DATA-DRIVEN APPROACHES FOR IMPROVING THE PHYSICS OF HEATING IN CLIMATE MODELS AND FIRE DANGER ASSESSMENT USING AI TECHNIQUES

A THESIS SUBMITTED TO
DEFENCE INSTITUTE OF ADVANCED TECHNOLOGY , PUNE
FOR THE SEMESTER FOURTH EVALUATION (2021-2023) OF
MASTER OF TECHNOLOGY
IN
DATA SCIENCE

BY
VISHNU DEV TRIPATHI
(Registration No. 21-27-12)

UNDER THE SUPERVISION OF
Dr. Yogeshwar Singh Dadwhal & Dr. Manmeet Singh

SCHOOL OF COMPUTER ENGINEERING
&
MATHEMATICAL SCIENCES
DEFENCE INSTITUTE OF TECHNOLOGY(DIAT), PUNE,
INDIA

APRIL 2023

CONFERENCE PUBLICATION

Presented my research on "**Improved Representation of Latent Heating in Weather and Climate Models for Improved Rainfall Predictions Over the Tropics**" at the **Indo-German conference on Computational Mathematics: Challenges and Opportunities towards scientific computing and machine learning (IGCM-2023)** held at **IISC Bangalore** on **27th March 2023**. The presentation focused on utilizing a data-science model to construct 3- dimensional fields of time-mean latent heating, trained using machine learning methods and subjected to extensive hyper parameter optimization, which has the potential to improve rainfall predictions over the tropics.

Dedicated

To each and everyone who supported and motivated

me.

C E R T I F I C A T E

This is to certify that the Thesis entitled "**Data-Driven Approaches For Improving The Physics Of Heating In Climate Models And Fire Danger Assessment Using AI Techniques**" submitted to Defence Institute of Advanced Technology, Girinagar, for the award of the degree of **Master of Technology**, is the bonafide research work done by **Mr. Vishnu Dev Tripathi** under my supervision. The contents of this thesis have not been submitted elsewhere for the award of any degree.

Dr. Manmeet Singh
External Supervisor
Scientist 'C'
IITM, Pune

Dr. Yogeshwar Singh Dadwhal
Internal Supervisor
Professor
School of Computer Engineering
& Mathematical Sciences
Defence Institute of Advanced Technology
Girinagar, Pune, INDIA

DECLARATION

This is to certify that the work presented in the Thesis entitled **"Data-Driven Approaches For Improving The Physics Of Heating In Climate Models And Fire Danger Assessment Using AI Techniques"**, is a bonafide work done by me under the supervision of **Dr. Manmeet Singh and Dr. Yogeshwar Singh Dadwhal** and has not been submitted elsewhere for the award of any degree.

Date:_____

Place:_____

Name: Vishnu Dev Tripathi

Roll No.: 21-27-12

Department. School of Computer Engineering & Mathematical Science

Defence Institute of Advanced Technology

Girinagar, Pune

COUNTERSIGNED

Dr. Manmeet Singh

Scientist C

Indian Institute of Tropical

Meteorology(IITM), Pune

Dr. Yogeshwar Singh Dadwhal

Professor

Defence Institute of Advanced

Technology(DIAT)

Girinagar, Pune

ACKNOWLEDGEMENTS

A B S T R A C T

IMPROVING THE PHYSICS OF HEATING IN CLIMATE MODELS USING DEEP LEARNING

Precipitation is vital for different sectors of societal development, including agriculture, infrastructure, water supplies, industries, transport, among others, and is an essential part of the Earth's weather and climate. Accurate and reliable predictions of precipitation distribution across space and time scales remains a major challenge in weather and climate models. Over the tropics, the release of latent heat by the condensation of water vapor into liquid or solid during the formation of clouds is a key process that drives thunderstorms, tropical cyclones and large-scale circulation. Latent heating must be accurately represented in weather and climate models for skillful precipitation forecasts, but this process is challenging due to its intricate and nonlinear connections among several meteorological variables. This research utilizes a data-science model to construct 3-dimensional fields of time-mean latent heating using meteorological fields. To provide the best possible results, the model is trained using cutting-edge machine learning methods and subjected to extensive hyperparameter optimization. Improved representation of latent heating in weather and climate models has the potential for improving rainfall predictions over the tropics.

FOREST FIRE DANGER ASSESSMENT USING MACHINE LEARNING

Wildfire risk assessment is challenging due to uncertainties around fuel flammability and ignite potential. Massive wildfires have the power to burn down homes and claim lives. Because it directly influences fuel flammability, fuel availability, and fire spread, live fuel moisture content (LFMC), or the amount of water per unit of dry biomass in plants, is an essential criterion for evaluating the risk of wildfires. The flammability of a fuel tends to decrease as its moisture content rises. This occurs as a result of a delayed ignition and a slower diffusion rate. Using information on the fuel's moisture content, one may predict the number of fire incidents and the area burned. In this work, we will forecast LFMC using models powered by machine learning and deep learning. For the Western American subregion, we modelled the relationship between spectral indices derived from satellite imagery and LFMC determined on the ground. In order to analyse the spatiotemporal distribution of fire risk, our LFMC model incorporates remote sensing data. The accompanying fire hazard map is an effective tool for speeding decision-making and organising early warning and firefighting operations.

Contents

List of Figures

N O M E N C L A T U R E

ANN   Artificial Nueral Network

NAN   Not A Number

ITCZ   Inter-Tropical Convergence Zone

NWP   Numerical Weather Prediction

CFSv2   Coupled Forecast System Model Version 2

MT   monsoon trough

MCSs   Mesoscale Convective Systems

LH   Latent Heat

ReLU   Rectified Linear Unit

RMSE   Root Mean Squared Error

NASA   National Aeronautics and Space Administration

MERRA   Modern-Era Retrospective analy- sis for Research and Applications

LFMC   Live Fuel Moisture Content

VV   VV Polarisation

VH   VH Polarisation

SWIR   Short Wave Infrared Band Reflectance

NIR   Near Infrared Band Reflectance

NDVI   Normalized Difference Vegetation Index

NDWI   Normalized Difference Water Index

NIRV   Near Infrared Vegetaion

USA   United States Of America

RMSE   Root Mean Squared Error

Chapter 1

Introduction and Literature Review

1.1 Latent Heat

1.1.1 Introduction

Definition

The Indian summer monsoon season, which lasts from June to September, is the principal rainy season of the region, providing 78 percent of India's yearly precipitation [5]. Despite its regularity, there is significant year-to-year variability in the rainfall distribution, resulting in periods of flood and drought. Diabatic heating over the tropics is a key forcing that drives the large-scale tropical atmospheric circulation, and the dominant contributor to diabatic heating is the release of latent heat from organised cloud structures [5]. Indian monsoon circulation is initiated by a land-sea thermal contrast, but is largely maintained by the positive feedback between the circulation and latent heat release.

The significance of latent heat in numerical weather prediction models is crucial for generating accurate precipitation forecasts. When water changes state, such as from liquid to vapour or from ice to liquid, latent heat refers to the energy that is either absorbed or released [15]. Precipitation is reproduced in numerical weather prediction models by following the movement of water vapor, cloud particles, and other atmospheric components [15]. This simulation is predicated on the physical rules that control the behaviour of these components, with latent heat playing a crucial role in these laws. As water vapour condenses into cloud droplets or ice crystals, it releases latent heat, which the surrounding air subsequently absorbs. This process causes the air to warm and alters its buoyancy, resulting in cloud formation and precipitation. In a similar manner, when cloud droplets or ice crystals re-evaporate into water vapor, they absorb latent heat, resulting in cooling of the surrounding air and changes in the stability of the atmosphere [15].

The effects of latent heat on atmospheric processes are represented in numerical weather prediction models using equations that account for the release and absorption of energy during phase shifts [15]. Based on fundamental thermodynamic principles, these equations

require precise measurements of atmospheric variables such as temperature, humidity, and pressure [15]. By incorporating the role of latent heat into their models, meteorologists can enhance the accuracy of their precipitation forecasts, especially for extreme weather events such as thunderstorms and hurricanes. Even minor variations in atmospheric conditions can have a substantial effect on precipitation patterns.
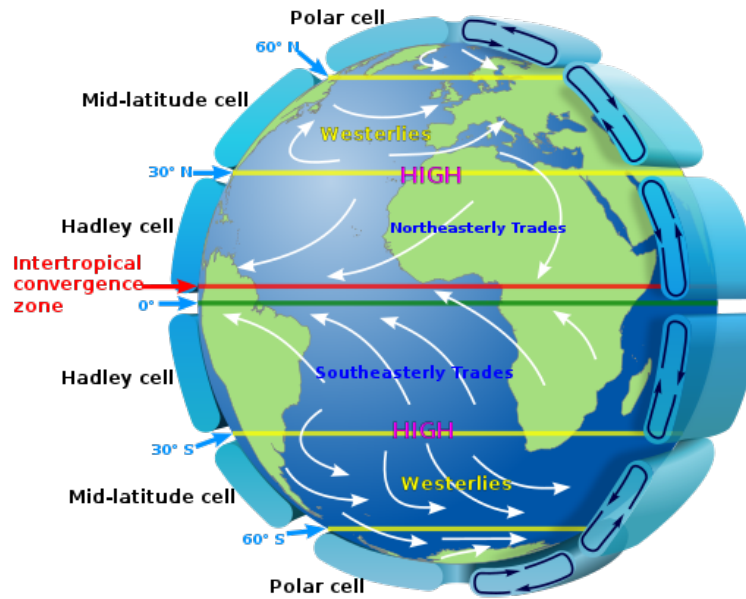
Figure 1.1: Idealised depiction (at equinox) of large-scale atmospheric circulation on Earth (Image from: $https : //en.wikipedia.org/wiki/Atmospheric\_circulation$)

To comprehend the Indian summer monsoon, we must comprehend the dynamics and notion of the ITCZ, the monsoon trough, and the offshore trough.The Inter-Tropical Convergence Zone, or ITCZ, is a low-pressure band that forms around the equator of the Earth due to the convergence of trade winds from the Northern and Southern hemispheres [16]. This convergence allows warm, humid air to rise, resulting in copious precipitation. The ITCZ plays a critical role in providing monsoon rains to India, which are essential for agriculture and the economy as a whole [16]. The ITCZ swings northward throughout the summer months, and monsoon rains follow [16]. The monsoon trough is a low-pressure zone that extends from the Himalayan foothills to the Bay of Bengal [11]. It is an important aspect of the South Asian monsoon and is responsible for the majority of monsoon-season rainfall in India. As the monsoon continues, the monsoon trough shifts northward, and its

movement is closely tied to that of the ITCZ [11]. A robust monsoon trough can result in broad and copious precipitation, whereas a weak or nonexistent monsoon trough might cause dry conditions.A near-coastal offshore trough is a low-pressure system that forms over the ocean. This phenomena is the result of the interaction between warm, wet air from the ocean and cooler air over land [11]. During the pre-monsoon and monsoon seasons, offshore troughs generally bring heavy precipitation to the coastal districts of India.
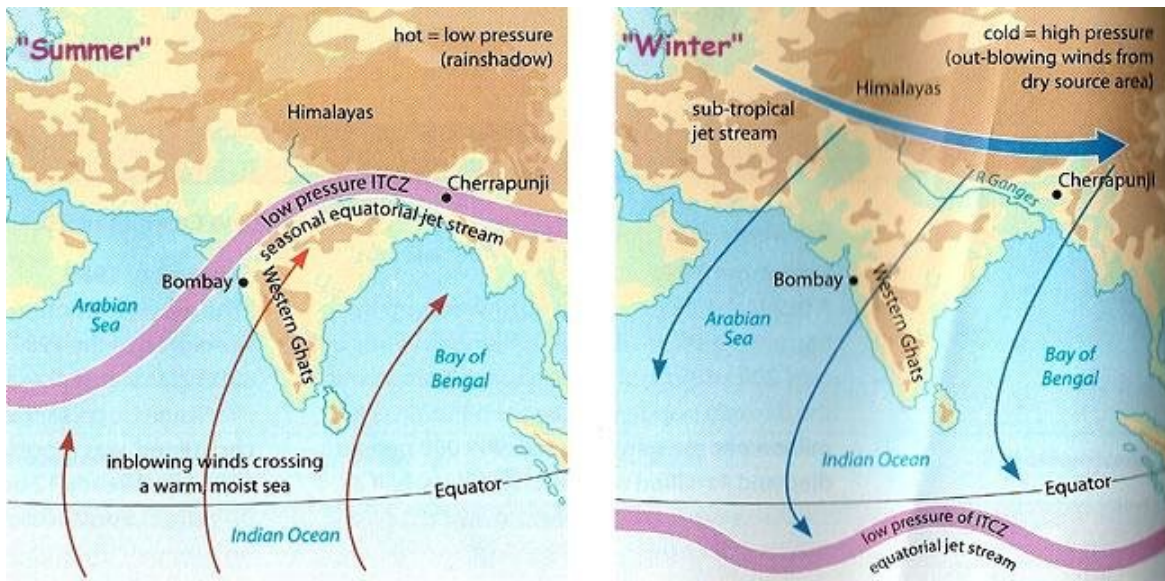
Figure 1.2: Comparison of the positions of the Inter-Tropical Convergence Zone (ITCZ) during the summer (left image) and winter (right image) in India [4]

Convective and stratiform precipitation systems are two unique precipitation patterns that occur in the atmosphere and are created by the dynamics of the monsoon trough and the offshore trough, respectively. These two systems differ in their vertical and horizontal structure, as well as in latent heat distribution [1]. Convective precipitation is caused by convective instability, which happens when warm and wet air rises and cools, resulting in cloud formation and precipitation. This form of precipitation is often severe and localized, and it is prevalent in regions with considerable atmospheric instability, such as tropical regions and regions with significant temperature extremes. The release of latent heat is restricted to convective cells in convective precipitation systems, resulting in a highly concentrated and localised distribution of latent heat [1]. Stratiform precipitation systems, on the other hand, have a more horizontal structure, with precipitation happening over a large area. Typically,

these systems are connected with large-scale weather patterns, such as frontal systems, in which a vast amount of warm and moist air is forced to ascend over a cooler, denser air mass. In stratiform precipitation systems, latent heat is distributed across a broader area, resulting in a more gradual and less intense dispersion of latent heat than in convective precipitation systems [1]. In conclusion, convective and stratiform precipitation systems differ in their vertical and horizontal structure, as well as their latent heat distribution. Convective precipitation is intense and concentrated, whereas stratiform precipitation is diffuse and extensive. In convective precipitation, the distribution of latent heat is extremely concentrated and confined, whereas it is more gradual and extensive in stratiform precipitation [1].
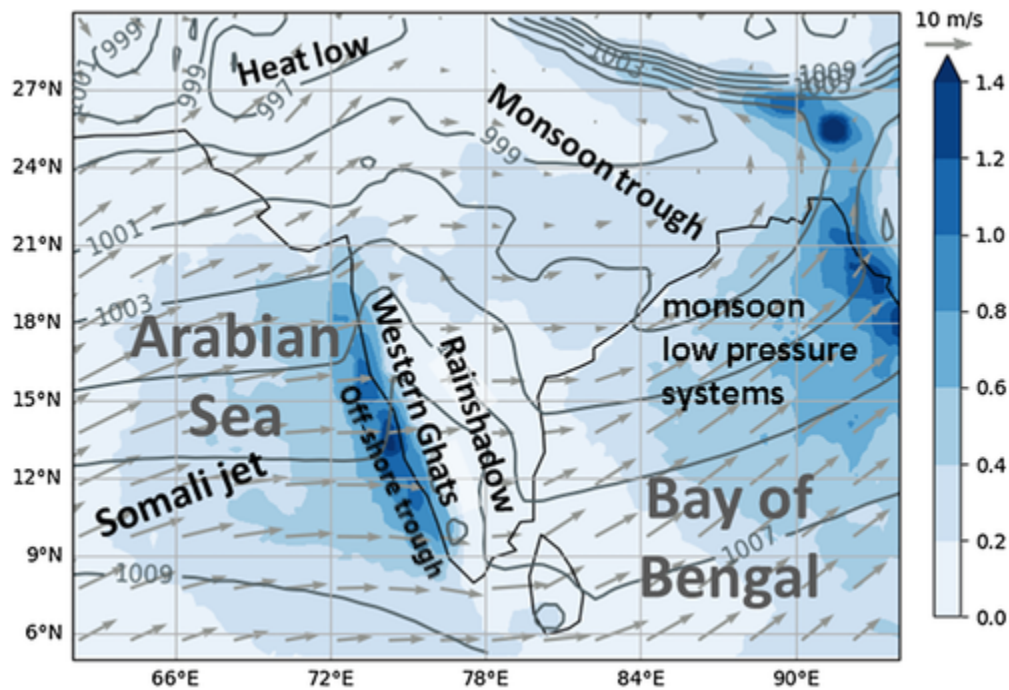
Figure 1.3: Mean rainfall (TRMM 3B42, 1998–2016, blue shading), sea level pressure (ERA-Interim, contours), and 950hPa horizontal winds (arrows) in South Asia in June. Some of the main features of the monsoon system are labelled [Colour figure can be viewed at *wileyonlinelibrary.com*] [6]

The importance of latent heat in numerical weather prediction models for precipitation forecasting cannot be overstated. It determines the time and intensity of precipitation by influencing the behaviour of water vapour and other atmospheric components [14]. In the tropics, the release of latent heat during cloud formation is a crucial mechanism that causes

numerous meteorological phenomena, such as thunderstorms, tropical cyclones, and large-scale circulation. To accurately and dependable precipitation forecasts, weather and climate models must accurately capture the latent heating process. This is a difficult undertaking due to the complex and nonlinear relationships between the numerous meteorological factors involved in this process [14]. Research is ongoing to understand these complex relationships [14].

1.1.2 Literature Review

Decades of scientific study into the improvement of global precipitation predictions have resulted in the state-of-the-art Numerical Weather Prediction(NWP) systems that underpin practical weather prediction across the world [14]. Although the NWP has been a success, there are still concerns in domains such as precipitation and soil moisture where the NWP models have a high level of bias [14]. NWP systems are fundamentally linked solutions to the governing equations of the atmosphere, ocean, and land. In order to solve these partial differential equations, it is assumed that the Earth is spherical and beginning conditions and boundary conditions are specified [14]. Equations are solved for prognostic fields including pressure, temperature, winds, and tracers like water vapour and ozone. Precipitation and soil moisture are diagnostic fields that are produced from prognostic fields using emperical relationships, often known as parameterization schemes [14]. These parameterization approaches are difficult to develop and enhance in NWP systems, and they are largely responsible for the diagnostic field issues. CFSv2 is a model utilised by operational nodal centres throughout the world for weather forecasting [14].

Ayantika Dey Choudhury and R. Krishnan [5] analyses how the South Asian monsoon trough (MT) responds to latent heating originating from mesoscale convective systems (MCSs). Examining the dynamics of the MT during active monsoons, the authors utilise a simplified general circulation model (GCM) of the atmosphere and further diagnostic studies of observations [5]. Over the Indo-Gangetic plains, the MT is distinguished by a deep cyclonic vorticity with organised MCSs spanning over 3500-4000 km. The study demonstrates

that a heating profile of stratiform type promotes the upward development of continental-scale cyclonic circulation well above the mid-troposphere in the MT area [5]. The results indicate that the horizontal scale of the midlevel vorticity reaction grows dramatically with stratiform population, but the midlevel response is more localised when deep convective clouds dominate the MCS [5]. The conclusion of the study is that the vertical deepening of the MT during active monsoons is a large-scale dynamical reaction to organised MCSs that exert a crucial influence on the upward development of cyclonic circulation well above the midtroposphere [5].

T. N. Krishnamurti, Arindam Chakraborty, and A. K. Mishra [10] discuss the significance of vertical profiles of diabatic heating rates for comprehending three-dimensional heating distributions. The convective–stratiform heating (CSH) product, made available by the Tropical Rainfall Measuring Mission (TRMM) project office of the National Aeronautics and Space Administration (NASA), provides observed estimates of heating derived from the TRMM satellite's microwave radiances and the precipitation radar [10]. This study investigates the performance of cumulus parameterization schemes in describing the amplitude and three-dimensional distributions of heating [10]. Three versions of the global atmospheric model from Florida State University (FSU) are executed, each with a unique cumulus parameterization scheme. The paper emphasises that some cumulus parameterization schemes overestimate the amplitude of heating, while others carry lower values, and that the models exhibit large errors in the placement of the vertical level with the highest heating [10]. The ensemble mean closely resembles the model characteristics and is subject to large errors, whereas the superensemble provides the most accurate heating forecasts for the next 60 hours [10]. This study demonstrates the importance of diagnosing and correcting heating model errors for individual member models [10].

Using data from the TRMM CSH and TMI 2A12 datasets, Manuel D. Zuluaga, Carlos D. Hoyos, and Peter J. Webster [20] analyse the four-dimensional latent heating structure over the Asian monsoon region. The study identifies interannual and intraseasonal variations in LH, with intraseasonal variability accounting for a significant amount of interannual variability [20]. During the summer active and repressed stages of the monsoon intraseasonal

oscillation, the analysis also identify a spatial dipole of LH anomalies between the equatorial Indian Ocean and the Bay of Bengal [20]. Even while satellite-based LH algorithms can reflect the geographical and temporal aspects of the vertically integrated heating in the Asian monsoon region, the vertical distribution of atmospheric heating during distinct convective phases is not recorded effectively [20]. The authors argue that satellite-derived radiative heating/cooling products are required to supplement the LH products in order to offer a more accurate global representation of atmospheric heating [20].

Novelty of work

As evidenced by the literature review section, numerous researchers have attempted to demonstrate that there are flaws in the parameterization of latent heat in climate models, particularly with its vertical profile. We have developed a first-of-its-kind ANN model based on deep learning that can better estimate latent heat using satellite data.

1.2   Live Fuel Moisture Content (LFMC)

1.2.1   Introduction

Definition

The moisture content of vegetation is one of the key factors governing fuel flammability and a strong predictor of fire danger [3]. As a result of delayed ignition and slow propagation, fuel flammability tends to decrease as fuel moisture content rises [3]. In addition, information on the moisture content of the fuel has been useful in estimating the volume and frequency of flames. Hence, anticipating fuel moisture content has been a primary focus of study on fire behaviour and fire hazards. In fact, some authors have proposed a classification of fire threats according to the LFMC (Living Combustible Fuel Moisture Content), with values for high fire dangers between 60 and 80 percent and extreme fire dangers below 60 percent [3].

$$LFMC = \frac{Wet\ Weight - Dry\ Weight}{Dry\ Weight}$$

Figure 1.4: The Equation for finding the LFMC [3]

LFMC Value	Fire Hazard Prone
Below 60%	Extreme
Between 80% to 60%	High
Above 80%	Low

Figure 1.5: Classification of Fire Hazard Prone on the basis of LFMC [3]

The amount of water in a fuel (plant) that is available for combustion is expressed as a percentage of the fuel's dry mass. For instance, if the gasoline is completely dry, its moisture content will be zero percent.

There are two main categories of fuel moisture [8]: 1)Living fuel moisture, which is present in living plants and ranges from 30 to 300 percent based on the species and season, is between 30 and 300 percent. 2) Dead fuel moisture: occurs in dead plants, forest litter, slash, etc., and ranges from 2% to 30%, with rapid temporal and spatial change. Fuel moisture influences combustion. [8]: 1)Before a fuel can burn, its moisture must be converted into vapour by the heat process. 2) The greater the heat temperature required to dry the fuel, the higher its moisture content. 3) The presence of moist fuel can influence the propagation of a wildfire.

The best method for determining LFMC is to survey the area and collect fuel samples. The sample must first be weighed to determine its fresh weight before being dried in an oven to get its dry weight. The relative humidity of fuel is stated as a percentage of its dry weight. Yet, LFMC field estimates are costly and time-consuming [3]. Hence, extrapolating over space and time is difficult. Environmental managers believe that fitting a model to estimate its LFMC using factors that are easy to assess and spatially generalizable is a beneficial technique [3].These LFMC estimates can also be used to feed the almost instantaneous firefighting system. LFMC can be estimated using remote sensing data because changes in vegetation water content influence the spectral fingerprints of the vegetation [3]. Fitting an

empirical model with field estimates of LFMC and spectral indices is one of the most typical ways for achieving this objective. The primary advantage of this method is that these models are straightforward to implement and generate accurate estimations [3].

However, these empirical relationships may vary according on sensor kind and sample location. This suggests that the empirical model is likely accurate just for the particular sensor used to calibrate the model and only at the local level [3]. Due to the fact that remote sensing satellites can give large-scale, multi-band, near-real-time imagery, remote sensing techniques are the method of choice for anticipating large-scale LFMCs. This approach has been the subject of the most research for forecasting LFMC using optical remote sensing data [18]. These strategies rely on the ability of leaf water to absorb light at near-infrared (NIR) or short-wave infrared (SWIR) wavelengths [18].

The wavelength of microwave remote sensing is four orders of magnitude longer than optical remote sensing. Microwave remote sensing is more effective than optical remote sensing at detecting dynamic changes in vegetation moisture because microwaves may travel through clouds and vegetation canopies [18]. Emerging evidence indicates that active microwave remote sensing techniques, such as Synthetic Aperture Radar (SAR), can reliably predict fire-related factors [18]. Using the linear terrestrial backscatter model and the vegetation backscatter water cloud model improved the Forest FMC retrieval using Landsat-8 data and empirical methodologies [18]. Since different remote sensing data have varying sensitivities to vegetation, water, and biomass, it is less efficient for LFMC to collect remote sensing data from a single source [18]. LFMC is estimated using remote sensing data from multiple sources, so avoiding the limits of data from a single source and providing a larger set for extracting the properties required for locating LFMC.

1.2.2   Literature Review

A spatial prediction model for the susceptibility to forest fires was proposed by Zhang, Wang, and Liu [19] utilising a CNN.Using a geographic information system, the locations of forest fires that occurred in Yunnan Province, China, between 2002 and 2010 were mapped, along

with a list of fourteen factors that influenced those fires [19]. The class imbalance was rectified through oversampling, and proportional stratified sampling was utilised to construct sample libraries for training and validation. To improve prediction accuracy, hyperparameters for a CNN architecture that is suited for predicting forest fire susceptibility were optimised. In order to generate the spatial prediction map of Yunnan Province's susceptibility to forest fires, the trained model was given the test dataset [19].The Wilcoxon signed-rank test, the receiver operating characteristic curve, and the area under the curve were used to evaluate the prediction performance of the suggested model (AUC). The results supported the better accuracy of the proposed CNN model (AUC 0.86 vs Random Forests, Support Vector Machine, Multilayer Perceptron Neural Network, and Kernel Logistic Regression Benchmark Classifiers) [19]. CNN is a viable alternative for the geographic prediction of forest fire susceptibility due to its superior fitting and classification abilities, as well as its capacity to fully utilise neighbourhood information [19].In this study, a novel use of CNN is explored: predicting forest fire vulnerability.

In the Chaco Serrano subregion, Juan Arganaraz, Marcos Landi, Sandra Bravo, Gregorio Gavier-Pizarro, Carlos Scavuzzo, and Laura Bellis (2016) [3] modelled the relationship between field estimates of LFMC and spectral indices derived from satellite imagery. The researchers next analysed the connection between fire hazard estimations derived from LFMC computations and fire activity. While the model corresponding to shrublands still need improvement (R2score = 0.57), empirical LFMC models fitted for grasslands, Chaco Serrano forests, and glossy privet woods can be regarded as highly realistic [3].Monthly maps of fire danger reliably depicted the incidence of fires during both years of high and low fire activity. It is obvious that there is a direct relationship between LFMC and fire activity in the Chaco Serrano subregion, as the bulk of fires occurred predominantly in locations with high or extreme fire hazard [3]. Using remote sensing data and LFMC models, it may be able to examine the spatial and temporal distribution of fire danger in the Chaco Serrano subregion [3]. The associated fire danger maps are a valuable tool for coordinating early warning and fire suppression activities and expediting decision-making processes.

Two ensemble models that combine deep learning models were proposed by Xie, Jiangjian,

Tao Qi, Wanjun Hu, Huaguo Huang, Beibei Chen, and Junguo Zhang in 2022 [18].The first is an Adaboost ensemble model based on LSTM-TCN, while the second is a stacking ensemble model based on LSTM, TCN, and LSTM-TCN. For their investigation, they selected measured LFMC data, remote sensing data from MODIS, Landsat-8, and Sentinel-1, as well as supplementary information about canopy height and land cover in fire-prone regions of the Western United States [18].Comparisons were made between the retrieval outcomes of several models utilising diverse groupings of remote sensing data. The results suggest that the use of multi-source data can combine the advantages of multiple types of remote sensing data, resulting in a greater level of LFMC retrieval accuracy than that of single-source data. Using all remote sensing data from MODIS, Landsat-8, and Sentinel-1, the stacking ensemble model generated the best LFMC retrieval results, with R 2 = 0.85, RMSE = 18.88, and ubRMSE = 17.99 [18].Ensemble models can extract the nonlinear link between LFMC and remote sensing data more precisely [18].

By combining the bare soil backscatter linear model with the vegetation backscatter water cloud model, Wang, Long, Xingwen Quan, Binbin He, Marta Yebra, Minfeng Xing, and Xiangzhuo Liu (2019) [17] sought to assess the effectiveness of C-band Sentinel-1A data for multi-temporal retrieval of forest FMC (WCM). In order to construct a look-up table (LUT) including the modelled VH backscatter and FMC, the coupled model that directly related the observed backscatter to the FMC was calibrated using field FMC measurements and the relevant synthetic aperture radar (SAR) backscatters (VV and VH) [17].As a cost function for determining the optimal FMC from the LUT, the absolute difference (MAEr) between the modelled and observed VH backscatters was used. The three-fold cross-validation method was utilised to evaluate the performance of the proposed methodology by dividing the total sample into three equal halves. This procedure was repeated three times, with two instances serving as model calibration and the other as model validation. With RMSEs of 19.53 percent for Sample 1, 12.64 percent for Sample 2, and 15.45 percent for Sample 3, the results revealed that the estimated and measured forest FMC were consistent across all three validation samples [17].In order to further test the efficacy of the C-band Sentinel-1A data for forest FMC estimation, the findings were compared to those obtained using the

optical Landsat 8 Operational Land Imager (OLI) data and the empirical partial least squares regression (PLSR) method. Compared to the stated Sentinel-1A data-based technique, the RMSE between estimated and measured forest FMC was greater by 20.11 percent (Sample 1), 26.21 percent (Sample 2), and 26.73 percent (Sample 3) for the latter (Sample 3) [17]. This work shown that C-band Sentinel-1A data have significant retrieval capabilities for forest FMC, which paves the way for the creation of a novel operational SAR data-based technique for estimating forest FMC [17].

Using microwave backscatter (from Sentinel-1) and optical reflectance, Krishna Rao, A. Park Williams, Jacqueline Fortin Flefil, and Alexandra G. Konings [12] offer a physics-assisted recurrent neural network model for mapping LFMC every 15 days at 250 m resolution over the western US (from Landsat-8). The model was aided by physics by incorporating manually generated inputs for radiative transfer processes and associated parameters. The National Fuel Moisture Database was utilised to train the deep learning model. When the model was cross-validated at 125 sites, $R2 = 0.63$, RMSE = 25.0 percent, and bias = 1.9 percent, it demonstrated an acceptable overall prediction [12].The model predicted site-means substantially more accurately than site-anomalies ($R2 = 0.71$ and RMSE = 14.3 percent vs. $R2 = 0.55$ and RMSE = 21.3 percent) [12].The model scored best in predicting LFMC in shrublands ($R2 = 0.69$) compared to closed broadleaf deciduous sites ($R2 = 0.49$) among the six land cover categories studied. The addition of microwave backscatter as a model input significantly enhanced performance ($R2$ increasing from 0.44 to 0.63, and RMSE decreasing from 31.8 percent to 25.0 percent) [12].Hence, microwave backscatter provides an additional data source for LFMC mapping in addition to optical remote sensing metrics. The high resolution dynamic LFMC maps described here could be utilised to more precisely characterise wildfire threat and to assess plant drought stress on a wide scale.

Novelty of work

As evidenced by the literature review, other researchers have attempted to forecast LFMC, but we have constructed a model for its improved prediction and included meteorological variables as predictors, which others have overlooked. Also, we have used more enriched

data capturing most of the locations of western USA.

Chapter 2

Data Analysis and Data Preprocessing

2.1 Latent Heat

2.1.1 Data Source

Target Data

The GPM DPR Spectral Latent Heating Profiles L3 1 day 0.5 degree x 0.5 degree V07 (GPM_3HSLH_DAY) dataset was used as a target data. It provides information on latent heating, Q1-QR, and Q2 profiles from DPR rain data. The dataset contains 80 layers at fixed heights of 0.00-0.25 km to 19.75-20.00 km, with a spatial resolution of 0.5 degrees by 0.5 degrees, a vertical resolution of 125-250 m, and a temporal resolution of 1 day. The dataset downloaded covers a date range from 09/03/2014 to 31/10/2022. The dataset includes several variables, including LHUnCndMean, which represents the latent heating unconditional mean, and precipPix, which is the number of precipitating pixels. The variable allPix represents all pixel counts. The Latent Heat variable is calculated as LHCndMean multiplied by precipPix and divided by allPix, and is expressed in units of K/hr. The dataset is available for access and download through the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) website.

Predictor Data

MERRA-2 inst3_3d_asm_Np: 3d,3-Hourly,Instantaneous,Pressure-Level,Assimilation,Assimilated Meteorological Fields V5.12.4 (M2I3NPASM) dataset was used as predictor data. M2I3NPASM (also known as inst3 3d asm Np) is a dataset including 3-hourly, 3-dimensional data from version 2 of the Modern-Era Retrospective analysis for Research and Applications (MERRA-2). The data include temperature, wind components, vertical pressure velocity, water vapour, ozone mass mixing ratio, and layer height for 42 pressure levels. Every three hours, beginning at 00:00 UTC, the data field is accessible. The collection contains data with a spatial resolution of 0.5 degrees by 0.625 degrees, a vertical resolution of 42 pressure levels, and a temporal resolution of 3 hours. The dataset downloaded covers a date range from

09/03/2014 to 31/10/2022. The dataset is available for access and download through the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) website.

2.1.2   Underlying variables

We have used 14 variables as predictors for the prediction of our target variable, latent heat. These variables are downloaded from the source of data previously mentioned. The table below contains the names of fourteen predictor variables, along with their abbreviated names and unit designations, as they appear in the MERRA-2 dataset (M2I3NPASM).

Variable	Shortname	Units
air_temperature	**T**	K
mass_fraction_of_cloud_liquid_water	**QL**	kg kg-1
vertical_pressure_velocity	**OMEGA**	Pa s-1
surface_pressure	**PS**	Pa
eastward_wind	**U**	m s-1
ertels_potential_vorticity	**EPV**	K m+2 kg-1 s-1
sea_level_pressure	**SLP**	Pa
mass_fraction_of_cloud_ice_water	**QI**	kg kg-1
edge_heights	**H**	m
northward_wind	**V**	m s-1
relative_humidity_after_moist	**RH**	1
surface geopotential height	**PHIS**	m+2 s-2
specific_humidity	**QV**	kg kg-1
ozone_mass_mixing_ratio	**O3**	kg kg-1

Figure 2.1: The list of 14 predictor variables, along with their abbreviated names and unit designations, as they appear in the MERRA-2 dataset (M2I3NPASM).

Air Temperature(T)

The air temperature is the measurement of the kinetic energy of the molecules in the atmosphere, which is derived from the assimilation of various observations such as surface weather stations, radiosondes, and satellite retrievals. The air temperature is expressed in Kelvin.

Mass fraction of cloud liquid water(QL)

Mass fraction of cloud liquid water refers to the mass of liquid water within a cloud per unit mass of the air containing the cloud. It is a variable used to describe the amount of cloud liquid water in the atmosphere. This variable is expressed in units of kg/kg (kilograms of water per kilogram of air) and is calculated based on the assimilation of various meteorological parameters such as temperature, pressure, and water vapor mixing ratio, as well as observational data from satellite and ground-based instruments.

Vertical pressure velocity(OMEGA)

Vertical pressure velocity is the rate at which air moves vertically in the atmosphere as a result of pressure differences. It is a unit of measurement for the force and direction of atmospheric motion, expressed in pascals per second (Pa/s).

Surface pressure(PS)

Surface pressure refers to the atmospheric pressure at the surface of the Earth, also known as barometric pressure. Generally, it is measured in Pascals (Pa). In weather forecasting, climate modelling, and air quality research, surface pressure is a crucial parameter because it provides information about atmospheric circulation patterns and the global movement of air masses.

Eastward wind(U)

Eastward wind refers to the wind component that blows in the direction of the east, one of the four cardinal directions on the horizontal plane. It is typically measured in metres per second (m/s) and is also known as the "u-wind" component. When the wind is blowing from the west to the east, the u-wind component is positive, and when the wind is blowing from the east to the west, it is negative.

Ertel's potential vorticity(EPV)

Ertel's potential vorticity quantifies the potential vorticity of atmospheric air parcels. Potential vorticity of Ertel is defined as the product of absolute vorticity and static stability of the air parcel, standardised by air density. It is primarily used to investigate the dynamics of the atmosphere, such as the creation and movement of weather systems such as cyclones and anticyclones. The potential vorticity of Ertel can also be used to identify regions of high or low potential vorticity that may be associated with atmospheric disturbances and to forecast their future evolution.

Sea level pressure(SLP)

Sea level pressure refers to the atmospheric pressure at sea level, which is the pressure that would be measured by a barometer placed at sea level. Sea level pressure is an important meteorological parameter used to describe the state of the atmosphere and is typically measured in units of Pascals (Pa) or millibars (mb).

Mass fraction of cloud ice water(QI)

Mass fraction of cloud ice water refers to the mass of ice particles in a cloud per unit mass of the cloud's contained air. It describes the amount of cloud ice water in the atmosphere. This variable is stated in kilogrammes per kilogramme (kilograms of ice per kilogramme of air). It is essential to note that the mass fraction of cloud ice water can vary greatly based on the air's temperature and moisture content, as well as cloud dynamics and precipitation processes.

Edge heights(H)

Edge heights refer to the height levels that form the vertical borders between distinct layers. The M2I3NPASM (or inst3 3d asm Np) dataset in MERRA-2 offers instantaneous, three-

dimensional, three-hourly meteorological parameters at 42 pressure levels. These pressure levels are not evenly distributed and change throughout time and space. To utilise the data, it is important to construct a series of predetermined height levels that separate the atmosphere into distinct layers. In MERRA-2, edge heights are specified as the fixed height levels that correspond to the vertical boundaries between consecutive layers.

Northward wind(V)

Northward wind denotes the component of wind speed that is headed northward. Wind speed is the horizontal velocity of air molecules in the atmosphere, and it is commonly measured in metres per second (m/s). The eastward wind component is the other orthogonal component describing the direction of the wind vector in the Earth's atmosphere. Combined, the northerly and easterly wind components indicate the direction and velocity of the wind at a certain location in the atmosphere.

Relative humidity after moist(RH)

Relative humidity after moist refers to the relative humidity of air after it has been moistened with water vapour. It is the ratio, stated as a percentage, of the partial pressure of water vapour in the air to the saturation vapour pressure at the same temperature. The phrase "after moist" alludes to the fact that this quantity is computed after the addition of water vapour, which can result from processes like as evaporation and transpiration. The relative humidity after moist is an essential meteorological measure used to characterise the amount of water vapour in the atmosphere and its connection to air temperature and pressure. It is especially helpful for comprehending the creation and behaviour of clouds, as well as for forecasting weather patterns and climate trends.

Surface geopotential height(PHIS)

Surface geopotential height is the height of a hypothetical surface of constant atmospheric pressure, known as the geopotential surface, compared to a reference level at the surface of the Earth. This variable is frequently used to characterise the topography of the Earth's surface and is a crucial factor in weather and climate prediction. Typically, surface geopotential height is measured in metres (m) or geopotential metres (gpm).

Specific humidity(QV)

Specific humidity is the amount of water vapour in a unit mass of moist air. It is commonly given in units of kilogrammes of water vapour per kilogramme of moist air (kg/kg). It is defined as the ratio of the quantity of water vapour to the total mass of moist air (including dry air and any other gases present). Specific humidity is a crucial meteorological quantity that influences the behaviour of the atmosphere, particularly in terms of cloud formation and precipitation.

Ozone mass mixing ratio(O3)

Ozone mass mixing ratio refers to the concentration of ozone gas in a unit mass of dry air at a specific point in the atmosphere. It is commonly measured in parts per million (ppm) or kilogrammes of ozone per kilogramme of dry air (kg/kg) and is a crucial atmospheric parameter used to research the Earth's ozone layer and its interactions with other atmospheric components. Moreover, alterations to the ozone layer can affect the chemistry and weather patterns of the atmosphere.

Latent heating conditional mean(LHCndMean)

Latent Heating Conditional Mean is a statistical quantity that represents the average rate of latent heat release (i.e., the energy released during the phase change of water vapour to liquid

or solid) per unit mass of precipitation, conditional on precipitation occurring within a given volume of atmosphere. Specifically, the Latent Heating Conditional Mean is computed by detecting all precipitation features inside the DPR swath (i.e., pixels containing precipitation echoes) and then calculating the average rate of latent heat release for each feature. These individual rates are then weighted by the area of each feature to generate the Latent Heating Conditional Mean for the entire swath.

The number of precipitating pixels(precipPix)

The number of precipitating pixels in GPM DPR (Dual-frequency Precipitation Radar) refers to the number of radar pixels that contain precipitation echoes. These echoes indicate the presence of precipitation in the atmosphere, such as rain, snow, or hail, and are used to assess the intensity and distribution of precipitation in a particular region. The number of precipitating pixels is a crucial statistic that is used to determine the areal coverage of precipitation over a particular region, and it is an essential metric for monitoring and forecasting weather trends.

All pixel counts(allPix)

All pixel counts refers to all pixels within the DPR swath, whether or whether they include precipitation echoes. Typically, these pixels are utilised to offer contextual data about the atmosphere, such as cloud cover, surface topography, and atmospheric moisture content.

Latent Heat(LH)

The Latent Heat variable is calculated as LHCndMean multiplied by precipPix and divided by allPix, and is expressed in units of K/hr.

2.1.3   Data Preparation & Preprocessing

Figure 2.2: The two selected region i.e. Central India [74.5°E-86.5°E, 16.5°N-26.5°N] in black box and Western Ghat region [70°E-80°E, 8°N-25°N] in red box. Left side and right side plots shows wind vector (one of predictor) at 1000 Hpa and 900 Hpa respectively. At 1000 Hpa there are NAN values present at those regions whereas at 900 Hpa there are none.

The data was downloaded for two specific regions i.e. Central India [74.5°E-86.5°E, 16.5°N-26.5°N] in black box and Western Ghat region [70°E-80°E, 8°N-25°N] as shown in figure 2.2.The downloaded data spans nearly eight years, from 09/03/2014 to 31/10/2022. Initially, data were available for 42 levels, ranging from 1000 Hpa to 0.1 Hpa, for the predictor variable. However, upon examination, we discovered that for our two selected locations, NAN values were present at four levels, from 1000 Hpa to 925 Hpa. Hence, we eliminated the levels with NAN values, leaving 38 levels, ranging from 900 to 0.1 hPa.
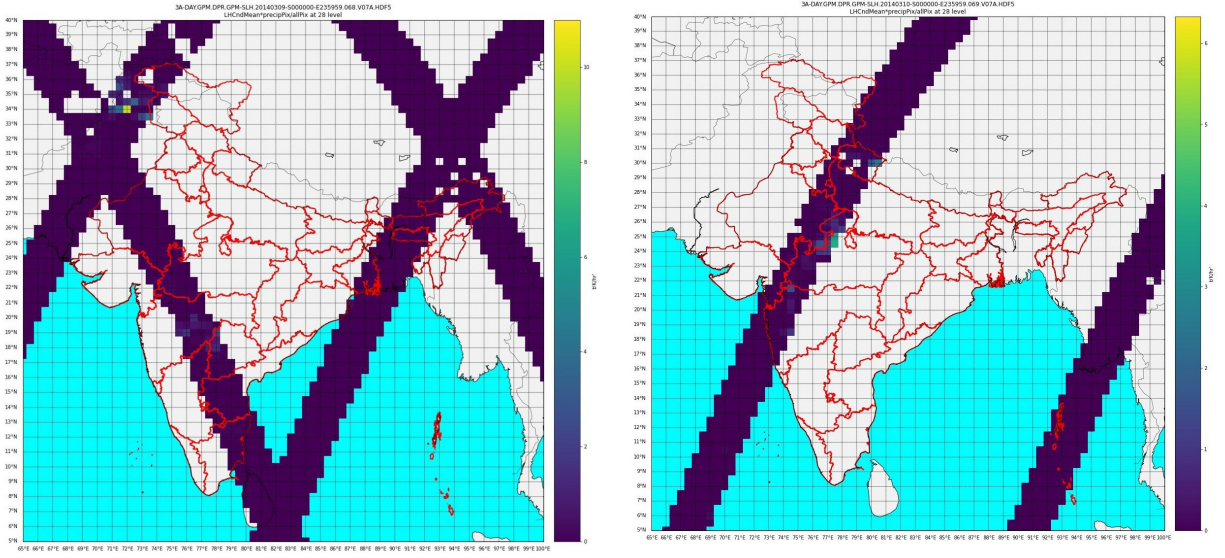
Figure 2.3: The graph shows swath nature of latent heat at two different dates i.e. 09/03/2014 (left one) and 10/03/2014 (right one)

Also, for target variable i.e. latent heat the satellite data available in swath. In figure 2.3 it exhibits the satellite data swath at two time stamps, 09/03/2014 (left) and 10/03/2014 (right), which demonstrates that data for a given region is not continuous on the following day. To account for the nature of the swath, we only extracted data from regions where the satellite passed and eliminated locations where NAN values were present. Similarly, we retrieved and tabulated the predictor data for the same places where the latent heat data was available at a particular time stamp.

To summarize we followed the steps given below for data preprocessing :-

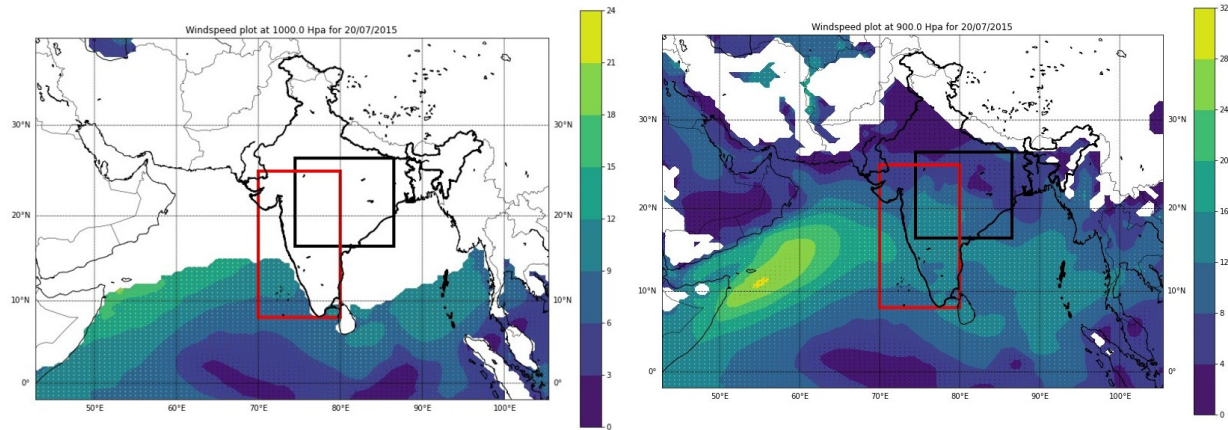1. Collect the data from different repositories mentioned. The data was downloaded for two specific regions i.e. Central India [74.5°E-86.5°E, 16.5°N-26.5°N] in black box and Western Ghat region [70°E-80°E, 8°N-25°N].

2. Convert all the predictor variables into a fixed resolution ,here it is 0.5 ° x 0.5 °, resolution just like target variable. We have used xESMF to convert the data.

3. Merge each variables along the time axis and their geometrical coordinates e.g. latitude and longitude. Also made each pressure level data as column for both predictor and

target variables.

4. Convert all the data into tabular format with .csv extension using pandas dataframe.

5. Replace all the nan value.

6. Load the csv file using Pandas library.

7. Normalize the data of predictor variables using Min Max Scaler. Equation used -

$$z = \frac{(x - xmin)}{(xmax - xmin)}$$

8. Create array of predictor variable and target variable.

9. Split the data into training, testing and validation.

2.2 Live Fuel Moisture Content (LFMC)

2.2.1 Data Source

LFMC Data

The National Fuel Moisture Database (NFMD) is accessible through a web-based querying interface.Since 1977, more than two hundred thousand measurements of fuel moisture have been taken.The database has 976 samples (mainly from the western United States), each of which spans 5 acres and is regularly updated. On a day without dew or precipitation, late afternoon measurements were obtained. For this thesis, 150 typical sample sites were selected. During the study period, LFMC values ranged from 15 to 320 percent, representing normal water conditions for residential fuels.

Sentinel-1 Data

Sentinel-1 is a 12-day-recurrent 5.4 GHz C-band Synthetic Aperture (SAR) radar in the western US. This study uses Sentinel-1 Ground Range Detector (GRD) data. Wide strip mode vertical-to-vertical (VV) interferometry and land-based VH polarisation were used to collect the data. Microwave signals have longer wavelengths, are less impacted by atmospheric conditions and cloud pollution, and can detect deeper vegetation canopies, making them suitable for continuous Earth monitoring. Microwave backscattering, largely driven by water content, influences microwave signal absorption and scattering by surfaces (including vegetation and soil), rendering microwave signals sensitive to vegetation moisture content. Hence, the microwave input for this model was VV and VH. As a result, in this work the microwave input for the model was VV and VH. VV and VH were retrieved from the Google Earth engine.

Landsat Data

Surface reflectance data from Landsat-8 for 16 days are used to create Landsat data. With a 30 m spatial resolution, this is a level 1T product. The normalised differential moisture index (NDWI) and LFMC have a significant positive connection. Because water absorbs energy largely in the near-infrared (NIR) and shortwave infrared (SWIR) spectral regions, the original band reflectance of the red, green, and blue, near-infrared, and shortwave infrared channels is now chosen to represent water change. The Normalized Vegetation Index (NDVI), a quick and accurate empirical gauge of surface vegetation, has a significant impact on LFMC projections. In order to distinguish the effects of biomass and LFMC on Sentinel-1 back scattering, the near-infrared vegetation index (NIRV), which is connected to photosynthetic carbon assimilation, serves as an indication of vegetation biomass levels. In conclusion, the Landsat-8 data were chosen to create three vegetation indices, the NDWI, NDVI, and NIRV, as well as original band reflections for the red, green, blue, near-infrared, and shortwave infrared channels. Google Earth Engine was used to download NDWI, NDVI, and NIRV as well as original band reflections for the red, green, blue, near-infrared, and shortwave

infrared channels.

Auxiliary Data

Seven static auxiliary data categories helped the model grasp the radiative transfer mechanism between the time-varying input and the LFMC. Three categories can categorise data: Soil data, such as silt, sand, and clay concentration, is used to modify microwave back scatter sensitivity to soil moisture to enable search models. Microwave back scatter provides vegetation data. The vegetation canopy's moisture content affects the near- and short-wave infrared bands' sensitivity to remote sensing data. Secondary auxiliary data included Litar's crown height measurement and GLOBCOVER's 300-meter land cover data. Terrain is third. Since the local incidence angle impacts vegetation water back scattering parameterization, the model must be calibrated for the local terrain using the National Elevation Data-slope set and elevation.

ERA5-Land Hourly - ECMWF Climate Reanalysis

The European Centre for Medium-Range Weather Forecasting created ERA5 as a global atmospheric reanalysis dataset (ECMWF). It gives hourly information on a variety of atmospheric factors, including temperature, pressure, wind speed and direction, and precipitation. ERA5 is the most advanced and precise reanalysis dataset available. It is the fifth generation of ECMWF's reanalysis datasets. It encompasses the years 1979 to the present and is updated monthly. Scientists, researchers, and meteorologists utilise ERA5 to study the environment and climate, as well as to enhance weather forecasting. Using Google earth engine, we have obtained 61 meteorology related predictor variables from the ERA5 repository.

2.2.2 Overview of the data

Data is downloaded for the 150-site Western America region as shown in fig 2.4. The region is well-suited for study on LFMC prediction techniques since it is rich in flora including deciduous woods, evergreen coniferous forests, shrub forests, grasslands, and forests with few trees. The chosen study period was from June 2015 to December 2020 due to the accuracy and generality of the data. because there was no satellite data before 2015 Target and predictor variables are shown as monthly time series in the data.

Figure 2.4: Details of 150 sites in Western USA with yearly averaged LFMC

Variables	Unit
Terrain slope	degrees
Altitude above mean sea level	meters
Height of canopy from ground level	meters
Land cover class of live fuel moisture content measurement site from GLOBCOVER dataset	category
Silt fraction of live fuel moisture content measurement site	fraction
Sand fraction of live fuel moisture content measurement site	fraction
Clay fraction of live fuel moisture content measurement site	fraction
Backscatter in VV polarization	decibels
Backscatter in VH polarization	decibels
Red band surface reflectance	x 10,000
Green band surface reflectance	x 10,000
Blue band surface reflectance	x 10,000
Shortwave infrared band surface reflectance	x 10,000
Near infrared band surface reflectance	x 10,000
Normalized difference vegetation index	fraction
Normalized difference water index	fraction
Near infrared vegetation	x 10,000
Backscatter in VV polarization divided by red band reflectance	decibels
Backscatter in VV polarization divided by green band reflectance	decibels
Backscatter in VV polarization divided by blue band reflectance	decibels
Backscatter in VV polarization divided by SWIR band reflectance	decibels
Backscatter in VV polarization divided by NIR band reflectance	decibels
Backscatter in VV polarization divided by NDVI	decibels
Backscatter in VV polarization divided by NDWI	decibels
Backscatter in VV polarization divided by NIRv	decibels
Backscatter in VH polarization divided by red band reflectance	decibels
Backscatter in VH polarization divided by green band reflectance	decibels
Backscatter in VH polarization divided by blue band reflectance	decibels
Backscatter in VH polarization divided by SWIR band reflectance	decibels
Backscatter in VH polarization divided by NIR band reflectance	decibels
Backscatter in VH polarization divided by NDVI	decibels
Backscatter in VH polarization divided by NDWI	decibels
Backscatter in VH polarization divided by NIRv	decibels
Backscatter in VH polarization divided by VV polarization	fraction
Dewpoint temperature 2m	K
Temperature 2m	K
Skin temperature	K
Soil temperature level 1	K
Soil temperature level 2	K
Soil temperature level 3	K
Soil temperature level 4	K
Lake bottom temperature	K
Lake mix layer depth	m
Lake mix layer temperature	K
Lake shape factor	
Lake total layer temperature	K
Snow albedo	
Snow cover	%
Snow density	kg/m^3
Snow depth water equivalent	m of
Snowmelt sum	m of
Temperature of snow layer	K
Volumetric soil water layer 1	1
Volumetric soil water layer 2	1
Volumetric soil water layer 3	1
Volumetric soil water layer 4	1
Forecast albedo	
Surface latent heat flux sum	J/m^2
Surface net solar radiation sum	J/m^2
Surface net thermal radiation sum	J/m^2
Surface sensible heat flux sum	J/m^2
Surface solar radiation downwards sum	J/m^2
Surface thermal radiation downwards sum	J/m^2
Evaporation from bare soil sum	m of
Evaporation from open water surfaces excluding oceans sum	m of
Evaporation from the top of canopy sum	m of
Evaporation from vegetation transpiration sum	m of
Potential evaporation sum	m
Runoff sum	m
Sub surface runoff sum	m
Surface runoff sum	m
Total evaporation sum	m of
U component of wind 10m	m/s
V component of wind 10m	m/s
Surface pressure	Pa
Total precipitation sum	m
Leaf area index high vegetation	1 (area
Leaf area index low vegetation	1 (area

Figure 2.5: The list of 78 variables used as predictor with description and their units.

2.2.3   Underlying variables

We have used 78 variables as predictor for our target variable which is latent fuel moisture content (LFMC) .This were variables downloaded from 4 different data sources as mentioned above. The table above in fig 2.1 shows the lists description about them and their units.

2.2.4   Data Preparation & Preprocessing

1. Collect the data from different repositories mentioned above and merge each variables along the time axis and their geometrical coordinates e.g. latitude and longitude for the whole 5 years for 150 sites of LFMC.

2. Convert all the variables into a fixed resolution ,here it is 0.25°by 0.25°resolution. We have used xESMF to convert the data.

3. Convert all the data into tabular format with .csv extension using Pandas Dataframe.

4. Replace all the NaN value.

5. Load the csv file using Pandas library.

6. Normalize the data of predictor variables using Min Max Scaler. Equation used -

$$z = \frac{(x - xmin)}{(xmax - xmin)}$$

7. Create array of predictor variable and target variable.

8. Split the data into training and testing. Duration for training is Jun, 2015 to Dec, 2019. And rest data (Jan, 2020 to Dec, 2020) is used for testing.

Chapter 3

Methodology

3.1 Latent Heat

3.1.1 Proposed Model

Model Architecture
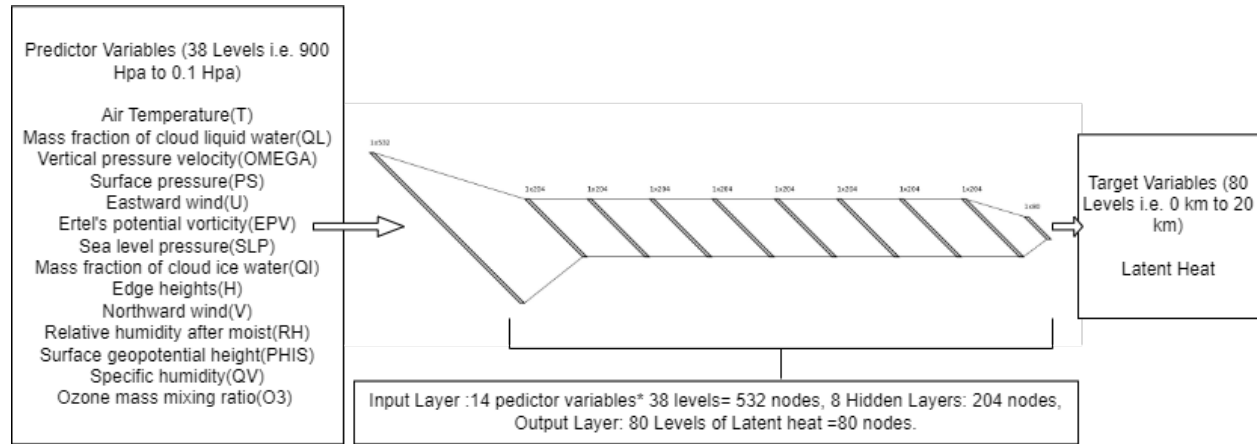
Figure 3.1: Model Architecture for Latent heat prediction

The above mentioned 14 predictor variables are the input variables for the model and latent heat is the target variable. The Deep Learning model will be trained using input and target variable. For the evaluation part the model is evaluated on test data based on its RMSE and R2Score since it is regression task.

Artificial Nueral Network

Artificial Neural Networks (ANNs) is used to predict a continuous output variable based on a set of input variables. ANNs are a form of machine learning model whose structure and function are inspired by the human brain [21].

The following is a summary of how an ANN does a regression task:

Input layer: The input layer of an artificial neural network (ANN) gets the input variables utilised to create predictions [21]. At the input layer, each input variable is represented by a node.

Hidden layers: The hidden layers of an ANN perform computations on the input variables to build a set of features needed for making predictions [21]. The number of hidden layers and the number of nodes within each hidden layer are dictated by the problem's complexity and the quantity of available data.

Output layer: The output layer of an artificial neural network predicts the value of the output variable. In a regression job, the output layer typically contains a single node representing the predicted value [21].

Activation function: Each hidden layer and output layer node is coupled with an activation function [21]. The activation function converts the node's input signal into its output signal [21]. The sigmoid function, the ReLU function, and the hyperbolic tangent function are typical activation functions for regression problems.

Loss function: The loss function calculates the difference between the predicted and actual output values [21]. The objective of ANN training is to reduce the loss function.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (3.1)$$

In this formula, n is the number of data points, y_i is the actual value of the i-th data point, and $\hat{y}_i$ is the predicted value of the i-th data point. The RMSE is calculated by taking the square root of the average of the squared differences between the actual values and the predicted values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (3.2)$$

In this formula, y_i represents the actual values of the dependent variable, $\hat{y}_i$ represents the predicted values of the dependent variable, and $\bar{y}$ represents the mean of the actual values of the dependent variable. The formula calculates the ratio of the variance explained by the model to the total variance in the data, with a maximum value of 1 indicating a perfect fit and a negative value indicating a poor fit.

Training: During training, the ANN's weights and biases are modified to minimise the loss function [21]. Typically, this is achieved using an optimisation approach like stochastic gradient descent.

Prediction: Once the ANN has been trained, it can be used to create predictions by feeding new input data through the network and producing a predicted output value [21].

In general, ANNs can be utilised for regression tasks by receiving input variables, processing them through hidden layers, and generating an output value through an output layer utilising activation functions. The ANN is then trained to reduce the gap between the expected and actual output values [21].

Hyper-parameter Tuning using Optuna

Optimization of hyperparameters is one of the most challenging aspects of machine learning model building. The relationship between model performance and good hyperparameter optimisation is direct. Optuna is an open-source Python module for optimising hyperparameters. Optuna operates in the background to balance sampling and pruning algorithms [2]. Optuna utilises sample techniques such as the Tree-Structured of Parzen Estimator (TPE), Gaussian Processes (GP), and Covariance Matrix Adaptation for relational parameter sampling, which attempts to capitalise on parameter correlation (CMA) [2].Optuna employs a variant of the Asynchronous Successive Halving (ASHA) algorithm to prune search areas. [2].

```python
def create_model(trial):
    # Define the hyperparameters to optimize
    n_layers = trial.suggest_int('n_layers', 1, 10)
    n_units = trial.suggest_int('n_units', 32, 512)
    dropout_rate = trial.suggest_float('dropout_rate', 0.0, 0.5)

    # Build the model architecture
    model = Sequential()
    for i in range(n_layers):
        if i == 0:
            model.add(Dense(n_units, activation='relu', input_shape=(X_train.shape[1],)))
        else:
            model.add(Dense(n_units, activation='relu'))
        model.add(Dropout(dropout_rate))
    model.add(Dense(80, activation='linear'))

    # Define the optimizer and compile the model
    lr = trial.suggest_float('lr', 1e-5, 1e-2, log=True)
    optimizer = tf.keras.optimizers.Adam(learning_rate=lr)
    model.compile(loss='mean_squared_error', optimizer=optimizer)

    return model

# Define the Optuna objective function
def objective(trial):
    # Build and compile the model
    model = create_model(trial)

    # Define the early stopping criterion
    es = EarlyStopping(monitor='val_loss', patience=10, verbose=1)

    # Train the model
    model.fit(X_train, y_train, epochs=100, batch_size=256, verbose=0, validation_split=0.2, callbacks=[es])

    # Predict on the test set and compute the R^2 score
    y_pred = model.predict(X_test)
    rmse = mean_squared_error(y_test, y_pred, squared=False)

    return rmse
```

Figure 3.2: Code snippet of Optuna based hyperparameter tuning of our ANN model

The figure 3.3 shows the parameters that are being hyper-parameterized in our ANN model using Optuna and the remaining parameters are taken as default. The algorithm was trained using Apollo High Processing Computer (HPC) of Indian Institute Of Tropical Meteorology,Pune. The specifications of HPC are :- 38,144 Intel Sandy Bridge processors in 2384 compute nodes,149 TB RAM,Infiniband FDR10 interconnect,6 PB of GPFS-based disk storage,1 PB of tape archive capacity.

3.2   Live Fuel Moisture Content (LFMC)

3.2.1   Proposed Model

Light GBM

The 78 predictor variables are the input variables for the model and LFMC percentage is the target variable. The lightGBM model will be trained using these input and target variable. For the evaluation part the model is evaluated on test data based on its RMSE and R2Score since it.

LightGBM is a framework for gradient boosting that employs tree-based learning algorithms [13]. It is optimised for both training and inference, making it ideal for large-scale machine learning applications [7]. In this response, we will discuss the key features and functionality of LightGBM.

Key Features of LightGBM

1. Lightweight and Fast: LightGBM is designed to be lightweight and fast, making it an excellent option for large-scale machine learning tasks [7]. It is optimised to utilise less memory and train faster than other conventional gradient boosting frameworks.

2. Gradient-Based One-Side Sampling (GOSS): GOSS is a technique employed by Light-GBM to reduce gradient calculation time by sampling the most crucial data points during gradient calculation [13]. This method reduces computation time while maintaining the gradient's precision.

3. Exclusive Feature Bundling (EFB): EFB is a technique used by LightGBM to group together highly correlated features [13]. This reduces the dataset's dimensionality, thereby reducing the memory requirements for training.

4. Histogram-Based Algorithm: LightGBM bins the data into discrete intervals using a histogram-based algorithm [9]. This method reduces the computational expense

associated with calculating the split gain during the tree-building procedure [9].

5. Leaf-Wise Tree Growth: In contrast to other gradient boosting frameworks, Light-GBM grows the tree leaf-wise, as opposed to level-wise [9]. This method improves the accuracy of the model by growing trees based on the most informative leaves.

Working of LightGBM

LightGBM operates by constructing a series of decision trees, with each tree enhancing the accuracy of the model. The process of constructing decision trees can be divided into the following steps:

1. Data Preprocessing: Using a histogram-based algorithm, LightGBM preprocesses the data by binning continuous features into discrete intervals [9]. Based on the split gain, the algorithm selects the best split point for each bin. The resultant binned data are utilised to construct the decision trees [9].

2. Gradient Calculation: LightGBM calculates the gradient for each sample in the training dataset during the training process [9]. It employs the GOSS technique to sample the most crucial data points during gradient calculation, thereby reducing the computation time [9].

3. Training: LightGBM trains its models using a gradient boosting algorithm [9]. It begins by fitting a decision tree to the data and then calculating the model's residual errors. The new decision tree is then fitted to the residual errors and added to the model. This procedure is repeated until the desired number of trees has been reached or the validation loss has ceased to decrease [9].

4. Building the Decision Trees: LightGBM constructs the decision trees in a leaf-by-leaf fashion, selecting the most informative leaves for data partitioning [9]. It employs the EFB technique to group highly correlated features, thereby reducing the dataset's dimensionality and training memory requirements [9].

$$H(T) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (3.3)$$

where T is the set of training examples, n is the number of classes, and p_i is the proportion of examples in T that belong to class i.

$$IG(T, X) = H(T) - \sum_{v \in X} \frac{|T_v|}{|T|} H(T_v) \qquad (3.4)$$

where T is the set of training examples, X is a feature being considered for splitting, $|T_v|$ is the number of examples in T that have the value v for feature X, and $H(T_v)$ is the entropy of the subset T_v.

5. Pruning the Trees: After the decision trees have been constructed, LightGBM prunes them by removing the branches that do not contribute to the model's overall accuracy [9]. This reduces the model's complexity and increases its efficiency.

6. Prediction:LightGBM uses the decision trees to predict the outcome of new data points after the model has been trained [9]. This is accomplished by traversing the decision trees from the root node to the leaf node, utilising the split criteria to determine which branch to take at each node [9].

LightGBM is a gradient boosting framework that makes predictions using a series of decision trees [9]. It is designed to be lightweight and quick, with features such as GOSS and EFB that reduce training's computational cost. By employing a histogram-based algorithm and leaf-wise tree growth, LightGBM is able to achieve high accuracy with less memory and processing time than other gradient boosting frameworks.

Hyper-parameter Tuning using Optuna

```python
def objective(trial):
    train_x, valid_x, train_y, valid_y = train_test_split(X_scaled, y, test_size=0.25,random_state=0)
    dtrain = lgb.Dataset(train_x, label=train_y)

    param = {
        "objective": "regression",
        "metric": "RMSE",
        "verbosity": -1,
        "boosting_type": "gbdt",
        "lambda_l1": trial.suggest_float("lambda_l1", 1e-8, 10.0, log=True),
        "lambda_l2": trial.suggest_float("lambda_l2", 1e-8, 10.0, log=True),
        "num_leaves": trial.suggest_int("num_leaves", 2, 256),
        "feature_fraction": trial.suggest_float("feature_fraction", 0.4, 1.0),
        "bagging_fraction": trial.suggest_float("bagging_fraction", 0.4, 1.0),
        "bagging_freq": trial.suggest_int("bagging_freq", 1, 7),
        "min_child_samples": trial.suggest_int("min_child_samples", 5, 100),
    }

    gbm = lgb.train(param, dtrain)
    preds = gbm.predict(valid_x)
    MSE = sklearn.metrics.mean_squared_error(valid_y, preds,squared=False)
    return MSE


if __name__ == "__main__":
    study = optuna.create_study(direction="minimize")
    study.optimize(objective, n_trials=1000)

    print("Number of finished trials: {}".format(len(study.trials)))

    print("Best trial:")
    trial = study.best_trial

    print("  Value: {}".format(trial.value))

    print("  Params: ")
    for key, value in trial.params.items():
        print("    {}: {}".format(key, value))
```

Figure 3.3: Code snippet of Optuna based hyperparameter tuning of our light GBM model

Parameters	Description	Tuned on
boosting_type	defines the type of algorithm you want to run like , default=gdbt	gbdt: traditional Gradient Boosting Decision Tree, rf: random forest, dart: Dropouts meet Multiple Additive Regression Trees, goss: Gradient-based One-Side Sampling .
lambda	lambda specifies regularization	value ranges between (1e-8, 10.0)
num_leaves	number of leaves in full tree, default: 31	value ranges between (2, 256)
feature_fraction	parameters taken randomly in each iteration for building trees	value ranges between (0.4, 1.0)
bagging_fraction	specifies the fraction of data to be used for each iteration and is generally used to speed up the training and avoid overfitting.	value ranges between (0.4, 1.0)
bagging_freq	specifies the frequency of data to be used for each iteration	value ranges between (1, 7)
min_child_samples	minimum sum of instance weight (Hessian) needed in a child (leaf)	value ranges between (5, 100)

Figure 3.4: List of parameters and their description that are being hyper-parameterized in our light GBM model

The table in 3.4 shows the list of parameters and their description that are being hyper-parameterized in our light GBM model using Optuna. The remaining parameters of light GBM are taken as default. The algorithm was trained using Apollo High Processing Computer (HPC) of Indian Institute Of Tropical Meteorology,Pune. The specifications of HPC are :- 38,144 Intel Sandy Bridge processors in 2384 compute nodes,149 TB RAM,Infiniband FDR10 interconnect,6 PB of GPFS-based disk storage,1 PB of tape archive capacity.

Chapter 4

Summary and Conclusions

4.1   Latent Heat

4.1.1   Results and Conclusions

Results

```
best_model.summary()

Model: "sequential_100"
_____
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_796 (Dense)            (None, 204)               108732

 dropout_696 (Dropout)        (None, 204)               0

 dense_797 (Dense)            (None, 204)               41820

 dropout_697 (Dropout)        (None, 204)               0

 dense_798 (Dense)            (None, 204)               41820

 dropout_698 (Dropout)        (None, 204)               0

 dense_799 (Dense)            (None, 204)               41820

 dropout_699 (Dropout)        (None, 204)               0

 dense_800 (Dense)            (None, 204)               41820

 dropout_700 (Dropout)        (None, 204)               0

 dense_801 (Dense)            (None, 204)               41820

 dropout_701 (Dropout)        (None, 204)               0

 dense_802 (Dense)            (None, 204)               41820

 dropout_702 (Dropout)        (None, 204)               0

 dense_803 (Dense)            (None, 80)                16400

=================================================================
Total params: 376,052
Trainable params: 376,052
Non-trainable params: 0
_____
```

Figure 4.1: Summary of best ANN model suggested by optuna trails

This artificial neural network (ANN) receives an input tensor of shape 14 predictor variables multiplied by 38 levels for a total of 532 input features. The ANN is intended to predict an output tensor with eighty latent heat values. The range of data is eight years. The training data accounts for 56% of the entire data, whereas the test data accounts for 24% and the

validation data accounts for 20%. The data has been divided into these three sets for training, evaluating, and testing the ANN. Even throughout optuna trials, the algorithm has never seen the validation data. This is a regression problem with several outputs. The outcomes are evaluated using evaluation parameters such as root mean squared error (RMSE) and R2-score.
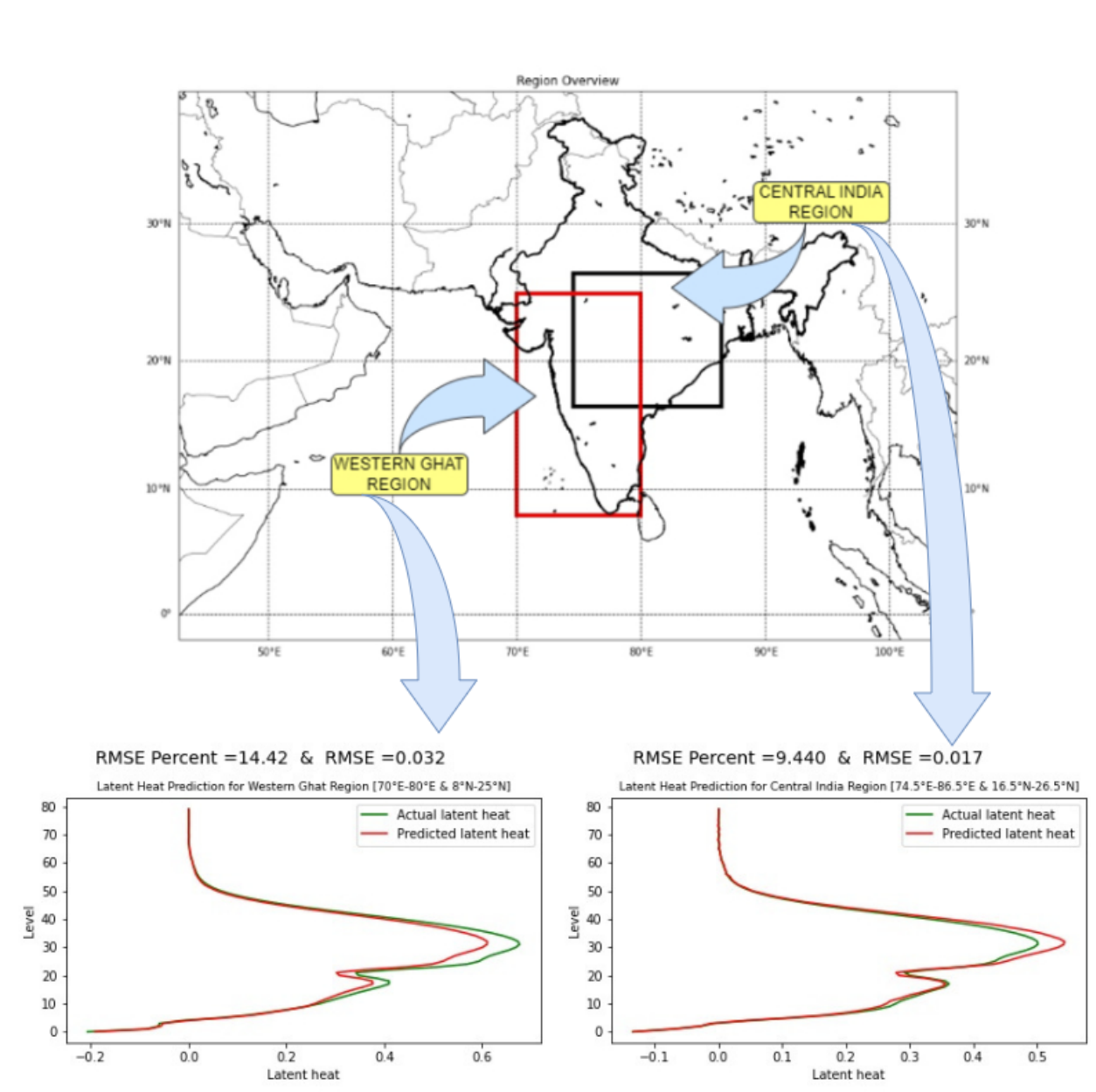

Figure 4.2: Climatology of latent heat predicted by our ANN model for both western ghats and central India regions
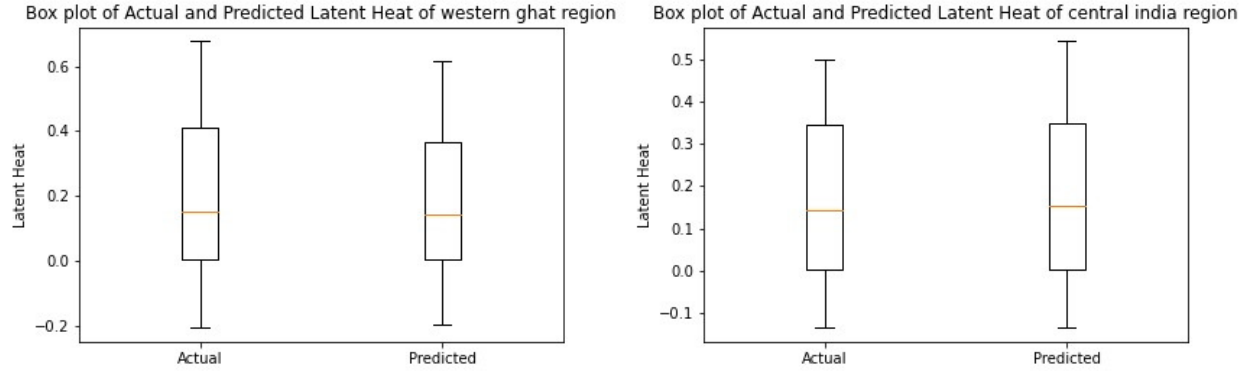
Figure 4.3: Box plot of Actual and Predicted Latent Heat of western ghats and central India region

We have trained and tested our ANN model on two specific regions i.e. central India and western ghat because both the regions share different precipitation systems. Monsoon in central India is mainly driven by convective precipitation system happening due to the formation of monsoon trough in that region. Whereas, monsoon in western ghats are mainly driven by strati-form precipitation system happening due to the formation of off-shore trough. So, prediction of latent heat in these two different precipitation systems by our ANN model plays a significant role. Our model predicted climatology of latent heat in western ghats with RMSE percentage of 9.44% and RMSE value of 0.017. For central India it predicted climatology of latent heat with RMSE percentage of 14.42% and RMSE value of 0.032. Even the box plot of actual and predicted Latent Heat climatology of western ghats and central India region shows significant similarity with less error.

Conclusion

It can be concluded that our Artificial Neural Network (ANN) model provides a more accurate estimation of latent heat at various precipitation systems controlling monsoon dynamics compared to traditional mathematical equations used for parameterization. By using this model-based approach, we can significantly reduce the error in estimating latent heat. Furthermore, the accurate depiction of latent heating in weather and climate models has the potential to improve tropical rainfall forecasts, which is essential for better understanding

the impacts of monsoon dynamics on the climate system. Therefore, our ANN model can be considered a superior alternative for estimating latent heat, which can contribute to improving the accuracy of weather and climate models.

4.2   Live Fuel Moisture Content (LFMC)

4.2.1   Results and Conclusions

Results

Best Parameters	Values
lambda_l1	0.48
lambda_l2	0
num_leaves	54
feature_fraction	0.53
bagging_fraction	0.95
bagging_freq	5
min_child_samples	13

Figure 4.4: The list of hyper-parameters for lightGBM that gives best accuracy for our data-set.

The train and test data divided into 75% and 25% respectively.Duration for training is Jun, 2015 to Dec, 2019. And rest data (Jan, 2020 to Dec, 2020) is used for testing. The train and test data were utilised for running optuna trails and determining the hyper-parameters for the lightGBM model that provide the highest accuracy for our data set.The list of hyper-parameters for lightGBM that gives best accuracy for our data-set is shown in fig 4.4.

Author	Model Used	RMSE
Our research	LightGBM	17.06
Xie, Jiangjian,Tao Qi, Wanjun Hu, Huaguo Huang, Beibei Chen, and Junguo Zhang in 2022 [18]	Ensemble model based on LSTM-TCN	18.88
Krishna Rao,A. Park Williams, Jacqueline Fortin Flefil, and Alexandra G. Konings [12]	LSTM	25

Figure 4.5: Comparison of our lightGBM model results with previous researches.

The fig 4.5 shows us the comparison of our lightGBM model results with previous researches and it clearly indicates that our model out performed.
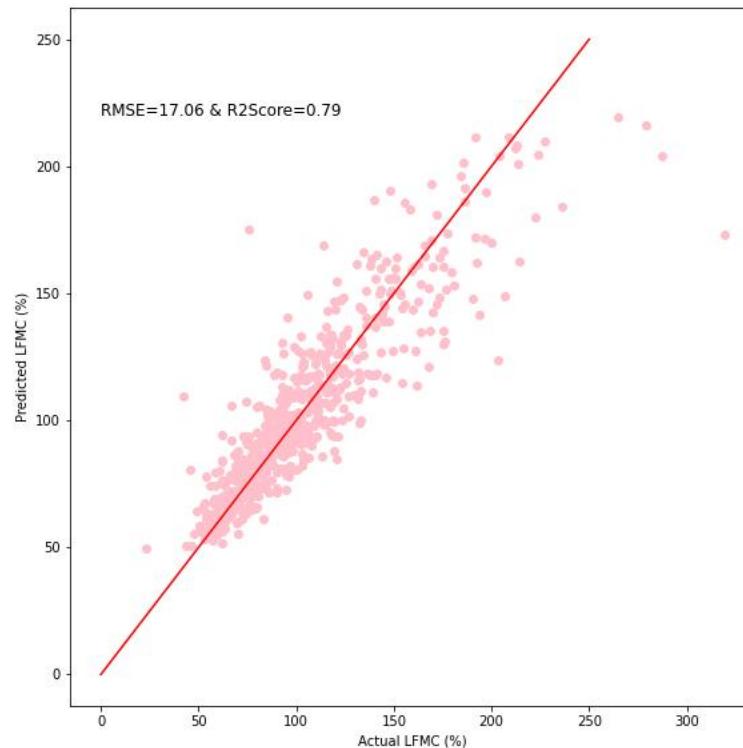

Figure 4.6: Graph between Predicted LFMC by LightGBM and Actual LFMC For Whole Test Period of Year 2020 for 150 sites

Figure 4.6 shows the scatter plot of predicted LFMC and actual LFMC is more confined

on the line of predicted LFMC and actual LFMC being equal. We got an Root Mean Squared Error (RMSE) of 17.06 percent and R2Score of 0.79.
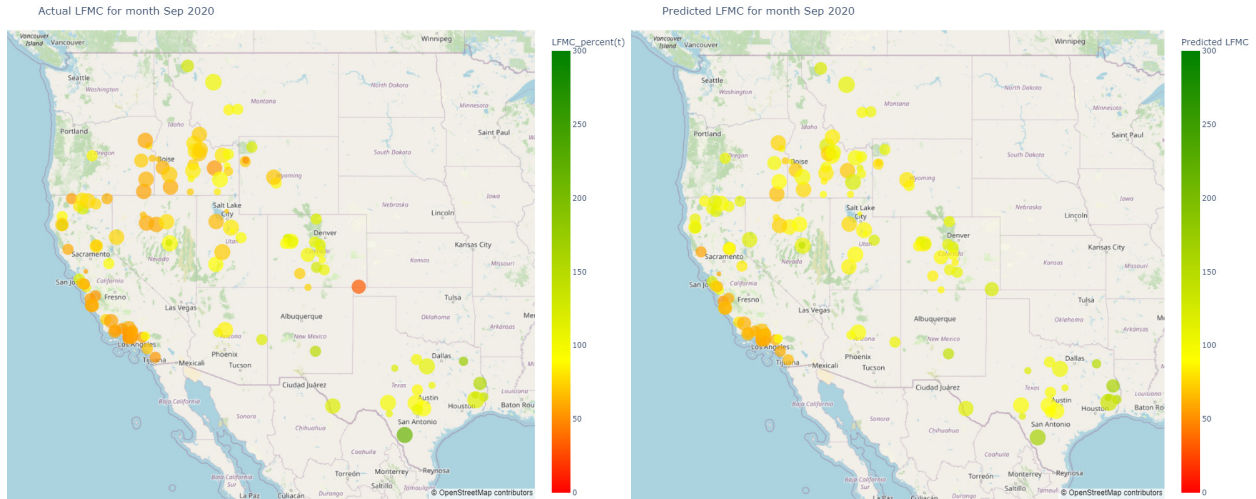

Figure 4.7: Actual and predicted LFMC at 150 sites for Sep 2020

Figure 4.7 shows the Actual and predicted LFMC at 150 sites for Sep 2020. We can easily classify from figure 4.7 the area which are having red spots are extreme fire prone areas with below 60 percent LFMC and Yellow one High fire prone areas with LFMC ranging from 60 to 100 percent. The green spotted areas are having LFMC greater than 100 percent and they are very less fire prone areas [3].

Conclusion

The model developed for mapping fire hazards based on Leaf Fuel Moisture Content (LFMC) ranging from 15 percent to 320 percent with a Root Mean Square Error (RMSE) of 17.06% is considered decent and can be used for mapping the rest of the USA or even India. Although there are not significant differences in predictions between the model and actual data, there is still room for improvement to enhance its accuracy. By accurately mapping fire hazard classification based on LFMC, we can classify areas into Low, High, and Extreme fire hazards and develop an early warning system to prevent losses of human life and property [3]. This research demonstrates the potential for satellite data to provide valuable insights into fire hazards and prevent future fires.

Bibliography

[1] Robert F. Adler and Andrew J. Negri. A satellite infrared technique to estimate tropical convective and stratiform rainfall. *Journal of Applied Meteorology and Climatology*, 27(1):30 – 51, 1988.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019.

[3] Juan P. Arganaraz, Marcos A. Landi, Sandra J. Bravo, Gregorio I. Gavier-Pizarro, Carlos M. Scavuzzo, and Laura M. Bellis. Estimation of live fuel moisture content from MODIS images for fire danger assessment in Southern Gran Chaco. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–11, 2016.

[4] David Brewer, Donna Hayes, Vincent Lyne, Anthea Donovan, Timothy Skewes, David Milton, and Nicole Murphy. Bay of bengal large marine ecosystem boblme-2015-ecology-13. 11 2014.

[5] Ayantika Dey Choudhury and R. Krishnan. Dynamical response of the south asian monsoon trough to latent heating from stratiform and convective precipitation. *Journal of the Atmospheric Sciences*, 68(6):1347 – 1363, 2011.

[6] Jennifer K. Fletcher, Douglas J. Parker, Andrew G. Turner, Arathy Menon, Gill M. Martin, Cathryn E. Birch, Ashis K. Mitra, G. Mrudula, Kieran M. R. Hunt, Christopher M. Taylor, Robert A. Houze, Stella R. Brodzik, and G. S. Bhat. The dynamic and thermodynamic structure of the monsoon over southern india: New observations from the

incompass iop. *Quarterly Journal of the Royal Meteorological Society*, 146(731):2867–2890, 2020.

[7] Ziliang He, Zijian He, Jiahong Wu, and Zhenguo Yang. Feature construction for posts and users combined with lightgbm for social media popularity prediction. pages 2672–2676, 10 2019.

[8] Piyush Jain, Sean C.P. Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Steve Taylor, and Mike D. Flannigan. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4):478–505, 2020.

[9] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc.

[10] T. N. Krishnamurti, Arindam Chakraborty, and A. K. Mishra. Improving multimodel forecasts of the vertical distribution of heating using the trmm profiles. *Journal of Climate*, 23(5):1079–1094, 2010.

[11] AK Mukherjee, MK Rao, and KC Shah. Vortices embedded in'the trough of low pressure off maharashtra-goa coasts-during the month of july. *Mausam*, 29(1):61–65, 1978.

[12] Krishna Rao, A. Park Williams, Jacqueline Fortin Flefil, and Alexandra Georges Konings. Sar-enhanced mapping of live fuel moisture content. *Remote Sensing of Environment*, 245:111797, 2020.

[13] Qiuyi Ren. Research on encrypted text classification based on natural language processing. *Journal of Physics: Conference Series*, 1792:012001, 02 2021.

[14] Manmeet Singh, Vaisakh S B, Nachiketa Acharya, Aditya Grover, Suryachandra A Rao, Bipin Kumar, Zong-Liang Yang, and Dev Niyogi. Short-range forecasts of global precipitation using deep learning-augmented numerical weather prediction, 2022.

[15] David J Stensrud. *Parameterization schemes: keys to understanding numerical weather prediction models.* Cambridge University Press, 2009.

[16] Duane E Waliser and Catherine Gautier. A satellite-derived climatology of the itcz. *Journal of climate*, 6(11):2162–2174, 1993.

[17] Long Wang, Xingwen Quan, Binbin He, Marta Yebra, Minfeng Xing, and Xiangzhuo Liu. Assessment of the dual polarimetric sentinel-1a data for forest fuel moisture content estimation. *Remote Sensing*, 11(13), 2019.

[18] Jiangjian Xie, Tao Qi, Wanjun Hu, Huaguo Huang, Beibei Chen, and Junguo Zhang. Retrieval of live fuel moisture content based on multi-source remote sensing data and ensemble deep learning model. *Remote Sensing*, 14(17), 2022.

[19] Guoli Zhang, Ming Wang, and Kai Liu. Forest fire susceptibility modeling using a convolutional neural network for yunnan province of china. *International Journal of Disaster Risk Science*, 10:386 – 403, 2019.

[20] Manuel D. Zuluaga, Carlos D. Hoyos, and Peter J. Webster. Spatial and temporal distribution of latent heating in the south asian monsoon region. *Journal of Climate*, 23(8):2010 – 2029, 2010.

[21] Jure Zupan. Introduction to artificial neural network (ann) methods: what they are and how to use them. *Acta Chimica Slovenica*, 41:327–327, 1994.