Development Of Atmospheric Formaldehyde Digital Twins For Air Pollution Applications

A THESIS SUBMITTED TO
DEFENCE INSTITUTE OF ADVANCED TECHNOLOGY, PUNE
FOR THE PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE OF
MASTER OF TECHNOLOGY
IN
DATA SCIENCE

BY
NEETIRAJ MALVIYA
(Registration No. **21-27-03**)

UNDER THE SUPERVISION OF

Dr. Bharath Ramkrishna
Assistant Professor
DIAT, Pune
(Internal Supervisor)

Dr. Manmeet Singh
Scientist 'D'
IITM, Pune
(External Supervisor)

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

APR 2023

DEDICATED

To each and everyone who supported and motivated me.

DEFENCE INSTITUTE OF ADVANCED TECHNOLOGY(DIAT)

(DEEMED TO BE UNIVERSITY)

SCHOOL OF COMPUTER ENGINEERING & MATHEMATICAL SCIENCES

M.TECH - DATA SCIENCES

2021-23

CERTIFICATE

This is to certify that the Thesis entitled “**Development Of Atmospheric Formaldehyde Digital Twins For Air Pollution Applications** ” 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. Neetiraj Malviya** under our supervision. The contents of this thesis have not been submitted elsewhere for the award of any degree.

Dr. Bharath Ramkrishna
Assistant Professor
SoCE&MS
DIAT, Pune
(Internal Supervisor)

Dr. Manmeet Singh
Scientist ”D”
IITM, Pune
(External Supervisor)

COUNTERSIGNED

Dr. Manisha J. Nene
Director
School of Computer Engineering &
Mathematical Sciences
DIAT, Pune

.

DECLARATION

This is to certify that the work presented in the Thesis entitled **“Development Of Atmospheric Formaldehyde Digital Twins For Air Pollution Applications”**, is a bonafide work done by me under the supervision of **Dr. Manmeet Singh and Dr. Bharath Ramkrishna** and has not been submitted elsewhere for the award of any degree.

Date:____________

Place:____________

Neetiraj Malviya
Roll No. 21-27-03.
School of Computer Engineering and
Mathematical Sciences
Defence Institute of Advanced Technology
Girinagar, Pune

COUNTERSIGNED

Dr. Bharath Ramkrishna
Internal Supervisor
Assistant Professor
Defence Institute of Advanced
Technology(DIAT)
Girinagar, Pune

Dr. Manmeet Singh
External Supervisor
Scientist ’D’
Indian Institute of Tropical
Meteorology(IITM), Pune

Approval Sheet

CERTIFICATE OF COURSE WORK

This is to certify that **Neetiraj Malviya (Reg No:21-27-03)** has successfully completed the necessary course work as required for the **M.Tech in Data Science**. The details of the coursework are as follows.

Sr. No.	Course Code	Course Name	Credits
1	AM603D	Computer-Oriented Optimization Methods	4
2	AM604D	Statistical Computing for Data Science	4
3	AM606D	Scientific Computing	4
4	AM607D	Data Structures and Algorithms with C	4
5	CE615A	Intelligent Algorithms	4
6	CE696A	Artificial Intelligence and DSS	4
7	AM623D	Machine Learning	4
8	AM624D	Data Science : Tools & Techniques	4
9	AM625D	Image and Video Analytics	4
10	CE631	Deep Learning	4
11	CE632	Computer Vision	4
12	CE694	Big Data Analysis & Algorithm	4
13	AM651D	M.Tech Dissertation Phase - I	14
14	AM652D	M.Tech Dissertation Phase - II	14

Place: DIAT, PUNE

Controller of Examinations

Date: / / 2023

DECLARATION FOR PLAGIARISM CHECK

This is to certify that M.Tech thesis entitled **"Development Of Atmospheric Formaldehyde Digital Twins For Air Pollution Applications"** submitted by **Mr. Neetiraj Malviya** bearing Registration No **21-27-03** under the supervision of Dr. Bharath Ramkrishna in the Department of **School of Computer Engineering & Mathematical Sciences** is the original research work done by the candidate.

We have read the provision of **DIAT (DU) Plagiarism Policy**, and it is certified that all the conditions prescribed in the policy above are complied with in respect of the abovementioned M Tech thesis.

The thesis has been checked for plagiarism, and the report is submitted along with the thesis for further processing.

- Originality content (including the contents from his own publications):

81 %

- Similarity Reproduction of the content from other sources:

19 %

We are aware that any issue related to plagiarism in future will have to be addressed by the candidate and the Supervisor (s) concerned.

Name of Candidate: **Neetiraj Malviya**
Signature:
Date: / / 2023

Name of Supervisor: **Dr. Bharath Ramkrishna**
Signature:
Date: / / 2023

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to **Dr. Bharath Ramkrishna**, School of Computer Engineering & Mathematical Sciences, for motivating me all through this work. I would like to thank for his timely guidance and support.

I would like to express my sincere gratitude to **Dr. Manmeet Singh**, Scientist 'D', Indian Institute of Tropical Meteorology, Pune for providing me the opportunity to take up this project and more importantly, his precious guidance and motivation, which helped me to conquer the difficulties in research and development of this project.

I also like to express my sincere gratitude to **Mr. Vaisakh S B**, Project scientist-l, Indian Institute of Tropical Meteorology, Pune for his valuable support and encouragement.

Above all, I offer my sincere thanks and gratitude to the merciful God Almighty for the providence and grace, He has given through this attempt.

Neetiraj Malviya
M.Tech in Data Science

ABSTRACT

One of the most significant trace gases in the atmosphere is formaldehyde (HCHO), as it is a contaminant that causes illnesses such as cancer problems. It also functions as a precursor to tropospheric ozone. It degrades human health and harms agriculture. Chemistry research on HCHO and long-term regarding human health, food security, air toxicity and environmental protection, satellite data monitoring is crucial. Simulating atmospheric formaldehyde is difficult using dynamic atmospheric chemistry models and it overestimate in comparison to satellite observations and reanalysis by up to two times. And also predicted HCHO distribution does not match satellite data. In this work, we use XGBoost model which is based on 'Embeddings for Numerical Features in Tabular Deep Learning' to quick and accurate atmospheric HCHO simulation. Almost four year of data (2018 - 2022) from the GEE and ERA5 repositories are used to target Predictions of HCHO based on SENTINEL-5P satellites. In total, seven meteorology variables are used. For model evaluation, we used the Pearson correlation coefficient and MSE. Without the use of chemical equations, we find that deep learning performs better than dynamical model simulations that simulate complex atmospheric chemistry. We decide to test our implementation in India because of the lack of in situ formaldehyde measurements and the requirement for higher-quality data across the area. In situ measurements require that the instrumentation be located directly at the point of interest and in contact with the subject of interest.

Contents

NOMENCLATURE

$HCHO$: Formaldehyde

w_j : Weight

J : Cost function

α : Learning rate

γ : L1-regularization

λ : L2-regularization

CH_4 : Methane

C_5H_8 : Isoprenes

O_3 : Ozone

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Introduction

In atmospheric chemistry, trace gases are extremely important since some of them have the potential to be pollutants (such as NO_2, SO_2, HCHO etc.). They influence the entire biosphere, impair agricultural productivity, and present health risks to people. Other trace gases alter the atmospheric radiation budget and contribute to climate change by acting as greenhouse gases (GHGs) and increasing global warming. This aldehyde is ample in amount but the problem is that it is a pollutant resulting in cancer problems and other health issues for mankind. The crucial function of HCHO plays a key role in the chemistry of the atmosphere by serving as a precursor to tropospheric ozone O_3. Unlike stratosphere O_3 shields all Earth's life by absorbing harmful ultraviolet (UV) rays and preventing their occurrence the tropospheric O_3 is a pollutant that should be prevented from entering Earth's atmosphere. Ozone in the troposphere is to blame resulting in overall losses for wheat, soybean, rice, and maize of 4–15%, 6–16%, 3–4%, and 2.2–5.5%, respectively, with worldwide agricultural damage estimates ranging from 11 to 26 billion. Respiratory and cardiovascular disease are connected to short-term O3 exposure ultimately resulting in early death. Consequently, it is crucial to research atmospheric HCHO because of the perspectives of atmospheric chemistry, air pollution, food security, and public health.

The major sources of tropospheric HCHO are higher volatile organic compounds (VOCs) and the oxidation of methane (CH_4). Precursors of formaldehyde (CH_4 and higher VOCs) can be pyrogenic, biogenic or both. Vegetation, primarily forests release a variety of VOCs at higher temperatures (30 degree C), such as isoprenes (C_5H_8) and monoterpenes. Different techniques are used by dynamic atmospheric chemistry models to determine VOC emission rates with regard to the temperature of the atmosphere. These VOCs oxidise in the presence of solar radiation to produce HCHO, which leads to the creation of tropospheric O_3. The other VOCs are in charge of the geographical and temporal variability of atmospheric HCHO, whereas the oxidation of C_5H_8 maintains the background concentration of HCHO. C_5H_8 is the biogenic VOC that contributes the most to the concentration of HCHO.

C_5H_8 is the biogenic VOC that contributes the most to the concentration of HCHO. Various earlier estimations of C_5H_8 emission from HCHO satellite observations have been published in research. HCHO is a product of the oxidation of VOCs in the environment and it has been used as an indicator for biogenic and pyrogenic VOC emission.

1.2 Sources of Formaldehyde

Formaldehyde (HCHO) can be produced from various sources, including biogenic, anthropogenic, and pyrogenic precursors. Here's a brief overview of each type:

1. Biogenic precursors: These are natural organic compounds that are emitted by plants and other living organisms. Some examples of biogenic precursors of HCHO include isoprene and terpenes, which are emitted by trees and other vegetation.

2. Anthropogenic precursors: These are organic compounds that are produced by human activities, such as transportation, industrial processes, and energy production. Some examples of anthropogenic precursors of HCHO include volatile organic compounds (VOCs), which are emitted by cars, factories, and other sources.

3. Pyrogenic precursors: These are organic compounds that are produced by combustion processes, such as wildfires, biomass burning, and fossil fuel combustion. Some examples of pyrogenic precursors of HCHO include acetaldehyde and other aldehydes, which are produced during the combustion of organic matter.

It's important to note that the sources and levels of HCHO precursors can vary depending on the location and time of year. For example, biogenic emissions may dominate in forested areas, while anthropogenic emissions may be more significant in urban areas. Understanding the sources and behaviour of HCHO precursors is crucial for assessing the health and environmental impacts of formaldehyde and developing effective mitigation strategies.

1.3 Effects of Formaldehyde

Formaldehyde (HCHO) exposure or emission can have various health effects and acts as a precursor for tropospheric ozone O3, crop loss etc.
Formaldehyde (HCHO) exposure can have various health effects, depending on the level and duration of exposure. Here are some potential effects of formaldehyde exposure:

1. Respiratory irritation: Formaldehyde is a strong irritant to the eyes, nose, and throat. Exposure to high levels of formaldehyde can cause burning, itching, and redness of the eyes, as well as sore throat, runny nose, and coughing. These symptoms can be particularly severe for individuals who are sensitive to formaldehyde or who have pre-existing respiratory conditions such as asthma.

2. Allergic reactions: Some individuals may develop an allergy to formaldehyde after repeated exposure. This can lead to symptoms such as skin rashes, itching, and hives, as well as difficulty breathing and wheezing.

3. Cancer: Formaldehyde has been classified as a known human carcinogen by the International Agency for Research on Cancer (IARC). Studies have linked formaldehyde exposure to an increased risk of several types of cancer, including nasopharyngeal cancer and leukaemia. The risk of cancer is highest for individuals who are exposed to formaldehyde for long periods, such as workers in industries where formaldehyde is used.

4. Neurological effects: Formaldehyde exposure has been linked to cognitive impairment and memory loss, particularly for individuals who are exposed to high levels of the chemical over long periods.

Studies have also suggested that formaldehyde exposure may be associated with an increased risk of Parkinson's disease.

5. Reproductive effects: Formaldehyde exposure may affect reproductive function and fetal development, potentially leading to birth defects and other adverse outcomes. Animal studies have suggested that formaldehyde exposure may be linked to decreased fertility and reduced fetal growth.

6. Environmental impacts: Formaldehyde is a highly reactive gas that can contribute to air pollution, particularly in urban areas with high levels of traffic and industrial activity. Formaldehyde can also damage ecosystems, particularly aquatic ecosystems, by disrupting the natural balance of bacteria and other microorganisms.

Formaldehyde exposure can potentially have negative effects on crops as well. Here are some potential impacts of formaldehyde on crops:

1. Reduced growth and yield: Formaldehyde exposure can stunt plant growth and reduce crop yields, particularly for sensitive plant species. This is because formaldehyde can interfere with photosynthesis and other essential plant processes.

2. Reduced quality: Formaldehyde exposure can also reduce the quality of crops by damaging the structure and function of plant cells. This can lead to reduced nutritional value and lower market value for crops.

3. Phytotoxicity: Formaldehyde can be toxic to plants at high concentrations, leading to symptoms such as leaf yellowing, necrosis (death of plant tissue), and reduced root growth.

4. Translocation to edible parts: Formaldehyde can be absorbed by plants and translocated to the edible parts, such as fruits and vegetables. This can potentially lead to health risks for consumers who consume formaldehyde-contaminated crops.

5. Environmental impacts: Formaldehyde emissions from agricultural practices, such as burning of crop residues, can contribute to air pollution and other environmental impacts.

It's important to note that the effects of formaldehyde on crops can vary depending on the level and duration of exposure, as well as other environmental factors such as temperature and humidity. To reduce the risks of formaldehyde exposure for crops, it's important to minimize emissions from industrial and agricultural sources, as well as implement measures to protect crops from exposure to formaldehyde and other air pollutants.

1.4 Research done in the past on atmospheric formaldehyde

Long-term research using satellites has documented the spatial variance and temporal development of HCHO. Smedt et al. have utilised SCIAMACHY satellite measurements, it was found that there was an annual rise of 1.6% across India. Later research calculated a rise in HCHO vertical column density (VCD) of 1.51±44% per year over India[9]. based on observations Studies have shown that human emissions dominate HCHO concentrations in Indian cities from automobile emissions, whereas HCHO concentrations were higher in rural and high-altitude areas predominated by biogenic emissions. GEOS-Chem model use HCHO seasonal cycle over India has been described by modelling to explain SENTINEL satellite data of HCHO to surface temperature correlation.

They came to the conclusion that mainly biogenic emissions cause HCHO concentrations in India. They concluded that biogenic VOC was found when GEOS-Chem simulation was compared to satellite (OMI, GOME-2a) HCHO. Emissions are overestimated by $30-60\%$. Studies from the past, however, have also revealed a serious mismatch between dynamic HCHO simulation models and satellite observations. The ECHAM5-HAMMOZ model overestimates, according to Mahajan et. al. We have shown a comparison between ECHAM5-HAMMOZ in Figure 1.1 SCIAMACHY satellite observations and the model. Compared to satellite measurements, the simulated HCHO is two times bigger.

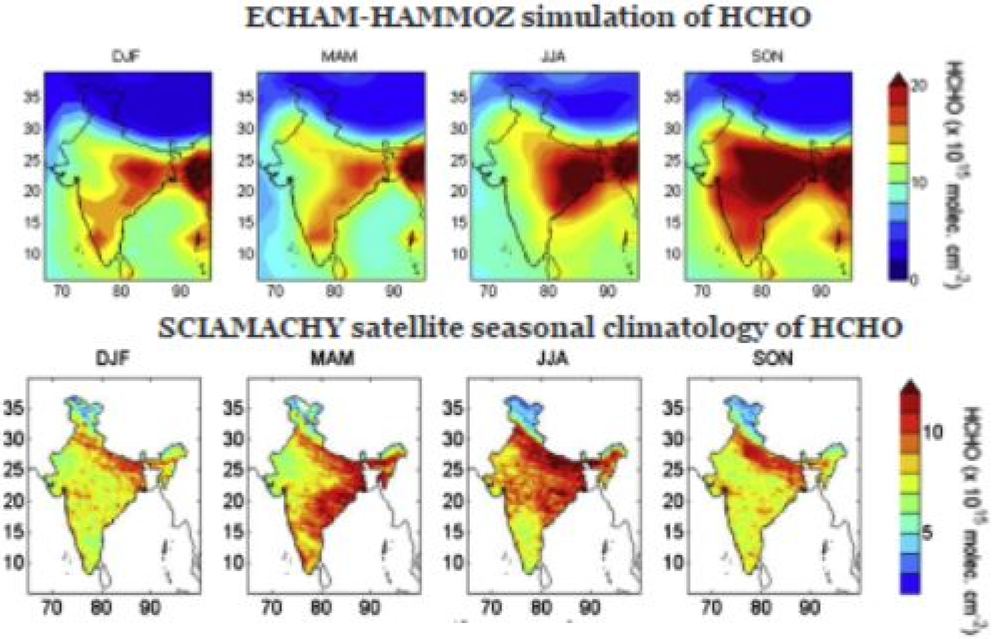

ECHAM-HAMMOZ simulation of HCHO
DJF
MAM
JJA
SON
35
30
25
20
15
10
70
80
90
20
10
0
HCHO (x 10^15 molec. cm^-2)
SCIAMACHY satellite seasonal climatology of HCHO
DJF
MAM
JJA
SON
35
30
25
20
15
10
70
80
90
10
5
HCHO (x 10^15 molec. cm^-2)

Figure 1.1: The dynamical model ECHAM-HAMMOZ simulation of HCHO and SCIAMACHY satellite. The HCHO ratio over India.[2]

Moreover, absolute values and geographical distribution can be shown to differ significantly. Modeled (WRF-Chem) HCHO overestimation, as opposed to MACC reanalysis, is up to 40%.

1.5 Previous work

1.5.1 Deep learning

Deep learning is the subfield of machine learning that deals with the artificial neural network (Figure 1.2) having multiple layers of neurons and being able to perform complex tasks such as image classification, NLP, and speech recognition. it consists of interconnected multiple layers of neurons where each neuron is getting some input from other neurons which are connected to it from the previous layer, it applies some mathematical operation on it and then sends it to another layer of the neuron. By stacking such layers together, the deep neural network learns more complex relationships of the data, finds patterns, and makes predictions with more accuracy.

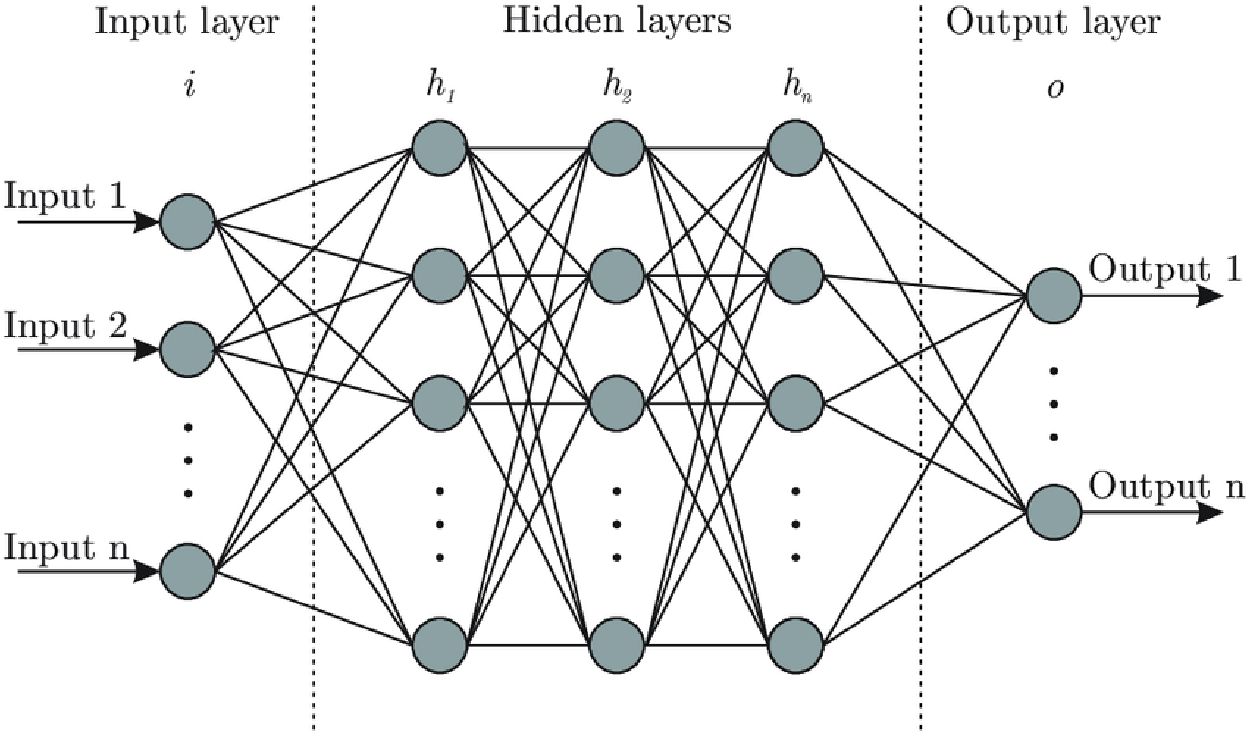


Figure 1.2: ANN-Architecture

1.5.2 Convolution Neural Network(CNN)

CNN(Figure 1.3) is one of the types of deep learning networks which is specially used for image analysis tasks. This network possesses the ability to recognize and classify features of the images. Their applications are widespread and used in domains such as medical image and video analysis, image and video recognition, and image classification.

Hyperparameters: Kernel size: is the height and width of the kernel. The number of kernels used for feature extraction.

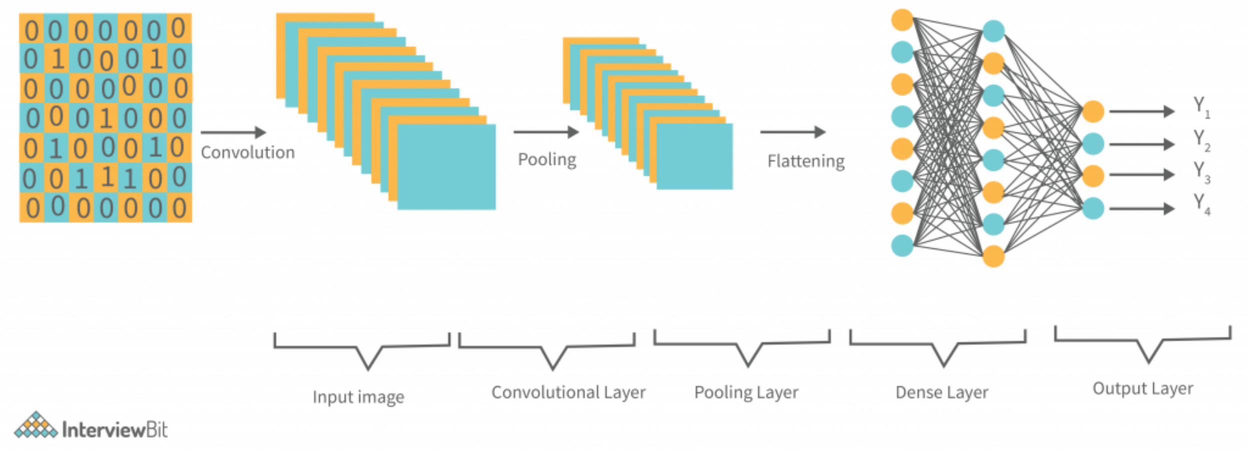


Figure 1.3: CNN-Architecture

Stride: The number of values for the kernel to move in each direction.

Padding: The output image of convolution is typically smaller than the input image. Padding describes the insertion of additional values to the image's border to modify the output's size. The padding options "Same" and "Valid" are two popular ones. While "Valid" does not employ any padding, "Same" makes sure that the output's size is the same as the input's. Usually, zero padding (values added at the boundary are all zeros) is done assuming no prior knowledge about the outside of the image. Convolution between an nxn image and k x k filter, with padding, p (zeros padded along a dimension at either end) and stride, s results in an image of size.

$$\text{size} = \left[\frac{n+2p-k}{s} + 1\right] \times \left[\frac{n+2p-k}{s} + 1\right]$$

Convolutional Layer

From Figure 1.3 The Convolutional Layer plays a crucial role in performing convolution operations. The filter moves across the image in a horizontal and vertical direction based on the stride given. Typically, the size of the kernel is smaller than the size of the image, but it has more depth than the image.

Pooling Layer (POOL)

The Pooling Layer is another critical component of a CNN(Figure 1.3) and is used to reduce the dimensionality of the convolved feature maps. This reduction in dimensionality leads to a decrease in the amount of computation required to process the data. Two common types of pooling techniques are used in CNN: maximum pooling and average pooling.

Fully Connected Layer (FC)

In a fully connected layer(Figure 1.3), the flattened input is connected to each and every neuron, through which mathematical expressions are performed. Subsequently, this flattened vector again passed through the fully connected layers of neurons. At this stage, the process of classification is started. It has been observed that two FC layers perform much better than a single FC layer.

1.5.3 Activation Functions

Activation is one of the most crucial things in machine learning or deep learning models. The activation function model learns the complexity as well as the relationship between the variables by deciding which information should be transmitted and which should not. Activation functions are responsible for introducing non-linearity to the network.

Sigmoid:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Tanh:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

ReLU :

$$\text{ReLU}(x) = \max(0, x)$$

1.5.4 Loss function

This is a function that is used to find the difference between the predictions given by the model and the actual output. The model is trained to find values of the weights which are responsible for minimizing the loss. The cost(error) function, J, which is generally the average of the loss function over all the training examples is then calculated. There are two steps associated with the computations involved in a neural network:

1. Forward Propagation: In this step, the output of the neural network is calculated.

The computations start from left to right(input layer to output layer).

2. The error at the output layer is back propagated to the hidden layers and consequently, the weights are updated by the optimizer. Here, the computations start from right to left(chain rule).

1.5.5 Optimizer

After the loss is calculated, weights need to be updated in order to lower the loss and move towards the minima with each iteration. This updation of weights is done by a specific algorithm called the optimizer. These are some of the optimizers:

1. Gradient descent: In the Gradient descent algorithm, With each iteration, the weights are updated as follows:

$$w_j \to w_j - \alpha \times \frac{\partial j}{\partial w_j}$$

where w_j is a weight
J is the cost function
α is called the learning rate.

2. Momentum: In Momentum optimization, the update of the parameters at each iteration is not only based on the gradient of the loss function but also on the momentum vector, which accumulates the previous gradients. This helps the optimizer to move faster in the direction of the minimum of the loss function and overcome the oscillations and noise that can arise from the stochastic nature of the gradients.

3. Adam(Adaptive moment estimation): Adam uses an adaptive learning rate that changes over time based on the average of the squared gradient values. It also includes momentum-like exponential decay of the gradient over time. The combination of these two techniques helps the algorithm to converge quickly and efficiently while avoiding oscillations and overshooting of the optimal solution.

Updating the weights in the steepest direction is how gradient descent operates. It is preferable if the algorithm reaches the minima more quickly. A momentum gradient descent works by computing an exponentially weighted average of the gradients (controlled by the parameter) and use that to update the weights in their place. When the gradient's sign changes, the velocity component, mj , dampens the weight update, whereas when the gradient is moving in the same direction as the velocity, it accelerates the weight update. As a result, the convergence occurs more quickly than with simple Gradient descent.

The most frequently used value is 0.9. Gradient descent can also be accelerated using RMSprop. It maintains the exponentially weighted average of the gradient's square (again, determined by). The velocity parameter, in this case, is u_j. In contrast to momentum, using the square of the gradient aids in dampening out the directions where there are significant oscillations.

This also makes it possible to use a higher learning rate, which speeds up the process. Adam optimizer is created by combining Momentum and RMSprop. Adam is a popular learning algorithm that has been shown to function well with a wide range of neural networks with a wide diversity of topologies. The bias exhibited in the moving average, primarily at the beginning of the sequence, is taken into account by the calculation of m_j and u_j. A little constant called ensures the stability of numbers.

1.5.6 Proposed Model

The SRCNN architecture is a fully-convolutional deep learning architecture. It learns to map the low-resolution images to the high-resolution ones with little pre or post-processing.

Patch extraction and representation: This is the process of extracting overlapping patches from low-resolution images. We will feed these as features to the SRCNN model and the corresponding high-resolution patches will act as the ground truth labels.

Non-linear mapping: The model will learn the mapping in the training phase. That is, it will learn to map the low-resolution patches to the high-resolution patches.

Reconstruction: This is the final operation where the model reconstructs (or outputs) the possible high-resolution image.

1.5.7 Model Architecture

Here we are using the model for the climate data which is just like of image. The latitude and longitude can be considered as rows and columns of an image and the value of each grid is like the intensity of the image.

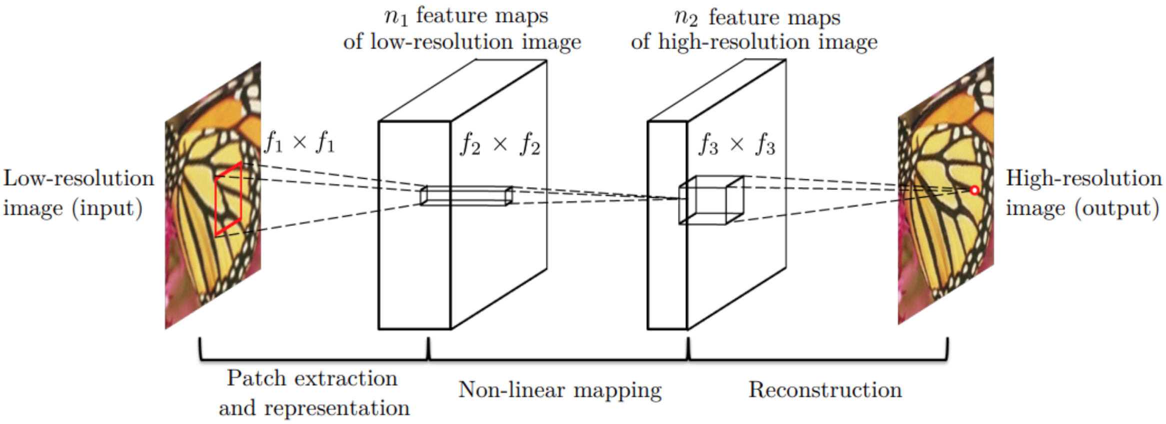


Figure 1.4: Model Architecture

This is how the climate data is similar to an image. Based on this assumption we are using the SRCNN model for the HCHO modeling.

1.5.8 Description of Data Variables

We have noted that plants as a natural reaction to heat stress VOCs are released at higher temperatures. VOCs are oxidised to produce HCHO, which is favoured at greater temperatures as well. As a result, we decide to include atmospheric temperature as one of the model's inputs. We considered solar radiation as another meteorological input.

We also took into account the leaf area index (LAI) with both high and low vegetation since, as a recent study reveals, biogenic emissions are a significant source of VOC emissions. CH_4 and c_5h_8 are regarded as chemical inputs. The main biogenic VOC responsible for the generation of atmospheric HCHO is isoprene (c_5h_8). Since all other higher VOCs contribute to the production of HCHO, we thought of them as additional inputs for the model.

Coming from the sun radiation, total precipitation, the atmospheric temperature at 2 meters above the surface, and leaf area index (high and low vegetation) are all daily. The study used total column CH_4, c_5h_8, and other VOCs from the CAMS reanalysis with 0.75 x 0.75-degree spatial resolution and three hourly temporal resolutions at 03:00, 06:00, and 09:00 UTC. Hourly files are used to produce a daily average dataset for each variable. To build a dataset with a uniform spatial resolution, the entire dataset is regridded.

The target variables are the SENTINEL-5P HCHO observations. The equatorial overpass time of the SENTINEL-5P satellite is 13:00 hours. Hence the HCHO signal is observed by the SENTINEL satellite at about midday. The concentration of HCHO's precursors will decide the HCHO concentration at noon.

As we have already mentioned that the lifespan of the precursor gases and the HCHO extends from a few minutes to a few hours, the HCHO concentration will also rely on the precursor concentration from a few hours ago. Because of this, we have used daily averaged input data.

All input variables are expected to affect HCHO concentration for a few hours prior to the observation period, according to our assumption.

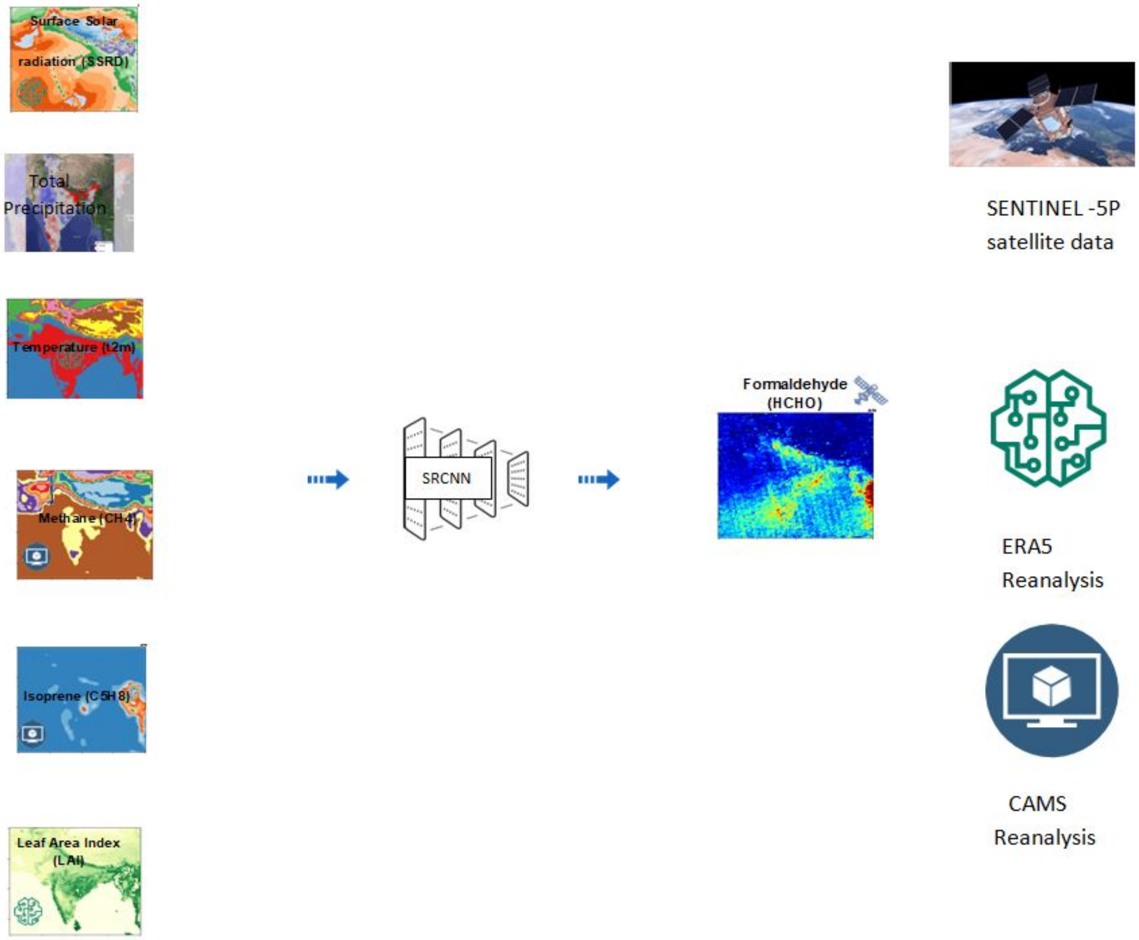


Figure 1.5: Schematic diagram of the present work. Using temperature, solar radiation, total precipitation, methane, isoprene and higher VOC concentrations input for the model.

1.5.9 IMPLEMENTATION

Data of HCHO from SENTINEL-5P is preprocessed data. We have removed the negative value at each grid point. We combine the ERA5 reanalysis, CAMS reanalysis, and HCHO data from the SENTINEL satellite to daily temporal resolution for the prediction of formaldehyde (HCHO) from diverse meteorological, chemical, and vegetative species.

Use the following equation to determine the total quantity of VOCs from CAMS reanalysis:

$$VOC = CH_3COCH_3 + C_2H_6 + C_2H_5OH + C_2H_4 + CH_3OH + C_3H_8$$

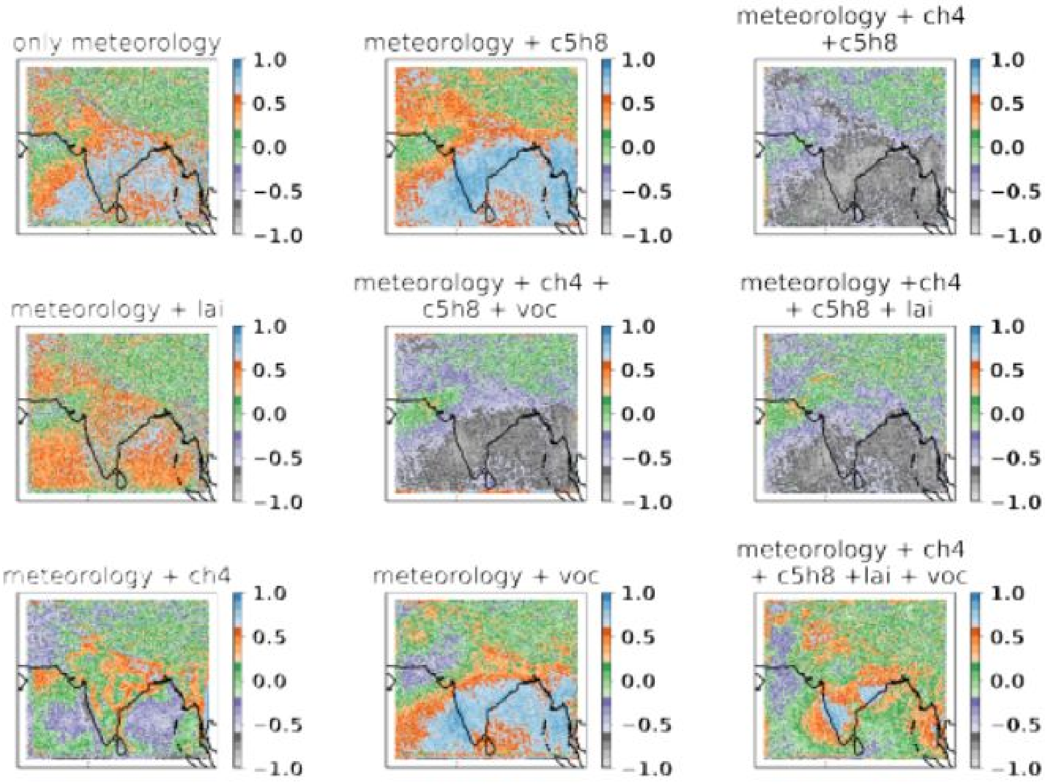


Figure 1.6: Pearson correlation coefficient between predicted and actualformaldehyde.

We extracted all the data from different sources such as ERA5, CAMS and SENTINEL-5P for latitude 5-35N, 70-90E. Then regrid all the datasets to 0.1 degree spatial resolutio by using xESMF. For getting a good idea of the dependencies of all the precursors we try different combinations of variables to target HCHO. We combine solar radiation, surface air temperature, and total precipitation as meteorology in our combination. The data we are using is from 2018 to 2022 so for training we have used two-year data from 2018 to 2020 for validation we have used six-month data from 2021 and 2021 to 2022 for testing.

We use a three-layer super-resolution convolutional neural network (SRCNN) to map from the inputs to the output. The kernel dimensions of the three layers are 9 x 9, 1 x 1, and 5 x 5, respectively. We use the Adam optimizer with a mean squared error loss function and a learning rate of 0.001. The input and target sizes of our target region sum to 284 x 223 pixels. To prevent overfitting, we used early stopping and preserved the best model after a validation loss during training. We have an NVIDIA A100 GPU. We examined the correlation between various combinations after receiving the forecast.

1.6 LITERATURE REVIEW

Guan et al.[1] this paper presents a method to generate the global surface distribution of formaldehyde (HCHO) using satellite observations and a neural network technique. The authors utilized data from the TROPOspheric Monitoring Instrument (TROPOMI) to train a deep neural network to estimate HCHO concentrations. The trained network was then used to generate a global map of surface HCHO distribution. The results were validated using ground-based measurements and compared with other HCHO products. The authors found that their method produced accurate and consistent results, with a high correlation coefficient between the estimated and observed HCHO concentrations. The paper concludes that this approach has the potential to provide valuable insights into air pollution and its impact on human health and the environment.

Mahajan et al. [2] this paper examine the inter-annual variations in satellite observations of nitrogen dioxide (NO2) and formaldehyde (HCHO) over India. The authors utilized data from the Ozone Monitoring Instrument (OMI) to investigate the changes in NO2 and HCHO concentrations over the period of 2005 to 2013. The study found a decreasing trend in NO2 concentrations over the period, which was attributed to the implementation of air pollution control measures in major Indian cities. In contrast, the study found an increasing trend in HCHO concentrations, which may be attributed to changes in the atmospheric chemistry and emissions from biomass burning. The paper concludes that satellite observations provide valuable insights into the changes in air pollution and can be used to assess the effectiveness of air pollution control measures.

Inness et al. [3] from this paper, we can learn about the CAMS (Copernicus Atmosphere Monitoring Service) reanalysis of atmospheric composition, which is a long-term and consistent record of atmospheric constituents such as ozone, nitrogen dioxide, carbon monoxide, and aerosols. The CAMS reanalysis is based on a combination of satellite observations, ground-based measurements, and atmospheric models. The study found that the CAMS reanalysis provides an accurate and consistent record of atmospheric composition over the past two decades, with particular improvements in the representation of aerosols and their impacts on air quality and climate. This paper highlights the importance of long-term and consistent records of atmospheric composition for studying the changes in air quality and their impact on human health and the environment. The CAMS reanalysis can be a valuable tool for researchers and policymakers to understand the long-term changes in atmospheric composition and their impacts on the environment.

Surl et al.[4] from this paper, we can learn about the variations of formaldehyde (HCHO) columns over India and the processes that drive these variations. The authors utilized satellite observations of HCHO columns from the Ozone Monitoring Instrument (OMI) to investigate the changes in HCHO concentrations over India. The study found significant spatial and temporal variability in HCHO columns over India, with higher concentrations observed during the summer monsoon season. The paper identifies several factors that drive these variations, including changes in biomass burning, atmospheric chemistry, and transport of air masses. The authors also highlight the importance of considering local and regional factors in understanding the variations of HCHO columns over India. This paper provides valuable insights into the sources and processes driving air pollution in India and emphasizes the need for improved monitoring and mitigation measures to address the impacts of air pollution on human health and the environment.

Stavrakou et al.[5] this paper tells about the evaluation of the performance of pyrogenic and biogenic emission inventories against a decade of space-based formaldehyde (HCHO) columns. The authors utilized satellite observations of HCHO columns from the Global Ozone Monitoring Experiment (GOME) and the Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY) to assess the accuracy of emission inventories of pyrogenic and biogenic sources of HCHO. The study found that the biogenic emissions inventories performed well in representing the seasonal and spatial variability of HCHO concentrations, while the pyrogenic emissions inventories showed less accuracy in representing the observed variations. The authors also identified the limitations of the HCHO column data for evaluating the inventories, including the influence of atmospheric transport and chemistry on the HCHO column measurements. This paper highlights the importance of accurate emission inventories for understanding the sources and impacts of air pollution and emphasizes the need for continued improvements in monitoring and modeling efforts to address the challenges of air pollution control.

Tianqi Chen et al.[6] describe about XGBoost, a scalable tree-boosting system that has been widely used in machine learning and data mining applications. The authors describe the design and implementation of XGBoost, which combines the strengths of tree-based models and gradient-boosting algorithms to achieve high accuracy and computational efficiency. The paper discusses the key features of XGBoost, including a novel regularization technique called ”tree pruning,” which helps prevent overfitting and improves generalization performance. The authors also highlight the advantages of XGBoost over other popular machine learning frameworks, such as speed, scalability, and accuracy. The paper presents several experiments demonstrating the effectiveness of XGBoost in a variety of real-world applications, including classification, regression, and ranking tasks. This paper provides valuable insights into the design and implementation of XGBoost and its practical applications in machine learning and data mining.

Takuya Akiba et al.[7] this paper aims to present Optuna, a next-generation hyperparameter optimization framework that aims to simplify and automate the process of tuning machine learning models. The authors describe the design and implementation of Optuna, which is based on the idea of using Bayesian optimization to efficiently search the hyperparameter space of a model. The paper discusses the key features of Optuna, including a flexible and extensible interface that allows users to define their own search spaces and objective functions, as well as a variety of optimization algorithms and pruning strategies that can improve the efficiency and effectiveness of the search. The authors also highlight the advantages of Optuna over other popular hyperparameter optimization frameworks, such as ease of use, scalability, and versatility. The paper presents several experiments demonstrating the effectiveness of Optuna in optimizing the hyperparameters of machine learning models in a variety of domains, including image classification, natural language processing, and reinforcement learning. This paper provides valuable insights into the design and implementation of

Optuna and its practical applications in machine learning research and development.

Mriganka Sekhar Biswa et al. [8] this article details observations of the effect of boundary layer evolution on nitrogen dioxide (NO2) and formaldehyde (HCHO) concentrations at a high-altitude observatory in western India. The authors describe the results of a study in which they analyzed the variations in NO2 and HCHO concentrations at the Hanle observatory, a remote high-altitude site in the Himalayas, over a period of three years. The study found that the concentrations of NO2 and HCHO were influenced by the boundary layer dynamics, which were in turn affected by the synoptic and mesoscale meteorological conditions. The paper discusses the key findings of the study, including the significant diurnal and seasonal variations in NO2 and HCHO concentrations, as well as the influence of long-range transport on the observed concentrations. The authors also highlight the importance of studying the boundary layer dynamics and their impact on air quality at high-altitude sites, which are often used as reference stations for air quality monitoring. This paper provides valuable insights into the complex interactions between atmospheric chemistry, meteorology, and boundary layer dynamics, and their implications for air quality monitoring and management.

I De Smedt et al. [9] this paper discusses the trend detection in satellite observations of formaldehyde (HCHO) tropospheric columns, which are an important indicator of air quality and atmospheric chemistry. The authors used data from the Global Ozone Monitoring Experiment (GOME) and the Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY) to study the trends in HCHO tropospheric columns from 1996 to 2008. They applied a statistical analysis method to detect the trends and evaluated the significance of the observed changes. The study found that the HCHO tropospheric columns increased significantly in many regions of the world, including South America, Africa, and South and Southeast Asia while decreasing in parts of North America and Europe. The authors suggest that the observed trends may be linked to changes in emissions of volatile organic compounds (VOCs) and other pollutants, as well as meteorological factors and long-range transport.

Narendra Ojha et al. [10] in this paper, the authors explore the potential of machine learning for simulating urban ozone variability. They trained a machine learning model called random forest regression to predict hourly ozone concentrations in an urban area in India based on meteorological and air quality data. The model was able to accurately predict ozone concentrations with a high degree of accuracy, indicating that machine learning could be a useful tool for simulating and predicting urban air pollution.

Chapter 2

Data Analysis and Data Preprocessing

In this Section, we will discuss the data variable we have used in our work. As we have discussed above that our area of interest is over India so we have collected the data for our work only for the Indian region. The scale we have used is 11km which is in degree is 0.10 °resolution. So here we have discussed every variable property and importance in our study. We also have mentioned their sources from where we collected the data and how we cleaned and preprocessed the data.

2.1 Data Source

The data is being collected by the Google Earth Engine. There are other data sources in GEE, however, we are choosing to use ERA5 and Sentinel-5P since they are better suited for our study.

2.1.1 ERA5

ERA5 (fifth generation of the European Re-Analysis) is a global atmospheric reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). It provides a comprehensive and consistent view of the Earth's atmosphere from 1979 to the present day, with a spatial resolution of approximately 31 km.

ERA5 incorporates observations from various sources, such as satellite data, ground-based measurements, and weather balloons, and uses a state-of-the-art data assimilation system to combine this information with a numerical model of the atmosphere. The resulting dataset includes a wide range of atmospheric variables, such as temperature, wind, humidity, pressure, and precipitation, at various levels in the atmosphere. ERA5 is widely used in atmospheric research and climate studies, as it provides a detailed and consistent view of the Earth's atmosphere over the past four decades. It is also used in applications such as weather forecasting, air quality monitoring, and renewable energy planning.

2.1.2 Sentinel-5P

The European Space Agency launched Sentinel-5 Precursor on October 13, 2017, as a satellite to track air pollution. A common name for the onboard sensor is Tropomi (TROPOspheric Monitoring Instrument). The SENTINEL-5P spacecraft's lone payload, TROPOMI, is designed to measure the composition and characteristics of the atmosphere. The device makes use of passive remote sensing techniques to accomplish its objective by determining the solar radiation reflected by and emitted from the Earth at the Top Of Atmosphere (TOA). The data which is collected by this satellite can be used for assessing air quality. It provides UV Aerosol Index data, Cloud, Carbon Monoxide, Formaldehyde, Nitrogen Dioxide, Ozone, Methane, and Sulphur Dioxide from 2018-07-10 to the Present. Every S5P dataset, with the exception of CH4, includes two versions: Near Real-Time (NRTI) and Offline (OFFL). CH4 is only accessible via OFFL. Despite being less substantial than OFFL assets, NRTI assets are acquired more quickly. The assets of OFFL comprise data from a single orbit, which only includes information for one hemisphere because half of the globe is dark.

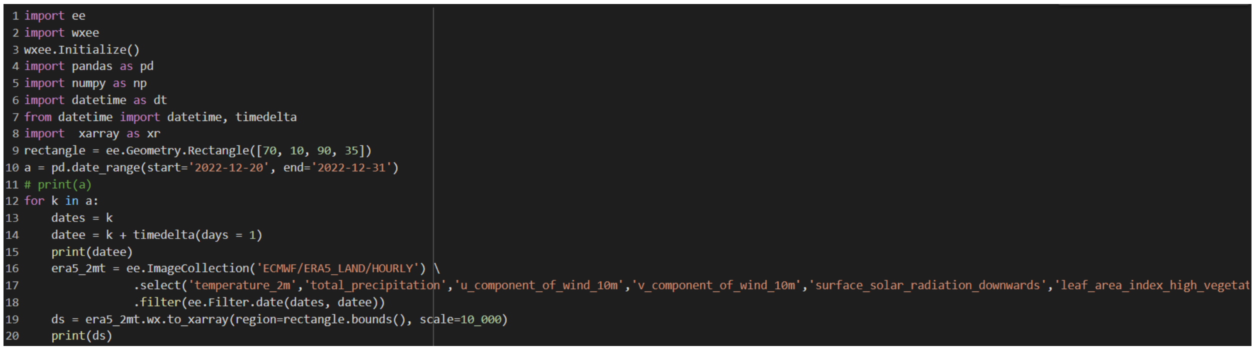
2.2 Description of Data Variables

We have used a total of seven input variables which we have downloaded from the google earth engine with the help of the wxee package which integrates the data catalogue and processing power of the Google Earth Engine. These variables are daily reanalyzed data starting from 2018 to 2022. The grid for the dataset was 10 °N to 35 °N latitude and 70 °E to 90 °E longitude. The target variable is the SENTINEL-5P HCHO observations. The equatorial overpass time of the SENTINEL-5P satellite is 13:00 hours. Hence the HCHO signal is observed by the SENTINEL satellite at about midday. Depending on the precursor concentration from a few hours prior, the HCHO concentration will change. According to our hypothesis, all input factors are predicted to influence HCHO concentration for a few hours previous to the observation period.

WXEE (short for Web-based Google Earth Engine Explorer) is an open-source web application developed by the University of Arizona that provides an interface for accessing and analyzing geospatial data using Google Earth Engine. It provides a user-friendly interface for exploring and visualizing geospatial data, and also allows users to run scripts and perform custom analyses using the Google Earth Engine API.

S.No	Input Variables Name(Units)	Min-Max Value
1	temperature_2m(K)	234-320
2	total_precipitation(m)	0-0.1843
3	u_component_of_wind_10m(m/s)	(-13.84314)-15.85161
4	v_component_of_wind_10m(m/s)	(-11.48421)-13.19966
5	surface_solar_radiation_downwards(J/m^2)	$0 - 3.020631e^{+07}$
6	leaf_area_index_high_vegetation(1 (area fraction))	0-6.605347
7	leaf_area_index_low_vegetation (1 (area fraction))	0-4.661133

Table 2.1: Details of Input Variables



```
import ee
import wxee
wxee.Initialize()
import pandas as pd
import numpy as np
import datetime as dt
from datetime import datetime, timedelta
import  xarray as xr
rectangle = ee.Geometry.Rectangle([70, 10, 90, 35])
a = pd.date_range(start='2022-12-20', end='2022-12-31')
# print(a)
for k in a:
    dates = k
    datee = k + timedelta(days = 1)
    print(datee)
    era5_2mt = ee.ImageCollection('ECMWF/ERA5_LAND/HOURLY') \
                .select('temperature_2m','total_precipitation','u_component_of_wind_10m','v_component_of_wind_10m','surface_solar_radiation_downwards','leaf_area_index_high_vegetat
                .filter(ee.Filter.date(dates, datee))
    ds = era5_2mt.wx.to_xarray(region=rectangle.bounds(), scale=10_000)
    print(ds)
```

Figure 2.1: Using google earth engine API to download the ERA5 data with the help of wxee package

2.2.1 Temperature

We have noted that plants as a natural reaction to heat stress VOCs is released at higher temperatures. VOCs are oxidized to produce HCHO, which is favoured at greater temperatures as well. As a result, we decide to include atmospheric temperature as one of the model's inputs. In Figure 3.2 the maximum temp is 320 K and the minimum temp is 234 K.

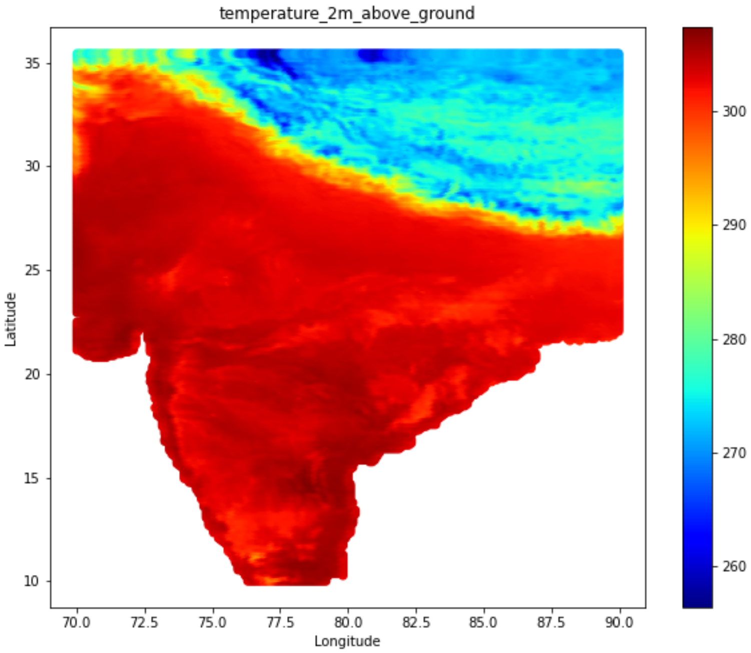


Figure 2.2: Plot for surface air temperature Input Variable

2.2.2 Precipitation

Water that has accumulated as rain or snow and that has frozen into liquid form. In Figure 3.2 the maximum precipitation value is 0.1843 m and the minimum precipitation is 0 m.

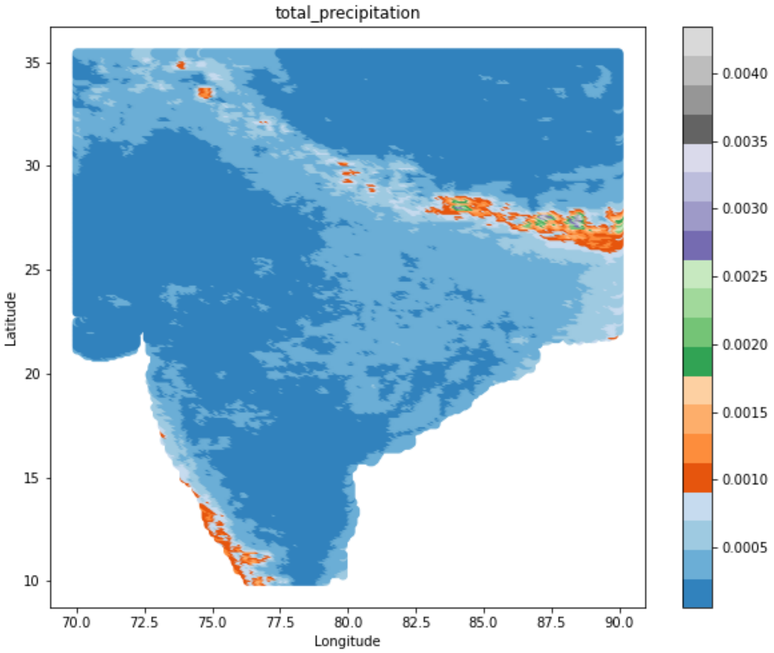


Figure 2.3: Plot for Total Precipitation Input Variable

2.2.3 Wind

We have considered two components of wind. One is the u component(Figure 3.4) which is the Eastward component at a height of ten meters above the Earth's surface, it is the horizontal speed of air moving in the east, measured in meters per second. Its maximum speed is 15.85161 m/s and the negative value shows the opposite direction of the wind.

The other component is the v component(Figure 3.5) which is the Northward component at a height of ten meters above the Earth's surface, it is the horizontal speed of air moving in the north, measured in meters per second. Its maximum speed is 13.19966 m/s and the negative value shows the opposite direction of the wind.

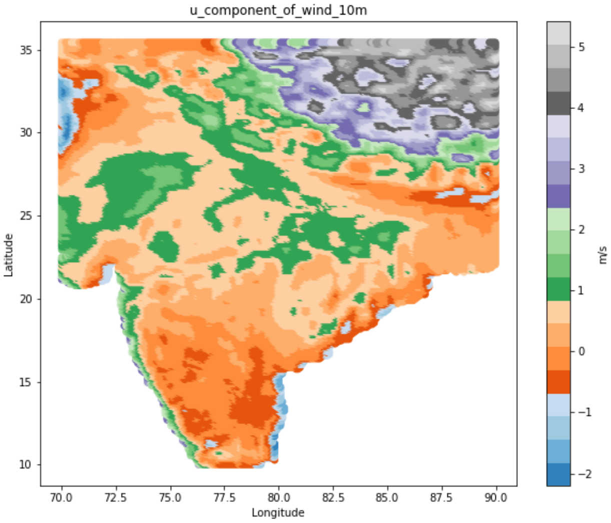


Figure 2.4: Plot for Eastward component of wind Input Variable

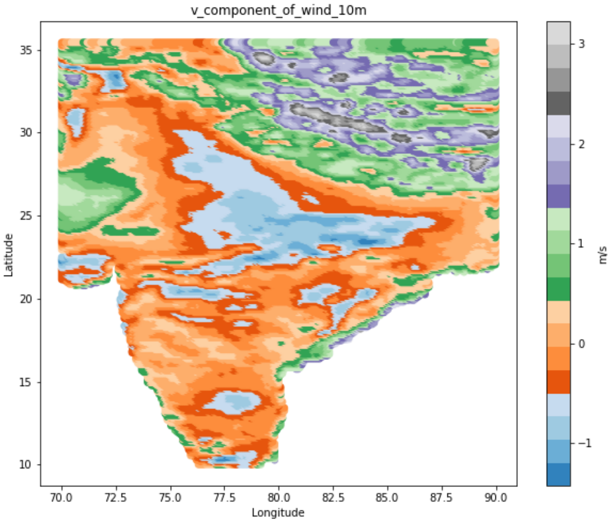


Figure 2.5: Plot for Northward component of wind Input Variable

2.2.4 SSRD

We considered solar radiation(Figure 3.6) as another meteorological input. The amount of solar radiation, sometimes referred to as shortwave radiation, reaches the Earth's surface.

Both direct and diffuse solar radiation are included in this variable. Solar, or shortwave, radiation from the Sun is partially reflected back to space by clouds and airborne particles (aerosols), and some of it is absorbed. The remainder occurs on Earth's surface (represented by this variable).
The maximum value is 3.020631^{+07} J/m^2 *minimum value is* $0J/m^2$.

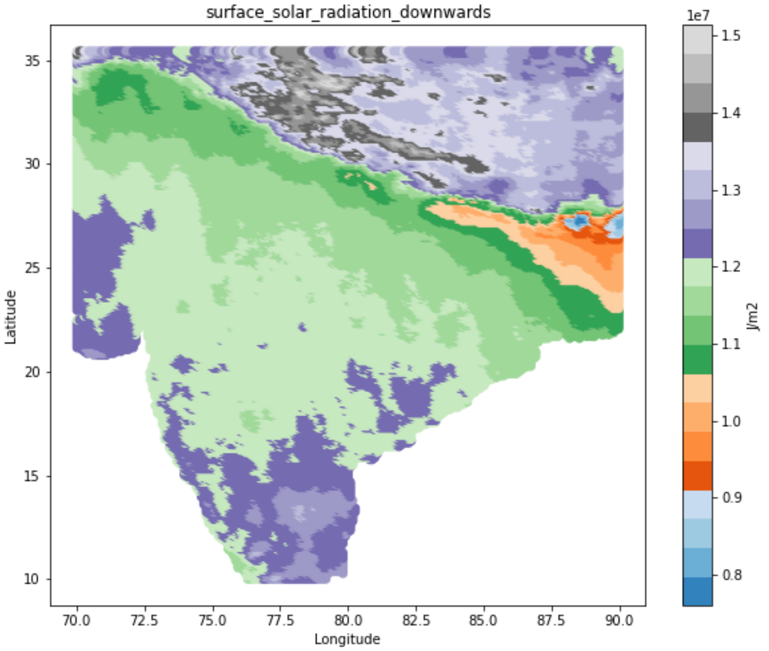


Figure 2.6: Plot for Surface Solar Radiation Downwards Input Variable

2.2.5 LAI

As a previous study suggests biogenic emissions are major contributors for VOCs emission that is why we have also considered leaf area index (LAI) with both high and low vegetation. LAI is a dimensionless quantity.

One-half of the total green leaf area per unit of horizontal ground surface area for the high vegetation type(Figure 3.7). One-half of the total green leaf area per unit of horizontal ground surface area for the low vegetation type(Figure 3.8). The amount of light that can be intercepted by plants is directly correlated to the total area of leaves per unit of ground area, or LAI.

$$LAI = \frac{\text{leaf area}}{\text{ground area}} \; \frac{\text{m}^2}{\text{m}^2}$$

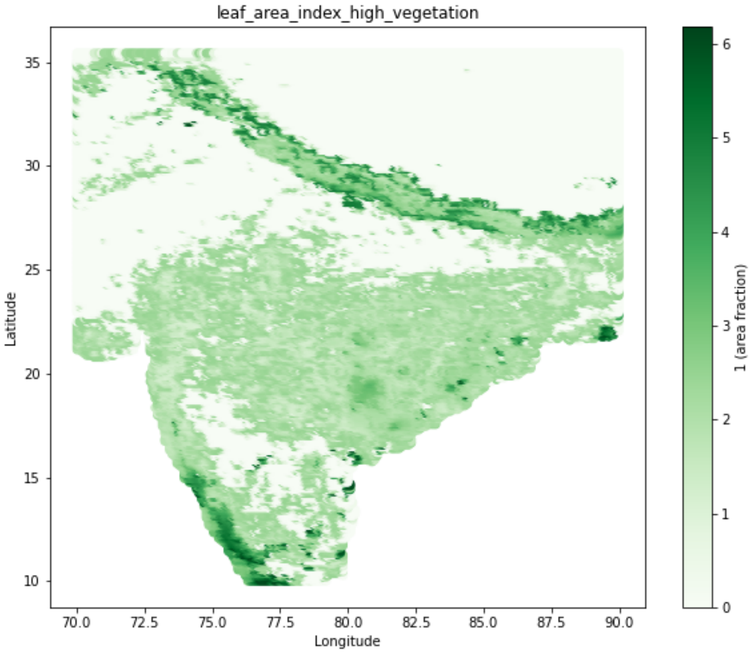


Figure 2.7: Plot for Leaf Area Index High Vegetation Input Variable

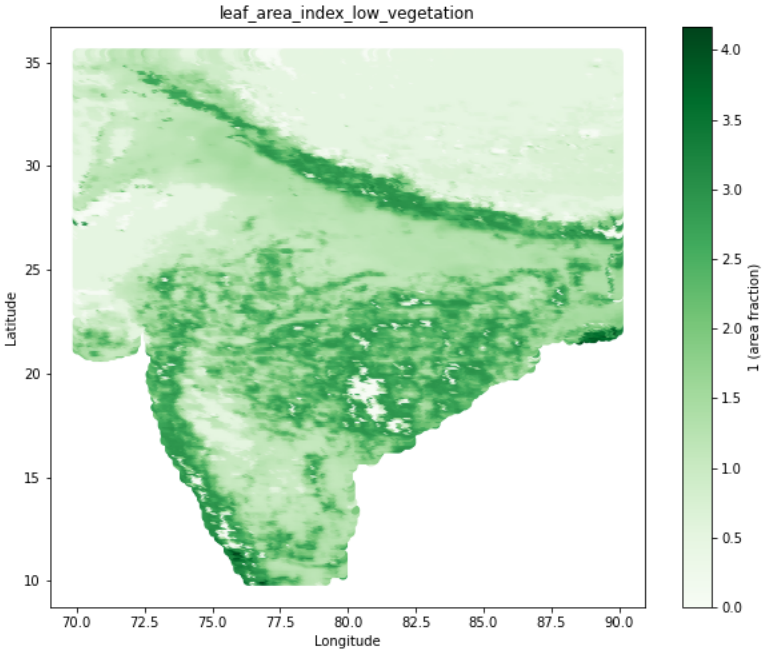


Figure 2.8: Plot for Leaf Area Index Low Vegetation Input Variable

2.2.6 Target data

We have used the Sentinel-5P satellite data as target data for our study. In our dataset the maximum value is 0.000269 mol/m^2 and minimum value is -0.000517 mol/m^2 In Figure 3.9 The five boxes on the map of the HCHO satellite observed data represent the significant hotspots of HCHO over India. The Indo-Gangetic Plain (IGP) is depicted in Box 'A'. Box 'B' depicts Bihar, while Box 'C' and Box 'E' depict the Indian states of Odisha and Chhattisgarh with their respective forest regions and the Western Ghats mountain range.

Mumbai and its environs are shown in Box 'D'. These areas are formaldehyde hotspots due to the biogenic, pyrogenic, and anthropogenic emissions they produce.

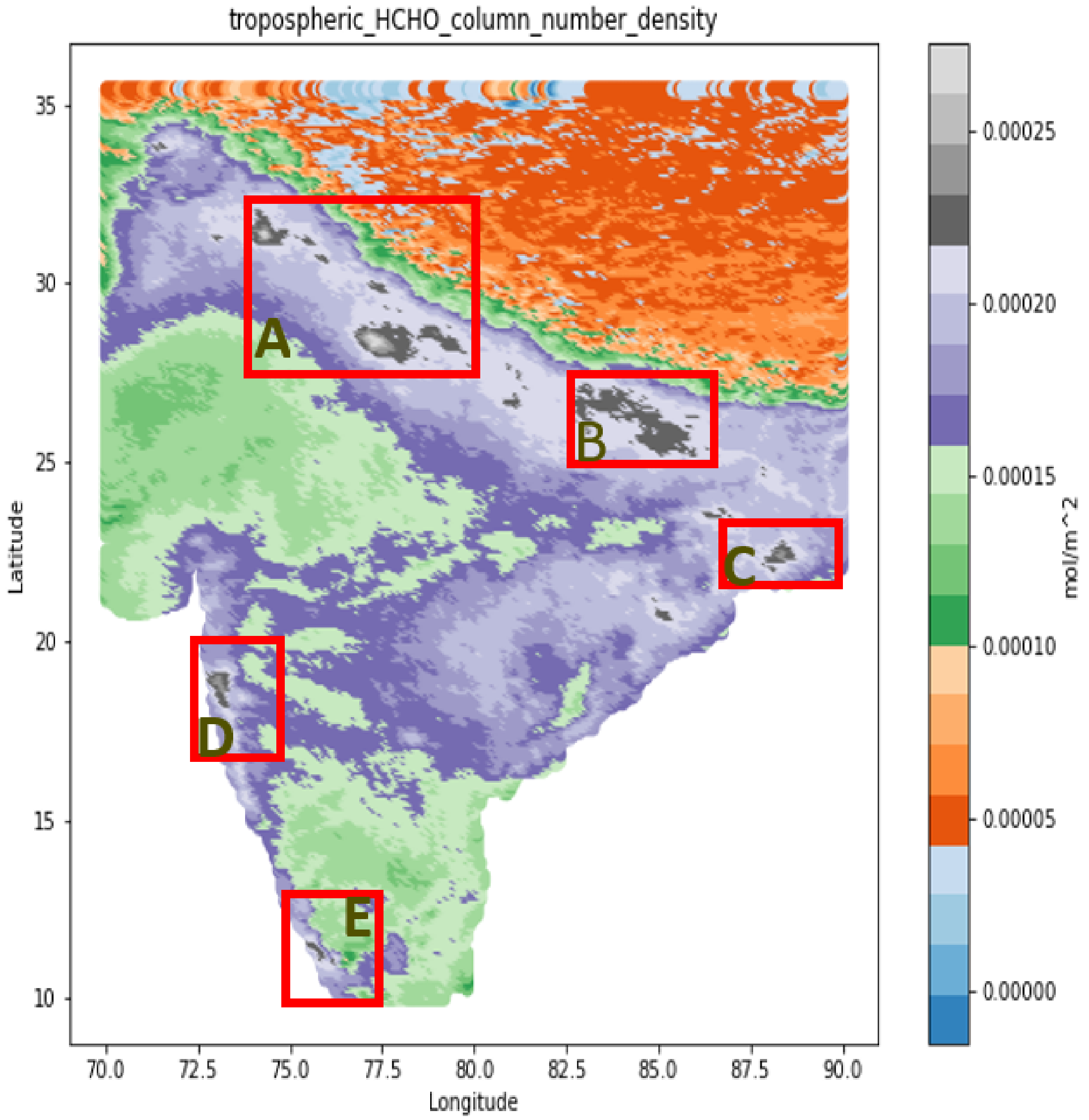


Figure 2.9: Spatial distribution of average Sentinel-5P satellite observed formaldehyde

2.3 Data Preparation

1. Download the input data from the ERA5 repository and the resolution is 284*223 for the period of 2018 to 2022

2. As we've already discussed, the HCHO's lifetime can be anything from a few minutes and

a few hours. So, we have used daily averaged input data from 12:00 to 14:00 local time.

3. collect the target data from the Sentinel-5P satellite from 2018 to 2022 with a resolution of 284*223.

4. It was also hourly data from 12:00 local time to 14:00 local time. Then Hourly files are used to produce a daily average dataset for each variable.

5. To build a dataset with a uniform spatial resolution, the entire dataset is regridded. We have used xESMF to regrid the data.

6. As there is no spatial relationship among data points, so converted to tabular data.

7. And we have developed an end-to-end tabular deep learning and machine learning model with optuna hyperparameter optimizations.

8. Created docker containers for tabular data.

Chapter 3

Methodology

In this chapter, we have discussed the problem statement, data preprocessing, model, algorithm flow, and training of the model.

3.1 Problem Statement

In comparison to satellite measurements and reanalysis, dynamic atmospheric chemistry models frequently exhibit an up to double overestimation of atmospheric formaldehyde. Additionally, the modelled HCHO's spatial distribution does not agree with satellite observations. These simulations make use of intricate atmospheric chemistry. GEOS-Chem simulation and satellite (OMI, GOME-2a) HCHO data were compared, and the results indicated that biogenic VOC emissions are 30–60% overestimated. The ECHAM5-HAMMOZ model, a dynamic atmospheric chemistry model, has been found to overestimate HCHO in comparison to satellite observations, according to previous studies. These studies also found a considerable discrepancy between dynamic model simulation and satellite observations of HCHO. As a result, our strategy provides an indirect technique of HCHO estimation that does not require chemical equations.

To do so we are taking the help of AI/ML. We train our model on collected data and make predictions. AI models can forecast the HCHO, but before that, we need to train the model for supervised learning. Here we have data for the last 3 years and using that we train a model.

3.2 Proposed Model

Based on the above-defined problem statement we can propose some deep learning and machine learning models like xgboost, catboost, transformer, resnet, mlp etc.

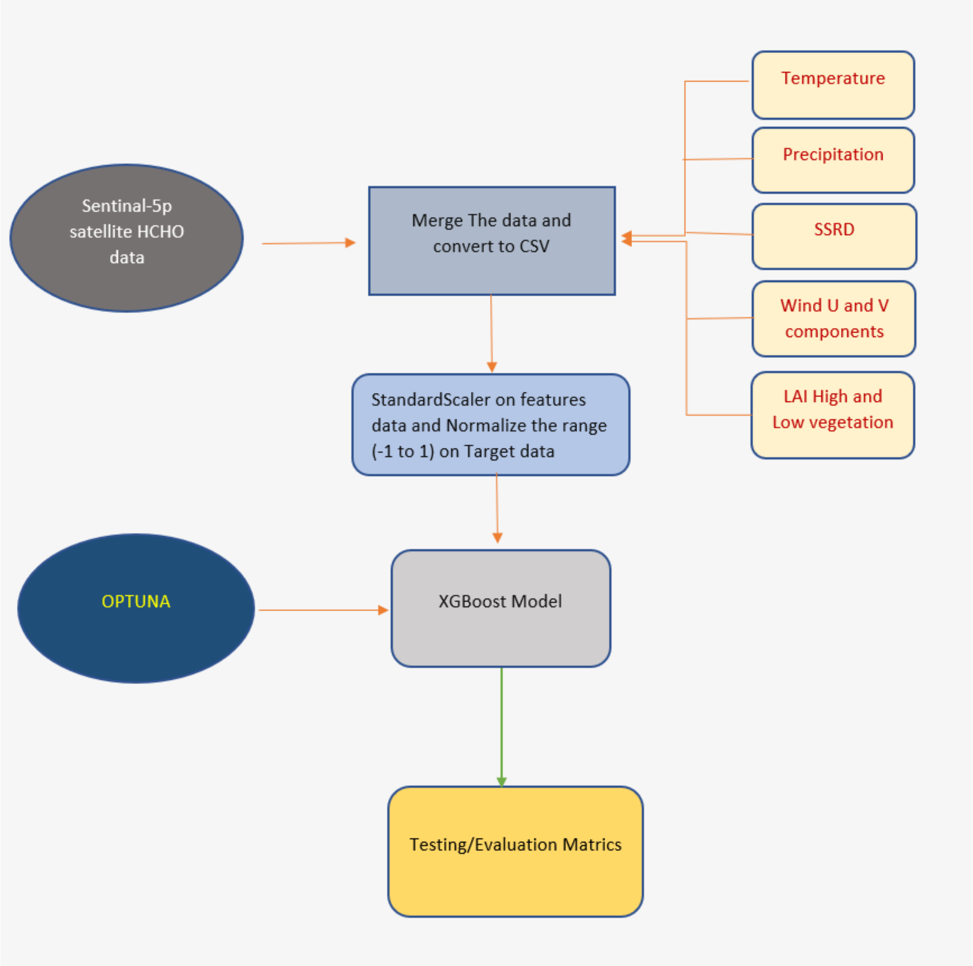

Sentinal-5p satellite HCHO data
Merge The data and convert to CSV
Temperature
Precipitation
SSRD
Wind U and V components
LAI High and Low vegetation
StandardScaler on features data and Normalize the range (-1 to 1) on Target data
OPTUNA
XGBoost Model
Testing/Evaluation Matrics

Figure 3.1: Methodology Used for forecasting of HCHO

In this study, we have used xgboost. In the xgboost model, we have used Optuna for hyperparameter optimization. It is an open-source framework for optimizing hyperparameter searches.

3.2.1 OPTUNA

As we know GridSearchCV and RandomizedSearchCV work well when we have a small number of parameters to be optimized. So when parameters get increased these optimization techniques become computationally expensive. So to optimize in a better way when a number of parameters are large we have OPTUNA. A number of algorithms are provided by the

well-known Python package Optuna for determining the best hyperparameters.
The process of choosing the appropriate hyperparameters for a machine learning model is known as hyperparameter optimisation. Hyperparameters are the parameters that must be defined before the model is trained because they cannot be learned from the data. Optuna effectively searches the hyperparameter space and minimizes the objective function via Bayesian optimization. Using earlier assessments, it builds a model of the objective function and offers fresh sets of hyperparameters based on model predictions.

You must build an objective function that accepts hyperparameters as input and outputs a scalar value that represents the model's performance in order to use Optuna for hyperparameter optimization. Following an iterative evaluation of the objective function with various hyperparameter values, Optuna updates the model and suggests further hyperparameters for evaluation based on the findings.

Optuna makes it simple to improve the hyperparameters of multiple machine learning models by offering a number of built-in integration with well-known machine learning libraries, including Scikit-learn, PyTorch, and TensorFlow. Furthermore, it offers distributed computation, allowing for effective hyperparameter optimization on substantial datasets or intricate models. It uses the Bayesian optimization technique.

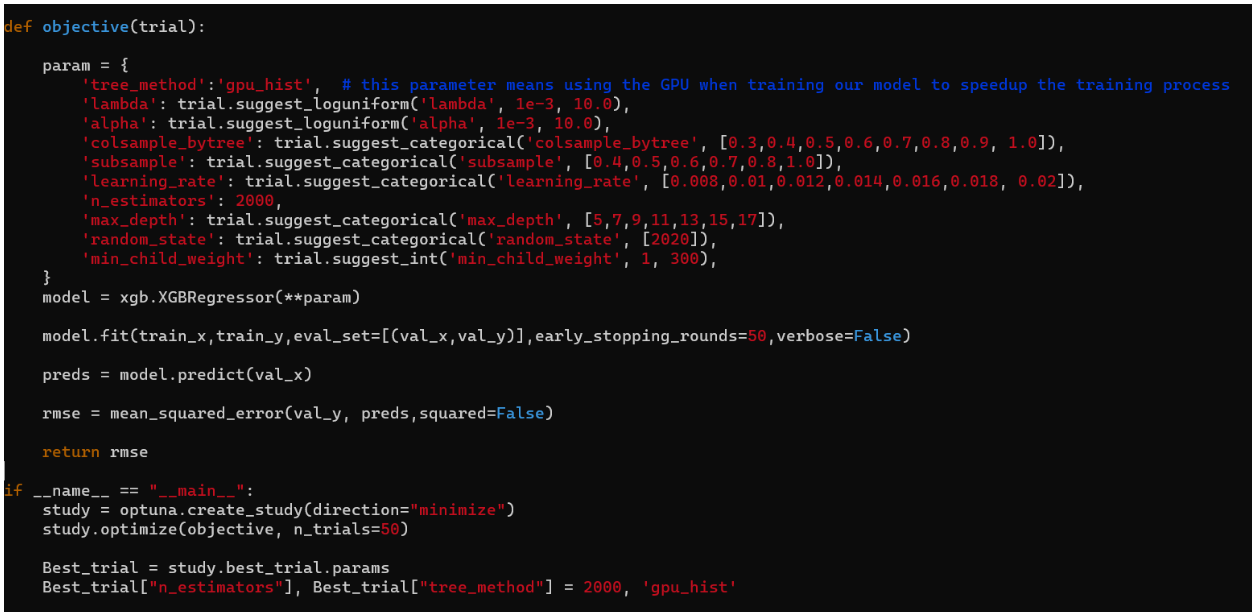
1. Build a surrogate probability model of the objective function.
2. Find the hyperparameters that perform best on the surrogate.
3. Apply these hyperparameters to the true objective function.
4. Update the surrogate model incorporating the new results.
5. Repeat steps 2-4 until max iterations or time is reached.

Surrogate probability model means it takes the parameter which we have given and prepares a probability model over it. The probability model means a function that it prepares by doing a hit and trial and it selects a build-up function and then it applies those hyperparameters which we have given to it over the surrogate probability model and the best-performing hyperparameters it will select on that surrogate probability model and apply the same probability to the true objective fun that has to maximize.

It uses the terms study and trial as follows:

Study: optimization based on an objective function

Trial: a single execution of the objective function



```
def objective(trial):

    param = {
        'tree_method':'gpu_hist',  # this parameter means using the GPU when training our model to speedup the training process
        'lambda': trial.suggest_loguniform('lambda', 1e-3, 10.0),
        'alpha': trial.suggest_loguniform('alpha', 1e-3, 10.0),
        'colsample_bytree': trial.suggest_categorical('colsample_bytree', [0.3,0.4,0.5,0.6,0.7,0.8,0.9, 1.0]),
        'subsample': trial.suggest_categorical('subsample', [0.4,0.5,0.6,0.7,0.8,1.0]),
        'learning_rate': trial.suggest_categorical('learning_rate', [0.008,0.01,0.012,0.014,0.016,0.018, 0.02]),
        'n_estimators': 2000,
        'max_depth': trial.suggest_categorical('max_depth', [5,7,9,11,13,15,17]),
        'random_state': trial.suggest_categorical('random_state', [2020]),
        'min_child_weight': trial.suggest_int('min_child_weight', 1, 300),
    }
    model = xgb.XGBRegressor(**param)

    model.fit(train_x,train_y,eval_set=[(val_x,val_y)],early_stopping_rounds=50,verbose=False)

    preds = model.predict(val_x)

    rmse = mean_squared_error(val_y, preds,squared=False)

    return rmse

if __name__ == "__main__":
    study = optuna.create_study(direction="minimize")
    study.optimize(objective, n_trials=50)

    Best_trial = study.best_trial.params
    Best_trial["n_estimators"], Best_trial["tree_method"] = 2000, 'gpu_hist'
```

Figure 3.2: Demonstration of Optune with the XGBoost model

3.2.2 Introduction to XGBoost

Extreme Gradient Boosting, also known as XGBoost, is a well-known open-source machine learning package that generates predictions using an ensemble of decision trees. It is a type of gradient-boosting method that Tianqi Chen and Carlos Guestrin initially described in a research paper in 2016. eXtreme Gradient Boost(XGBoost) is a sub-class of Supervised Machine Learning and uses a Tree-based Ensemble algorithm.

An example of a supervised learning method that can be applied to both classification and regression problems is the XGBoost model.

It functions by building a sequence of decision trees, with each tree trying to fix the flaws of the one before it. This is accomplished by changing the weights of the training examples so that the subsequent tree concentrates more on the cases that the preceding tree incorrectly identified. Gradient boosting is one of its key characteristics which helps to optimize the model's parameters. Gradient boosting involves gradually incorporating weak learners into the model, each of whom attempts to correct the flaws of the previous one. By using this technique, XGBoost can build complex models that can depict nonlinear relationships between the input variable and the target variable.

In this study, we are using the model for the climate data which is just like a sequence of images. Latitude and longitude can be considered as rows and columns of an image and the value of each grid is like the intensity of the image. This is how the climate data is similar to an image. Based on this assumption we are using XGBoost model for forecasting.

Step 1:Initial Prediction and Residuals

It is an ensemble technique that uses boosting technique. It creates a sequential decision tree. So it calculates the average of our target data and based on that average it calculates the residual. So our model is trained based on that particular residual values.

$$Residual = actual\ y\ value - predicted\ y\ value$$

Step 2: Build an XGBoost Tree

In XGBoost binary tree is created. So Every tree begins with a single leaf and all the residuals go into that leaf. Now Similarity Score of this leaf is calculated.

$$Similarity\ score = \frac{(\sum Residuals)^2}{N+\lambda}$$

Regularization is a machine-learning technique that avoids overfitting a model. Overfitting is a condition where a model performs poorly on new, unforeseen data because it is overly complex and fits the training data too closely. The loss function of the model is modified by regularisation to include a penalty term that prevents the model from learning excessively intricate patterns from the training set.

The strength of the penalty term is typically controlled by a hyperparameter known as lambda, which also controls regularisation. While a smaller value of lambda enables the model to fit the data more closely, potentially leading to overfitting, a larger value of lambda produces stronger regularisation and a simpler model.

$$GAIN = Left\ Similarity + Right\ Similarity - Root\ Similarity$$

Now it quantifies how much root residuals and leaves residual are similar. Calculating the Gain of dividing the residuals into two groups will help us achieve this. Splitting makes sense if the Gain is positive; otherwise, it does not.

Step 3: Prune the Tree

Another way to prevent overfitting the data is through pruning. To determine whether a split is valid or not, we start at the bottom of our tree and proceed to move up. We use (gamma) to establish validity. If Gain — γ is positive, the split is kept; otherwise, it is eliminated. The value of gamma is 0 by default.

Step 4: Calculate the Output of Leaves and make New Predictions

It can no longer have a leaf node giving us multiple outputs, so all it needs to do is calculate a single value in its leaf nodes. $\lambda = 1$

$$Output\ Value = \frac{\sum Residuals}{Number\ of\ Residuals + \lambda}$$

The only difference between this and the similarity score calculation procedure is that we are not squaring the residuals.

Step 5: Make New Predictions

Now comes the part where evaluating how much our forecasts improve as a result of the new model. We can make predictions using this formula:

$$\text{predicted value} = \text{Initial prediction} + \epsilon(\text{output value}) \tag{3.1}$$

Our residuals will continue to shrink, a sign that the predicted values are approaching the observed values.

Now, we just keep repeating the same procedure, creating a new tree, generating predictions, and figuring out residuals with each iteration. We continue doing this until the residuals are extremely minimal or we reached the maximum number of iterations we set for our algorithm.

3.2.3 Loss function

Perhaps the most basic and widely used loss function is the Mean Squared Error (MSE). The MSE is calculated by taking the difference between the predictions made by the model and the actual data, squaring it, and averaging it across the whole dataset.

Mean Squared Error can be expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The model is trained to find values of the weights which are responsible for the minimization of the loss. Root Mean Squared Error (RMSE) is one common metric used to evaluate the accuracy of predictions. RMSE measures the Euclidean distance between predicted values and actual values, it provides the measure that how far the prediction is from the actual or true values.

It is calculated by finding the square of the residuals (i.e. the differences between the predicted and true values) at every data point, then the average of the residual is computed, and finally, the square root of the average is calculated to obtain the RMSE. The formula for RMSE is given below.
Root mean square error can be expressed as:

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

where n is the number of data points, y_i is the i (th) measurement, and $\hat{y}$ is its corresponding prediction. Standard RMSE = RMSE/STD. . . . . (standard deviation from actual dataset)

3.3 Implementation

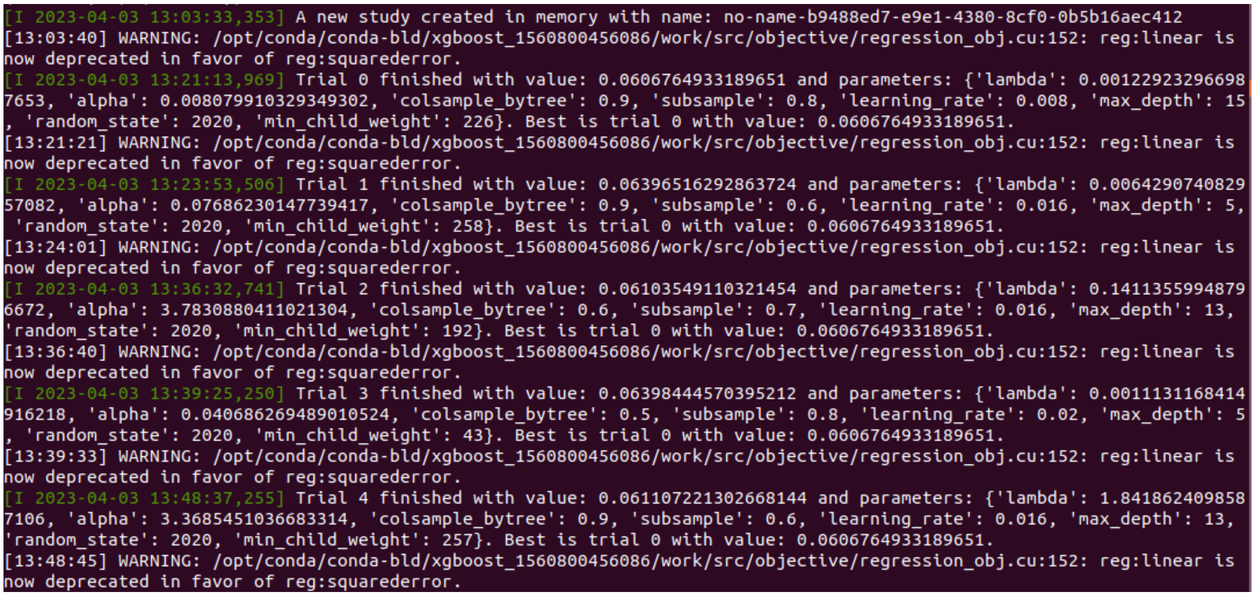
3.3.1 Data Preprocessing

Data preprocessing is the step before the dataset is sent to the model. This technique is used to make all the variable values uniformly distribute the data over all records and fields. Data processing make data suitable for machine learning algorithm and helps to increase the accuracy and efficiency of the machine learning model.

After collecting the data from the different repositories, we converted the data to tabular form in CSV format because there is no spatial relationship among data points and drop Nan values. The HCHO data from the Sentinel-5p satellite is first quality controlled. Each grid point's formaldehyde column is cleaned if its amount is less than the uncertainty. The next step is to concat the input variable and target variable and then merge all daily data files into a single CSV file.

3.3.2 Training

So for training the XGBoost model we first split the data for training and testing. So the training data is taken for the years 2018-2021 and for validation, we have used the first six-month data of 2022, and for testing used the last six months of data of 2022. We have given 100 trials and in order to prevent overfitting, we have used early stopping and saved the pickle file of the best model. We use Optuna optimizer with a loss function as the mean squared error. We have used an NVIDIA A100 GPU which took 2 Days to complete 100 trials.



```
[I 2023-04-03 13:03:33,353] A new study created in memory with name: no-name-b9488ed7-e9e1-4380-8cf0-0b5b16aec412
[13:03:40] WARNING: /opt/conda/conda-bld/xgboost_1560800456086/work/src/objective/regression_obj.cu:152: reg:linear is
now deprecated in favor of reg:squarederror.
[I 2023-04-03 13:21:13,969] Trial 0 finished with value: 0.0606764933189651 and parameters: {'lambda': 0.00122923296698
7653, 'alpha': 0.008079910329349302, 'colsample_bytree': 0.9, 'subsample': 0.8, 'learning_rate': 0.008, 'max_depth': 15
, 'random_state': 2020, 'min_child_weight': 226}. Best is trial 0 with value: 0.0606764933189651.
[13:21:21] WARNING: /opt/conda/conda-bld/xgboost_1560800456086/work/src/objective/regression_obj.cu:152: reg:linear is
now deprecated in favor of reg:squarederror.
[I 2023-04-03 13:23:53,506] Trial 1 finished with value: 0.06396516292863724 and parameters: {'lambda': 0.0064290740829
57082, 'alpha': 0.07686230147739417, 'colsample_bytree': 0.9, 'subsample': 0.6, 'learning_rate': 0.016, 'max_depth': 5,
 'random_state': 2020, 'min_child_weight': 258}. Best is trial 0 with value: 0.0606764933189651.
[13:24:01] WARNING: /opt/conda/conda-bld/xgboost_1560800456086/work/src/objective/regression_obj.cu:152: reg:linear is
now deprecated in favor of reg:squarederror.
[I 2023-04-03 13:36:32,741] Trial 2 finished with value: 0.06103549110321454 and parameters: {'lambda': 0.1411355994879
6672, 'alpha': 3.7830880411021304, 'colsample_bytree': 0.6, 'subsample': 0.7, 'learning_rate': 0.016, 'max_depth': 13,
'random_state': 2020, 'min_child_weight': 192}. Best is trial 0 with value: 0.0606764933189651.
[13:36:40] WARNING: /opt/conda/conda-bld/xgboost_1560800456086/work/src/objective/regression_obj.cu:152: reg:linear is
now deprecated in favor of reg:squarederror.
[I 2023-04-03 13:39:25,250] Trial 3 finished with value: 0.06398444570395212 and parameters: {'lambda': 0.0011131168414
916218, 'alpha': 0.040686269489010524, 'colsample_bytree': 0.5, 'subsample': 0.8, 'learning_rate': 0.02, 'max_depth': 5
, 'random_state': 2020, 'min_child_weight': 43}. Best is trial 0 with value: 0.0606764933189651.
[13:39:33] WARNING: /opt/conda/conda-bld/xgboost_1560800456086/work/src/objective/regression_obj.cu:152: reg:linear is
now deprecated in favor of reg:squarederror.
[I 2023-04-03 13:48:37,255] Trial 4 finished with value: 0.061107221302668144 and parameters: {'lambda': 1.841862409858
7106, 'alpha': 3.3685451036683314, 'colsample_bytree': 0.9, 'subsample': 0.6, 'learning_rate': 0.016, 'max_depth': 13,
'random_state': 2020, 'min_child_weight': 257}. Best is trial 0 with value: 0.0606764933189651.
[13:48:45] WARNING: /opt/conda/conda-bld/xgboost_1560800456086/work/src/objective/regression_obj.cu:152: reg:linear is
now deprecated in favor of reg:squarederror.
```

Figure 3.3: Training Process for HCHO by XGBoost

3.3.3 Testing

The testing is done with the help of pickle files generated while training. So for testing, we have daily data for six months of 2022 from July to December. So we have tested our trained model on this test data and plotted it with the help of the matplotlib library. To evaluate the performance or effectiveness of the model we calculated Pearson correlation coefficient.

Chapter 4

Results and Conclusions

4.1 Results

In this section, we have discussed the final result that we got from the model. The model has predicted the HCHO for six months of 2022. This prediction is compared with the ground truth (actual values) and comparison is done on the basis of the Pearson correlation coefficient and Standard Root Mean Square Error. Figure 4.1 represents the spatial distribution of average HCHO over India from satellite dataset. The plot shows HCHO hotspots over different regions and the spatial variation is quite non-homogeneous.

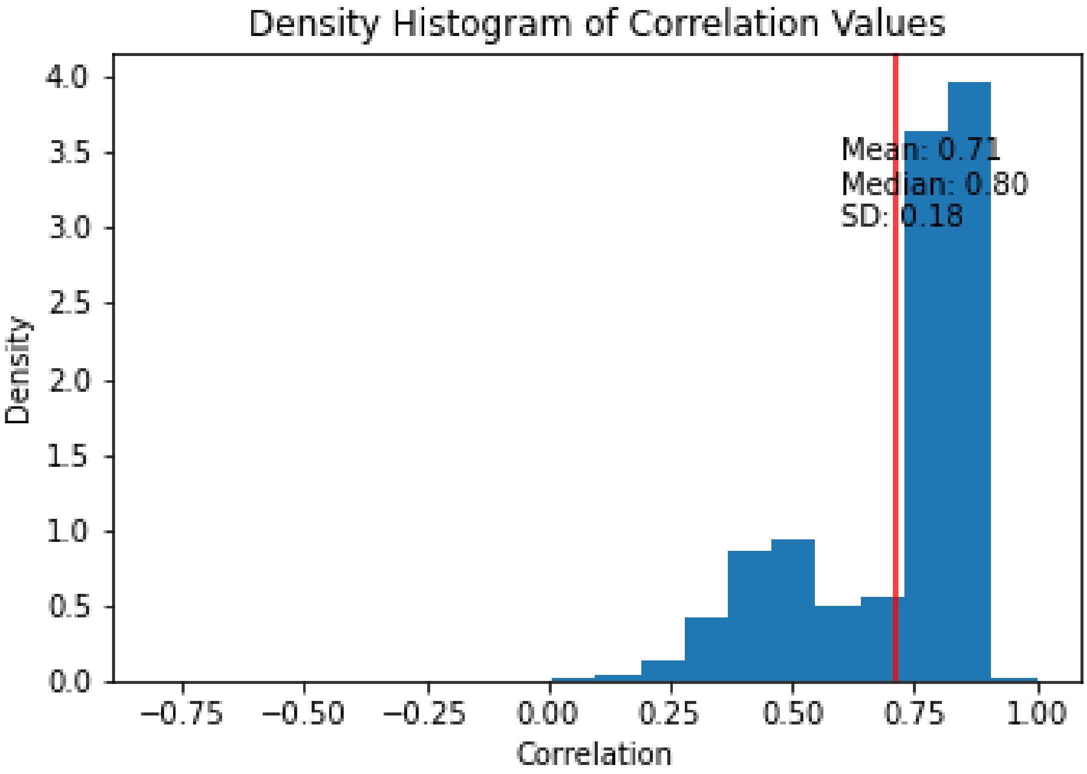
The Equation of Pearson Correlation value :-

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Here, x_i represents the values of the x-variable in a sample,
$\bar{x}$ represents the mean of the values of the x-variable,
y_i represents the values of the y-variable in a sample,
and $\bar{y}$ represents the mean of the values of the y-variable.

In Figure 2.9 we can see five hot-spot regions where the HCHO value is high. So we find a correlation between those hot-spot regions. Box'A' is The Indo-Gangetic Plain (IGP) area of India, which includes numerous large cities, including the country's capital, Delhi, which has the country's greatest population density. So in this region, HCHO emission is quite high because of the industries, automobiles, and crop residue. See Table 4.1 the correlation over this region is above 0.79.

Box'B' is Bihar and the nearby region which has a correlation value of 0.81. Box "C" displays an additional forested area from the Indian states of Odisha and Chhattisgarh; as a result, the correlation value is 0.80 and HCHO hotspot. Box'D' shows Mumbai and Pune region which are big cities and hcho emission is quite high because of the industries, automobiles and the correlation value over this region is 0.80.



The densely forested Western Ghats mountain ranges in Kerala, India, and other locations make up Box ”E,” which is a significant source of biogenic emissions. So in this region, the correlation is above 0.78.

Region coordinate(xmin,ymin,xmax,ymax)	Box	Pearson correlation coefficient
26.49,82.35,30.09,76.46	Box’A’	0.79
24.83,87.45,26.84,83.63	Box’B’	0.81
18.58,83.63,22.69,85.74	Box’C’	0.80
18.35,72.91,19.85,73.47	Box’D’	0.80
8.98,76.28,12.65,76.32	Box’E’	0.78

Table 4.1: Pearson’s correlation coefficient over different hot-spot regions.

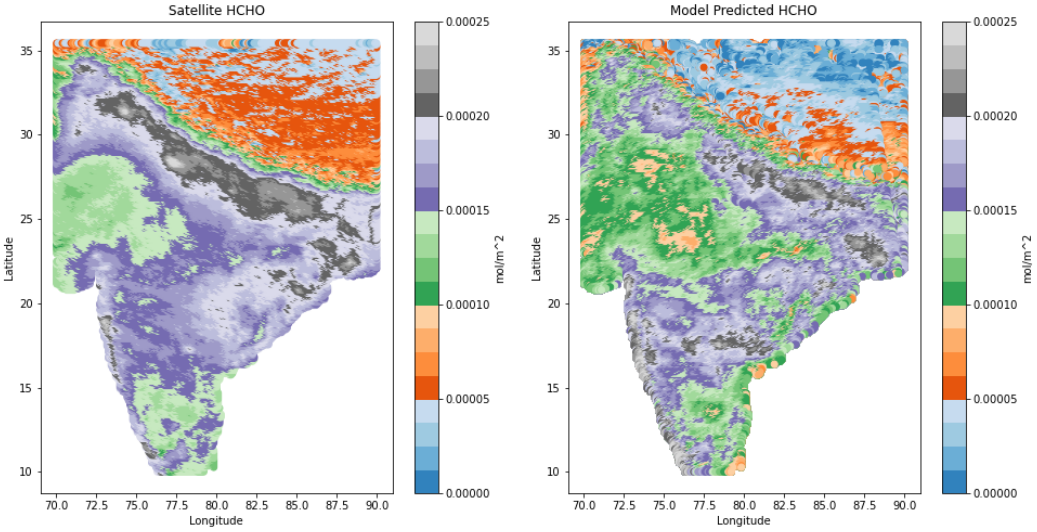


Figure 4.1: satellite and model measured formaldehyde.

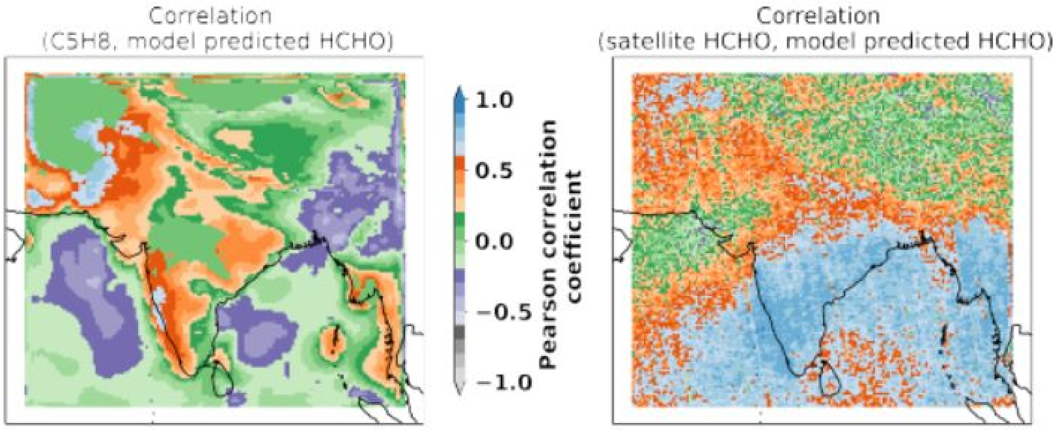

Correlation
(C5H8, model predicted HCHO)
Correlation
(satellite HCHO, model predicted HCHO)
1.0
0.5
0.0
−0.5
−1.0
Pearson correlation coefficient

Figure 4.2: Correlation between modelled formaldehyde and isoprene (left side) and satellite measured formaldehyde (right side) with SRCNN model

If we compare the result from deep learning-based architecture SRCNN models result with XGBoost we can see clearly that xgboost is giving far better results.

Variable combinations	Pearson correlation coefficient over India
meteorology + c_5h_8	0.30
only meteorology	0.21
meteorology + voc	0.23
meteorology + lai	0.14
meteorology + ch_4	-0.03
meteorology + $ch_4 + c_5h_8$ +lai + voc	0.08
meteorology + $ch_4 + c_5h_8$ + voc	-0.23
meteorology + ch_4 +c_5h_8	-0.50
meteorology + $ch_4 + c_5h_8$ + lai	-0.39

Table 4.2: Pearson's correlation coefficient values w.r.t. different variable combinations.

There is one more plot, Figure 4.4 which shows the average of every grid point of six-month data and as we can see actual and predicted are very much similar.

Variable	Pearson correlation coefficient	R$\hat{2}$ Value
meteorology	0.71	0.504

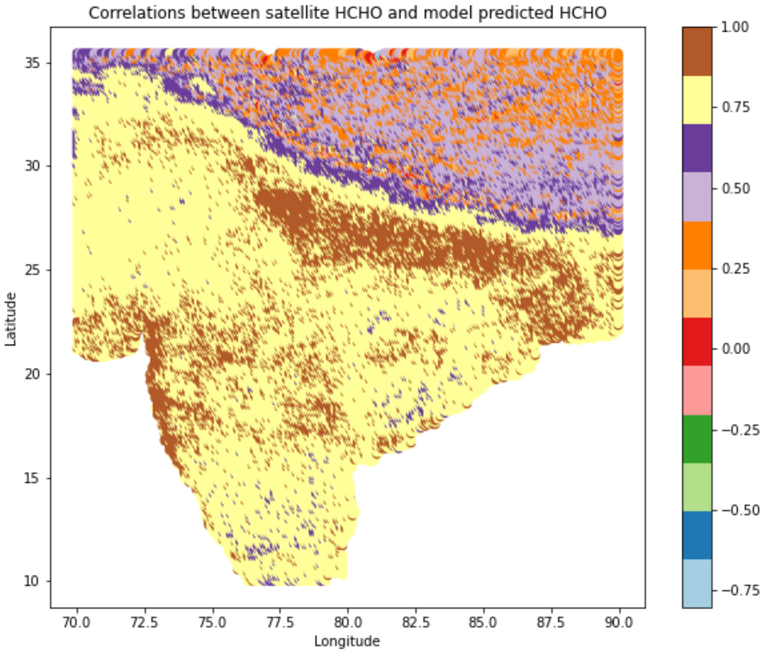


Figure 4.3: Pearson correlation coefficient between modelled and satellite measured formaldehyde.

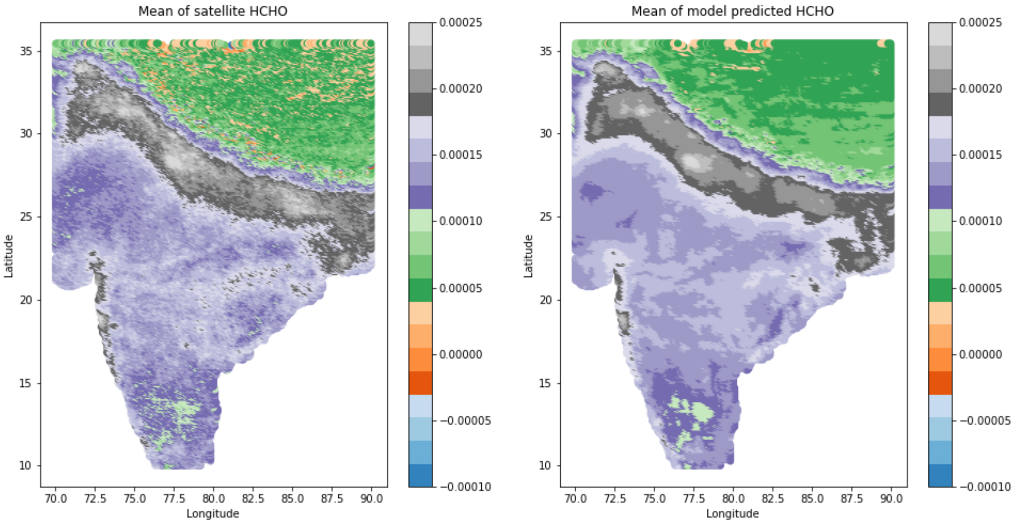


Figure 4.4: Mean of satellite HCHO and Mean of model predicted HCHO.

4.2 Conclusions

The results clearly show that the correlation varies a lot for the input variables. Based on the region there is a change in the value of correlation. We can see some regions have a very good correlation of above 0.80. It can be because of so many factors that affect HCHO. We can also see the mean of both actual and predicted are very much similar. The results of simulations that meteorology is the most promising have a global correlation value of 0.71.

In the Indian region, our work reveals the tremendous potential of ML modelling for computationally cheap simulations of HCHO fluctuation. The periodicity in HCHO and meteorological parameters resulting from India's systematic seasonal cycle tends to make it possible for ML models to reasonably accurately recreate data. The effects of HCHO on health and agriculture in this area may be evaluated using ML simulations in the absence of high-resolution observations. The simulations performed here might also be used as a paradigm for future uses of AI/ML-based modelling to supplement traditional Earth system models. Further work is required on the model to make it more efficient.

4.3 Future Outlook

- We will work Globally similarly to India with the same set of variables and with the same Model.
- To improve the performance of the model further we can replace certain variables which have less impact on HCHO. And similarly, add new variables.
- We will use high-resolution data for good results.
- We will try other models instead of XGBoost.
- Try some combination or hybrid models to get better performance.

Bibliography

[1] Jian Guan, Bohan Jin, Yizhe Ding, Wen Wang, Guoxiang Li, and Pubu Ciren. Global surface hcho distribution derived from satellite observations with neural networks technique. *Remote Sensing*, 13(20):4055, 2021.

[2] Anoop S Mahajan, Isabelle De Smedt, Mriganka Sekhar Biswas, Sachin Ghude, Suvarna Fadnavis, Chaitri Roy, and Michel van Roozendael. Inter-annual variations in satellite observations of nitrogen dioxide and formaldehyde over india. *Atmospheric Environment*, 116:194–201, 2015.

[3] Antje Inness, Melanie Ades, Anna Agusti-Panareda, J erˆome Barr e, Anna Benedictow, Anne-Marlene Blech-schmidt, Juan Jose Dominguez, Richard Engelen, Henk Eskes, Johannes Flemming, et al. The cams reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, 19(6):3515–3556, 2019.

[4] Luke Surl, Paul I Palmer, and Gonzalo Gonz alez Abad. Which processes drive observed variations of hcho columns over india? *Atmospheric Chemistry and Physics*, 18(7):4549–4566, 2018.

[5] T Stavrakou, J-F M uller, I De Smedt, M Van Roozendael, GR Van Der Werf, L Giglio, and A Guenther. Evaluating the performance of pyrogenic and biogenic emission inventories against one decade of space-based formaldehyde columns. *Atmospheric Chemistry and Physics*, 9(3):1037.1060, 2009.

[6] Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *arXiv:1603-02754*, 2016.

[7] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv:1907.10902*, 2019.

[8] Mriganka Sekhar Biswas, G Pandithurai, MY Aslam, Rohit D Patil, V Anilkumar, Shrikant D Dudhambe, Christophe Lerot, Isabelle De Smedt, Michel Van Roozendael, Anoop S Mahajan, et al. Effect of boundary layer evolution on nitrogen dioxide (NO2) and formaldehyde (HCHO) concentrations at a high-altitude observatory in western India. *Aerosol and Air Quality Research*, 21(3):200193, 2021.

[9] I De Smedt, T Stavrakou, J-F Muller, RJ Van Der A, and M Van Roozendael. Trend detection in satellite observations of formaldehyde tropospheric columns. *Geophysical Research Letters*, 37(18), 2010.

[10] Narendra Ojha, Imran Girach, Kiran Sharma, Amit Sharma, Narendra Singh, and Sachin S Gunthe. Exploring the potential of machine learning for simulations of urban ozone variability.*Scientific reports*, 11(1):1–7, 2021.

[11] Meinrat O Andreae and Pedro Merlet. Emission of trace gases and aerosols from biomass burning. Global biogeochemical cycles, 15(4):955–966, 2001.

[12] A Balzarini, G Pirovano, L Honzak, R Zabkar, G Curci, R Forkel, M Hirtl, R San Jose, P Tuccella, and GA Grell. Wrf-chem model sensitivity to chemical mechanisms choice in reconstructing aerosol optical properties. Atmospheric Environment, 115:604–619, 2015.

[13] Mriganka Sekhar Biswas, Anoop S Mahajan, et al. Year-long concurrent max-does observations of nitrogen dioxide and formaldehyde at Pune: Understanding diurnal and seasonal variation drivers. Aerosol and Air Quality Research, 21(6):200524, 2021.

[14] J Brown, C Bowman, et al. Integrated science assessment for ozone and related photochemical oxidants. Washington, DC: US Environmental Protection Agency, 2013.

[15] William PL Carter and Roger Atkinson. An experimental study of incremental hydrocarbon reactivity. Environmental science technology, 21(7):670–679, 1987.

[16] Lei Zhu, Loretta J Mickley, Daniel J Jacob, Elo¨ıse A Marais, Jianxiong Sheng, Lu Hu, Gonzalo Gonzalez Abad, and Kelly Chance. Long-term (2005–2014) trends in formaldehyde (HCHO) columns across North America as seen by the OMI satellite instrument: Evidence of changing emissions of volatile organic compounds. Geophysical Research Letters, 44(13):7079–7086, 2017.

[17] Yang Zhang, Yaosheng Chen, Golam Sarwar, and Kenneth Schere. Impact of gas-phase mechanisms on weather research forecasting model with chemistry (WRF/Chem) predictions: Mechanism implementation and comparative evaluation. Journal of Geophysical Research: Atmospheres, 117(D1), 2012.

[18] GM Wolfe, J Kaiser, TF Hanisco, FN Keutsch, JA De Gouw, JB Gilman, M Graus, CD Hatch, J Holloway, LW Horowitz, et al. Formaldehyde production from isoprene oxidation across NOx regimes. Atmospheric chemistry and physics, 16(4):2597–2610, 2016.

[19] David Simpson, Alex Guenther, C Nicholas Hewitt, and Rainer Steinbrecher. Biogenic emissions in Europe: 1. Estimates and uncertainties. Journal of Geophysical Research: Atmospheres, 100(D11):22875–22890, 1995.

[20] Paul I Palmer, Daniel J Jacob, Arlene M Fiore, Randall V Martin, Kelly Chance, and Thomas P Kurosu. Mapping isoprene emissions over North America using formaldehyde column observations from space. Journal of Geophysical Research: Atmospheres, 108(D6), 2003.