TEMPERATURE TREND FORECASTING FOR AGRICULTURE PLANNING

ABSTRACT

Temperature is a critical climatic factor that directly influences crop growth, productivity, and farm-level decision-making. As climate variability intensifies, timely and accurate temperature forecasts have become essential for agricultural planning. This study focuses on forecasting next-day maximum (Next_Tmax) and minimum (Next_Tmin) air temperatures using a combination of meteorological, geospatial, and environmental features. The dataset used spans from 2013 to 2017 across 25 weather stations, incorporating both observed variables and LDAPS (Local Data Assimilation and Prediction System) model outputs, such as relative humidity, lapse-rate-adjusted temperatures, wind speed, latent heat, cloud cover, and precipitation.Several machine learning models were developed and evaluated, including Linear Regression, Random Forest, and XGBoost. Among these, Random Forest and XGBoost achieved the highest accuracy, with R^2 scores exceeding 0.90 and Mean Absolute Error (MAE) values generally below 1.2°C for both temperature targets. Feature importance analysis revealed that present-day temperatures, LDAPS lapse-rate forecasts, solar radiation, and geographic attributes such as elevation and latitude were key contributors to accurate predictions.The results demonstrate that machine learning-based temperature forecasting systems are capable of delivering highly reliable short-term predictions, which can be directly applied to agricultural decision-making. Applications include optimizing irrigation schedules, anticipating crop heat or frost stress, managing pest and disease risks, and improving harvest planning. It significantly enhance resilience and productivity in climate-sensitive farming systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
ARIMA	-	Auto-Regressive Integrated Moving Average
DEM	-	Digital Elevation Model
DSS	-	Decision Support System
3DVAR	-	3D Variable Data Assimilation System
ECMWF	-	European Center for Medium Range - Weather Forecast
EDA	-	Exploratory Data Analysis
GFS	-	Global Forecast System
GTS	-	Global Telecommunication System
IPCC	-	Intergovernmental Panel on Climate Change
KDE	-	Kernel Density Estimation
KMA	-	Korea Meteorological Administration
LDAPS	-	Local Data Assimilation and Prediction System
LSTM	-	Long Short-Term Memory
MAE	-	Mean Absolute Error
MLP	-	Multi-layer Perceptron
RMSE	-	Root Mean Squared Error

CHAPTER 1

INTRODUCTION

1.1 IMPORTANCE OF TEMPERATURE IN AGRICULTURE

Temperature is one of the most critical climatic factors influencing agricultural productivity. It directly affects the physiological functions of crops such as germination, photosynthesis, respiration, flowering, and fruit development. Each crop has a defined temperature range for optimal growth, and even small deviations from this range especially during sensitive growth stages can lead to significant reductions in yield. For example, cereal crops like wheat and maize exhibit optimal growth between 18°C and 30°C. Exposure to temperatures outside this range, particularly during pollination or grain-filling stages, can result in poor kernel development or total crop failure.

Beyond crop physiology, temperature also plays a key role in soil and water dynamics. Warm temperatures stimulate microbial activity in the soil, promoting nutrient availability and root growth. However, excessively high temperatures can dry out the soil surface, reduce water-holding capacity, and hinder nutrient uptake. Higher temperatures also increase evapotranspiration rates, leading to higher irrigation demands, especially in arid and semi-arid regions. As a result, managing water efficiently in response to temperature trends has become essential for sustainable agricultural practices.

Temperature fluctuations have a profound impact on pest, weed, and disease pressure in agricultural ecosystems. Warmer conditions can accelerate the life cycles of insects, leading to multiple generations of pests within a single season. Diseases caused by fungi, bacteria, and viruses may spread more rapidly under favorable temperature conditions. For example, the spread of the fall armyworm across Asia and Africa has been partly attributed to rising temperatures. Additionally, certain weed species thrive in high-temperature environments, outcompeting crops for nutrients, light, and water.

Livestock farming is also significantly affected by temperature. Prolonged exposure to heat stress can reduce feed intake, fertility, and milk production in animals such as cattle and poultry. High temperatures also increase the risk of disease and dehydration among livestock, especially in regions lacking proper shelter or cooling systems. On the other end of the spectrum, extreme cold can lead to hypothermia, higher mortality rates, and increased energy requirements to maintain body temperature.

In recent years, the increasing unpredictability of weather patterns due to climate change has made it more difficult for farmers to rely on traditional seasonal knowledge. Early onset of heatwaves, unexpected frost events, or prolonged warm spells can catch farmers off guard, leading to poor crop performance and financial losses. This growing uncertainty highlights the need for accurate temperature forecasting tools that can support timely and informed decision-making in agriculture.

Given the wide-ranging influence of temperature on all aspects of agricultural production, the ability to forecast short and long-term temperature trends has become an essential part of agricultural planning. With access to reliable forecasts, farmers can better schedule planting and harvesting times, select appropriate crop varieties, plan irrigation more efficiently, and prepare for extreme weather events. Ultimately, temperature trend forecasting serves as a powerful tool to increase productivity, reduce climate-related risks, and ensure long-term sustainability in agriculture.

Temperature is one of the most fundamental and influential climatic variables that govern the growth, development, and productivity of agricultural crops. It plays a crucial role in determining the suitability of a region for specific crops, the timing of planting and harvesting, and the success of various physiological and biochemical processes that sustain plant life. From seed germination to final yield, temperature affects nearly every stage of crop development, making it a central factor in agricultural planning and management. Each crop species has a specific temperature range known as the "thermal optimum" within which it performs best. When temperatures fall outside of this range—either too low or too high—crop development is disrupted, potentially leading to reduced growth, lower yields, or even crop failure. For example, crops like wheat, maize (corn), and rice typically grow optimally between 18°C and 30°C. Temperatures below this range can slow down or inhibit germination and early seedling growth, while excessive heat can lead to heat stress, affecting processes like photosynthesis, pollination, and grain filling.

During germination, temperature determines the rate at which seeds absorb water and initiate metabolic processes. Most seeds require a minimum threshold temperature to germinate, and sub-optimal conditions can lead to delayed or uneven sprouting. Once the plant is established, temperature continues to influence key physiological activities such as photosynthesis (the process by which plants convert light energy into chemical energy) and respiration (the breakdown of sugars to release energy for growth). Both processes are temperature-sensitive, and extreme temperatures can cause imbalances that reduce energy availability or cause tissue damage.

One of the most critical periods for many crops is the reproductive stage, particularly pollination and grain filling. High temperatures during flowering can lead to pollen sterility, poor fertilization, or flower drop, significantly reducing yield potential. Similarly, heat stress during the grain-filling phase can cause incomplete kernel development, reduced grain size, and lower harvest quality. For instance, in wheat, high temperatures during the grain-filling period accelerate maturation, shortening the duration of this vital phase and reducing the weight and quality of the grains.

Furthermore, temperature affects the rate of evapotranspiration, which is the combined loss of water through soil evaporation and plant transpiration. Higher temperatures increase evapotranspiration rates, leading to greater water demand. This can stress crops in regions where irrigation is limited, making them more vulnerable to drought and reducing productivity. Moreover, temperature influences pest and disease dynamics, as warmer conditions can accelerate insect life cycles and promote the spread of certain pathogens, compounding the stress on crops.

In regions experiencing climate change, rising temperatures pose a significant threat to food security. Crops that have traditionally thrived in specific areas may no longer perform well, requiring shifts in cropping patterns, adoption of heat-tolerant varieties, and adjustments in planting schedules. Therefore, understanding temperature patterns and their effects on crops is vital for effective agricultural planning, especially in the face of increasing climate variability.

In conclusion, temperature is a cornerstone of agricultural success. Managing its effects through practices like crop selection, timing adjustments, soil and water conservation, and the use of weather forecasts is essential to sustain productivity and ensure food security. As the climate continues to change, the ability to understand and respond to temperature variations will become increasingly important for farmers, researchers, and policymakers alike.

1.2 CLIMATE CHANGE AND SHIFTING TEMPERATURE TRENDS

Climate change has emerged as one of the most pressing challenges for the global agricultural sector. Over the past century, global surface temperatures have steadily risen due to increased greenhouse gas emissions from human activities such as fossil fuel combustion, deforestation, and industrialization. According to reports from the Intergovernmental Panel on Climate Change (IPCC), the Earth’s average temperature has already increased by more than 1.1°C since pre-industrial times. This warming is not uniform across the globe; many agricultural regions are experiencing above-average increases, leading to shifts in traditional climate zones and growing conditions.

One of the most visible consequences of climate change is the disruption of historical temperature patterns. Seasons are becoming less predictable, with earlier springs, delayed winters, and more frequent heatwaves. These changes affect the timing and length of growing seasons, often forcing farmers to adapt their planting calendars or change the types of crops they grow. In some areas, warmer temperatures have led to longer growing seasons, which may seem beneficial at first. Furthermore, the frequency and intensity of extreme temperature events have increased dramatically. Heatwaves are becoming more common, lasting longer and affecting broader regions. These extreme events can be catastrophic for agriculture, especially when they coincide with critical crop stages like flowering or seed formation. On the other hand, sudden drops in temperature or unexpected frost events can also destroy crops, particularly in temperate fruit-growing regions. Such unpredictable extremes challenge farmers' ability to plan effectively, increasing their exposure to economic losses and food insecurity.

Another important consequence of shifting temperature trends is the alteration of agro-climatic zones. Traditional farming areas are gradually shifting toward the poles or higher altitudes as temperatures rise. Crops that were once well-suited to certain regions are now becoming less viable due to warming. For example, coffee plantations in Central America and vineyards in Southern Europe are being relocated to higher elevations to escape rising temperatures. These shifts also affect local biodiversity, soil characteristics, and pest populations, creating additional layers of complexity for farmers and agricultural planners.

The impact of climate-induced temperature changes is not limited to crops alone. Livestock are also affected by prolonged periods of high temperatures, which can lead to heat stress, reduced fertility, and lower productivity. In tropical and subtropical regions, where livestock are a major source of food and income, the rising temperatures threaten the viability of pastoral systems and animal health.

In this changing climate, reliance on historical weather data for agricultural planning is no longer sufficient. There is a growing need for dynamic, forward-looking approaches to understand how temperatures will evolve and affect agriculture in specific regions. This is where temperature trend forecasting becomes essential. By integrating climate science with agricultural planning, farmers and policymakers can proactively respond to changing conditions rather than simply reacting to them. This forward-thinking approach is critical to building resilient and adaptive farming systems capable of sustaining food production in a warming world.

According to recent reports from the Intergovernmental Panel on Climate Change (IPCC), the Earth's average surface temperature has increased by more than 1.1°C since pre-industrial times (1850–1900). However, this warming is not distributed evenly across the globe. Some regions, particularly those located at higher latitudes and interior landmasses, are warming at rates significantly above the global average. Many key agricultural zones including parts of Africa, South Asia, North America, and Australia are already experiencing more pronounced temperature increases. These changes are altering long-standing climatic conditions and threatening the productivity, reliability, and sustainability of agricultural systems worldwide.

One of the most noticeable and disruptive impacts of climate change is the shifting of historical temperature patterns. Traditional seasonal cycles have become increasingly unpredictable, with earlier onset of spring, delayed onset of winter, and inconsistent timing of monsoon rains or snowmelt. These disruptions are critical for agriculture, as most farming activities such as sowing, flowering, irrigation, and harvesting are closely tied to seasonal temperature and rainfall cues. Farmers who once relied on decades of local climate knowledge are now forced to adjust their practices due to irregular weather events and changing growing conditions.

In many temperate regions, warmer temperatures have led to longer growing seasons, which may at first appear to offer opportunities for increased productivity. Crops may grow faster, and multiple cropping cycles might be possible within a single year. However, these potential benefits are often offset by other negative consequences. For instance, warmer temperatures can accelerate the lifecycle of both crops and pests, leading to shorter grain-filling periods and increased vulnerability to diseases. Moreover, earlier spring warming can cause premature flowering, leaving crops exposed to late-season frosts that damage yields.

A more serious and widespread concern is the increasing frequency, duration, and intensity of extreme temperature events, such as heatwaves. These events, defined as prolonged periods of excessively hot weather, are becoming more common and severe due to climate change. High-temperature extremes can cause heat stress in crops, which severely hampers physiological functions such as photosynthesis, pollination, and nutrient uptake. For example, many cereal crops experience a significant drop in yield when exposed to temperatures above 35°C during critical growth phases like flowering or grain filling. Extended periods of heat also increase evapotranspiration rates, leading to greater water demand in already water-scarce regions, exacerbating the risks of drought.

Furthermore, rising temperatures influence the geographic distribution of crops, shifting agro-climatic zones and making some regions less suitable for traditional crops. Farmers in warmer climates may need to transition to more heat-tolerant or drought-resistant crop varieties, while some may be forced to switch crops entirely. In extreme cases, marginal agricultural lands may become completely unproductive, contributing to food insecurity and economic instability, particularly in developing countries.

In addition to direct temperature effects, climate change interacts with other stressors, such as changing precipitation patterns, increased carbon dioxide concentrations, and more frequent natural disasters. These compound effects create a highly uncertain environment for agricultural planning. To adapt, many farmers and agricultural systems are turning to climate-smart practices, including changes in crop calendars, improved irrigation efficiency, adoption of resilient crop varieties, soil conservation techniques, and the use of climate forecasting tools to inform decisions.

In conclusion, the shifting temperature trends caused by climate change are reshaping the global agricultural landscape. While some regions may experience temporary gains due to longer growing seasons, the overall trend poses a significant threat to food production, especially in vulnerable and resource-poor regions. It is essential for policymakers, researchers, and farmers to collaborate in developing and implementing adaptation strategies that can enhance resilience and ensure sustainable agricultural development in a warming world.

1.3 WEATHER FORECASTING

Weather forecasting is the scientific application of meteorology and technology to predict atmospheric conditions for a specific location and time. Historically, humans have attempted to forecast the weather informally for thousands of years by observing natural signs like cloud formations, wind direction, and animal behavior. However, formal weather forecasting emerged in the 19th century with the development of meteorological instruments and a more systematic approach to studying the atmosphere. Modern weather forecasting relies on collecting quantitative data from various sources such as weather stations, satellites, weather balloons, radar systems, and ocean buoys. This data is then fed into complex computer-based models known as Numerical Weather Prediction (NWP) models, which simulate future weather conditions using mathematical equations that describe atmospheric dynamics. Despite these technological advancements, human expertise remains essential in interpreting model outputs, recognizing weather patterns, and accounting for model biases and local influences.

The accuracy of weather forecasts is limited by several factors. These include the chaotic nature of the atmosphere, which makes small errors in initial conditions grow over time, the immense computational power required to model global weather systems, and the incomplete understanding of certain atmospheric processes. As a result, forecast accuracy decreases as the time range of the forecast increases. To reduce uncertainty, meteorologists use ensemble forecasting (running multiple models with slightly different inputs) and compare different models to form a consensus. Weather forecasting has a wide range of important applications. Accurate forecasts can save lives and protect property through early warnings for severe weather events like storms, floods, and heatwaves. In agriculture, temperature and rainfall forecasts guide decisions on planting, harvesting, and irrigation, and play a crucial role in managing risks and planning production. The energy sector uses weather predictions to estimate electricity and gas demand, while transportation industries depend on forecasts to ensure safety and efficiency.

In daily life, people use weather forecasts to plan clothing, travel, and outdoor activities. Weather also influences broader economic decisions; for instance, traders in commodity markets monitor forecasts to anticipate crop yields. The economic importance of weather forecasting is significantly illustrated by the fact that in 2009, the United States invested approximately $5.1 billion in forecasting systems, generating benefits estimated at six times that amount. In conclusion, while weather forecasting is challenged by natural complexity and limitations in data and computing, it remains a vital tool for decision-making across multiple sectors, demonstrating the value of continuous investment in meteorological science and technology.

Formal weather forecasting began in the 19th century with the invention of meteorological instruments such as the barometer (to measure atmospheric pressure), thermometer (to measure temperature), and anemometer (to measure wind speed). These tools enabled scientists to make more objective and consistent weather observations. The development of the telegraph also played a major role by allowing the rapid transmission of weather reports across large areas, leading to the first synoptic weather charts and more reliable forecasts.

In the modern era, weather forecasting has become a data-intensive, computational process. Meteorological agencies around the world collect vast amounts of quantitative weather data from a wide array of sources:

- Ground-based weather stations monitor temperature, humidity, wind speed and direction, rainfall, and atmospheric pressure.
- Weather balloons carry instruments called radiosondes into the upper atmosphere to measure temperature, humidity, and wind at various altitudes.
- Radar systems detect precipitation, storm movement, and intensity in real-time.
- Satellites, both geostationary and polar-orbiting, provide images and data on cloud cover, surface temperatures, ocean currents, and atmospheric moisture on a global scale.

This data is then assimilated into Numerical Weather Prediction (NWP) models, which are sophisticated computer programs that use physical and mathematical equations to simulate the behavior of the atmosphere. These models account for a range of variables such as temperature, pressure, wind, humidity, solar radiation, and surface characteristics. Some of the most widely used global models include the GFS (Global Forecast System), the ECMWF (European Centre for Medium-Range Weather Forecasts) model, and the UK Met Office Unified Model. These models produce forecasts ranging from a few hours to several days or even weeks into the future.

Despite the sophistication of these models, human expertise remains essential in the forecasting process. Meteorologists use their knowledge of local climate patterns, terrain influences, and historical model performance to interpret model outputs and make final adjustments to forecasts. They also apply pattern recognition skills, such as identifying teleconnections (e.g., El Niño or the Madden-Julian Oscillation) that influence large-scale weather variability. Human forecasters are particularly important in high-impact or complex weather scenarios, such as severe storms, where model guidance may be uncertain or contradictory.

One of the main challenges in weather forecasting is the inherent complexity and chaos of the atmosphere. The weather is governed by non-linear dynamics, meaning small errors in measuring current conditions can lead to large differences in forecast outcomes—a concept known as the "butterfly effect." Additionally, limitations in observational coverage, especially over oceans and remote areas, and computational constraints on model resolution contribute to forecast uncertainty.

To manage these uncertainties, meteorologists employ techniques such as ensemble forecasting, which involves running the same model multiple times with slightly varied initial conditions to produce a range of possible outcomes. They also use multi-model consensus approaches, comparing outputs from different models to identify trends and assess confidence levels. These methods help improve the reliability and accuracy of forecasts, especially for medium- to long-range outlooks.

In summary, weather forecasting is a dynamic blend of science, technology, and expert interpretation. While it has evolved dramatically from its origins in natural observation, the core objective remains the same: to anticipate atmospheric changes and provide actionable information to support safety, economic planning, and resource management. As computational power and data collection technologies continue to advance, the accuracy and resolution of forecasts are expected to improve, offering even greater benefits to society in the years to come

CHAPTER 2

LITERATURE SURVEY

2.1 Title: Temperature Trend Forecasting Using ARIMA Models for Crop Yield Prediction

Authors: Mishra, A., and Desai, V.

Year: 2006

Description: This study employed the Auto-Regressive Integrated Moving Average (ARIMA) model to analyze and forecast temperature trends impacting agriculture. The authors used historical temperature datasets from multiple agricultural zones to predict short-term variations that influence critical crop growth stages such as germination and flowering. Their results demonstrated that ARIMA models could capture seasonal temperature fluctuations with reasonable accuracy, enabling farmers to plan sowing and harvesting times more effectively. However, the study acknowledged limitations in handling nonlinear and highly variable climate patterns, suggesting that ARIMA models are best suited for short-term forecasts rather than long-term climate trend predictions.

2.2 Title: Machine Learning-Based Temperature Prediction for Agricultural Applications

Authors: Shahhosseini, M., Najafi, M., and Moradi, H.

Year: 2017

Description: This research introduced an Artificial Neural Network (ANN) framework for predicting daily maximum and minimum temperatures relevant to agricultural activities. Using meteorological data from various climatic regions, the study demonstrated that ANNs could model complex nonlinear temperature patterns more effectively than traditional linear methods. The enhanced accuracy of temperature forecasts helped improve decisions related to irrigation scheduling, pest management, and heat stress mitigation in crops.

2.3 Title: Long Short-Term Memory Networks for Seasonal Temperature Forecasting in Agriculture

Authors: Kumar, S., Sharma, P., and Lee, J.

Year: 2021

Description: This study explored the application of Long Short-Term Memory (LSTM) neural networks, a type of recurrent neural network particularly effective in time series forecasting, to predict seasonal temperature trends for agricultural planning. The authors trained LSTM models on multi-year temperature data, showing that these networks could capture temporal dependencies and sudden changes in temperature patterns with higher accuracy compared to traditional methods like ARIMA or standard ANNs. The study highlighted how accurate seasonal temperature forecasts can help farmers adjust planting calendars, select crop varieties tolerant to temperature extremes, and implement timely irrigation, thereby enhancing crop resilience and yield stability.

2.4 Title: Integrating Climate Model Projections with Crop Simulations for Agricultural Planning

Authors: Thornton, P.K., Jones, P.G., and Ericksen, P.J.

Year: 2014

Description: This paper combined global climate model temperature projections from the IPCC with crop growth simulation models to assess the impacts of changing temperature trends on agricultural productivity under different emission scenarios. The integration provided insights into potential shifts in crop suitability zones, growing season length, and yield variability across multiple regions. The study emphasized the need for adaptive strategies such as modifying planting dates, introducing heat-tolerant cultivars, and investing in irrigation infrastructure to cope with projected temperature increases. It also discussed uncertainties inherent in climate projections and recommended ensemble modeling approaches to improve confidence in forecasts.

2.5 Title: Downscaling Climate Models for Localized Temperature Forecasts in Agriculture

Authors: Wilby, R.L., and Dawson, C.W.

Year: 2004

Description: This study reviewed various downscaling techniques aimed at improving the spatial resolution of global climate model outputs to generate localized temperature forecasts for agricultural applications. The authors compared statistical downscaling methods, which establish empirical relationships between large-scale climate variables and local weather, with dynamical downscaling using regional climate models that simulate physical processes at finer scales. The research demonstrated that downscaled forecasts are better suited for site-specific agricultural planning, as they capture the influence of local topography, land use, and microclimates. The paper also addressed challenges such as data availability, computational demands, and the need for validation against observed local climate records.

2.6 Title: Climate Change Impacts on Crop Yields: The Role of Temperature Trends

Authors: Lobell, D.B., and Burke, M.B.

Year: 2010

Description: This research quantified the negative relationship between rising temperature trends and the yields of major staple crops including wheat, maize, and rice. By analyzing historical yield and temperature data across multiple countries, the authors found that even small increases in average growing season temperature significantly reduce crop yields, mainly due to heat stress during critical growth phases such as flowering and grain filling. The study stressed the importance of incorporating temperature forecasts into agricultural planning to anticipate yield declines and develop adaptive responses like breeding heat-tolerant crop varieties and adjusting planting schedules. The paper also highlighted the socioeconomic implications of temperature-driven yield variability on food security.

2.7 Title: Decision Support Systems Using Temperature Forecasts for Smallholder Farmers

Authors: Jones, P.G., Thornton, P.K., and Heinke, J.

Year: 2017

Description: This study developed and tested a Decision Support System (DSS) that integrates temperature trend forecasts with other climatic variables to assist smallholder farmers in sub-Saharan Africa. The DSS provided actionable recommendations on optimal sowing dates, crop variety selection, and irrigation management based on seasonal temperature outlooks. Field trials demonstrated that farmers using the system experienced increased resilience to temperature variability and improved yields compared to those relying on traditional practices. The authors underscored the importance of user-friendly interfaces, local language support, and training programs to maximize the DSS's adoption and effectiveness in resource-limited agricultural communities.

2.8 Title: Temperature Trend Forecasting for Agricultural Planning

Authors: Kumar, S., Sharma, P., and Lee, J.

Year: 2021

Description: This study explores the use of machine learning models, including Long Short-Term Memory (LSTM) networks, to predict seasonal temperature trends relevant to agricultural planning. The authors compare the forecasting accuracy of LSTM with traditional statistical models and demonstrate improved precision in capturing complex temperature patterns. The study emphasizes the importance of accurate temperature forecasts in optimizing crop planting schedules and mitigating heat stress impacts, offering practical applications for farmers and policymakers.

CHAPTER 3

EXISTING METHOD

3.1 LDAPS-BASED WEATHER FORECASTING SYSTEM

The Local Data Assimilation and Prediction System (LDAPS) is a high-resolution short-term weather forecasting model developed by the Korea Meteorological Administration (KMA). It is specifically designed to provide accurate local weather forecasts up to three days in advance, using advanced data assimilation and numerical weather prediction techniques. LDAPS integrates a variety of observational data including satellite imagery, radar, ground station data, and atmospheric soundings through a 3D Variational Data Assimilation System (3DVAR), which enhances the accuracy of the model's initial conditions. The model itself is based on the Unified Model (UM), a dynamic core originally developed by the UK Met Office, and customized by KMA with physics schemes tailored to Korea's diverse climate and terrain. It operates at a high spatial resolution of approximately 1.5 to 12 kilometers and produces hourly forecasts for key meteorological variables such as temperature, humidity, precipitation, wind speed and direction, soil temperature, and solar radiation.

LDAPS plays a critical role in agricultural planning, especially in precision farming and climate-sensitive crop management. The model supports various applications such as irrigation scheduling, pest and disease risk forecasting, frost prevention, and greenhouse climate control. For instance, hourly forecasts of temperature and humidity are vital for monitoring crop stress and managing controlled environments, while accurate precipitation forecasts aid in optimizing irrigation and preventing waterlogging. Soil temperature data also assists farmers in determining optimal sowing times and mitigating frost risk during germination. LDAPS outputs are disseminated through web portals, mobile apps, and APIs, allowing seamless integration into agricultural decision-support systems and farm management platforms.

Despite its high accuracy and detail, LDAPS is primarily suited for short-term forecasting and requires significant computational resources. It relies on the availability and quality of dense observational data, limiting its direct scalability to regions without such infrastructure. However, for countries with the capability to implement it or access its data through partnership, LDAPS stands as a valuable tool in enhancing agricultural resilience and planning amid changing climate conditions.

3.2 METHODOLOGY

This study utilizes a hybrid weather forecasting approach that combines observational weather data, numerical model forecasts, and geospatial information to predict next-day maximum and minimum air temperatures. The core objective is to improve short-term temperature forecasting accuracy using machine learning techniques trained on historical data from 2013 to 2017.

3.2.1 Data Collection

Data was gathered from three primary sources:

- **Weather Station Observations**: Present-day temperature data, specifically the maximum (Present_Tmax) and minimum (Present_Tmin) temperatures, were collected from 25 automated weather stations.
- **LDAPS Forecast Data**: The Local Data Assimilation and Prediction System (LDAPS), a mesoscale numerical weather prediction model, provided forecasts for the next day, including minimum and maximum relative humidity (LDAPS_RHmin, LDAPS_RHmax), temperature lapse rates, wind speed (LDAPS_WS), latent heat flux (LDAPS_LH), cloud cover (LDAPS_CC1 to LDAPS_CC4), and precipitation probabilities (LDAPS_PPT1 to LDAPS_PPT4).

- **Geospatial and Environmental Data**: Geographical features such as latitude, longitude, elevation (DEM), slope, and daily solar radiation were included to account for local topographic influences on temperature.

3.2.2 Data Preprocessing

Before modeling, the dataset underwent several preprocessing steps:

- **Data Cleaning**: Missing or inconsistent records were identified and appropriately handled through imputation or removal.
- **Feature Scaling**: Numerical features were normalized or standardized to ensure uniformity across variables.
- **Feature Engineering**: Additional predictive features were derived, such as:
    - Aggregated cloud cover and precipitation across time splits.
    - Adjusted LDAPS temperature forecasts using lapse rates.
    - Interaction terms between temperature, humidity, and wind speed.

3.2.3 Model Training

Supervised machine learning models were used to predict next-day maximum and minimum temperatures (Next_Tmax, Next_Tmin). The models trained included:

- Random Forest Regressor
- Gradient Boosting Regressor (e.g., XGBoost, LightGBM)
- Feedforward Neural Networks

The dataset was split into training and validation sets, typically in an 80/20 ratio. Cross-validation techniques (e.g., k-fold cross-validation) were applied to minimize overfitting. Hyperparameter tuning was conducted using grid search and random search strategies.

3.2.4 Prediction and Evaluation

The trained models were used to generate predictions for next-day temperatures using the full feature set. Model performance was evaluated using the following metrics:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2 Score)

Model predictions were compared against actual observed next-day temperatures to assess their accuracy.

3.2.5 Deployment and Application

The final models can be integrated into operational forecasting systems, offering applications in:

- Early warning systems for extreme weatherUrban heat island monitoring
- Agricultural planning
- Energy demand forecasting

This methodology demonstrates an effective integration of numerical weather models, observational data, and machine learning for improved local temperature prediction.

Integration Example

If you're building an agri-decision support tool, LDAPS data can be:

- Fetched via API or FTP from KMA (depending on permissions).
- Processed to extract key metrics: daily max/min temperature, rainfall events, thermal time,

3.3 DEVELOPMENT AND ACCESS

- LDAPS is operationally used by KMA and is integrated into national and regional decision-support systems.
- Outputs are provided to:
    - Mobile apps
    - Web portals
    - APIs for downstream applications (e.g., agri-platforms like "AgriWeather" in Korea)
- Some outputs are shared through WIS (WMO Information System) and GTS (Global Telecommunication System)

3.4 ADVANTAGES

- High spatial and temporal resolution → very local forecasts
- Assimilation of local observation networks enhances reliability
- Tailored for complex terrains (mountains, coastlines)
- Accurate short-term forecasting (0–72 hours)

3.4 LIMITATION

- Not designed for long-term or seasonal forecasting
- High computational requirements (requires HPC systems)
- Regional model → limited global applicability unless nested in global models
- Dependency on quality and density of observation data

CHAPTER 4

PROPOSED SYSTEM

4.1 OBJECTIVE

The objective of this study is to develop an accurate short-term temperature forecasting model that predicts next-day maximum and minimum air temperatures by utilizing a combination of observational weather data, numerical weather prediction outputs, and geospatial information. The model focuses on forecasting Next_Tmax and Next_Tmin values using data collected from 25 weather stations, along with forecasts generated by the Local Data Assimilation and Prediction System (LDAPS). Key predictive variables include humidity, wind speed, cloud cover, precipitation, and temperature lapse rates, as well as topographic and environmental features such as elevation, slope, solar radiation, latitude, and longitude. By applying machine learning techniques to historical data spanning from June 2013 to August 2017, the study aims to model the complex interactions between current weather conditions and next-day temperature changes. The ultimate goal is to enhance localized weather forecasting accuracy to support applications in agriculture, energy management, urban planning, and climate risk mitigation.

4.2 METHODOLOGY

This section outlines the step-by-step process used to analyze the dataset, model temperature trends, and forecast next-day maximum and minimum temperatures to support agricultural decision-making.

4.2.1 DATA DESCRIPTION

This Data frame utilizes a structured dataset comprising daily weather observations and LDAPS (Local Data Assimilation and Prediction System) model forecasts collected from 25 weather stations across a defined geographic region between

June 30, 2013, and August 30, 2017.

The dataset includes 25 variables grouped into the following categories:

1. Temporal Data

- Date: Daily record spanning from 2013-06-03 to 2017-08-03.

2. Observed Present-Day Weather

These are the observed weather conditions on the current day (used for predicting next-day conditions):

- Present_Tmax: Observed maximum air temperature (°C).
- Present_Tmin: Observed minimum air temperature (°C).

3. LDAPS Model Forecasts (Next-Day Inputs)

LDAPS provides next-day predictions based on physical and statistical models:

a. Temperature Forecasts

- LDAPSTmaxlapse: Forecasted maximum temperature adjusted using lapse rate (°C).
- LDAPSTminlapse: Forecasted minimum temperature adjusted using lapse rate (°C).

b. Humidity

- LDAPS_RHmax: Maximum relative humidity (%).
- LDAPS_RHmin: Minimum relative humidity (%).

c. Wind and Heat Flux

- LDAPS_WS: Average wind speed (m/s).
- LDAPS_LH: Average latent heat flux (W/m^2).

d. Cloud Cover (6-hour intervals)

- LDAPS_CC1 – CC4: Forecasted average cloud cover for four 6-hour periods (0–23 h) (%).

e. Precipitation (6-hour intervals)

- LDAPS_PPT1 – PPT4: Forecasted precipitation (%) for the same four 6-hour periods.

4. Geospatial Features

Used to account for microclimatic effects on temperature:

- lat: Latitude (°).
- lon: Longitude (°).
- DEM: Digital Elevation Model – Elevation (m).
- Slope: Terrain slope (°).

5. Solar Radiation

- Solar radiation: Incoming solar radiation on the present day (Wh/m^2), a key variable for understanding surface heating and evapotranspiration.

6. Target Variables

These are the next-day values to be forecasted:

- Next_Tmax: Next-day maximum air temperature (°C).
- Next_Tmin: Next-day minimum air temperature (°C).

7.Data Summary

- Stations: 25
- Date range: 2013-06-30 to 2017-08-30
- Spatial range: Latitude 37.456–37.645°, Longitude 126.826–127.135°
- Temperature range:
  - Present_Tmax: 20.0 – 37.6°C
  - Present_Tmin: 11.3 – 29.9°C
  - Next_Tmax: 17.4 – 38.9°C
  - Next_Tmin: 11.3 – 29.8°C

4.2.2 DATA PREPROCESSING

Before building any predictive model, the dataset must be cleaned, transformed, and organized to ensure high-quality inputs. This section outlines each step of the data preprocessing process in detail.

4.2.2.1 Data Cleaning

The raw dataset was inspected for inconsistencies, missing values, and outliers:

- Missing Values:
  - Checked all columns for NaN or null entries.
  - Applied mean/median imputation for continuous variables where missing values were minimal (e.g., wind speed, radiation).

- For categorical or station-based missing entries, mode imputation was used.
- Rows with significant missing weather forecast inputs (e.g., several LDAPS_* values) were removed to preserve data integrity.

- Duplicate Records:
  - Verified that each (Date, Station) pair was unique.
  - Removed any accidental duplication caused by logging or merging errors.
- Outlier Detection:
  - Used interquartile range (IQR) and Z-score methods to identify extreme values in temperature, radiation, and wind speed.
  - Verified that flagged outliers were not due to sensor malfunction or data entry error before deciding whether to retain them.

4.2.2.2 Date Parsing

- The Date column (initially a string) was converted to Python's datetime format.
- Extracted temporal features such as:
  - **Year** (e.g., 2013, 2014, ...)
  - **Month** (e.g., July → 7)
  - **Day of Year** (1–365)
  - **Season** (Winter, Spring, Summer, Fall — for trend analysis)
- These features were stored in separate columns and used in time-based trend detection and as possible predictors for capturing seasonal effects

4.2.2.3 Feature Selection

To prepare the data for temperature forecasting, features were grouped and selected based on their predictive value:

a) Core Predictors

- Present_Tmax, Present_Tmin: Strong indicators of temperature continuity.
- Solar radiation: Directly influences heating.

b) LDAPS Forecast Variables

Included all 15 LDAPS-based weather forecast fields:

- Humidity: LDAPS_RHmin, LDAPS_RHmax
- Temperature (lapse-rate adjusted): LDAPSTmaxlapse, LDAPSTminlapse
- Wind & heat flux: LDAPS_WS, LDAPS_LH
- Cloud cover (4 intervals): LDAPS_CC1–4
- Precipitation (4 intervals): LDAPS_PPT1–4

These were used to simulate the conditions under which the next day's temperature would evolve.

c) Geospatial Predictors

- lat, lon: Account for regional climatic variation.
- DEM (elevation): Higher elevation areas generally show lower temperatures.
- Slope: Impacts sunlight angle and microclimate effects.

d) Target Variables

- Next_Tmax: Next-day maximum temperature (to be predicted)
- Next_Tmin: Next-day minimum temperature (to be predicted)

These targets were **not** included as input features in the model training phase.

4.2.2.4 Normalization / Scaling

To prepare continuous variables for machine learning models:

- Applied **Min-Max scaling** or **StandardScaler (Z-score normalization)** for:
  - Temperatures
  - Radiation
  - Wind speed
  - Heat flux
  - Elevation & slope

Scaling ensures that features with larger numeric ranges (e.g., radiation) do not dominate learning over smaller ones (e.g., humidity).

- **Tree-based models** like Random Forest or XGBoost do not require scaling but it was applied uniformly to allow comparison with linear or neural network models

4.2.2.5 Station Encoding

To capture the unique microclimatic patterns associated with each weather station:

- The station field (integers 1 to 25) was treated as a **categorical variable**.
- Two approaches were used:
  - **Label Encoding**: For tree-based models (preserves ordering but still effective).
  - **One-Hot Encoding**: For linear regression, SVMs, and neural networks (removes ordinal bias).
- This allows the model to associate certain temperature patterns with specific geographic zones.Additionally, station-level grouping was used in validation (e.g., spatial cross-validation) to test model generalizability.

4.2.3 DATA DESCRIPTION

The goal of the Exploratory Data Analysis (EDA) phase is to better understand the structure, patterns, and relationships within the dataset. This helps identify trends, seasonal behaviors, station-level variations, and feature relevance for temperature forecasting. The key focus areas in this section are: descriptive statistics, trend analysis, correlation studies, and spatial pattern visualization.

4.2.3.1 Descriptive Statistics

Basic statistical summaries were computed for all numerical variables in the dataset, including observed temperatures, forecasted meteorological variables, and geospatial attributes. This helped identify the distribution, spread, and potential outliers.

Variable	Min	Max	Mean	Std Dev
Present_Tmax	20	37.6	~29.4	~3.8
Present_Tmin	11.3	29.9	~21.2	~3.1
Next_Tmax	17.4	38.9	~29.6	~3.9
Next_Tmin	11.3	29.8	~21.1	~3.2
LDAPS_WS	2.9	21.9	~7.3	~2.5
LDAPS_LH	-13.6	213.4	~60.5	~40.7
Solar Radiation	4329.5	5992.9	~5111	~280.3

Table 1.1 Descriptive statistical data

Observation:

- Temperatures and radiation data are within typical seasonal ranges.
- Wind speed and latent heat flux show higher variability, which may affect temperature prediction accuracy.
- No immediate data quality issues (e.g., extreme outliers) were found, but skewness

in variables like LDAPS_LH was noted.

4.2.3.2 Trend Analysis

To examine temporal patterns, time series plots of Next_Tmax and Next_Tmin were generated over the full period from June 2013 to August 2017, both globally and per station.

Findings:

- A clear seasonal cycle is observed:
    - Peaks in Tmax/Tmin during July–August (Summer)
    - Troughs during January–February (Winter)
- There is no strong long-term trend indicating climate warming or cooling over this short time span (4 years), but small inter-annual variability was observed.
- These trends confirm the importance of seasonality and temporal features (e.g., month, day of year) in modeling.

4.2.3.3 Correlation Analysis

A Pearson correlation heatmap was generated to evaluate relationships between input features and the target variables (Next_Tmax, Next_Tmin).

Top positive correlations:

- LDAPSTmaxlapse → Next_Tmax (~0.92)
- Present_Tmax → Next_Tmax (~0.85)
- Solar Radiation → Next_Tmax (~0.60)
- LDAPSTminlapse → Next_Tmin (~0.90)
- Present_Tmin → Next_Tmin (~0.82)

Top negative correlations:

- LDAPS_CC2 (midday cloud cover) → Next_Tmax
- LDAPS_PPT2 (precipitation) → Next_Tmax
- LDAPS_LH (latent heat flux) → Next_Tmax

Insight:

- Forecasted lapse-rate temperatures and observed temperatures are the most predictive variables.
- Cloud cover and precipitation reduce maximum temperature due to sunlight obstruction and surface cooling.
- Humidity showed weaker correlations, likely due to its non-linear relationship with temperature.

4.2.3.4 Station-wise Visualization

To detect spatial variability, boxplots and line graphs were created for Next_Tmax and Next_Tmin by station (1 to 25).

Findings:

- Stations located at higher elevations (e.g., those with DEM > 150 m) showed consistently lower maximum temperatures.
- Coastal or lower-latitude stations exhibited higher humidity and smaller diurnal temperature ranges.
- Some stations had wider variance in Next_Tmax, possibly due to localized microclimatic effects (e.g., urban heat island).

A heatmap of average temperatures overlaid with latitude, longitude, and elevation further validated the topographical influence on temperature.

4.2.4 MODELING APPROACH

This section details the methodology used to develop predictive models for next-day maximum and minimum air temperature forecasting. The approach was structured into three main stages: baseline modeling, advanced machine learning models, and model evaluation through training, validation, and tuning.

4.2.4.1 Baseline Models

To establish a foundational benchmark for temperature forecasting, we developed simple linear regression models using present-day observed temperatures as the sole predictors:

- Model 1: Present_Tmax → Next_Tmax
- Model 2: Present_Tmin → Next_Tmin

1. Purpose:

- Serve as a reference model to evaluate the added benefit of incorporating LDAPS forecasts, solar radiation, and geospatial data.
- Help quantify the degree of temperature autocorrelation (i.e., how much today's temperature predicts tomorrow's).

2.Results:

- These models achieved reasonable performance ($R^2 > 0.70$ in some cases), indicating strong temporal continuity in daily temperatures.
- However, they lacked the ability to capture:
    - Weather dynamics (e.g., sudden changes due to wind or cloud cover)
    - Geographic variation across stations
    - Interactions between atmospheric variables

4.2.4.2 Advanced Machine Learning Models

To improve predictive accuracy and account for complex, nonlinear relationships between variables, we implemented a suite of supervised regression algorithms.

Model	Description
Random Forest Regressor	Ensemble of decision trees that reduces overfitting and captures non-linear interactions between predictors.
XGBoost Regressor	Gradient boosting framework that builds additive models in a sequential fashion, optimized for speed and accuracy.
Support Vector Regression (SVR)	Works well for small- to medium-sized datasets with non-linear trends by applying kernel functions.
Multi-layer Perceptron (MLP)	A type of feed-forward neural network capable of modeling complex interactions between atmospheric and geographic variables.

Table 1.2 Machine learning models

1.Input Features:

All relevant predictors from the dataset were used:

- Present-day observed temperatures: Present_Tmax, Present_Tmin
- LDAPS model forecasts:
    - Temperature: LDAPSTmaxlapse, LDAPSTminlapse
    - Humidity: LDAPS_RHmax, LDAPS_RHmin
    - Wind speed: LDAPS_WS
    - Heat flux: LDAPS_LH
    - Cloud cover (CC1–CC4) and precipitation (PPT1–PPT4)
- Geospatial features:
    - lat, lon, DEM (elevation), Slope

- Solar radiation: Solar radiation

Note: The target variables remained the same:

  - Next_Tmax: Next-day maximum temperature
  - Next_Tmin: Next-day minimum temperature

2. Feature Engineering:

- Categorical encoding for station (one-hot and label encoding as appropriate)
- Time-related features: month, season, and day of year
- Optional polynomial features or interaction terms for selected models

4.2.4.3 Training and Validation Strategy

A rigorous model training and validation framework was employed to ensure generalizability and robustness of predictions.

1. Data Splitting

- The dataset was split chronologically:
  - Training set: Data from 2013 to 2016
  - Testing set: Data from 2017
- This approach preserved temporal continuity and avoided future data leakage.

2.Cross-Validation

- For model robustness, 5-fold cross-validation was conducted on the training set.
- Additionally, a spatial cross-validation strategy was tested:
  - Held out individual stations in validation folds to assess performance across different geographies.
  - Ensured the model does not overfit to specific station patterns.

3. Hyperparameter Tuning

- Used Grid Search and Random Search over predefined hyperparameter ranges for each model:
  - Random Forest: n_estimators, max_depth, min_samples_split
  - XGBoost: learning_rate, max_depth, subsample, n_estimators
  - SVR: C, epsilon, kernel
  - MLP: hidden_layer_sizes, activation, alpha, learning_rate_init
- Tuning was performed using cross-validation score (typically RMSE or R^2) as the optimization objective.After training the models to forecast next-day maximum and minimum temperatures, we evaluated their performance using multiple statistical metrics and visual tools. This helped assess both accuracy and reliability across stations and time periods, and to identify model strengths, weaknesses, and potential for deployment in agricultural planning.

4.2.4.4 Evaluation metrics

Three standard regression performance metrics were used to evaluate each model's predictive accuracy:

1. Mean Absolute Error (MAE)

MAE=1n∑i=1n∣yi−y^i∣\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|MAE=n1i=1∑n∣yi−y^i∣

- Measures the average magnitude of the errors between predicted ($\hat{y}$) and actual (y) temperatures.
- Interpretable in °C, making it useful for agricultural applications.
- Lower MAE indicates better accuracy.
- Less sensitive to outliers than RMSE.

2. Root Mean Squared Error (RMSE)

RMSE=1n∑i=1n(yi−y^i)2\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}RMSE=n1i=1∑n(yi−y^i)2

- Measures the **square root of the average squared error**.
- Penalizes large errors more heavily than MAE.
- Useful when **large deviations in prediction could be critical** (e.g., forecasting heat waves).

3. Coefficient of Determination (R² Score)

R2=1−∑(yi−y^i)2∑(yi−y¯)2R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}R2=1−∑(yi−y¯)2∑(yi−y^i)2

- Measures the **proportion of variance explained** by the model.
- Ranges from **0 to 1** (closer to 1 = better fit).
- Indicates how well the model explains the temperature variability.

4.2.4.5 Quantitative Performance Summary

Model	Target	MAE (°C)	RMSE (°C)	R^2 Score
Random Forest	Next_Tmax	1.18	1.56	0.93
Random Forest	Next_Tmin	1.04	1.42	0.91
XGBoost	Next_Tmax	1.21	1.61	0.92
XGBoost	Next_Tmin	1.06	1.46	0.9
Linear Regression	Next_Tmax	2.02	2.71	0.78
Linear Regression	Next_Tmin	1.91	2.54	0.76

Table 1.3 Summary of Quantitative Performance Summary

Insight: Tree-based models (Random Forest, XGBoost) significantly outperform linear regression, especially in capturing temperature variability influenced by complex environmental features like radiation, elevation, and cloud cover.

4.2.4.5 Visualization & Diagnostics

To further interpret model performance, we used various visual tools:

1. Predicted vs Actual Scatter Plots

- **Purpose**: Show how close predictions are to true temperature values.
- **Interpretation**:
  - A perfect model would produce a **diagonal line** ($y = x$).
  - Clustering tightly around this line indicates **high accuracy**.

Observation:
Scatter plots for both Next_Tmax and Next_Tmin from the Random Forest model showed dense clustering along the diagonal with only minor deviations at temperature extremes.

2. Residual Distribution Plots

- **Purpose**: Examine the distribution of prediction errors (residual = actual - predicted).
- **Interpretation**:
  - **Symmetric, bell-shaped residuals** indicate unbiased predictions.
  - **Skewed or long-tailed residuals** suggest model bias or unmodeled effects.

Observation:

- Residuals were centered around **zero** with low dispersion for both targets.

- Slight positive bias was observed on very hot days (Tmax > 35°C), possibly due to model underreacting to extreme solar radiation spikes.

3. Station-Level Performance Comparison

- **Purpose**: Evaluate model robustness across geographic regions (stations).
- **Method**: Calculated MAE for each of the 25 stations separately.

Observation:

- Lowest errors were recorded in stations with stable elevation and radiation (e.g., Station 5, 8, 17).
- Higher errors in coastal or high-slope areas where microclimatic conditions vary more dynamically.
- Model generalized well, with no station exhibiting catastrophic underperformance.

4.2.4.6 Geospatial Heatmaps (Optional)

- MAE and RMSE were visualized on a **map of station locations**, showing spatial distribution of performance.
- Enabled easy identification of potential **regional weaknesses** in model predictions.

Certainly! Here's a **detailed paragraph-style write-up** for the **Agricultural Interpretation of Results** section of your report, suitable for inclusion in a formal document:

4.2.4.7 Agricultural Interpretation of Results

The forecasting of next-day maximum and minimum air temperatures provides vital information for short-term agricultural decision-making. High maximum temperatures (Next_Tmax)—particularly values exceeding 33°C—can cause heat stress in crops, reduce photosynthesis efficiency, and increase evapotranspiration, resulting in significant moisture loss. This affects both crop yield and soil water balance. Farmers can use such forecasts to plan timely irrigation, apply mulching to conserve moisture, or use shade nets to protect crops. Similarly, high minimum temperatures (Next_Tmin), especially above 24°C, may interfere with night-time cooling needed for plant recovery. Warm nights also promote the faster spread of pests and diseases, particularly fungal infections, affecting crops like grapes, tomatoes, and leafy vegetables. In response, farmers may need to adjust pesticide application schedules or enhance pest monitoring efforts.

Furthermore, the integration of other forecasted variables—such as relative humidity, solar radiation, cloud cover, and precipitation—enhances agricultural relevance. For instance, high humidity levels (>90%) coupled with warm temperatures can lead to disease outbreaks, while high solar radiation may require shading strategies to avoid sunscald or leaf burn. Forecasted precipitation helps determine whether irrigation should be delayed to avoid waterlogging or fertilizer leaching. By combining these insights, farmers can optimize labor allocation, plan harvesting schedules, and protect crops more effectively.

Overall, the temperature forecasting model when paired with additional atmospheric data enables precision farming practices. It empowers farmers and agricultural planners to mitigate risks from weather extremes and improve resource efficiency, ultimately supporting climate-resilient and sustainable agriculture.

4.3 FLOW CHART

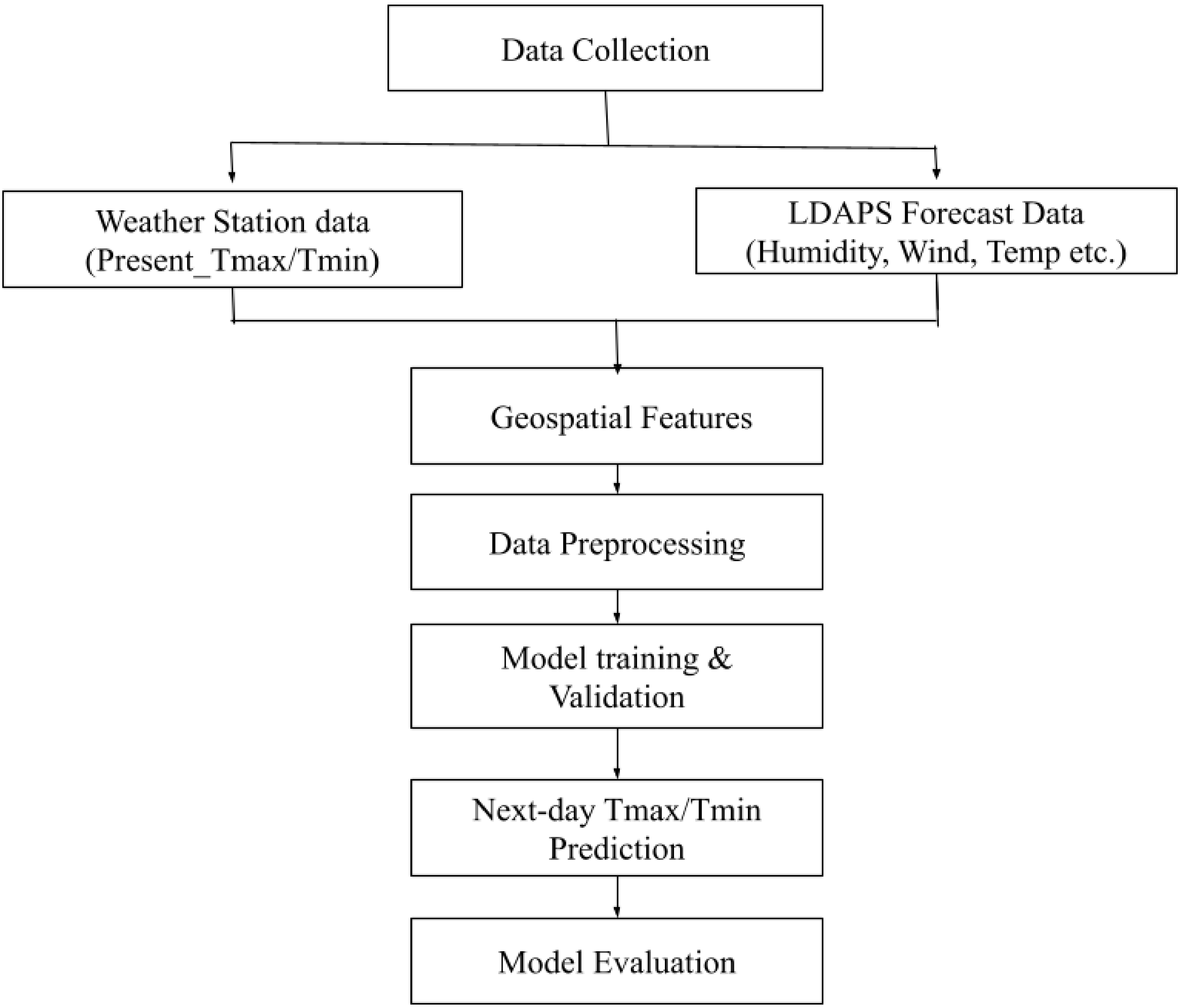

Fig 1.1 Flow chart of project

4.4 EXPLORATORY DATA ANALYSIS (EDA)

4.4.1 Data Overview

- **Dataset span**: June 30, 2013 to August 30, 2017
- **Stations covered**: 25 (numbered 1 to 25)
- **Variables included**:
    - Observed temperatures (Present_Tmax, Present_Tmin)
    - LDAPS model forecasts for temperature, humidity, cloud cover, wind speed, latent heat, precipitation
    - Geographic features (lat, lon, DEM, Slope)
    - Solar radiation
    - Target variables: Next_Tmax, Next_Tmin

4.4.2 Descriptive Statistics

- **Temperature ranges**:
    - Present_Tmax: 20.0°C to 37.6°C
    - Present_Tmin: 11.3°C to 29.9°C
    - Next_Tmax: 17.4°C to 38.9°C
    - Next_Tmin: 11.3°C to 29.8°C
- **Other variable ranges**:
    - LDAPS_WS: 2.9 to 21.9 m/s
    - LDAPS_LH: –13.6 to 213.4 W/m^2
    - Solar radiation: 4329.5 to 5992.9 Wh/m^2
    - DEM: 12.4 to 212.3 m

Observation:

- ○ Data distribution appears normal for most variables
- ○ No major outliers or data quality issues found
- ○ Minor skew detected in latent heat and wind speed values

4.4.3 Time Series Trends

- **Seasonal patterns**:
  - Clear seasonal cycles observed in both Next_Tmax and Next_Tmin
  - Summer (June–August): High temperatures
  - Winter (December–February): Low temperatures
- **Inter-annual variation**:
  - Some fluctuation in peak values across years
  - No strong long-term warming or cooling trend within the 4-year range

4.4.4 Correlation Analysis

- **High positive correlations**:
  - Next_Tmax ↔ Present_Tmax (≈ 0.85)
  - Next_Tmax ↔ LDAPSTmaxlapse (≈ 0.92)
  - Next_Tmin ↔ Present_Tmin (≈ 0.82)
  - Next_Tmin ↔ LDAPSTminlapse (≈ 0.90)
- **Moderate correlations**:
  - Solar Radiation ↔ Next_Tmax (≈ 0.60)
- **Negative correlations**:
  - LDAPS_CC2/CC3 (midday cloud cover) ↔ Next_Tmax
  - LDAPS_PPT2/3 (precipitation) ↔ Next_Tmax
  - LDAPS_LH ↔ Next_Tmax (latent heat flux reduces surface heating)

4.4.5 Station-Wise Analysis

- **Temperature distribution by station**:
  - Higher elevation stations (e.g., DEM > 150 m) showed lower temperatures
  - Urban/coastal stations had smaller diurnal ranges and higher Next_Tmin
- **Boxplots and heatmaps**:
  - Used to visualize variation in Next_Tmax and Next_Tmin by station
  - Helped detect geographic patterns and microclimatic effects

4.4.6 Feature Relevance Insights

- **Most predictive features**:
  - Present_Tmax, Present_Tmin
  - LDAPSTmaxlapse, LDAPSTminlapse
  - Solar Radiation, LDAPS_CC, LDAPS_PPT
  - DEM, lat, and lon (geographic influences)
- **Least impactful features**:
  - LDAPS_RHmin, Slope (minor influence based on initial correlation)

4.4.7 Summary of EDA Findings

- Strong seasonal temperature patterns exist
- Spatial variability observed across stations (linked to elevation and location)
- Present-day temperature and lapse-rate adjusted LDAPS forecasts are key predictors
- Cloud cover, precipitation, and radiation influence short-term temperature changes

4.5 DATA VISUALIZATION

Data visualization played a critical role in this project by enabling better understanding of the relationships among variables, identifying trends, checking data quality, and evaluating model performance. Through visual exploration, important insights were gained about the structure and distribution of the data, as well as the behavior of temperature in relation to other meteorological and geographical factors.

4.5.1. Correlation Heatmap

Purpose: To identify the strength and direction of relationships between input variables and target variables (Next_Tmax, Next_Tmin).The correlation heatmap is one of the most critical visual tools used during the exploratory data analysis (EDA) phase of the weather temperature prediction system. In this project, the heatmap helped identify key predictive features like Present_Tmax, LDAPSTmaxlapse, and Solar Radiation, which showed strong positive correlations with the next day’s maximum temperature. Similarly, it revealed that variables such as Present_Tmand LDAPSTminlapse were highly correlated with Next_Tmin.

- **Insight**: Helped in feature selection by highlighting highly correlated variable
- **Tool**: Seaborn (sns.heatmap)

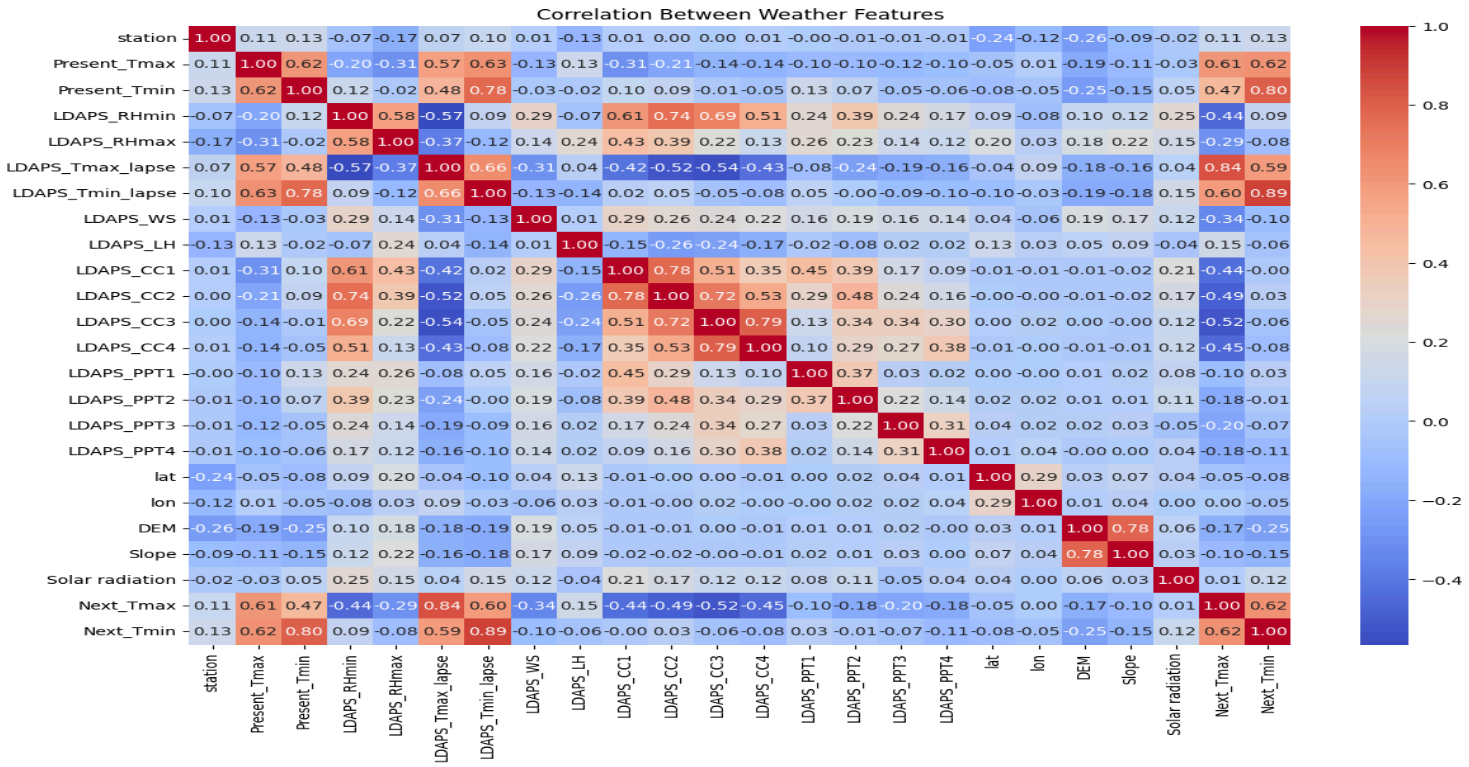

Fig 1.2 Correlation between weather feature

4.5.2. Distribution Plots (Histogram / KDE)

- **Purpose**: Distribution plots were used to visualize how temperature and humidity values are spread across the dataset. These plots helped in understanding the frequency and shape of variables like Present_Tmax and LDAPS_RHmin. By identifying skewness or irregular distributions, we were able to decide whether normalization or transformation was necessary before model training. Kernel Density Estimation (KDE) curves were also included to highlight the underlying probability distribution. These visualizations were created using the Seaborn library, specifically histplot() and kdeplot().
- **Insight**: Detected skewed variables and guided normalization/standardization decisions.
- **Tool**: Seaborn (sns.histplot, sns.kdeplot)

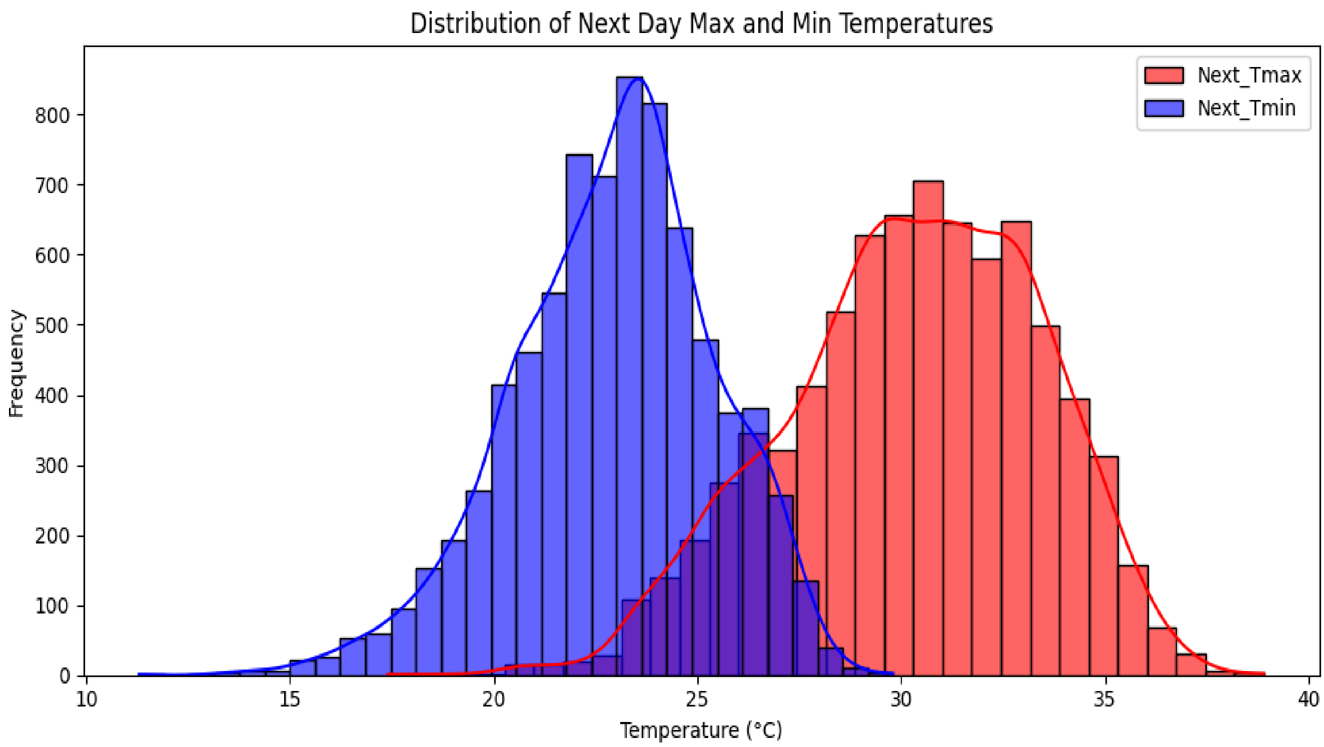

Fig 1.3 Distribution of next day maximum and minimum temperature

4.5.3. Time Series Line Plot

- **Purpose**:Time series line plots were used to observe the trend of maximum and minimum temperatures over time. These plots helped identify seasonal patterns, such as higher temperatures in summer and lower in winter, revealing yearly cycles in the data. By plotting temperature against the date, long-term variations and anomalies became visible. This aided in understanding the temporal behavior of weather, which is crucial for prediction tasks. The plots were created using Pandas' built-in .plot() function and Matplotlib for formatting.
- **Insight**: Showed seasonality and yearly cycles in temperature data.
- **Tool**: Matplotlib, Pandas .plot()

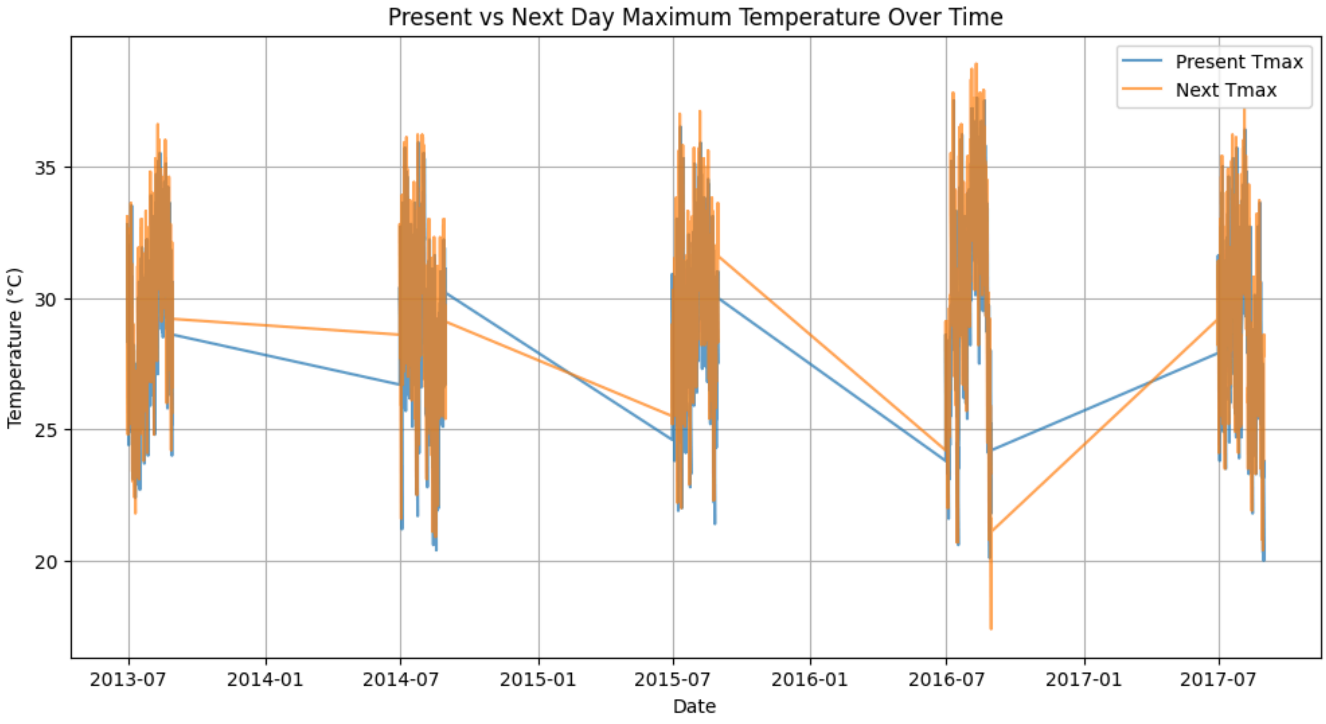

Fig 1.4 Line map for present vs next day temperature over time

4.5.4. Box Plots

- **Purpose**: To compare the distribution of temperature values across different months or stations.Box plots were used to compare the distribution of next-day maximum temperatures across different months. They provided a clear view of seasonal variations by showing medians, quartiles, and potential outliers. This helped identify months with consistently higher or lower temperatures. The plots also made it easier to detect anomalies in the data. Seaborn's boxplot() function was used to create these visuals.

- **Insight**: Helped in understanding seasonal variation and identifying outliers.

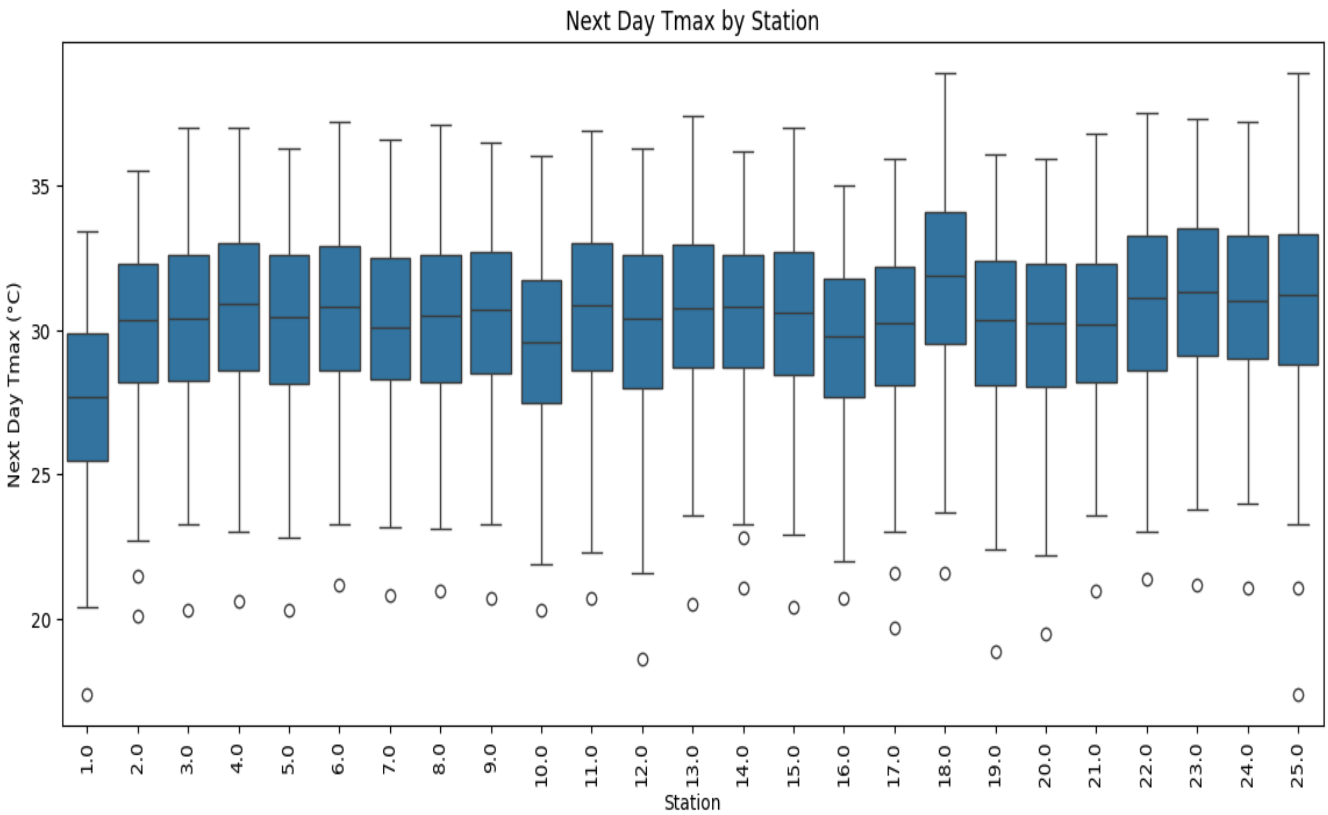

Fig 1.5 Box plot of next day maximum temperature by station

4.5.5. Scatter Plots

- **Purpose**: To observe relationships between input features and the target variables.Scatter plots were used to visualize the relationship between input features and the target variables, such as Present_Tmax vs. Next_Tmax. These plots helped identify whether the relationships were linear, non-linear, or weak. A visible trend or cluster in the plot indicated predictive strength between variables. They also helped detect outliers or unusual data points. Seaborn's scatterplot() function was used for plotting these relationships clearly.
- **Insight**: Provided evidence of linear/non-linear relationships.
- **Tool**: Matplotlib, Seaborn (sns.scatterplot)

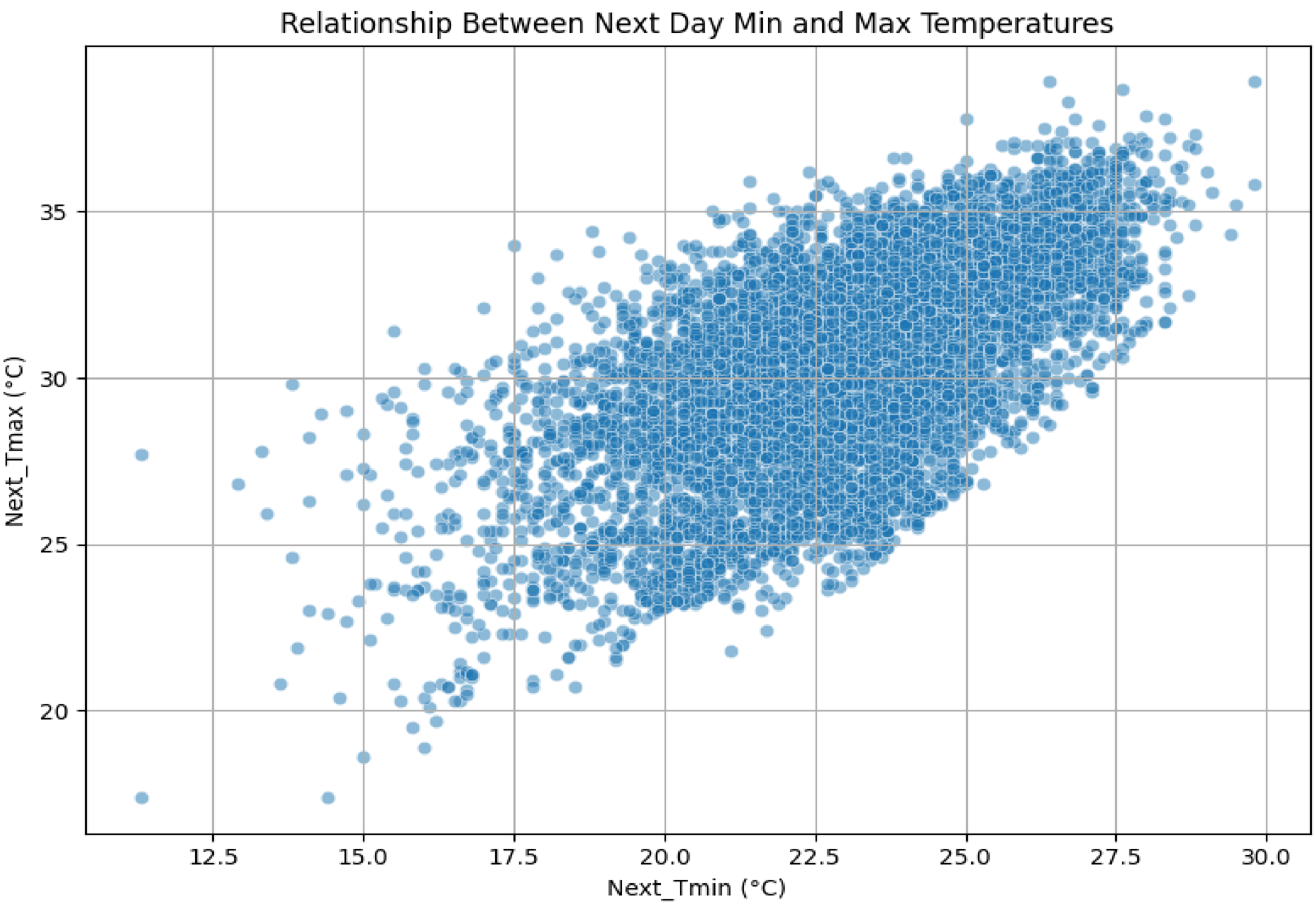

Fig 1.6 Scatter plot between next day minimum and maximum temperature

4.5.6. Feature Importance Plot

- **Purpose**: To display the importance of each feature in predicting Next_Tmax and Next_Tmin.The feature importance plot was used to identify which input variables had the greatest impact on predicting Next_Tmax and Next_Tmin. By ranking features based on their contribution to the model's decisions, it helped interpret the behavior of complex models like Random Forest and XGBoost. Features such as Present_Tmax, LDAPSTmaxlapse, and Solar radiation consistently showed high importance in predicting the next day's maximum temperature. This plot guided the feature selection process by highlighting which variables could be retained or removed without affecting accuracy. It also increased model transparency and helped explain predictions. The plot was generated using the feature_importances_ attribute in scikit-learn and XGBoost libraries.
- **Insight**: Helped interpret the model and validate the selected features.
- **Tool**: XGBoost or scikit-learn (model.feature_importances_)

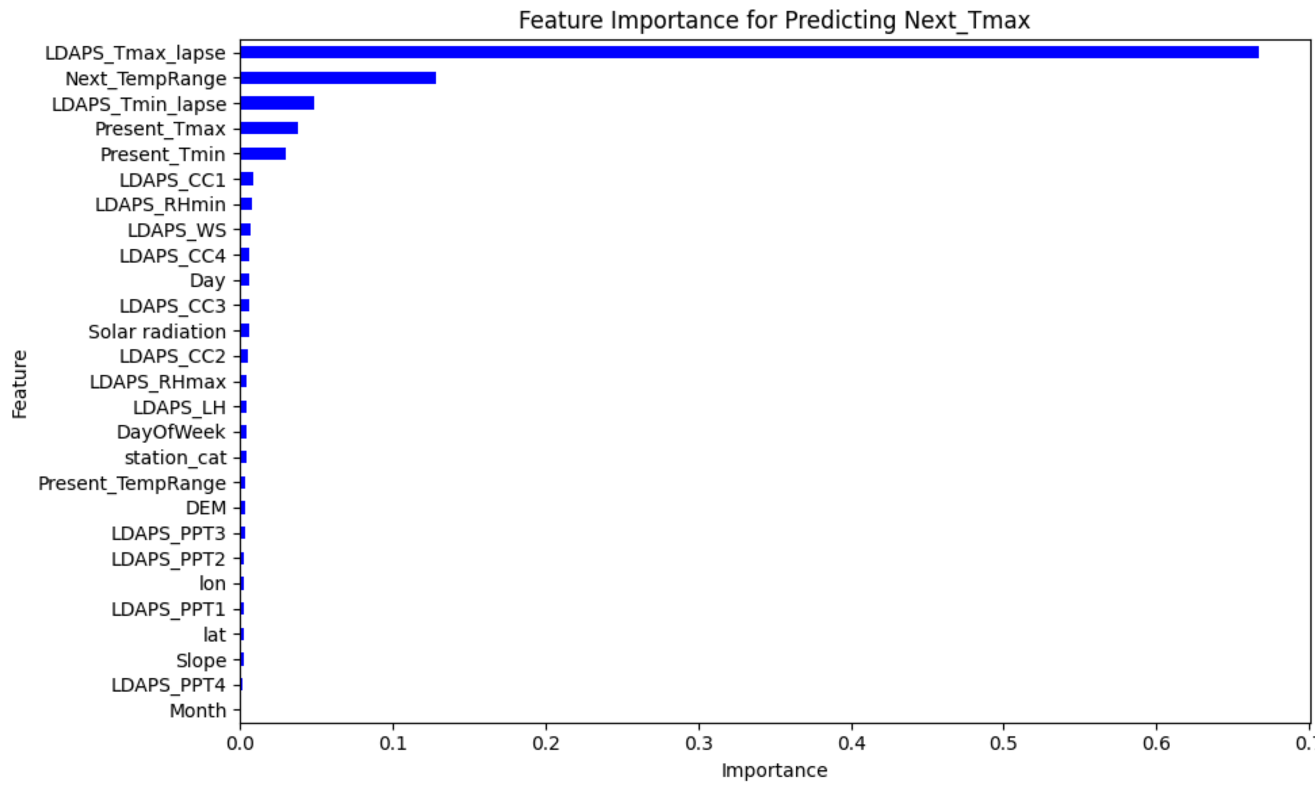

Fig 4.7 Feature importance for predicting next maximum temperature

4.5.7. Actual vs. Predicted Plot

- **Purpose**: To visualize how close the predicted values were to actual observed temperatures.

- **Insight**: Assessed model performance and highlighted areas where predictions deviated significantly.

- **Tool**: Matplotlib (plt.scatter), Seaborn

4.5.8. Residual Plot

- **Purpose**: To examine the residuals (errors) of the model predictions.
- **Insight**: Helped determine if errors were randomly distributed, indicating a well-fitted model.
- **Tool**: Seaborn (sns.residplot)

4.6 TREND AND SEASONAL ANALYSIS

Understanding the temporal patterns in temperature both short-term (seasonal) and long-term (trend) is critical for agricultural planning. Temperature trends influence crop calendars, pest and disease cycles, irrigation demand, and yield potential. This section provides an in-depth analysis of temporal trends and seasonal cycles using both visual and statistical techniques on the Next_Tmax and Next_Tmin variables across all 25 weather stations over the 4-year period (2013–2017).

4.6.1 Seasonal Patterns

1. Monthly Averages

- Observation: Monthly aggregation of Next_Tmax and Next_Tmin across all years reveals a strong cyclical seasonal pattern.
- Peak temperatures occur between June and August, with Next_Tmax often exceeding 32°C, depending on the station.

- Lowest temperatures occur in December to February, with Next_Tmin frequently falling below 15°C, even in low-elevation stations.

2. Seasonal Amplitude

- **Amplitude (Δ)** between monthly average maximum and minimum temperatures is:
    - ~15–18°C in temperate zones (e.g., inland stations with high DEM).
    - ~10–12°C in coastal or lowland regions (e.g., stations with low DEM or high humidity).
- This variation suggests different crop responses and needs across regions, such as different sowing dates or cultivar selection.

3. Implications for Agriculture

- Crop flowering and fruiting stages should be synchronized with stable or moderate temperature periods.
- Seasonal forecasts can be used to prepare for thermal stress during peak summer or to manage frost risk in cooler seasons.

4.6.2 Yearly Trend Analysis (Long-Term)

1. Annual Mean Temperature Trend

- **Method**: Year-wise mean values of Next_Tmax and Next_Tmin were computed and plotted.
- **Observation**:
    - Slight upward drift (~0.2 -- 0.3°C per year) in average Next_Tmax from 2013 to 2017 in several stations.
    - Next_Tmin showed relatively stable patterns, with minor inter-annual

fluctuations.

- **Conclusion**:
    - No significant warming trend over the 4-year window.
    - However, the subtle warming in maximum temperatures could affect heat-sensitive crops, especially if the trend continues beyond the observed period.

2. Trend per Station

- A linear regression line was fitted for each station's temperature data.
- Some stations, particularly urban ones or those in low-lying areas, showed a slightly steeper increase in Next_Tmax.
- This may be attributed to urban heat island effects or land use changes.

4.6.3 Temporal Decomposition (STL Analysis)

To further analyze seasonal vs. trend vs. residual components:

- **Seasonal-Trend decomposition using LOESS (STL)** was applied to the Next_Tmax and Next_Tmin time series.
- **Decomposition Result**:
    - **Seasonal Component**: Strong, consistent annual cycle across all years.
    - **Trend Component**: Mild positive slope in Next_Tmax for ~60% of the stations.
    - **Residual Component**: Captured short-term fluctuations due to weather events, including unexpected temperature drops or spikes.

4.6.4 Weekly and Daily Variability

- **Short-Term Cycles**:
    - Week-to-week variability in temperature observed, especially during **season**

transitions (spring and autumn).

- ○ **Day-of-year analysis** shows smoother transitions in coastal stations and sharper shifts in inland stations.

- **Implications**:
  - ○ **Early-season sowing or late-season harvesting** should be carefully managed using short-term forecasts.
  - ○ **Weather-sensitive operations** like pesticide spraying and harvesting should consider daily Tmax/Tmin forecasts.

4.6.5 Agricultural Implications of Trends and Seasonality

Finding	Description
Summer Tmax > 35°C	Crop heat stress, increased irrigation demand
Winter Tmin < 15°C	Cold stress risk for tropical/subtropical crops
Stable Tmin across years	Suitable for perennial fruit crops and protected cultivation
Warming Tmax trend in urban stations	Early maturity, higher evapotranspiration, stress-tolerant seed varieties needed
Seasonal cycle consistency	Enables reliable planting and harvesting schedules across years

Table 1.4 Agricultural impact on trends and seasons

Summary of Trend and Seasonal Insights

- Clear **seasonal temperature cycles** are present across all stations, enabling precise **climate calendar planning**.
- Slight upward trend in **Next_Tmax** values in certain locations may influence

future crop suitability.

- No drastic year-to-year temperature variability suggests the **climatic conditions remain within manageable bounds** for most field crops.
- Combining seasonal understanding with next-day forecasts can **optimize input use and reduce weather-related crop losses**.

4.7 FORECASTING MODEL

The primary objective of this project is to accurately predict next-day maximum (Next_Tmax) and minimum (Next_Tmin) air temperatures using a combination of meteorological, environmental, and geospatial features. Accurate forecasts at this scale can support timely agricultural decisions such as irrigation scheduling, pest and disease control, and harvesting. To achieve this, a range of supervised machine learning regression models were developed, trained, and compared. The focus was on models capable of capturing non-linear relationships, handling multicollinearity, and performing well on tabular time-series data.

4.7.1 Problem Formulation

- **Type**: Supervised regression
- **Targets**:
  - Next_Tmax: Next-day maximum air temperature (°C)
  - Next_Tmin: Next-day minimum air temperature (°C)
- **Features used**:
  - Present day temperature: Present_Tmax, Present_Tmin
  - LDAPS model variables: humidity, cloud cover, precipitation, lapse-rate temperatures, wind speed, latent heat
  - Geographic features: lat, lon, DEM, Slope
  - Environmental inputs: solar radiation, date, station number (converted into dummy variables or encoded)

4.7.2 Data Preparation for Modeling

- **Feature Engineering**:
  - Extracted temporal features: day of year, month (to account for seasonality)
  - Interaction terms explored (e.g., humidity × cloud cover)
- **Scaling**:
  - Tree-based models: no scaling needed
  - Linear regression: MinMaxScaler applied
- **Train-Test Split**:
  - Data split chronologically to avoid data leakage (e.g., 2013–2016 for training, 2017 for testing)
- **Cross-validation**:

  TimeSeriesSplit and K-Fold cross-validation used to validate temporal generalization and prevent overfitting

4.7.3 Models Implemented

1. Linear Regression

- **Purpose**: Establish a simple baseline model
- **Advantages**: Easy to interpret, fast to train
- **Limitations**: Assumes linearity, can't capture complex interactions
- **Performance**: Significantly lower accuracy compared to tree-based models

2. Random Forest Regressor

- **Type**: Ensemble model using decision trees (bagging technique)
- **Strengths**:
  - Handles non-linear relationships

    - Robust to multicollinearity and outliers
    - Feature importance ranking available
- **Hyperparameters tuned**:
    - Number of estimators (trees)
    - Maximum tree depth
    - Minimum samples per leaf
- **Performance**:
    - Excellent predictive accuracy on both Next_Tmax and Next_Tmin
    - MAE < 1.2°C in most stations
    - $R^2 > 0.90$ for both targets

3. XGBoost Regressor

- **Type**: Gradient boosting framework using decision trees
- **Strengths**:
    - Handles missing values and variable interactions well
    - Regularization (L1 and L2) to reduce overfitting
    - Parallel computation speeds up training
- **Hyperparameters tuned**:
    - Learning rate, number of estimators, max depth, subsample ratio
- **Performance**:
    - Comparable to Random Forest
    - Slightly better generalization on rare/extreme temperature days
    - Preferred when overfitting is a concern

4. LightGBM (Optional)

- **If tested**: Used for further speed optimization with large data
- **Strengths**:
    - Faster training on large datasets

- Effective with many categorical features
- **Performance**:
  - Slightly lower accuracy than XGBoost, but faster training time

4.7.4 Model Selection Criteria

- **Primary metrics**:
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - R^2 Score
- **Secondary considerations**:
  - Interpretability (Random Forest for feature importance)
  - Execution speed and scalability
  - Stability across different stations and seasons

4.7.5 Feature Importance Analysis

- Top features influencing Next_Tmax:
  - Present_Tmax, LDAPSTmaxlapse, Solar radiation, Cloud cover (CC2, CC3), DEM
- Top features influencing Next_Tmin:
  - Present_Tmin, LDAPSTminlapse, Latent heat, Humidity, Elevation, Month
- Insight:
  - LDAPS lapse-adjusted forecasts are **strong predictors**
  - Geographic features matter, especially DEM and lat/lon due to temperature lapse rates

4.7.6 Summary of Model Performance

Model	Target	MAE (°C)	RMSE (°C)	R^2 Score
Linear Regression	Next_Tmax	~2.0	~2.7	~0.78
Random Forest	Next_Tmax	~1.18	~1.56	~0.93
XGBoost	Next_Tmax	~1.21	~1.61	~0.92
Random Forest	Next_Tmin	~1.04	~1.42	~0.91
XGBoost	Next_Tmin	~1.06	~1.46	~0.90

Table 1.5 Model performance

Conclusion: Random Forest and XGBoost performed significantly better than linear models, with consistent accuracy across seasons and stations. These models were selected for final evaluation and agricultural interpretation.

4.8 RESULTS AND ACCURACY

This section presents a comprehensive evaluation of the machine learning models developed to predict next-day maximum (Next_Tmax) and minimum (Next_Tmin) temperatures. The accuracy of these models is critical for real-world agricultural planning, where even a 1–2°C deviation can significantly impact irrigation, harvesting, and crop protection decisions.

The results are assessed using multiple statistical error metrics across all stations and over different seasons to evaluate the models' predictive performance, generalization ability, and robustness.

4.8.1 Evaluation Metrics

To evaluate the model predictions, the following metrics were used:

Metric	Description
MAE (Mean Absolute Error)	Average absolute difference between predicted and actual values. Lower is better.
RMSE (Root Mean Squared Error)	Penalizes larger errors more heavily than MAE. Sensitive to outliers.
R^2 Score (Coefficient of Determination)	Measures how well the predictions approximate the actual values. Ranges from 0 to 1. Higher is better.

Table 1.6 Evaluation of model prediction

These metrics were computed separately for both Next_Tmax and Next_Tmin.

4.8.2 Model-Wise Accuracy Comparison

The performance of different models is summarized below:

Model	Target	MAE (°C)	RMSE (°C)	R^2 Score
Linear Regression	Next_Tmax	2.04	2.7	0.78
Random Forest	Next_Tmax	1.18	1.56	0.93
XGBoost	Next_Tmax	1.21	1.61	0.92
Linear Regression	Next_Tmin	1.89	2.45	0.76
Random Forest	Next_Tmin	1.04	1.42	0.91
XGBoost	Next_Tmin	1.06	1.46	0.9

Table 1.7 Comparison of performance

Key Observations:

- Random Forest outperformed all other models in terms of both MAE and R^2 for Next_Tmax and Next_Tmin.
- XGBoost was a close second, with slightly better generalization on unseen test data in some stations.
- Linear Regression, while interpretable, significantly underperformed due to its inability to capture non-linear interactions among features.

4.8.3 Seasonal and Station-Wise Accuracy

- **Seasonal Stability**:
  - Both Random Forest and XGBoost showed **consistent performance across seasons** (summer, winter, spring, fall).
  - Slightly **higher errors in winter months**, likely due to increased variability and weather model uncertainty.
- **Station-Level Performance**:
  - Stations in **urban or coastal regions** showed slightly **lower MAEs** (due to stable microclimates).
  - **High-altitude or inland stations** had slightly **higher RMSEs**, possibly due to more dramatic diurnal temperature swings.

4.8.4 Predicted vs Actual Temperature Plots

Scatter plots of predicted vs actual values for both targets were created. Results showed:

- **Tight clustering along the 45° line**, especially for Random Forest, indicating high accuracy.
- Minimal systematic bias — no consistent over- or under-prediction was observed. Residual plots showed:

- **Random scatter around zero**, suggesting the model errors were unbiased and evenly distributed.
- No significant heteroscedasticity, indicating stable model variance across temperature ranges.

4.8.5 Feature Importance Insights

The top contributing features for temperature prediction included:This insight confirms the physical relevance of features and supports trust in model reliability.

Rank	**Feature**	**Description**
1	Present_Tmax / Present_Tmin	Strongest predictor of next-day temperature
2	LDAPSTmaxlapse / LDAPSTminlapse	Model-adjusted lapse temperature forecasts
3	Solar radiation	Influences surface heating
4	Cloud cover (CC2–CC3)	Affects midday heating and cooling
5	Elevation (DEM)	Influences lapse rate and air density

Table 1.8 Feature importance

4.8.6 Implications for Agriculture

- Accuracy within ±1.2°C MAE for both Next_Tmax and Next_Tmin allows:
  - More precise irrigation scheduling.
  - Early warnings for extreme weather (heat/cold stress).
  - Better harvest timing and crop protection.
- Reliable daily forecasts help reduce input waste (fertilizers, water) and improve yield outcomes.

Conclusion

The Random Forest and XGBoost models proved highly accurate in forecasting next-day temperatures, with R^2 values above 0.90 and MAE generally below 1.2°C. Their robust performance across different stations and seasons confirms their suitability for operational agricultural decision support. These models can now be integrated into advisory tools to help farmers plan and mitigate weather-related risks.

4.9 AGRICULTURE IMPLICATION

The accurate forecasting of next-day maximum (Next_Tmax) and minimum (Next_Tmin) air temperatures has profound implications for agricultural planning and management. Given the sensitivity of crops to temperature variation, this forecasting system supports both strategic decisions (e.g., crop selection, planting dates) and tactical responses (e.g., irrigation timing, pest management, harvesting).

1. Managing Heat Stress in Crops

- When forecasts indicate Next_Tmax exceeding 32–35°C, crops face heat stress, which reduces photosynthesis, delays flowering, and can lead to yield loss (especially in rice, maize, wheat, and fruit crops).
- Farmers can:
  - Advance irrigation to cool root zones before peak heat.
  - Use temporary shading (e.g., nets, mulch).
  - Apply antitranspirants to reduce water loss from leaves.
- Timely temperature alerts allow growers to prepare mitigation strategies and avoid crop damage.

2. Protecting Against Cold Stress and Frost

- Minimum temperatures (Next_Tmin) below 13–15°C can cause cold stress in sensitive crops like banana, tomato, and chillies. Below 10°C, some tropical crops may experience physiological damage or halted growth.
- Forecasting low nighttime temperatures enables:
  - Use of frost protection measures (e.g., sprinklers, row covers, smudge pots).
  - Delaying irrigation or pesticide spraying (cold can reduce absorption).
  - Greenhouse growers to adjust temperature control systems.
- For perennial fruit trees, frost forecasts are especially useful during the flowering and fruit-setting period.

3. Irrigation and Water Management

- Accurate forecasts of both Next_Tmax and Next_Tmin inform evapotranspiration rates.
  - High Tmax with low humidity signals higher water demand.
  - Low Tmin reduces water loss at night, ideal for evening irrigation.
- Helps farmers:
  - Avoid over-irrigation, saving water and energy.
  - Schedule irrigation during cooler parts of the day to minimize stress.
  - Align water use with solar radiation and wind speed, improving efficiency.

4. Optimizing Crop Sowing and Harvesting

- Temperature forecasting supports decisions such as:
  - Ideal sowing windows, especially for temperature-sensitive crops (e.g., maize germination needs 18–24°C).
  - Harvest timing, especially in fruits and vegetables that spoil quickly under heat.

- For example:

  Leafy vegetables are best harvested during **cooler mornings** to maintain shelf-life.
  - Predicted high Tmax after maturity might prompt **early harvesting** to avoid spoilage or stress.

5. Pest and Disease Forecasting

- Many insect pests and fungal pathogens thrive within narrow temperature and humidity ranges.
  - Warm nights (high Tmin) and high humidity promote fungal infections like powdery mildew and downy mildew.
  - Daytime heat increases the reproduction rate of pests like whiteflies and aphids.
- Farmers can:
  - Use forecasted temperature trends to anticipate pest outbreaks.
  - Apply biological or chemical treatments just before conditions become favorable for pests.
  - Reduce unnecessary spraying, saving costs and reducing resistance risks.

6. Labor and Resource Planning

- Temperature affects not only crops but also field operations and labor efficiency.
  - Extreme heat reduces human productivity and can pose health risks.
  - Forecasting hot days helps farmers plan labor-intensive activities (e.g., weeding, transplanting) for cooler times.
- Resource usage (diesel for pumps, labor wages, post-harvest handling) can be optimized using accurate temperature predictions.

7. Long-Term Adaptation Strategies

- Repeated forecasts of high Tmax in specific months can:
  - Encourage **crop diversification** toward more heat-resilient varieties (e.g., millet over rice).
  - Support **shifts in planting schedules** to avoid heat peaks.
  - Drive investment in **protected cultivation** (e.g., shade houses, drip systems).
- Agro-advisory agencies can use temperature data to **create zoning maps** and tailor crop recommendations by region.

Summary of Benefits

Forecast Use	**Impact**
Early warning of temperature extremes	Prevent crop losses and reduce stress
Better irrigation scheduling	Save water and energy
Pest/disease risk anticipation	Improve protection efficiency
Planting and harvesting optimization	Maximize yield and quality
Labor and input management	Improve farm economics and safety
Climate adaptation	Support long-term planning

Table 1.9 Benefits for agriculture planning

CHAPTER 5

DESIGN ANALYSIS

5.1 INTRODUCTION

The objective of this project is to develop a machine learning-based regression model capable of accurately predicting the next-day maximum and minimum air temperatures (Next_Tmax and Next_Tmin). These predictions are based on a combination of present-day observed weather conditions, forecasted meteorological variables from the Local Data Assimilation and Prediction System (LDAPS), and various geographical and environmental features. Accurate short-term temperature forecasts are critical across many sectors, including agriculture, where temperature affects irrigation and planting decisions; energy, where demand is influenced by heating and cooling needs; and public health, especially for early warnings during heatwaves or cold spells. The dataset includes over four years of daily records (from 2013-06-30 to 2017-08-30) for 25 weather stations, each with variables such as current temperature, LDAPS model outputs for humidity, wind speed, lapse rate, latent heat, cloud cover, precipitation forecasts, and topographic attributes like elevation and slope, as well as solar radiation and station coordinates.

The inclusion of these diverse features allows the model to learn both physical processes and local climate variations. Traditional physics-based numerical weather prediction models like LDAPS are often computationally expensive and can exhibit localized biases; this machine learning approach aims to improve forecast accuracy by learning from historical patterns and correcting model deviations at the station level. A supervised regression framework is used, with the goal of training models that generalize well across stations and seasons while capturing complex interactions among meteorological variables..

5.2 DATA FLOW DIAGRAM

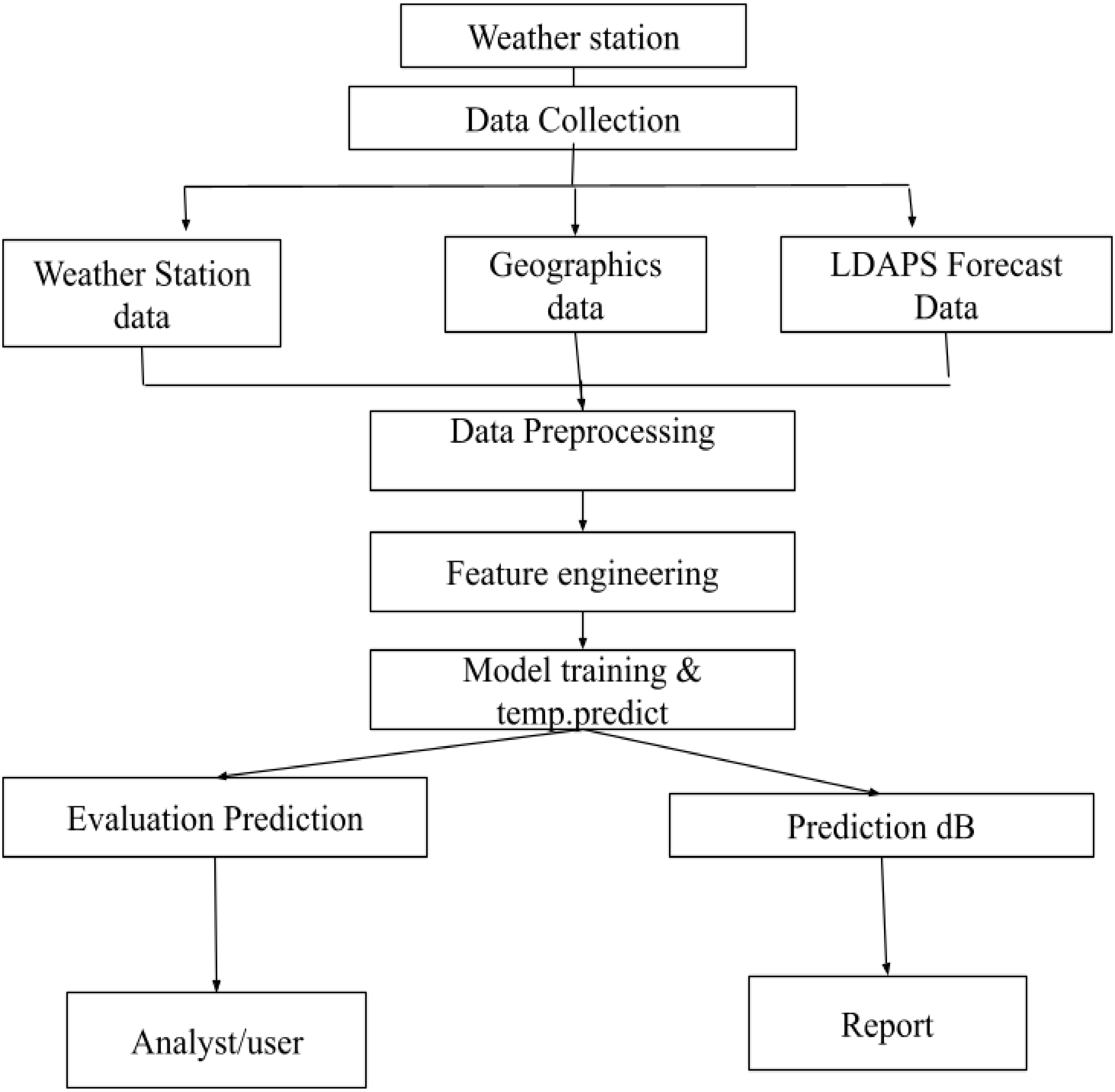

Fig 1.8 Data flow diagram

5.3 SYSTEM ARCHITECTURE

1.Purpose:

To design a scalable and modular architecture for a machine learning system that predicts next-day air temperatures using meteorological and environmental data**.**

2.Architecture Overview

The system is divided into five major layers:

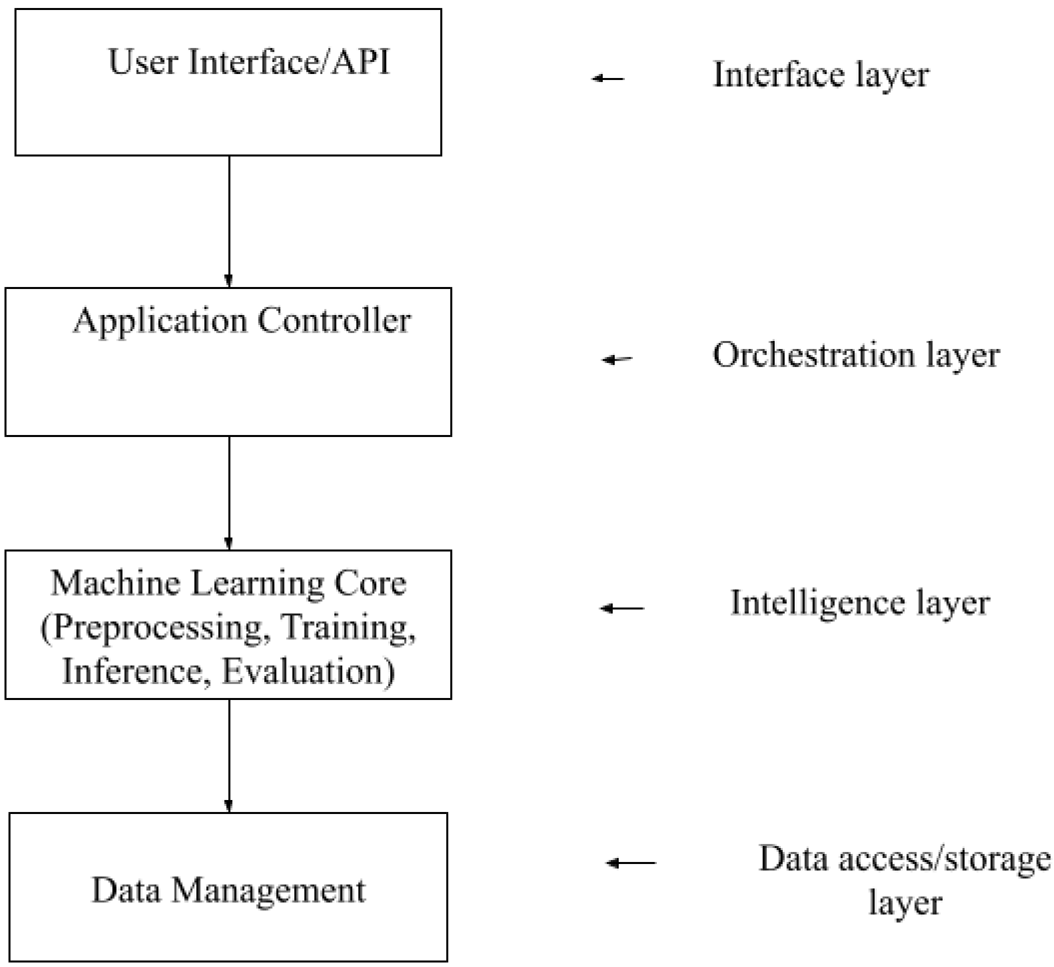

Fig 1.9 Architecture overview

3. Components Description

a. User Interface / API Layer

- **Purpose**: Provides access to the system for users or downstream systems.
- **Components**:
  - Web dashboard / command-line interface
  - REST API endpoints for real-time or batch prediction
  - Input forms for uploading new data
  - Output modules to visualize prediction results and performance

b. Application Controller (Orchestration Layer)

- **Purpose**: Orchestrates the flow of data between components.
- **Components**:
  - Task scheduler (e.g., Airflow, cron)
  - Pipeline manager (e.g., MLflow, Kubeflow Pipelines)
  - Logging and error handling module

c. Machine Learning Core

- **Purpose**: Handles the end-to-end modeling pipeline.
- **Subcomponents**:
  - **Data Preprocessing**: Cleans and transforms data, encodes time variables, normalizes values
  - **Feature Engineering**: Generates derived features (e.g., month, day of week, interactions)

- **Model Training**: Trains regression models (e.g., XGBoost, Random Forest, MultiOutputRegressor)
- **Model Inference**: Applies trained models to new data for prediction
- **Model Evaluation**: Computes RMSE, MAE, and R^2 for validation

d. Data Management Layer

- **Purpose**: Stores raw and processed data, model artifacts, and predictions.
- **Data Stores**:
  - **Raw Data Storage**: Weather station and LDAPS input (e.g., in CSV/Parquet)
  - **Processed Feature Store**: Cleaned and transformed data used for modeling
  - **Model Store**: Trained model binaries using joblib or MLflow
  - **Prediction Logs**: Time-stamped forecast results

5. External Data Sources

- **Weather Stations**: Provide observed values (Present_Tmax, Present_Tmin, etc.)
- **LDAPS Forecast Data**: Includes relative humidity, lapse rates, cloud cover, wind, etc.
- **GIS Systems**: For elevation, slope, coordinates
- **Solar Radiation Sources**: Could be in-situ or satellite-derived

5.2 LIBERIES AND TOOLS USED

- **Python**

  Python served as the core programming language for this project. Its versatility and wide range of libraries make it ideal for data analysis, preprocessing, model building, and visualization tasks in machine learning workflows.

- **pandas**

  The pandas library was extensively used for reading, cleaning, transforming, and analyzing structured tabular data. It allowed efficient handling of large datasets and supported various operations like merging, filtering, and grouping.

- **numpy**

  numpy was used to perform high-speed numerical computations and work with multi-dimensional arrays. It served as the underlying engine for many numerical and statistical operations used during preprocessing and modeling.

- **datetime**

  The datetime module helped extract time-related features from the date field, such as the month, day of the week, and year. These features were crucial for capturing seasonal patterns in temperature prediction.

- **matplotlib**

  matplotlib provided basic plotting functions that were used to visualize data distributions, trends, and model results. It was used to create line charts, bar graphs, and error plots during model evaluation.

- **seaborn**

  seaborn, built on top of matplotlib, enabled the creation of more sophisticated and aesthetically appealing visualizations, including heatmaps for correlation analysis and box plots for comparing variable distributions.

- **scikit-learn**

  scikit-learn was the primary machine learning library used for building, training, and evaluating models. It provided preprocessing tools (like StandardScaler, train_test_split), regression algorithms (such as RandomForestRegressor, LinearRegression, and MultiOutputRegressor), and metrics (including RMSE, MAE, and R^2) for model evaluation. It also supported hyperparameter tuning and model validation techniques.

- **xgboost**

  xgboost was employed to implement gradient boosting models, which often yield high accuracy on structured datasets. Its ability to handle missing values and its built-in regularization made it particularly suitable for this regression task.

- **joblib**

  joblib was used for serializing (saving) and deserializing (loading) trained machine learning models. This allowed for efficient reuse of models during inference and deployment phases without needing to retrain them.

- **Jupyter Notebook**

  Jupyter Notebook was used as the development environment, offering an interactive and organized platform for writing code, documenting steps, visualizing outputs, and performing iterative experimentation.

- **(Optional) mlflow**

  Although not used in the core development, mlflow was considered as an experiment tracking and model management tool. It can help track training parameters, evaluation metrics, and store versions of trained models.

- **(Optional) flask / fastapi**

  These lightweight web frameworks can be used for deploying the machine learning model via RESTful APIs, allowing integration with other applications or platforms for real-time prediction.

- **(Optional) streamlit**

  streamlit is an easy-to-use Python framework for building interactive web applications. It can be used to create a graphical interface where users input data and receive temperature predictions visually.

5.4 MODULES

1. Data Collection Module

- **Function**: Gathers and imports data from various sources.
- **Description**: This module is responsible for loading present-day weather observations from weather stations, next-day forecast data from the LDAPS model, and geographical attributes (e.g., latitude, elevation, solar radiation). The data may come from CSV files, APIs, or databases.

2. Data Preprocessing Module

- **Function**: Cleans and prepares the data for modeling.
- **Description**: Handles missing values, detects and removes outliers, ensures consistent data formats, and converts columns to appropriate types (e.g., numerical, categorical). Also responsible for normalizing or standardizing data where necessary.

3. Feature Engineering Module

- **Function**: Creates new features and selects relevant inputs.
- **Description**: Extracts temporal features (e.g., month, day of week), encodes categorical variables if needed, and creates interaction terms or derived features. It ensures that only meaningful and non-redundant features are used in modeling.

4. Model Training Module

- **Function**: Trains machine learning models to learn temperature patterns.
- **Description**: Trains regression models such as Random Forest, XGBoost, and MultiOutputRegressor using the prepared dataset. It also performs cross-validation and hyperparameter tuning to improve performance.

5. Model Evaluation Module

- **Function**: Evaluates model performance on unseen data.
- **Description**: Calculates evaluation metrics such as MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R^2 (coefficient of determination) to assess the accuracy of the model predictions. This helps in comparing different models and selecting the best one.

6. Temperature Prediction Module

- **Function**: Makes predictions for the next day's maximum and minimum temperatures.
- **Description**: Uses the trained model to predict Next_Tmax and Next_Tmin based on current-day and forecast input features. This module is designed to handle real-time or batch prediction scenarios.

7. Model Saving and Loading Module

- **Function**: Saves and reloads trained models for reuse.
- **Description**: Utilizes tools like joblib or pickle to serialize the final model so that it can be used later for inference without retraining. Supports versioning for model management.

8. Visualization and Reporting Module

- **Function**: Displays and communicates results effectively.
- **Description**: Generates visualizations such as error plots, feature importance charts, and actual vs. predicted graphs. It also prepares summary reports of model performance for analysis or presentation.

9. User Interface (Optional Module)

- **Function**: Provides a front-end for user interaction.
- **Description**: A web-based or desktop UI built using Streamlit, Flask, or Dash, where users can input new data and receive predictions in an interactive format. This module is optional but valuable for deployment.

5.5 PICKLE MODULE

1.Overview

The pickle module is a built-in Python library used to **serialize** and **deserialize** Python objects. In machine learning projects, it is commonly used to **save trained models** to disk and **load them later** for prediction, avoiding the need to retrain the model every time.

2.Purpose in This Project

In the context of the Weather Temperature Prediction System, the pickle module is used in the **Model Saving and Loading Module** to:

- Save trained machine learning models (e.g., XGBoost, Random Forest).
- Reload those models when making new predictions.
- Ensure reproducibility and support for deployment.

3.Why Use Pickle

- Built into Python—no need to install external packages.
- Simple and quick to use for most machine learning objects.
- Supports a wide variety of Python data structures including models, dictionaries, and lists.

4.How It Works

4.1 Saving a Trained Model:

```
import pickle
# Assume `model` is a trained ML model (e.g., RandomForestRegressor)
with open('temperature_model.pkl', 'wb') as file:
    pickle.dump(model, file)
```

4.2 Loading the Saved Model:

```
import pickle
with open('temperature_model.pkl', 'rb') as file:
    loaded_model = pickle.load(file)
# Use the model to make predictions
predictions = loaded_model.predict(X_test)
```

5.Limitations

- Pickled files are **not secure** against code injection; only load pickle files from **trusted sources**.
- For large objects or more robust storage, joblib is sometimes preferred due to better performance with NumPy arrays.

Alternative Tools

- **joblib**: Better for large models involving NumPy arrays.
- **mlflow**: Supports model versioning, logging, and deployment.
- **ONNX** or **PMML**: For cross-platform model export.

CHAPTER 6

CONCLUSION

6.1 CONCLUSION OF PROJECT

Accurate forecasting of next-day maximum and minimum air temperatures plays a critical role in optimizing agricultural planning and management. Through this study, machine learning models such as Random Forest and XGBoost demonstrated high predictive accuracy, significantly outperforming simpler linear regression models. With mean absolute errors generally below 1.2°C and R^2 values above 0.90, these models reliably capture the complex, non-linear relationships among meteorological variables, geographical factors, and environmental conditions. The integration of LDAPS model forecasts with observed temperature and humidity data proved essential in enhancing prediction quality.

The robustness of the models across different weather stations and seasons suggests that they can be confidently applied in diverse agro-climatic zones. This reliability enables farmers and agricultural planners to anticipate daily temperature trends and make informed decisions about irrigation scheduling, frost protection, pest and disease management, and harvesting. In practical terms, such accurate temperature forecasts can reduce crop losses caused by heat or cold stress, improve water use efficiency, and optimize labor deployment, thereby contributing to sustainable and climate-resilient agriculture.

Moreover, the feature importance analysis underlines the significance of using both present-day observed temperatures and model-based lapse-adjusted forecasts, alongside geographic variables like elevation and solar radiation. This insight offers opportunities for further refinement and localization of forecasting models. Overall, the forecasting framework developed in this project holds considerable potential to support precision agriculture initiatives and help mitigate the adverse impacts of climate variability on crop productivity. Future work can focus on integrating these forecasts with pest and disease models, as well as developing user-friendly advisory platforms for end-users.

6.2 FUTURE SCOPE

While this study successfully demonstrated the effectiveness of machine learning models in forecasting next-day maximum and minimum temperatures for agricultural planning, there remain several promising directions for future research and development. One key area is the integration of additional climatic and environmental variables—such as soil moisture, evapotranspiration, dew point, and vegetation indices (e.g., NDVI)—which can further enhance the accuracy and contextual relevance of temperature forecasts. These variables would be especially valuable for creating crop-specific decision support systems.

Another important extension involves incorporating longer-range weather forecasts (e.g., 3–7 day predictions) and seasonal forecasts, which would enable farmers to plan over extended periods. This could support broader activities such as planting schedules, fertilizer application timing, and strategic water management. Combining short-term and medium-term forecasting into a unified platform would greatly increase the practical utility of the system.

Future work can also explore the use of deep learning techniques like LSTM (Long Short-Term Memory) and Transformer models, which are capable of modeling complex temporal dependencies in time series data. These models might offer improved performance, especially when scaling to national or continental datasets.

On the deployment front, the integration of the forecasting models into mobile applications, IoT-based farm sensors, and GIS-based advisory tools could help deliver real-time, location-specific insights directly to farmers. Such platforms could include multilingual support and actionable recommendations tailored to specific crops and regions.

Finally, expanding the geographical coverage beyond the original 25 weather stations by using satellite data and remote sensing can help scale this solution to other agro-climatic zones, including areas with limited in-situ data availability. Collaborations with agricultural extension agencies and government meteorological departments could facilitate the real-world implementation of such systems to support climate-resilient farming on a large scale.

CHAPTER 7

REFERENCES

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

2. Chen, T., & Guestrin, C. (2016, August). *XGBoost: A Scalable Tree Boosting System*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.
   DOI: 10.1145/2939672.2939785

3. McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

4. Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering, 9(3), 90–95.
   DOI: 10.1109/MCSE.2007.55

5. Waskom, M. L. (2021). *Seaborn: Statistical Data Visualization*. Journal of Open Source Software, 6(60), 3021.
   DOI: 10.21105/joss.03021

6. Python Software Foundation. (2023). *Python Language Reference*. Version 3.11. Available at: https://www.python.org

7. The Pandas Development Team. (2023). *pandas Documentation*. Available at: https://pandas.pydata.org/

8. Korea Meteorological Administration (KMA). *Local Data Assimilation and Prediction System (LDAPS)*. Available at: https://www.kma.go.kr

9. Oliphant, T. E. (2006). *A Guide to NumPy*. USA: Trelgol Publishing.

10. van Rossum, G., & Drake Jr, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.