model, the variogram parameters are estimated, in the same way as earlier, using MLE. Estimated Variogram parameters are shown here

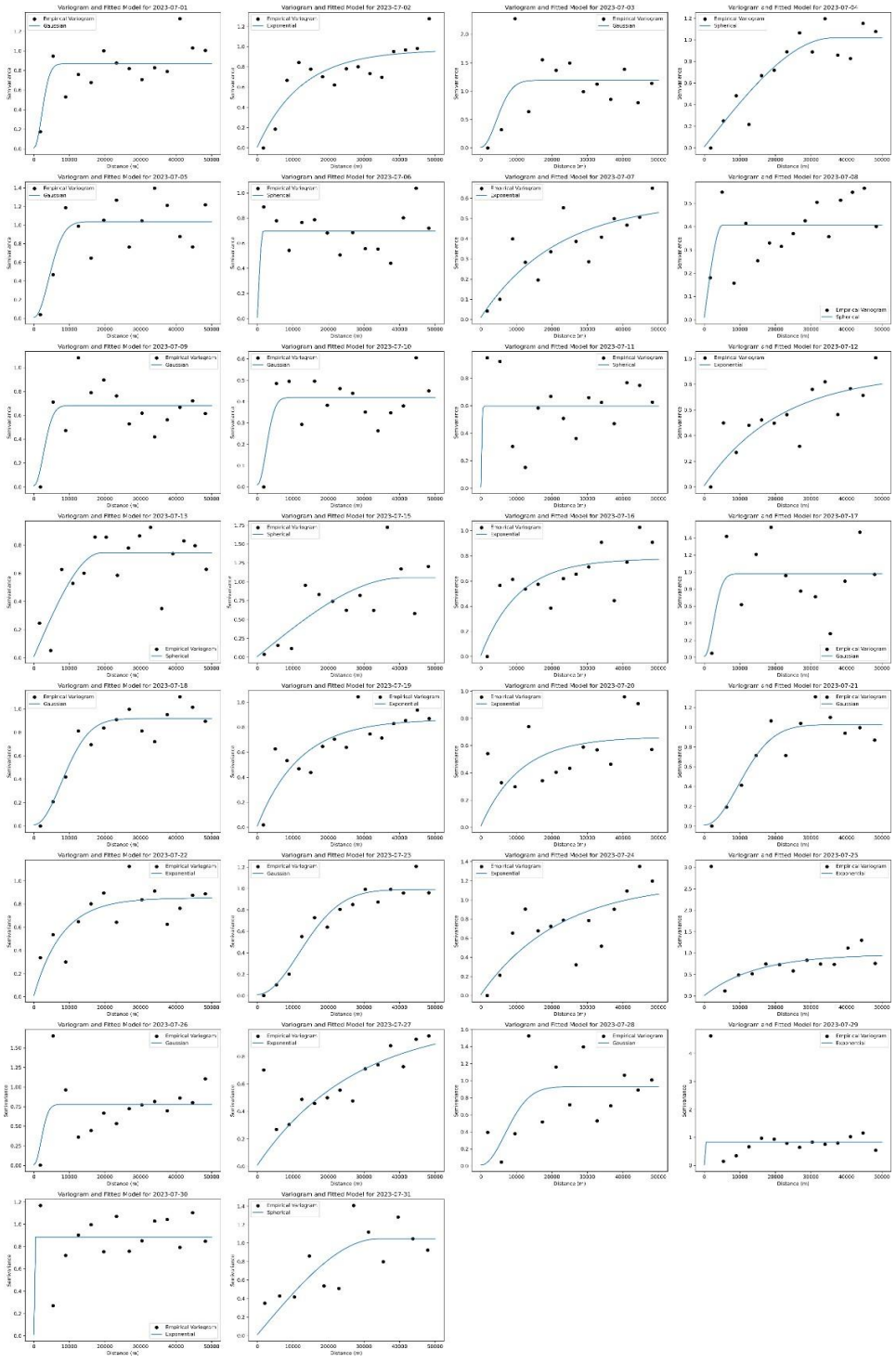


Figure 13 *Fitted Variograms, used in Regression based Kriging*

The modelled parameters of these variograms are :

Table 3: Variogram Parameters and Covariance model, used in Regression based Kriging

Date	Nugget	Partial Sill	Practical Range	Covariance Model
01-07-2023	1.04	1.94	23.6	Exponetial
02-07-2023	0.09	4.52	29.13	Exponetial
03-07-2023	0.07	1.59	10.6	Exponetial
04-07-2023	0.037	1.44	28.3	Gaussian
05-07-2023	0.052	5.59	30.22	Exponetial
06-07-2023	4.38	2.81	98.7	Exponetial
07-07-2023	0.46	0.96	78.7	Spherical
08-07-2023	0.49	1.019	82.73	Spherical
09-07-2023	16.3	1.078	86.76	Spherical
10-07-2023	0.647	0.419	103.4	Exponetial
11-07-2023	2.015	1.966	85.5	Exponetial
12-07-2023	3.383	3.513	92.53	Exponetial
13-07-2023	7.751	0.86	27.23	Spherical
15-07-2023	0.481	2.28	48.14	Spherical
16-07-2023	0.846	2.51	48.39	Exponetial
17-07-2023	0.703	7.145	27.43	Exponetial
18-07-2023	1.029	3.02	31.59	Exponetial
19-07-2023	0.383	2.29	28.63	Gaussian
20-07-2023	0.341	2.62	49.71	Exponetial
21-07-2023	0.868	3.88	46.07	Exponetial
22-07-2023	0.687	1.56	41.46	Exponetial
23-07-2023	1.104	2.39	34.56	Exponetial
24-07-2023	0.114	3.39	32.55	Exponetial
25-07-2023	0.275	4.06	39.63	Exponetial
26-07-2023	0.163	3.66	32.38	Exponetial
27-07-2023	1.013	1.93	45.23	Exponetial
28-07-2023	0.475	1.74	33.08	Exponetial
29-07-2023	0.924	1.46	46.56	Gaussian
30-07-2023	0.174	4.04	47.48	Gaussian
31-07-2023	0.579	1.25	41.95	Exponetial

The variograms generated from the residual values of the regression model, are used for the kriging model for the spatial prediction. The daily rainfall surfaces generated from the regression analysis are shown here:

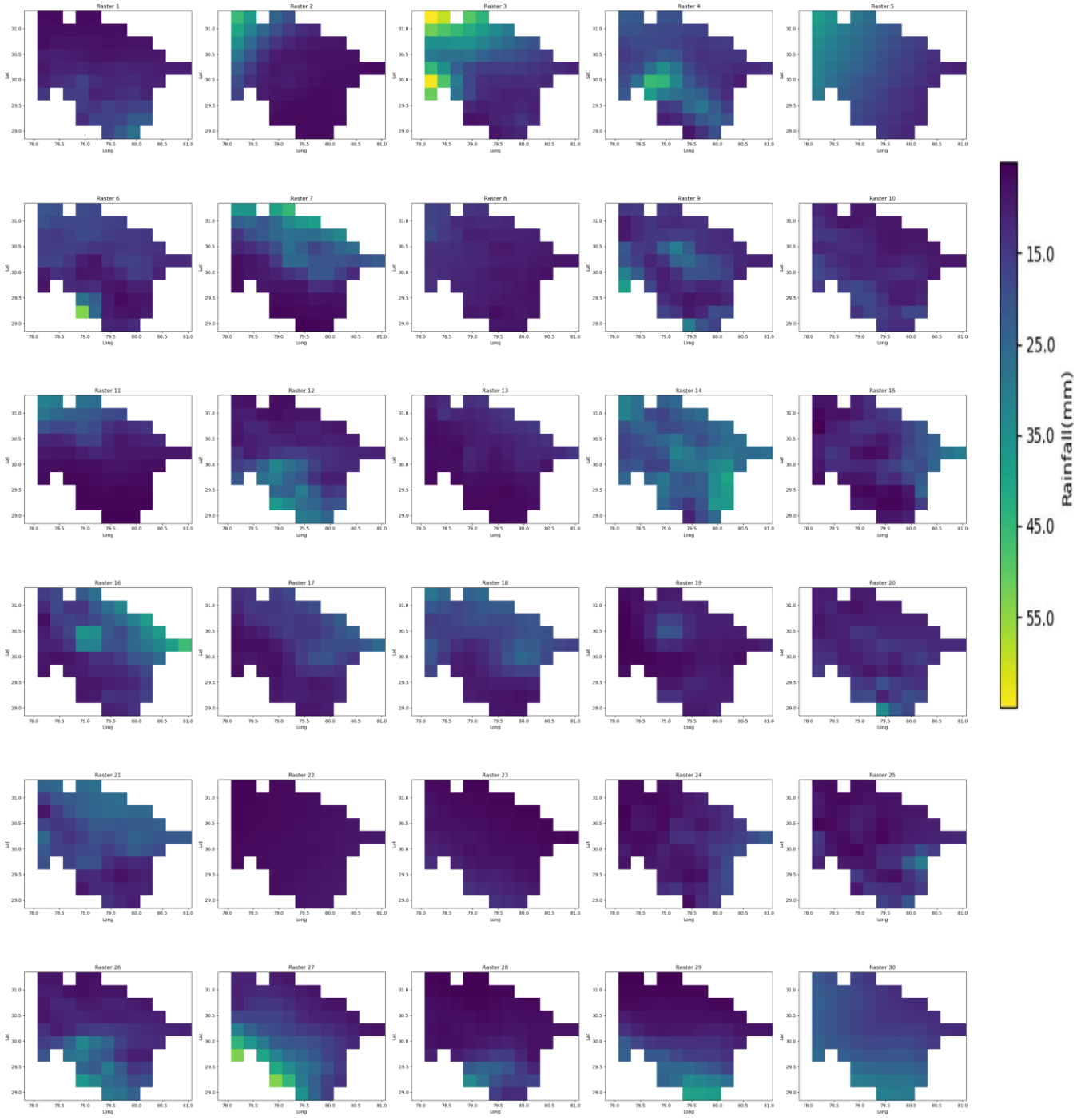


Figure 14 *Spatial Predicted Rainfall maps from regression-based kriging*

5.5. Deep Learning based Interpolation

The MoNET model, is trained and validated over the input datasets. This step is essential to understand the performance of the model. The loss curve, which is representation of model's loss or error changes over time with training and validation.

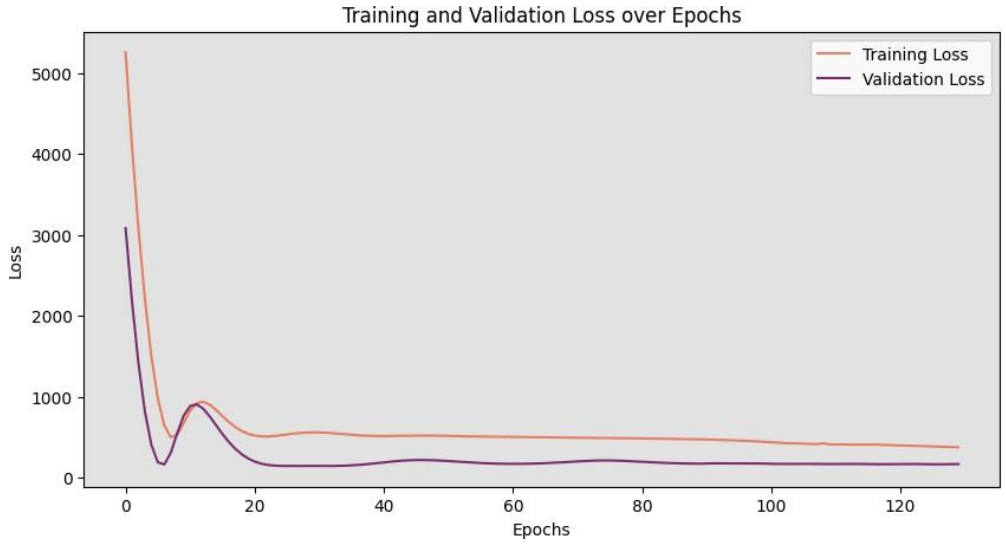


Figure 15 *Loss Curve*

The R^2 score, which also helps to evaluate the model's performance over time. This represents how well the predicted value matches with the actual values.

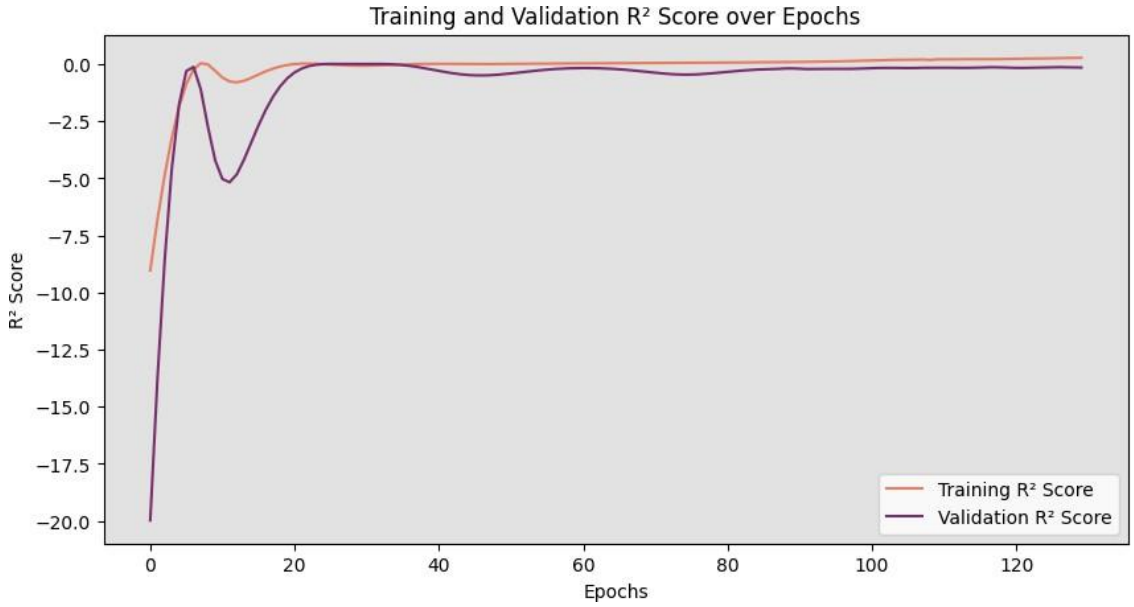


Figure 16 *: R^2 curve*

The model is applied over the distributions of each day's rainfall observations, to get the daily rainfall surfaces.

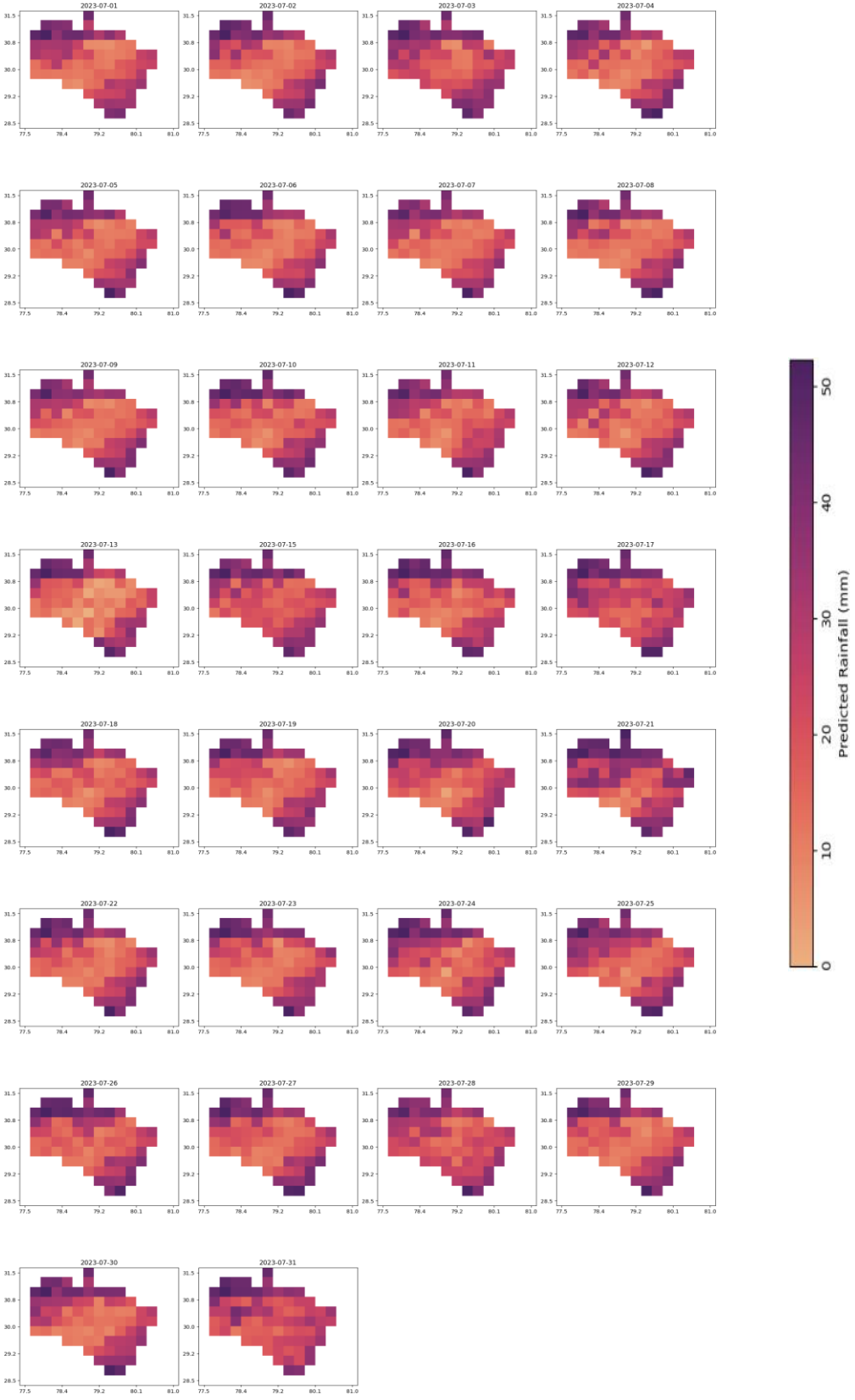


Figure 17 *Predicted Rainfall Maps using MoNET*

5.6. Daily Grided Maps from IMD

As mentioned in the dataset section, the daily grided data, which are freely available from the IMD, is used here as a result of the deterministic spatial interpolation.

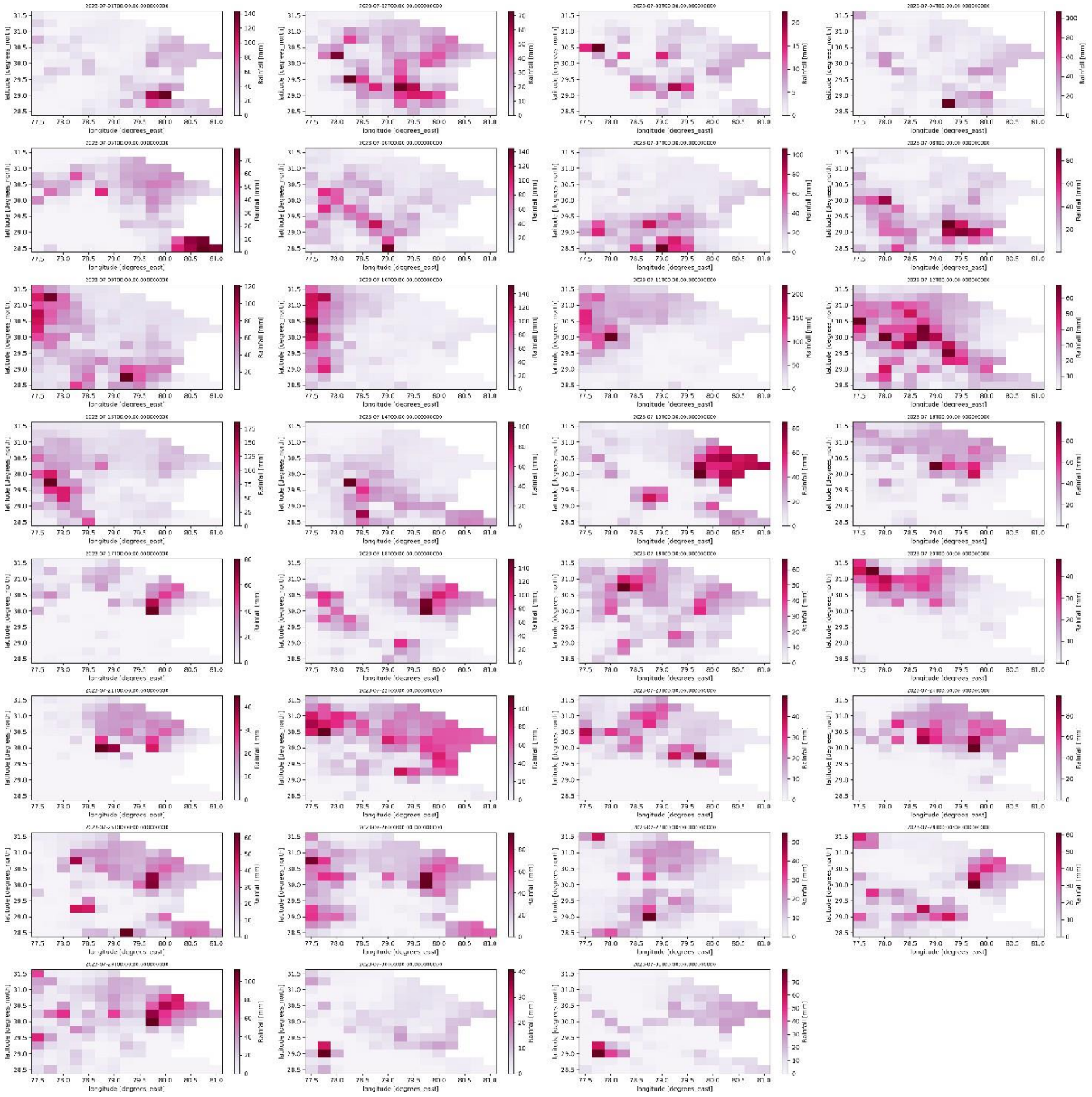


Figure 18 *IMD's Daily Rainfall Maps*

5.7. Comparison of the Results

The results from the 2 geostatistical spatial interpolation methodologies and MoNET, the deep learning interpolation technique, along with the grided data of IMD, are compared with the actual rainfall observations from the AWSs. There are 7 different AWS only, which had the continuous records over the whole month. Location of those stations are:

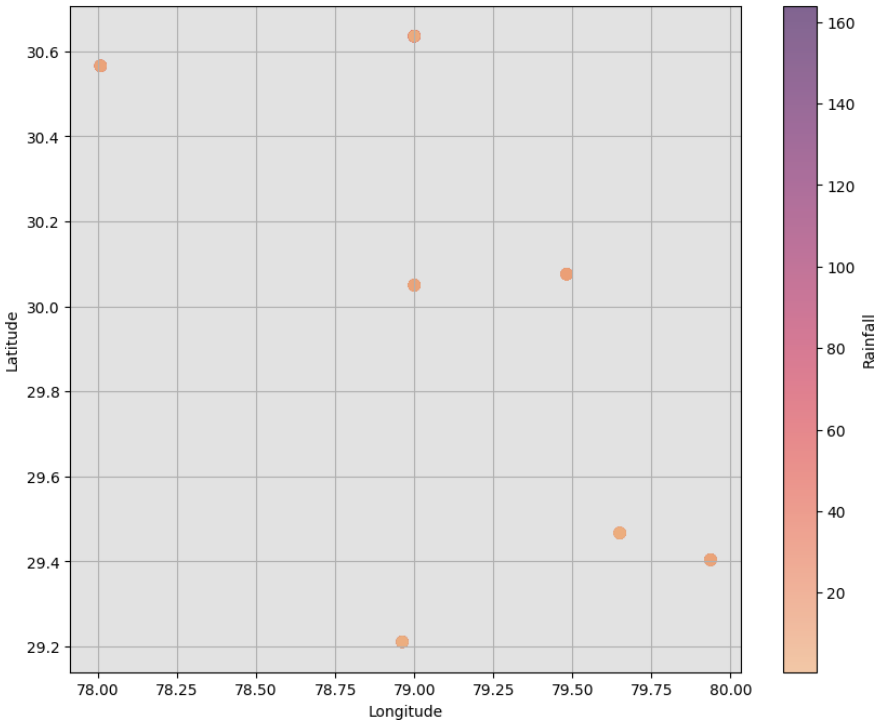


Figure 19 *Distinct Locations with continuous records*

The results are compared with the values of those stations. The plot of RMSE values is shown here.
The red lines represent RMSE associated with MoNET model, green is with Kriging model, orange is with Kriging with Regression model and blue is with the IMD data, which represents the IDW model.

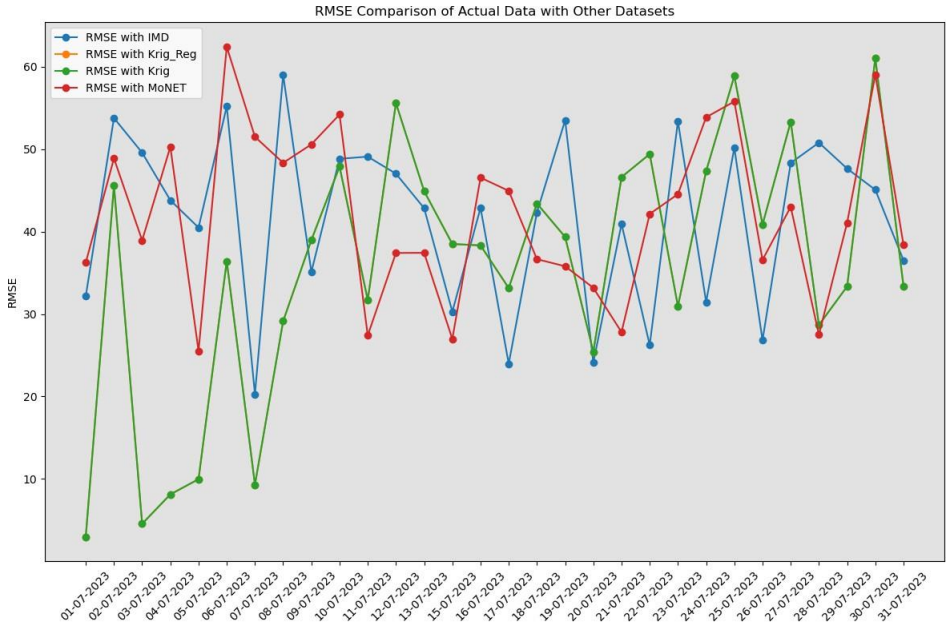


Figure 20 *RMSE values of Comparisons between Deterministic, Stochastic and Deep Learning Methods*

6. DISCUSSION

This chapter is based on the interpretation and discussion of the results of the 3 different models and the results of the comparison. Following sections are to be framed to discussion of results to help to answer the different subobjectives. Final section is to highlight the common findings.

6.1. Sub-Objective 1

Overall, analysis of almost all days shows very small amounts of nugget, which indicates there are very small amounts of variability in the very short ranges. From the Table 1 values of partial sill and ranges, it can be understood that the spatial auto correlation is low and quickly diminishing, from the values between 20 to 40 km of the range, as both the partial sill and range values are less. In few cases, the range values are significantly high, with a low partial sill value, which, also, shows the very low autocorrelation between 2 distant pairs. The effect of the weak spatial structure can be visible on the kriging results, where it can be seen, the portion, with very less and almost no observations, predicted results are not proper.

6.2. Sub-Objective 2

From the figure 16, The R^2 values of the training and validation phase, over time, shows that the model has a good convergence, as the validation R^2 values keep increasing and ideally follows the stable training curve. But the model can be still improved, as the validation R^2 eventually should reach to the positive side of the plot. From the figure 15, The loss curve shows continuous pattern in training and validation curve and ultimately, they converge, showing the use of Learning Rate and Decay Function values with the Optimizer is justified. With the K neighbour value of 5 and the 3 gaussian kernels, it effectively carried the detailed relationship, where the observations are relatively closed. But the model is not able the show the effect of higher individual rainfall observations, on a few days. Overall, the outputs make the model applicable enough to be an additional computational methodology for spatial interpolation.

6.3. Sub-Objective 3

From the significance analysis, it has been found that the Elevation is a highly significant predictor of rainfall, as the p- values lies between 0.001 to 0.01 in the different days of analysis. However, from Table 2 the coefficient of regression is very less, which make the trend, that added by the regression, is not strong enough, and that's why the values of variogram parameters in Table 1 and Table 2 are almost similar, similarity can be seen in the figures 12 and 14 also.

6.4. Sub-Objective 4

In the figure 20, the green and orange lines are overlapped, as the RMSE values of both Kriging and Kriging with regression has very less differences. For the first seven days, these both models have the least RMSE values, showing better representation of the actual data. However, for the rest of the days, all the models show approximately similar patterns of RMSE values, ranging from 30 to 60. Which implies that all the models are moderately able to capture the accurate rainfall values. Out of 106 unique locations, where the rainfall values were observed the month, only 7 stations were available to perform the comparison with the accurate data. Which is not capturing the whole workability and assessment of the individual models over the study area.

6.5. Overall Discussion

From the visual inspection, it can be seen that the results of the MoNET tends to have a similar pattern over some part of the study area over some certain days, which resembles with those regions where the observed data were having the similar values on those days, which shows the lack of capturing different spatial patterns, with less no of use of the kernels. The IMD data has very less visual resemblance with the results of Kriging methods as well as with MoNET results.

In terms of daily predictions, incorporating a model like MoNet with GMMConv layers is helpful since it allows for training a model that can be applied consistently to fresh daily data as it becomes available. Compared to Kriging, which lacks a generalisable variogram model appropriate for all days and necessitates evaluating the spatial structure of each day and fitting variograms appropriately, this is far less complex. The findings of this work illustrate the usefulness of GMMConv-based MoNet in spatial prediction tasks, a unique approach that offers a beneficial alternative to the frequently utilised methods. However, while the pre-training of such models can be highly beneficial, their success heavily depends on the availability of a robust and sufficient training dataset. In the present work, the sparse availability of data poses a challenge, limiting the potential of the model to fully capture the spatial patterns across different days.

7. CONCLUSION

This chapter is divided into 3 sections. Firstly, the Research Questions related to each sub-objectives have been reviewed and discussed with the probable answers. Next, is about the Recommendation about the future development of the current thesis work. Lastly, the conclusion of the project has been presented.

7.1. Discussion on Research Questions

From the 1st Sub Objective:

1. **Which model of Interpolation going to be used?**

Ans: This work has been done using Ordinary Kriging and Universal Kriging models. Ordinary Kriging was used with only rainfall values. Universal Kriging was used with the calculated residuals of the regression model between rainfall and elevation.

2. **How well these models are performing in the prediction?**

Ans: The discussion about the performance of those model has been highlighted in the 6.1 and 6.3 sections.

From the 2nd Sub Objective:

1. **How to implement Deep Learning for Interpolation?**

Ans: In this study, a framework of Graph based neural networks is used, where the distribution of rainfall is treated as individual nodes over the axes between Latitude and Longitude. The relationship between each node is tried to find out using K Neighborhood mechanism. These relationships have been used in the learnable Gaussian kernels, to identify spatial patterns in the following data. This model is GMMConv based MoNET model, which can be used in the specified spatial grid, generate interpolated surface over the learned spatial patterns.

2. **How effectively can Rainfall be predicted using DL?**

Ans: This has been discussed in the 6.2, 6.4 and 6.5 sections.

From the 3rd Sub Objective:

1. **Is it possible to establish a relation between the elevation and Rainfall, in rainfall prediction analysis? Can this relation influence the predictive performance of the Rainfall Surface Generation**?

Ans: From the present study, based on the significance score, it is clarified that Elevation can be a significant predictor of rainfall, in the regression work. However, the influence of elevation over the rainfall in the present study is very less, so ultimately, it does not impact the overall rainfall prediction is a great way.

From the 4th Sub Objective:

1. **How the newly developed grids are performing alongside with the IMD, for rainfall prediction?**

Ans: This has been visually represented in the figure 20 and discussed in the 6.4 section.

2. **Does the application of deep learning provide any significance?**

Ans: Yes, for the present study, the Deep Learning based model gave significant results. The results were moderately accurate for almost every day, with RMSE values between 30 to 50, in some days it has values

around 20, where it gave the best prediction of the actual data. However, evaluation of the predicted maps by the model needs to be done with a significant amount of actual data, which are not available in this work.

7.2. Recommendations and Future Scope

- This model needs to be compared with more continuous datasets, to know more about the predictive capabilities of the models. Satellite based rainfall product can be good option for validating. However, area of the study may need to be extended based on the availability of the satellite-based rainfall products. Also, the incorporation of satellite-based rainfall products, ground data-based rainfall prediction models can be a great addition to a finer predictive rainfall dataset, from regional to global scale.
- The Mo-NET model can be modified with a good computing facility to use it the web platforms directly, so that it can work in real time with daily input of the rainfall data.
- As the Elevation parameter, doesn't have any serious impact on the prediction in the current study, it is highly recommended to add the local climate variables in the prediction, along with elevation, specially in the mountain regions, like Uttarakhand, where the micro climatic effects are significant.
- Increase in the training dataset is a need as in this study only 106 data are used to train the model. With increasing the study area, training samples, changing the parameters like no of kernels and no of K neighbourhood values, the more detailed analysis of the performance of the model can be deduced.

7.3. Conclusion

The study conducted a comparative analysis of rainfall surface generation using deterministic, stochastic, and deep learning approaches, focusing on Uttarakhand, India, with the primary objective of developing a high-resolution rainfall surface model that accurately reflects the region's spatial patterns using techniques such as Inverse Distance Weighting (IDW), kriging, regression kriging, and the deep learning-based MoNET (Mixture Model Network). The methodologies encompassed data preprocessing, variogram modeling, regression analysis, and the application of MoNET, followed by a statistical comparison of results using Root Mean Square Error (RMSE) analysis. The findings revealed that while kriging and regression kriging captured spatial variability moderately well, they struggled in areas with sparse observational data, resulting in less accurate predictions. The MoNET model, despite its innovative use of graph-based neural networks and Gaussian Mixture Models, demonstrated only moderate predictive accuracy, with RMSE values mostly between 30 to 60, due to the limited training dataset and weak spatial structure of the rainfall data, particularly in regions with high elevation variability. Although elevation was statistically significant, its impact on improving rainfall prediction accuracy was minimal within the study area. The study's limitations,

including the sparse availability of continuous observational data and the lack of integration of additional climate variables, constrained the models' predictive capabilities. The comparison with IMD's gridded rainfall data, derived using the deterministic IDW method, further emphasized the challenges in achieving high accuracy across diverse topographical regions. Nonetheless, the research underscores the potential of deep learning models like MoNET in rainfall surface generation, particularly in regions with complex spatial patterns, while highlighting the necessity of larger and more diverse datasets, and the incorporation of additional environmental variables to fully realize the benefits of these advanced techniques. Future research should focus on expanding the dataset by incorporating satellite-based rainfall products and additional climate variables, especially for mountainous regions where microclimatic effects are significant. Enhancing the MoNET model's architecture, including adjustments to the number of kernels and neighborhood values, could improve its ability to capture spatial patterns more effectively. Moreover, developing web-based applications that leverage MoNET for real-time rainfall prediction could offer practical benefits for disaster management and water resource planning in vulnerable regions like Uttarakhand. Ultimately, while this study advances the understanding of rainfall surface generation, it also highlights ongoing challenges and opportunities for improving spatial interpolation techniques in complex environments.

7.4. Use of AI in this Project

"During the preparation of this work, the author(s) used [Grammarly / Sentence Correction and Grammar correction] to [Present and maintain the structure of meaningful sentences]. After using this tool/service, the author(s) reviewed and edited the content as needed and took (s) full. responsibility for the content of the work." [1]
Students using AI, without the explicit consent of the instructor and acknowledgement of the tool in an appendix should therefore be considered to have committed academic misconduct.

---

[1]. Elsevier. The use of AI and AI-assisted writing technologies in scientific writing. Accessed 02- 05-2023. Available at: https://www.elsevier.com/about/policies/publishing-ethics/the-use-of-ai- and- ai assisted-writing-technologies-in-scientific-writing

LIST OF REFERENCES

AWS, ARG , AGRO AWS AND ASG (Surface Instrument Division). (n.d.).

AWS and ARG Automatic Weather Stations. (n.d.).

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). *Geometric Deep Learning Grids, Groups, Graphs, Geodesics, and Gauges*.

Cecinati, F., Moreno-Ródenas, A. M., Rico-Ramirez, M. A., ten Veldhuis, M. C., & Langeveld, J. G. (2018). Considering rain gauge uncertainty using kriging for uncertain data. *Atmosphere*, *9*(11). https://doi.org/10.3390/atmos9110446

Chen, F. W., & Liu, C. W. (2012). Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy and Water Environment*, *10*(3), 209–222. https://doi.org/10.1007/s10333-012-0319-1

Chen, M., Sun, Z., Davis, J. M., Liu, C., & Gao, W. (2018). *Spatial interpolation of surface ozone observations using deep learning*. https://doi.org/10.1117/12.2320755

Chen, Y. C., Wei, C., & Yeh, H. C. (2008). Rainfall network design using kriging and entropy. *Hydrological Processes*, *22*(3), 340–346. https://doi.org/10.1002/hyp.6292

Delbari, M., Afrasiab, P., & Jahani, S. (2013). Spatial interpolation of monthly and annual rainfall in northeast of Iran. *Meteorology and Atmospheric Physics*, *122*(1–2), 103–113. https://doi.org/10.1007/s00703-013-0273-5

Feki, H., Slimani, M., & Cudennec, C. (2012). Incorporation de l'altitude pour l'interpolation des pluies en Tunisie en utilisant les méthodes géostatistiques. *Hydrological Sciences Journal*, *57*(7), 1294–1314. https://doi.org/10.1080/02626667.2012.710334

GlobalSIP : 2014 IEEE Global Conference on Signal and Information Processing : 3-5 December 2014, Atlanta, GA, USA. (2014). Institute of Electrical and Electronics Engineers.

Goovaerts, P. (n.d.). *Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall*. www.elsevier.com/locate/jhydrol

Johnson, M. J., Duvenaud, D., Wiltschko, A. B., Datta, S. R., & Adams, R. P. (n.d.). *Composing graphical models with neural networks for structured representations and fast inference*.

Kirkwood, C., Economou, T., Pugeault, N., & Odbert, H. (2022). Bayesian Deep Learning for Spatial Interpolation in the Presence of Auxiliary Information. *Mathematical Geosciences*, *54*(3), 507–531. https://doi.org/10.1007/s11004-021-09988-0

Liang, F., Qian, C., Yu, W., Griffith, D., & Golmie, N. (2022). Survey of Graph Neural Networks and Applications. In *Wireless Communications and Mobile Computing* (Vol. 2022). Hindawi Limited. https://doi.org/10.1155/2022/9261537

Martorell, Sebastian., Barnett, Julie., & Soares, C. Guedes. (2009). *Safety, reliability and risk analysis : theory, methods and applications : ESREL 2008, proceedings of the European Safety and Reliability Conference and the 17th SRA-Europe [Society for Risk Analysis Europe], Valencia, Spain, September 22-25, 2008*. Taylor & Francis.

Meuer, J., Bouwer, L. M., Kaspar, F., Lehmann, R., Karl, W., Ludwig, T., & Kadow, C. (n.d.). *Infilling of Missing Rainfall Radar Data with a Memory-Assisted Deep Learning Approach*. https://doi.org/10.5194/egusphere-2024-1392

Mitra, A. K., Bohra, A. K., Rajeevan, M. N., & Krishnamurti, T. N. (2009). Daily indian precipitation analysis formed from a merge of rain-gauge data with the TRMM TMPA satellite-derived rainfall

estimates. *Journal of the Meteorological Society of Japan*, *87 A*, 265–279. https://doi.org/10.2151/jmsj.87A.265

Monti, F., Boscaini, D., Masci, J., Rodolà, E., Svoboda, J., & Bronstein, M. M. (2016). *Geometric deep learning on graphs and manifolds using mixture model CNNs*. http://arxiv.org/abs/1611.08402

Mukhopadhaya, S. (n.d.). *Rainfall Mapping using Ordinary Kriging Technique: Case Study: Tunisia*. *3*(1), 1–5. http://www.krishisanskriti.org/Publication.html

Navalgund, R. R., Kumar, A. S., & Nandy, S. (2018). Remote Sensing of Northwest Himalayan Ecosystems. In *Remote Sensing of Northwest Himalayan Ecosystems*. Springer Singapore. https://doi.org/10.1007/978-981-13-2128-3

Pai, D. S., Sridhar, L., Rajeevan, M., Sreejith, O. P., Satbhai, N. S., & Mukhopadhyay, B. (2014). *Development of a new high spatial resolution (0.25° × 0.25°) long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region* (Vol. 65, Issue 1).

Papacharalampous, G., Tyralis, H., Doulamis, N., & Doulamis, A. (n.d.). *Uncertainty estimation in spatial interpolation of satellite precipitation with ensemble learning*. https://orcid.org/0000-0001-5446-954X

Ribeiro, A. M. T., Ribeiro Junior, P. J., & Bonat, W. H. (2022). A Kronecker-based covariance specification for spatially continuous multivariate data. *Stochastic Environmental Research and Risk Assessment*, *36*(12), 4087–4102. https://doi.org/10.1007/s00477-022-02252-9

Rodriguez-Ramirez, M. A., & Fuentes-Mariles, Ó. A. (2023). Daily rainfall assimilation based on satellite and weather radar precipitation products along with rain gauge networks. *Journal of Hydroinformatics*, *25*(6), 2354–2368. https://doi.org/10.2166/hydro.2023.104

Sati, V. P. (n.d.). VERTICAL AND HORIZONTAL DISTRIBUTION OF FORESTS IN UTTARAKHAND HIMALAYA: A GEOGRAPHICAL ANALYSIS. In *Turkish Journal of Forest Science* (Vol. 4, Issue 2). http://orcid.org/0000-0001-6423-3119

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605

Veličkovi´veličkovi´c, P., Cucurull, G., Casanova, A., Romero, A., Lì, P., & Bengio, Y. (n.d.). *GRAPH ATTENTION NETWORKS*.

Xiang, Z., & Demir, I. (n.d.). *Fully distributed rainfall-runoff modeling using spatial-temporal graph neural network*.

Zareifard, H., Mahbod, M., & Mohammadi, Z. (2023). Geostatistical modelling of rainfall in Fars Province of Iran using non-Gaussian spatial process. *Theoretical and Applied Climatology*, *153*(1–2), 57–72. https://doi.org/10.1007/s00704-023-04415-2

Zhang, M., Yu, D., Li, Y., & Zhao, L. (2022, November 1). Deep geometric neural network for spatial interpolation. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. https://doi.org/10.1145/3557915.3561008

Zsolt Farkas, J., Hoyk, E., & Rakonczai, J. (2017). Geographical analysis of climate vulnerability at a regional scale: The case of the southern great plain in Hungary. *Hungarian Geographical Bulletin*, *66*(2), 129–144. https://doi.org/10.15201/hungeobull.66.2.3