UNIVERITÀ DEGLI STUDI DI PALERMO

Dottorato di Ricerca in Scienze Fisiche e Chimiche

Dottorati e contratti di ricerca su tematiche dell'innovazione

Dottorati su tematiche Green

Dipartimento di Fisica e Chimica "Emilio Segrè"

Dottorati PON - Ciclo XXXVII

Machine Learning-Driven Solutions for Enhancing Agricultural Sustainability: The Case of Mango Farms in Sicily

DOTTORANDO

Mohsen Pourmohammad Shahvar

TUTOR

Prof. Giovanni Marsella

COORDINATORE

Prof. Marco Cannas

ANNO CONSEGUIMENTO TITOLO 2025

Abstract

This thesis presents a novel interdisciplinary framework that transfers machine learning models originally developed for astroparticle physics to agricultural forecasting, with a specific application to mango farming in coastal Sicily. Techniques such as XGBoost, Long Short-Term Memory (LSTM) networks, and Residual Networks (ResNet), previously applied to reconstructing inclined muon events in water-Cherenkov detectors, were reconfigured to address agro-meteorological challenges including temperature prediction, wind component forecasting, and climate risk assessment.

The models were trained and validated using a combination of ground sensor data, MODIS satellite-derived indices (e.g., Normalized Difference Vegetation Index (NDVI), Land Surface Temperature (LST), Aerosol Optical Depth (AOD)), and topographic features from Digital Elevation Models (DEM). A hybrid architecture combining ResNet and XGBoost was developed for high-resolution temperature forecasting, while a Bayesian Network was used to integrate probabilistic risk scenarios related to drought, wind, and heat stress. These tools were evaluated against unseen data from 2022–2024, achieving high predictive accuracy and robustness.

The findings demonstrate that machine learning models optimized for spatial-temporal analysis in astrophysics can be successfully adapted for precision agriculture under climate stress. This interdisciplinary approach improves predictive decision-making, resource allocation, and crop resilience. The methodology offers a transferable blueprint for applying domain-agnostic ML to broader sustainability challenges, including environmental monitoring and food security.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, **Professor Giovanni Marsella**, for his invaluable guidance, encouragement, and unwavering support throughout this journey. His expertise and mentorship have been instrumental in the successful completion of this thesis.

I would also like to extend my heartfelt thanks to **Professor Markus Roth** and **Dr. David Schmidt** from the Karlsruhe Institute für Technologie (KIT) for their generous help and insightful discussions during my international activities in Germany as part of the Pierre Auger Observatory project. Their contributions enriched my research and provided me with invaluable experiences in the field of astrophysics.

A special thanks to my coordinator, **Professor Marco Cannas**, for his continuous support and guidance throughout my academic journey. His encouragement has been a vital part of my success.

I am deeply thankful to **Professor Vittorio Farina** and his team for their collaboration on this work, which significantly enhanced the agricultural applications of this research. Their partnership has been a cornerstone of the interdisciplinary advancements achieved in this thesis.

This work would not have been possible without the support of the **PNRR Project Sicilian MicronanoTech Research and Innovation Center - SAMOTHRACE - ID ECS00000022 - CUP B73C22000810001**, whose contributions are gratefully acknowledged.

Lastly, and most importantly, I want to thank my wife, **Melika**, for her unwavering love, patience, and encouragement. Her support has been my anchor throughout this journey. I am also profoundly grateful to my family, who have always stood by me with their unconditional love and support, inspiring me to persevere and achieve my goals.

To all of you, I owe my deepest gratitude. Thank you.

Table of contents

List of Figures

List of Tables

List of Acronyms

ABM: Agent-Based Model

ADM: Agent-Decision-Machine framework

AOD: Aerosol Optical Depth

ARIMA: AutoRegressive Integrated Moving Average

BN: Bayesian Network

CNN: Convolutional Neural Network

DEM: Digital Elevation Model

FAO: Food and Agriculture Organization

FNN: Feedforward Neural Network

GB: Gradient Boosting

GIS: Geographic Information System

HYBRID: Hybrid Model

ICT: Information and Communication Technology

IEEE: Institute of Electrical and Electronics Engineers

IPCC: Intergovernmental Panel on Climate Change

KE: Kinetic Energy

LST: Land Surface Temperature

LSTM: Long Short-Term Memory

MAE: Mean Absolute Error

MISAR: Multi-Input Satellite-Aided Resilience

ML: Machine Learning

MLP: Multi-Layer Perceptron

MODIS: Moderate Resolution Imaging Spectroradiometer

MSE: Mean Squared Error

NDVI: Normalized Difference Vegetation Index

OF: Objective Function

PLOS: Public Library of Science

PMT: Photomultiplier Tube

RESNET: Residual Network

RF: Random Forest

RH: Relative Humidity

RMSE: Root Mean Square Error

SMAP: Soil Moisture Active Passive

SP: Surface Pressure

SR: Shortwave Radiation

TH: Threshold (Trigger)

VEM: Vertical Equivalent Muon

WD: Wind Direction

WS: Wind Speed

Chapter 1

1. Transforming agriculture through innovation

1.1 Problem Statement

Agriculture is undergoing unprecedented pressure due to the accelerating impacts of climate change, especially in vulnerable Mediterranean regions like Sicily. Over the past two decades, Sicily has experienced a measurable increase in extreme climatic events, including a 20–30% rise in summer heatwaves, more frequent drought periods, and wind anomalies disrupting seasonal predictability (Abd-Elmabod et al., 2020; Pulighe et al., 2024). These shifts pose a direct threat to the viability of high-value crops such as mangoes and avocados, which are sensitive to temperature fluctuations, water scarcity, and wind-induced stress during flowering and fruit-setting stages.

In particular, mango cultivation is susceptible to temperature drops below 5°C during winter, and extreme heat events exceeding 40°C in summer both of which are becoming increasingly common in coastal Sicily (M. Pourmohammad Shahvar et al., 2025). Moreover, coastal winds (Sirocco and Tramontana), often amplified by local topography, induce turbulence that impairs pollination and damages young fruit. At the same time, efficient irrigation is challenged by unpredictable rainfall and soil evaporation patterns (Pourmohammad Shahvar et al., 2025).

While the use of machine learning (ML) and artificial intelligence (AI) is gaining traction in agriculture, most applications are generic and lack the precision modeling required for Mediterranean microclimates. Furthermore, these approaches often fall short in capturing non-linear, stochastic, and spatially heterogeneous dynamics that characterize both climate variability and plant responses.

In contrast, the field of astroparticle physics has developed highly specialized ML techniques to manage similar complexities. For example, the detection of inclined muon showers at the Pierre Auger Observatory involves processing massive, noisy, and temporally irregular datasets from distributed sensor networks (Abdul Halim et al., 2023a). These include XGBoost regressors, LSTM networks, and ResNet models, capable of capturing spatial-temporal dependencies and optimizing predictions under uncertainty. This thesis proposes a transformative solution: **transferring machine learning frameworks from astrophysics to agriculture**, specifically mango farming in Sicily. By repurposing models originally built to decode cosmic signals, we aim to address environmental monitoring, wind forecasting,

temperature prediction, and yield dynamics under stochastic conditions. This cross-disciplinary adaptation provides a novel framework to overcome the limitations of current agrotechnologies and contributes to sustainable agricultural transformation under climate stress.

1.2 Interdisciplinary Approach

Modern agricultural systems demand innovative paradigms capable of managing a wide array of dynamic, often unpredictable environmental challenges. Traditional agronomic techniques fall short in providing high-resolution, real-time insights into multi-factorial stressors like extreme temperatures, fluctuating wind patterns, and irregular rainfall conditions that are now the norm in Mediterranean agroecosystems. In response to these challenges, this thesis adopts a bold interdisciplinary strategy: the transfer of machine learning frameworks developed in the high-complexity domain of astroparticle physics into agricultural applications.

Astroparticle physics, particularly in projects like the Pierre Auger Observatory, has long tackled the problem of extracting structured information from massive, noisy, and temporally unstructured sensor data. One notable example is the detection of inclined muon showers, which requires analyzing particle traces across spatially distributed Cherenkov detectors. These detectors generate terabytes of data, and the need to isolate rare, directionally inclined events amidst background noise led to the adoption of advanced deep learning architectures such as LSTM networks for temporal modeling, ResNet for spatial trace classification, and hybrid combinations with XGBoost and Random Forest for precision enhancement.

These tools are not just technically advanced; they are optimized for uncertainty, noise resilience, and distributed real-time environments the very same constraints faced in smart agriculture. Yet, their application outside astrophysics remains largely untapped. This thesis argues that these models are not domain-bound but fundamentally data-driven, and therefore can be reconfigured to serve agriculture's complex prediction tasks.

The agricultural case study focuses on mango cultivation in Sicily, a region that combines environmental richness with climatic volatility. Here, sensor data from meteorological stations, MODIS satellite products (NDVI, LST, AOD), digital elevation models (DEM), and soil moisture indices form a distributed environmental sensing network analogous to astrophysical detector arrays. By employing the same hybrid architectures used for muon detection such as XGBoost-LSTM for temporal-spatial prediction and ResNet-based convolutional classifiers this research builds a robust pipeline for forecasting crop stress, temperature anomalies, and wind vectors in real time.

The interdisciplinary value is not merely technical. It reflects a new scientific logic: that methods optimized for discovering cosmic signals can also detect subtle changes in Earth's environmental systems. In doing so, this thesis sets a precedent for technological transfer across domains, demonstrating that solutions for data-heavy, high-uncertainty systems can and should be adapted beyond their original context.

Furthermore, this approach is strengthened by its integration of Bayesian Networks for probabilistic risk modeling, and Agent-Based Modeling (ABM) for simulating plant-level decision responses components that reinforce the multi-scale, multi-method strategy characteristic of astrophysical research but rarely applied in agricultural modeling at this level of fidelity.

In essence, the research does not merely adapt algorithms, it imports a design philosophy: one that values interpretability, resilience, and predictive power in complex, stochastic environments. This interdisciplinary adaptation is both methodologically rigorous and socially urgent, offering a new direction for sustainable agricultural innovation in the face of global climate volatility.

1.3 Research Motivations and Objectives

1.3.1 Motivation:

The accelerating impacts of climate change manifested through increased heatwaves, erratic precipitation, and high-velocity winds pose significant threats to agriculture in Southern Europe, particularly in the Mediterranean basin. Mango cultivation in Sicily, while promising due to warming trends, is acutely vulnerable to temperature extremes (<5°C in winter, >40°C in summer), storm-induced wind damage, and seasonal water deficits. These risks reduce fruit set, cause flower abortion, and diminish yield quality.

Meanwhile, global food security demands smarter, adaptive farming solutions capable of anticipating environmental stressors and optimizing inputs. However, conventional agricultural decision-support tools often rely on deterministic models that fail to capture the non-linear, stochastic, and spatio-temporal complexity of the modern climate-agriculture interface.

In contrast, the domain of astroparticle physics has developed robust machine learning (ML) methodologies to detect rare events such as inclined muons and air showers within large, uncertain, and high-dimensional datasets. These models are optimized for sensor network integration, real-time inference, and error tolerance, features directly applicable to the challenges faced by modern agriculture.

This thesis is motivated by the hypothesis that such models, if carefully adapted, can radically improve agricultural forecasting and decision-making systems, particularly for crops like mango that are highly sensitive to climatic disruptions. The research aims to demonstrate not only the feasibility of this technological transfer but its superior performance over traditional methods in predictive accuracy and practical resilience.

1.3.2 Objectives:

The primary objective of this thesis is to establish a cross-disciplinary framework that applies machine learning models from astroparticle physics to precision agriculture. Specific goals include:

1.3.2.1 Development of Machine Learning Algorithms for Agriculture:

Design and implement data-driven models (ResNet, LSTM, XGBoost, Transformer) capable of capturing temporal dynamics, spatial heterogeneity, and uncertainty in agricultural datasets derived from both remote sensing and in-situ sensors.

1.3.2.2 Application of Astrophysical Techniques in Agriculture:

Repurpose and adapt hybrid deep learning models originally used for muon detection such as XGBoost-LSTM and ResNet-XGBoost to forecast variables critical to mango farming: temperature, wind vectors, and probabilistic yield proxies.

1.3.2.3 Optimization of Resource Use and Sustainability:

Employ predictive modeling outputs to inform irrigation scheduling, fertilizer application, and harvest timing, aiming to reduce resource waste and increase yield stability under climate stress.

1.3.2.4 Real-Time Monitoring and Predictive Analysis:

Integrate real-time data streams from weather stations, MODIS satellite imagery, and topographic models to create a multi-resolution decision-support system, capable of early detection of adverse events and extreme conditions.

1.3.2.5 Integration of Agent-Based Modelling and Bayesian Networks:

Simulate mango tree behavior and farm-level dynamics using agent-based modeling (ABM), while implementing Bayesian Networks to quantify multi-risk scenarios involving temperature, precipitation, and wind hazards.

1.3.2.6 Collaboration with Agricultural Enterprises:

Validate the proposed methodology through collaboration with mango growers in coastal Sicily, ensuring practical relevance and scalability beyond experimental conditions.

1.4 Thesis Structure

This thesis is structured to reflect a progressive development from conceptual motivation to methodological innovation and real-world implementation. The chapters are designed to emphasize the logical transfer of machine learning techniques from astrophysics to agriculture, culminating in an integrated, scalable framework for climate-resilient mango farming.

1.4.1 Transforming agriculture through innovation (Chapter 1)

1.4.2 Machine Learning in Astrophysics and Agriculture (Chapter 2)

Provides a detailed review of machine learning applications in both astrophysics and agriculture. It quantifies existing gaps and highlights methodological overlaps between the two domains. This chapter also contextualizes the specific models used in Chapters 3 and 4, emphasizing the novelty of their cross-domain adaptation.

1.4.3 Development of Machine Learning Models in Astrophysics (Chapter 3)

Details the original use case of ML models in the detection of inclined muons at the Pierre Auger Observatory. It includes model design (XGBoost, LSTM, ResNet), hybrid architectures, and performance validation, serving as the technical foundation for their subsequent repurposing in agriculture.

1.4.4 Application of Machine Learning Models in Agriculture (Chapter 4)

Describes the reconfiguration of these astrophysical ML methods for agro-environmental prediction. It includes climate risk assessment using Bayesian Networks, temperature forecasting with hybrid deep learning models, wind component prediction, and agent-based simulation of mango farm dynamics in Sicily.

1.4.5 Bridging Astrophysics and Agriculture: A Transformative Framework (Chapter 5)

Synthesizes insights from previous chapters to propose a generalizable framework for interdisciplinary model transfer. It evaluates the performance, robustness, and scalability of the

approach and reflects on broader implications for sustainable development and data-driven environmental management.

Chapter 2

Machine Learning: A Comprehensive Review

2.1 Summary

The rise of ML has transformed the way scientists interpret complex systems across disciplines. In astroparticle physics, ML techniques are deployed to identify rare cosmic events such as inclined muon showers, reconstruct particle trajectories, and manage real-time signal streams across massive distributed detector networks (Abdul Halim et al., 2023a). Similarly, agriculture particularly in climate-sensitive regions like Sicily is increasingly turning to ML to interpret satellite data, forecast environmental stress, and support precision farming decisions under conditions of uncertainty (Pulighe et al., 2024). Despite their distinct scientific objectives, astrophysics and agriculture share fundamental computational challenges: multivariate data with noise, spatio-temporal correlations, uncertainty in measurement, and limited ground-truth supervision. Both domains rely on large, heterogeneous datasets generated by sensor networks whether photomultiplier tubes in water-Cherenkov detectors or satellite-derived indices such as NDVI, LST, and soil moisture content.This convergence of computational demands offers an unprecedented opportunity for cross-domain knowledge transfer (Zhang et al., 2021). In this thesis, we investigate how advanced ML methods originally developed in astrophysics such as XGBoost, Long Short-Term Memory (LSTM) networks, and Residual Networks (ResNet) can be adapted and reconfigured for use in climate-resilient mango cultivation. Unlike previous studies that treat each domain in isolation, this chapter aims to:

- Present a unified analysis of ML methods used in both domains,
- Provide quantitative and technical comparisons, and
- Justify the model selection and hybridization strategies used in Chapters 3 and 4.

To that end, this chapter is structured into two main reviews**,** ML in astrophysics and ML in agriculture followed by a comprehensive comparative table that summarizes model characteristics, strengths, and limitations across both contexts. A concluding section critically evaluates the models selected for implementation and discusses the rationale behind excluding others.

This interdisciplinary synthesis sets the foundation for the technical developments in subsequent chapters and illustrates how ML architectures, originally designed for decoding cosmic signals, can also decode climatic stress signals in agriculture.

2.2 Machine Learning in Astrophysics

2.2.1 XGBoost: Tree-Boosted Regression for Tabular Inputs

XGBoost (eXtreme Gradient Boosting) is an ensemble learning method that builds a series of decision trees, each one minimizing residual errors from the previous tree while applying regularization to avoid overfitting (Chen & Guestrin, 2016a; He et al., 2016).

Objective Function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum j = 1^{T} w_j^2$$

Where:

- l is the loss function (e.g., squared error),
- **γ** penalizes tree complexity (number of leaves T),
- **λ** is the L2 regularization term.

Application in Astrophysics:

In this thesis, XGBoost was applied to the regression of zenith angles based on engineered PMT signal features collected from the water-Cherenkov detectors, demonstrating high accuracy and robustness under noisy conditions.

Strengths:

- High interpretability
- Native handling of missing values
- Fast, scalable training with regularization

Weaknesses:

- Relies on pre-engineered features
- No built-in memory for temporal correlations

2.2.2 LSTM: Long-Term Dependency Modeling in Temporal Signals

Long Short-Term Memory (LSTM) networks are a specialized form of Recurrent Neural Networks (RNNs) designed to overcome the vanishing gradient problem through the use of memory gates that regulate the flow of information across time steps. At each step t, the LSTM processes the current input x_t and previous hidden state h_{t-1} through three primary gates:

- The forget gate, which determines which past information to discard,
- The input gate, which selects new information to store, and
- The output gate, which controls what information contributes to the current output.

The figure 1 illustrated that the cell state C_t serves as a persistent memory that is updated using the previous cell state and a newly generated candidate state. This architecture enables LSTMs to manage long-range dependencies and learn temporal patterns over extended sequences. They have been widely applied in domains such as time series forecasting, natural language processing, and audio signal interpretation (Dara et al., 2023; Graves & Schmidhuber, n.d.; Shiri et al., 2023).

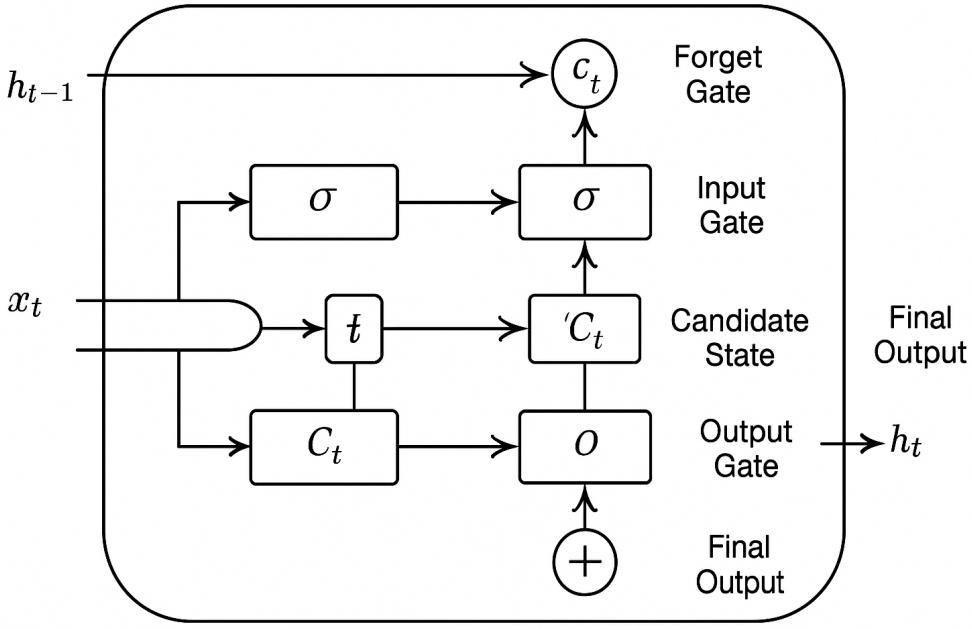


Figure 1 LSTM Architecture (original figure generated by author)

Application in Astrophysics:

In this thesis, LSTM networks were used to model temporal dependencies in photomultiplier tube (PMT) signal traces recorded by water-Cherenkov detectors. These signals, particularly

from inclined muon events, exhibit irregular timing shifts and amplitude drifts across detector stations. The LSTM's ability to capture delayed signal structure and sequence dynamics proved instrumental in improving zenith angle regression, particularly in noisy or temporally misaligned datasets (Chapter 3, Section 3.4).

Strengths:

- Excellent for irregular and delayed signals
- Preserves sequence structure
- Learns temporal relationships automatically

Weaknesses:

- Sensitive to hyperparameters (e.g., sequence length)
- Overfitting risk if sequence is short or noisy
- Slower to train than tree-based models

2.2.3 ResNet: Deep Spatial Feature Extraction with Skip Connections

Residual Networks (ResNets) solve vanishing gradient problems in deep neural networks using skip connections that enable the model to learn identity mappings (He et al., 2016).

Residual Equation:

$$y = F(x, \{W_i\}) + x$$

Where:

- F(x) represents convolutional transformations (Conv → ReLU → BN),
- x is the original input to the layer.

F(x) represents the sequence of transformations applied to the input, typically consisting of a convolutional layer (Conv) for feature extraction, followed by a Rectified Linear Unit (ReLU) activation function to introduce non-linearity, and a Batch Normalization (BN) layer to stabilize and accelerate training. This structure allows the network to learn complex spatial hierarchies while mitigating internal covariate shifts.

Application in Astrophysics:

In this thesis it used to extract spatial correlations across tanks and waveform patterns. Captured geometric information vital for distinguishing vertical vs. inclined events.

Strengths:

- Effective in deep architectures
- Learns spatial structure without handcrafted features
- Mitigates vanishing gradient issue

Weaknesses:

- Requires large datasets
- Computationally intensive
- Less interpretable than tree models

2.2.4 Hybrid Models

To harness complementary strengths of multiple models, hybrid architectures were explored:

LSTM + XGBoost

- LSTM modeled raw signal sequence.
- XGBoost refined outputs with engineered features.
- Achieved better generalization and accuracy than either model alone .

ResNet + XGBoost

- ResNet captured spatial waveform patterns.
- XGBoost used final activations to regress zenith angles.
- Outperformed standalone ResNet in cases with limited training data .

Strengths:

- Combine sequence memory, spatial learning, and interpretability.
- Improved robustness under noise and irregular sampling.

Weaknesses:

- Require coordination between models
- More hyperparameters and training stages
- Slightly harder to interpret than single-model outputs

2.3 Machine Learning in Agriculture (Lesson learn from Mango and Avocado)

Modern agriculture, particularly under climate stress, increasingly depends on machine learning (ML) to support predictive decision-making. This need is especially pronounced in

regions like Sicily, where extreme temperature variability, erratic rainfall, and high wind conditions jeopardize crops like mango and avocado.

The complexity of agro-meteorological systems lies in their spatiotemporal heterogeneity, non-linearity, and stochastic behavior all of which necessitate models that can handle missing data, sequence variability, and non-trivial interactions between environmental variables.

2.3.1 Random Forest (RF)

Random Forest is an ensemble learning technique based on bootstrap aggregation (bagging) of decision trees. Each tree is trained on a subset of data and features, and predictions are averaged to reduce variance (Asha et al., 2020).

Prediction:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

Where:

- T_b: prediction of the b-th decision tree

Applications:

Used in yield proxy estimation and temperature forecasting (Chapter 4), and as part of a hybrid network for wind prediction.

Strengths:

- Robust to overfitting
- Handles categorical and continuous variables
- Minimal tuning required

Weaknesses:

- Not sequence-aware
- Slow with high-dimensional data
- Limited extrapolation beyond training range

2.3.2 Bayesian Networks (BNs)

Bayesian Networks are graphical models representing probabilistic dependencies among variables. They are effective for reasoning under uncertainty and encoding expert knowledge (Pham et al., 2024).

Joint Distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\big(X_i \mid \text{Parents}(X_i)\big)$$

Applications:

Used to assess multi-risk scenarios (temperature, precipitation and wind) under climate change using historical sensor data and expert-informed structure.

Strengths:

- Intuitive graphical structure
- Captures uncertainty
- Works with limited data

Weaknesses:

- Sensitive to graph structure
- Requires expert domain knowledge
- Not suitable for high-frequency forecasting

2.3.3 Transformer Models

Transformers are deep learning models designed for long-sequence modeling using attention mechanisms rather than recurrence (Vaswani et al., 2017).

Scaled Dot-Product Attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where Q, K, V are the query, key, and value matrices; d_k is the dimension of keys.

Applications:

Used in Chapter 4 for long-term temperature forecasting with Fourier feature encoding and residual correction using ARIMA.

Strengths:

- Captures long-range dependencies
- Parallelizable and scalable
- Performs well on long time series

Weaknesses:

- Data-hungry
- Needs normalization and tuning
- Harder to interpret

2.3.4 Feedforward Neural Networks (ANNs)

ANNs, particularly feedforward networks, are utilized to capture intricate patterns in temperature data. By learning from historical temperature records and related variables, ANNs can forecast future temperatures with notable accuracy.

A study by Kumar et al. (2023) demonstrates the application of ANNs in temperature forecasting. The authors developed a model that predicts air temperature using a three-layer backpropagation ANN combined with meteorological data, achieving a coefficient of determination value of over 99% across different seasons (Kumar & Elumalai, 2023).

2.3.5 Multilayer Perceptron (MLPs)

Basic multilayer perceptrons (MLPs) used in this thesis serve as nonlinear regression models for temperature and yield proxy prediction (Robson et al., 2017).

Forward Pass:

$$h^{(l)} = \sigma\left(W^{(l)}h^{(l-1)} + b^{(l)}\right)$$

Where σ is a non-linear activation (ReLU), and $W^{(l)}$ and $b^{(l)}$ are the weights and biases of layer l.

Strengths:

- Captures complex nonlinear relationships
- Flexible structure
- Fast training

Weaknesses:

- No memory (unlike RNNs)
- Susceptible to overfitting
- Requires careful feature scaling

2.4 Comparative Summary of Selected ML Models in Astrophysics and Agriculture

Note: This table includes only a representative subset of models relevant to the thesis scope; many other ML techniques are used in astrophysical and agricultural research.

Model	*Domain(s)*	*Input Type*	*Primary Use Case*	*Strengths*	*Limitations*	*Reference(s)*
XGBoost	Both	Tabular, structured	Regression (zenith angle, temp, wind)	Fast, regularized, interpretable, handles missing data	Requires feature engineering, no sequence memory	Abdul Halim et al., 2023a, Chen & Guestrin, 2016b
LSTM	Both	Time series	Temporal pattern learning (PMT, climate)	Long-memory, noise tolerant, learns sequence relationships	Sensitive to tuning, risk of overfitting, slow to train	Abdul Halim et al., 2023a, Dara et al., 2023
ResNet	Both	Spatial sequences	Feature extraction (muon, temperature)	Deep spatial structure, stable gradient flow	Needs large data, harder to interpret	He et al., 2016
Transformer	Agriculture	Long sequences	Long-term temperature forecasting	Attention-based memory, strong on long dependencies	Complex, data-hungry, lower interpretability	Vaswani et al., 2017
Random Forest (RF)	Agriculture	Tabular	Temp, wind, and yield proxy estimation	Robust, interpretable, works on small datasets	No temporal memory, slower with large feature sets	Liakos et al., 2018
Bayesian Network	Agriculture	Multivariate	Risk assessment (climate uncertainty)	Probabilistic, interpretable, handles missing data	Needs expert design, poor scalability	Yet et al., 2020, Pham et al., 2024
ANN (Feedforward NN + MLP)	Agriculture	Tabular	Temperature/yield proxy prediction	Captures non-linear relationships, simple training	No memory, prone to overfitting	Zhang et al., 2021, Robson et al., 2017, Breiman 1996,2001

2.5 Conclusion

This chapter has presented a comparative review of machine learning (ML) methods employed in both **astrophysics** and **agriculture**, highlighting their strengths, limitations, and relevance to this thesis. Through a side-by-side analysis, it becomes clear that while the two domains differ in their scientific objectives, they share fundamental computational challenges, namely, high-dimensional data, environmental noise, temporal dependencies, and incomplete information.

In the astrophysical context, models such as **XGBoost**, **LSTM**, and **ResNet** were developed to process massive volumes of sensor-generated data with high temporal resolution and complex spatial correlations. Their use in detecting inclined muons where signal traces span multiple photomultiplier tubes with variable delay and intensity demonstrated these models' ability to learn from noisy, distributed, and non-linear data. Importantly, these models were validated in real-world observatories such as Pierre Auger, providing both theoretical and empirical justification for their robustness.

Recognizing these capabilities, this thesis strategically transferred and adapted these models to agro-meteorological applications. In agriculture, the focus shifts to **temperature forecasting**, **wind component prediction**, **proxy yield estimation**, and **climate risk analysis**. However, the underlying data challenges are similar: multi-modal, uncertain, and evolving across space and time.

As such, the models implemented in Chapters 3 and 4 were not selected arbitrarily. Their selection followed a principled logic:

- **XGBoost** was chosen for its interpretability, speed, and strong performance on structured climate datasets.
- **LSTM** was introduced to model time-dependent patterns in temperature and wind dynamics, particularly under seasonality and stochastic disturbance.
- **ResNet** was employed to learn spatial hierarchies in both Cherenkov traces and agricultural variables derived from topography and satellite imagery.
- **Transformer models**, introduced later in the thesis, extended sequence modeling beyond what LSTM could achieve, particularly in forecasting.
- **Random Forest** and **Bayesian Networks** were integrated due to their resilience to noise and ability to quantify uncertainty, which are critical in environmental systems.

In summary, the review and evaluation presented in this chapter not only supports the **methodological integrity** of this thesis but also underscores the **scientific novelty** of applying

advanced ML models originally trained to decode cosmic phenomena to forecasting agricultural outcomes in a Mediterranean climate. This interdisciplinary transfer, as demonstrated in Chapters 3 and 4, forms the core innovation and contribution of this work.

Chapter 3

Machine Learning Models for Astrophysics

Part of this report has been published in PoS Journal- https://doi.org/10.22323/1.484.0115

Neural network identification of highly inclined muons in water-Cherenkov particle detectors

Mohsen Pourmohammad Shahvar[1,2], Giovanni Marsella[1,2], Markus Roth[3] and David Schmidt[3]

[1] *INFN sezione di Catania*

[2] *Università degli studi di Palermo*

[3] *Karlsruher Institut für Technologie (KIT)*

PIERRE AUGER OBSERVATORY

PROCEEDINGS OF SCIENCE

3.1 Abstract:

This study focuses on identifying highly inclined muons in water-Cherenkov detectors similar to those used by the Pierre Auger Observatory using neural networks. Highly inclined muons, which are distinctive signatures of air showers induced by neutrinos or cosmic rays arriving at significant inclinations, offer a lower background rate compared to less inclined atmospheric particles. We explore the transition from conventional statistical approaches to machine learning methodologies to identify highly inclined muons by leveraging their unique signatures in the temporal signal distributions of three photosensors that uniformly observe the volume of a water-Cherenkov detector. By adopting machine learning, particularly neural network techniques, we aim to enhance the identification of highly inclined muons, thus improving triggering schemas designed for detecting neutrino primaries. This study not only advances the identification of highly inclined muons but also investigates the optimization of machine learning models for their efficient recognition within the water-Cherenkov detector setup.

3.2 Introduction

The Pierre Auger Observatory, a pioneering institution in astroparticle physics, is dedicated to investigating cosmic rays and neutrinos of extraordinary energies exceeding 10^{19} eV.

The Pierre Auger Observatory, designed for an extensive study of cosmic rays at ultra-high energies, employs a combination of surface water-Cherenkov detectors and air fluorescence telescopes to investigate air shower phenomena comprehensively (Abdul Halim et al., 2023a). These instruments, working in unison, create a robust platform for determining the energy, direction, and composition of the most energetic particles in the universe (Allekotte et al., 2008; The Pierre Auger Collaboration, 2005). The air fluorescence telescopes, operational under dark moonless conditions, capture the electromagnetic showers generated by primary particles interacting with the upper atmosphere. Meanwhile, the surface array gauges particle densities upon impact, while the fluorescence telescopes provide a near-calorimetric energy assessment, transferable to the surface array for enhanced data collection (Abdul Halim et al., 2023a; Abreu

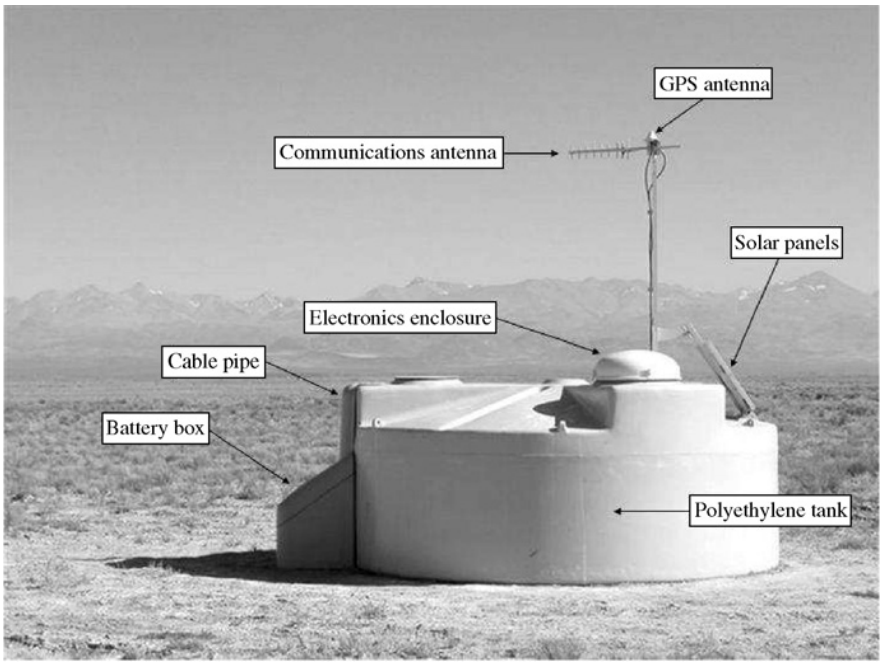


Figure 2 A schematic view of a surface detector station in the field showing (Allekotte et al., 2008)

et al., 2023; Roth, 2007).The selection of water-Cherenkov detectors for the surface array was based on their durability, cost-effectiveness, and uniform exposure to a wide range of zenithal angles (Lawrence et al., 1991; The Pierre Auger Collaboration, 2005). These detectors offer sensitivity to charged particles and energetic photons that convert to pairs within the water medium, showcasing success in prior experiments. The Pierre Auger Observatory comprises 1660 surface detector stations, each featuring a 3.6 m diameter water tank containing 12,000

liters of purified water. These detectors form a regular hexagonal grid across ~3,000 km², with an additional denser infill array designed to capture lower-energy events (Allekotte et al., 2008; P. A. Collaboration, 2022).

The detector spacing, determined by a balance between cost-effectiveness and energy threshold considerations, facilitates efficient data sampling across the array. Differential GPS systems enable precise station positioning for accurate shower reconstruction, essential for comprehensive data analysis (Allekotte et al., 2008).

3.2.1 Tank System

The tank system consists of cylindrical water-Cherenkov detectors designed for simplicity and cost-effectiveness. The tanks have a complex top structure to provide rigidity for external components and to accommodate internal photomultiplier assemblies and cabling. Standing at a maximum height of 1.6m, the tanks are transportable within regulations and are colored beige to blend with the environment (Allekotte et al., 2008).

3.2.2 Distinguishing Muonic Showers

The distinction between inclined and vertical muonic showers is fundamental to understanding the origins and characteristics of cosmic rays and neutrinos (Abreu et al., 2023; Valiño et al., 2010). Muonic showers are categorized based on the zenith angle of their trajectories relative to the vertical direction. Vertical showers, characterized by small zenith angles, originate from cosmic ray interactions occurring near the zenith, while inclined showers result from interactions at larger zenith angles (Abdul Halim et al., 2023b; T. P. A. Collaboration, 2013). This categorization is essential as inclined muons offer a distinct signature of air showers induced by neutrinos or cosmic rays arriving at substantial inclinations, presenting an opportunity to minimize background noise and enhance the sensitivity of astroparticle measurements (Abraham et al., 2010; Kampert et al., 2019).

3.2.3 Triggering and Event Selection in the Pierre Auger Observatory Surface Detector Array:

Traditionally, identifying highly inclined muons within water-Cherenkov detectors has relied on empirical criteria and statistical analyses applied to temporal signal distributions recorded by PMTs. However, these methods often struggle to distinguish subtle muon signatures from background noise effectively. Thus, exploring alternative approaches to improve identification accuracy is crucial (The Pierre Auger Collaboration et al., 2009).

The Unified Readout Board (UUB) in the Pierre Auger Observatory serves as the interface between water-Cherenkov detectors (WCDs) and the data acquisition system, ensuring efficient signal processing and transmission. Each UUB processes signals from multiple PMTs within a WCD, detecting Cherenkov radiation generated by charged particles like muons. After initial amplification and shaping by the front-end components integrated within the Unified Readout Board (UUB), signals from the PMTs are processed and transmitted to the central acquisition system (Sato et al., 2023; The Pierre Auger Collaboration et al., 2023).

According to the Pierre Auger Experiment study (Abraham et al., 2011), the trigger system for the surface detector array at the Pierre Auger Observatory is hierarchical, consisting of T1, T2, and T3 triggers. The T1 triggers initiates data acquisition in each water Cherenkov detector, utilizing two independent trigger modes. The first mode, known as the "Threshold" trigger (TH), requires the coincidence of the three PMTs each above 1.75 times the peak Vertical Equivalent Muon (VEM) value. This threshold effectively selects large signals, particularly from highly inclined showers dominated by muons, reducing the rate due to atmospheric muons from approximately 3 kHz to about 100 Hz. The second mode, the "Time-over-Threshold" trigger (ToT), selects sequences of small signals spread in time, requiring at least 13 bins (i.e., >325 ns) in a sliding window of 3 μs to be above a threshold of 0.2 times the peak VEM value in coincidence in 2 out of 3 PMTs. This trigger is optimized for detecting near-by, low-energy showers dominated by the electromagnetic component or high-energy showers with a distant core (Huege, 2023).

Figure 2 presents the trigger efficiency curve P(s), which represents the probability of detecting an event as a function of the signal strength measured in Vertical Equivalent Muon (VEM) units. Each data point corresponds to experimental measurements, and the curve illustrates how

the probability of triggering increases with signal amplitude, plateauing near full detection above ~8 VEM (Allard et al., 2005). While this curve reflects the performance under the electronics described by Allard et al. (2005), the upgraded UUB system introduced in AugerPrime is expected to improve time resolution and dynamic range. Updated efficiency studies for these new electronics are still being evaluated and are not yet fully published.

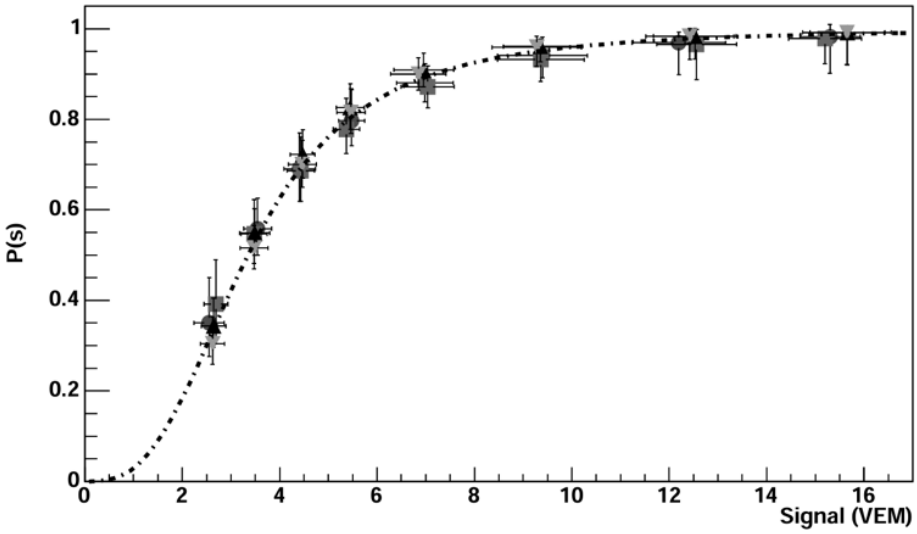


Figure 3 Trigger efficiency curve showing the probability P(s) of triggering as a function of signal strength (in VEM units). The plot demonstrates the combined performance of the Threshold (TH) and Time-over-Threshold (ToT) triggers. Each point represents aggregated measurements from the detector array. Source: Adapted from Allard et al., 2005

At the T2 level, triggers are applied in the station controller to further reduce the event rate per detector to about 20 Hz. All ToT-T1 triggers are promoted to the T2 level, while TH-T1 triggers are required to pass a higher threshold of 3.2 times the peak VEM value in coincidence among the three PMTs. The T3 trigger, formed at the CDAS based on the spatial and temporal combination of T2 triggers, initiates the central data acquisition from the array (Abraham et al., 2011; Huege, 2023; The Pierre Auger Collaboration et al., 2009).

The trigger of the array is realized in two modes: "ToT2C1&3C2" and "2C1&3C2&4C4". The former requires the coincidence of at least three detectors passing the ToT condition, while the latter requires a four-fold coincidence of any T2 with moderate compactness. Both modes utilize timing criteria based on the time spread of signals, which depends on distance and zenith angle, effectively selecting events with varying characteristics (Abraham et al., 2011; Huege, 2023; The Pierre Auger Collaboration et al., 2009).

Overall, the trigger system, incorporating VEM values and time-over-threshold measurements, enhances the array's capability to identify and reconstruct cosmic ray showers accurately, contributing significantly to astroparticle physics (Abraham et al., 2011).

3.2.4 Machine Learning Paradigm:

In response to this challenge, our study embraces the paradigm shift from conventional statistical methods to machine learning techniques, particularly focusing on neural network architectures. By harnessing the power of neural networks, we aim to uncover intricate patterns and correlations in the temporal signal distributions recorded by PMTs, enabling more accurate and efficient identification of highly inclined muons. Additionally, we explore the potential of long short-term memory (LSTM) networks and gradient boosting algorithms, such as XGBoost, to further enhance the discriminatory capabilities of our models. Moreover, we employed a hybrid ResNet + XGBoost model, leveraging the feature extraction strengths of ResNet with the robust classification power of XGBoost, to achieve superior performance.

In summary, this study represents a multifaceted endeavor to advance the field of astroparticle physics through the integration of machine learning methodologies for muon identification in water-Cherenkov detectors.

3.2.5 Muon Simulation and Feature Extraction

In the pursuit of enhancing muon identification within water-Cherenkov detectors (WCDs), we commence our exploration by simulating single muon events injected into a WCD tank. As showon in figure 4, this tank, with a diameter of 1.8 meters and a height of 1.2 meters, houses three photomultiplier tubes (PMTs) placed at equal distances from the tank center, forming an equilateral triangle at the base to ensure isotropic light collection. The muon events were simulated using the official Pierre Auger Offline Framework, which enables high-fidelity modeling of particle interactions and Cherenkov light propagation within the WCD geometry.

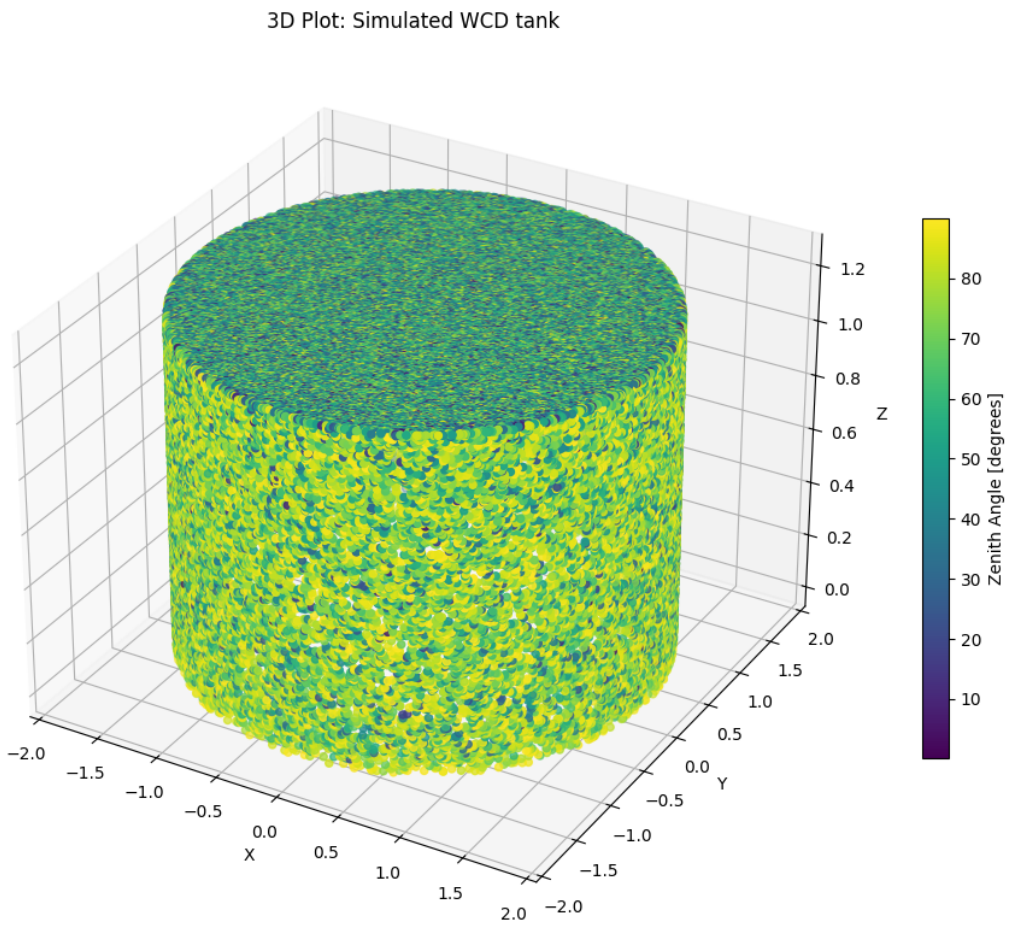


*Figure 4 Simulated WCD tank showing Cherenkov photon hit distribution under muon injection. The color scale corresponds to the **zenith angle** of injected muons, with higher angles shown in yellow and lower angles in blue. (Originally generated by the author.)*

By simulating muon injections from various coordinates and directions, we generated >>1000K events, each characterized by its x, y, z vector, enabling the determination of zenith and azimuth angles.

The distribution of zenith angles, ranging from 0 to 90 degrees (10 degree bin), elucidates the transition from vertical to inclined muons. As our goal is to distinguish between vertical and inclined muons, we focus on zenith angles, where 0-60 degrees represent vertical trajectories, and 60-90 degrees denote inclined paths (Example in Figure 6).

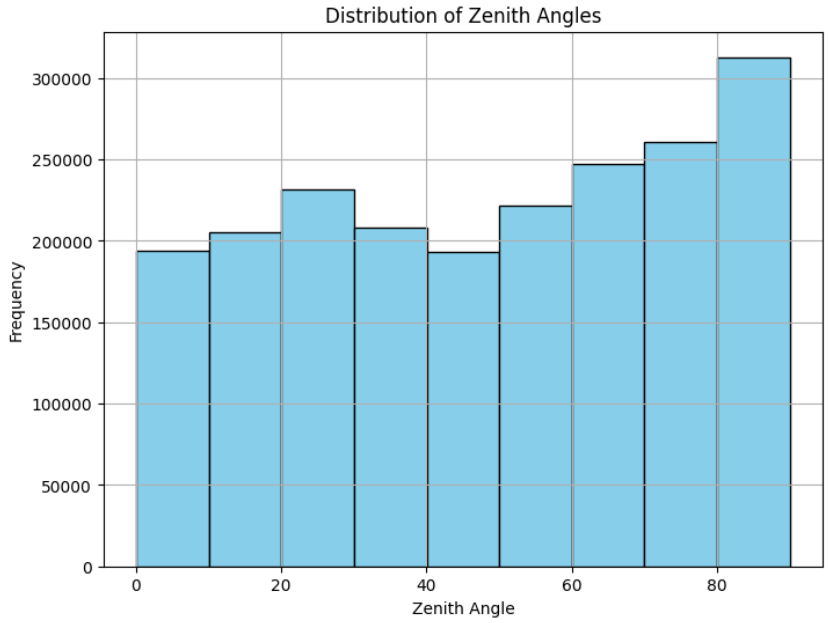


Figure 5 Distribution of Zenith Angle in Our Dataset

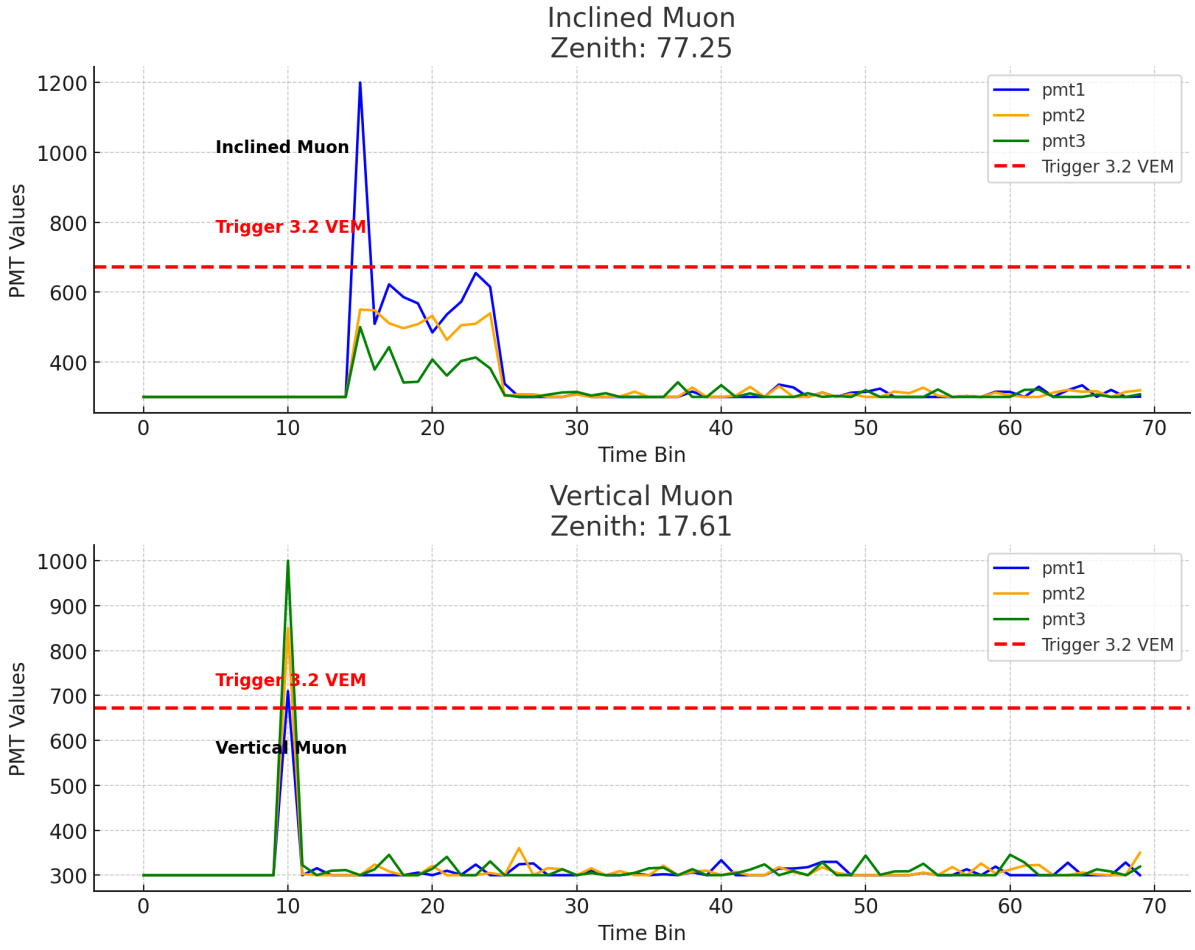


Figure 6 Examples of PMTs Traces for both Inclined and Vertical Events from the simulation data by Author

3.2.6 Analytical Insights

An analytical examination of the absolute values of relative differences in highest peaks among the three PMTs reveals a correlation with zenith angles. This insight suggests that mean values of relative peak differences can serve as thresholds to differentiate between vertical and inclined events. However, as it is shown in figure 7, fluctuations and errors in this metric highlight the challenges in achieving accurate muon classification solely based on peak differences (~63% Accuracy- This value was evaluated using a confusion matrix derived from binary classification of muon events (vertical vs. inclined) based solely on peak difference thresholds. Accuracy was calculated as the ratio of correctly classified events (true positives and true negatives) to the total number of events). Moreover, the rise time metric exhibits variability between vertical and inclined muons, further emphasizing the importance of comprehensive feature extraction for robust classification.

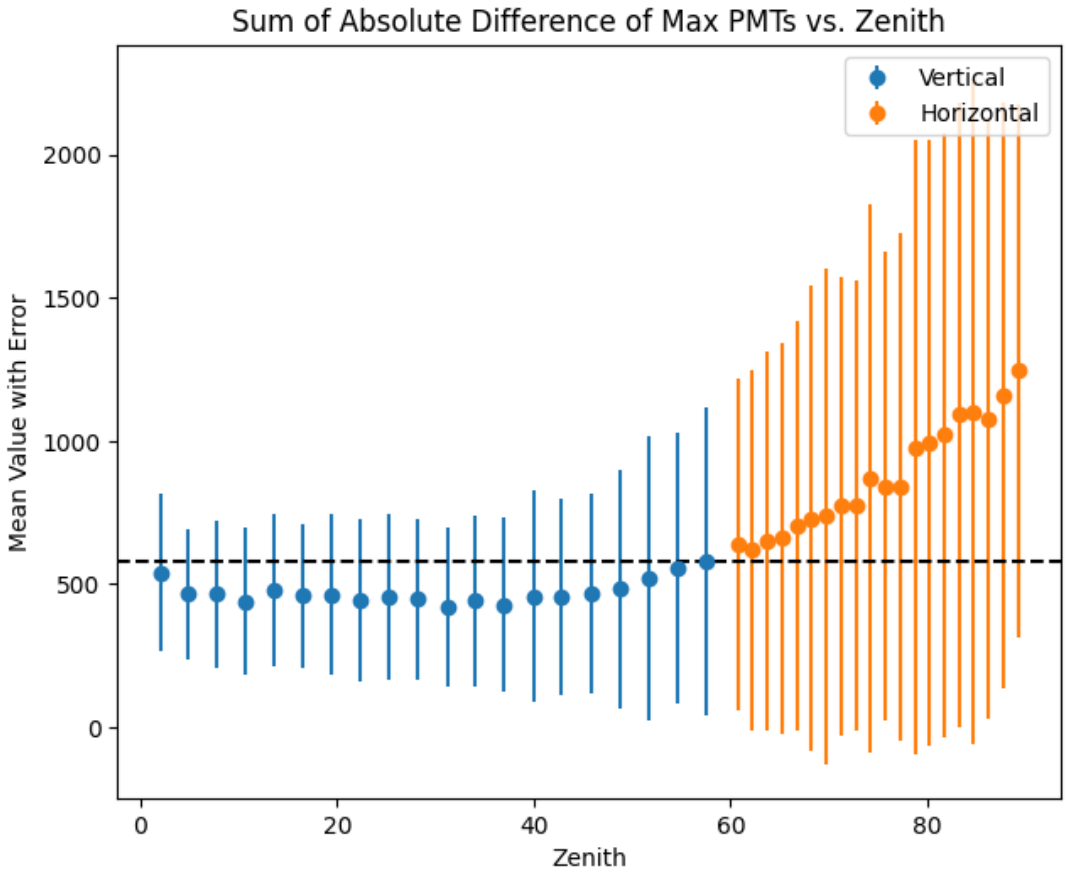


Figure 7 Function of absolute values of differences in highest peaks vs. Zenith

Preprocessing involves extracting key features from the raw PMT signals to enhance the accuracy of muon identification. These features include the **highest peak values**, which represent the maximum signal amplitude and help in distinguishing strong particle interactions; the **integrals of PMT traces**, which provide a measure of the total signal energy over time, capturing the overall intensity of the Cherenkov light; and the **rise times**, which reflect the speed at which the signal reaches its peak, offering insights into the timing and dynamics of the muon's passage through the detector. Together, these features encapsulate critical information to forming the basis for effective signal classification (See figure 8).

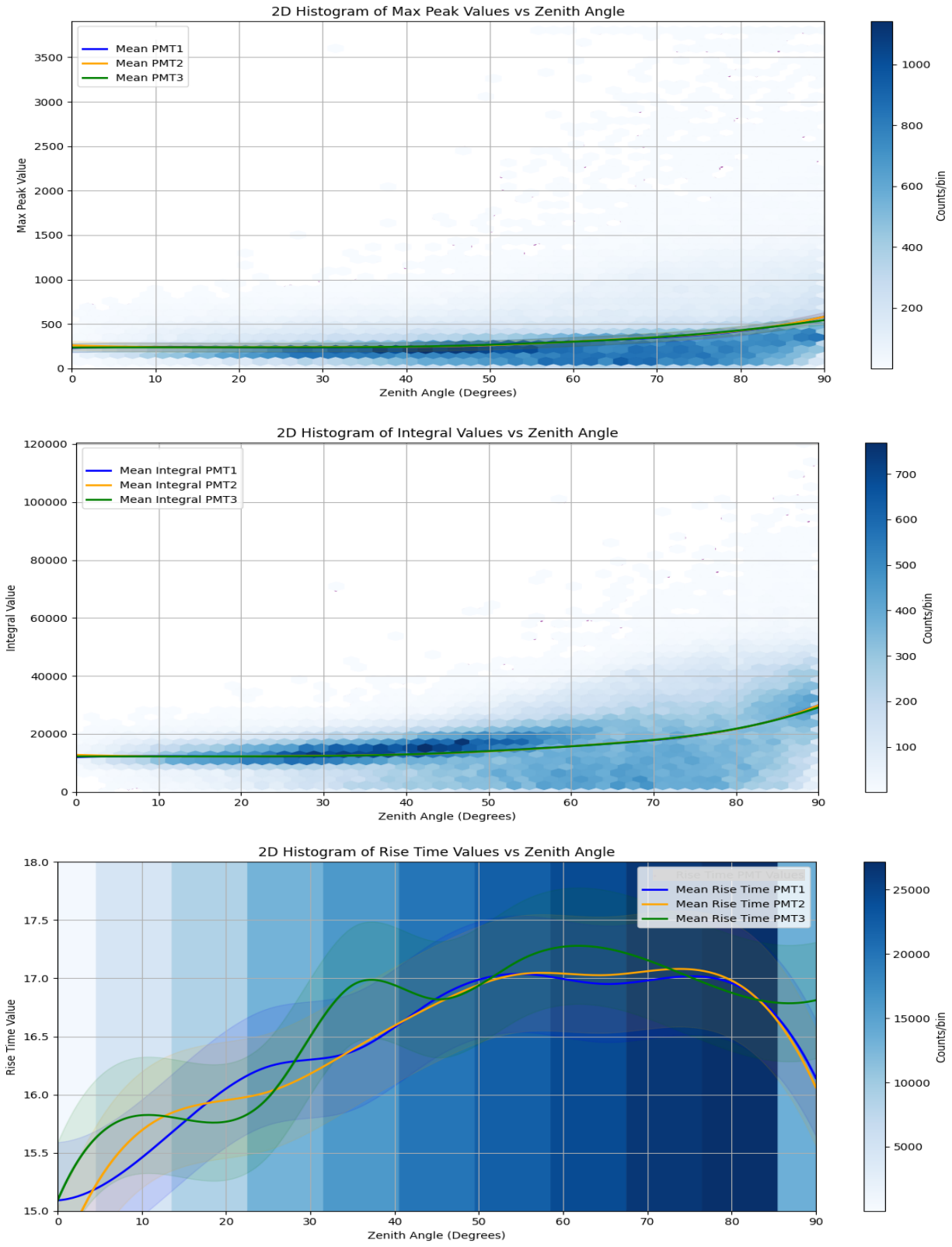


Figure 8 Relation of selected features versus zenith angle. Each subplot shows a 2D histogram overlaid with mean trends for PMTs 1–3. The background color density represents the overall distribution of the raw feature (e.g., max peak, integral, rise time), while the overlaid lines represent the mean trend per PMT

3.3 XGBoost Regression:

XGBoost, widely recognized for its strength in classification, is also highly effective in tackling regression problems, where the aim is to predict continuous values. In this section, we'll explore how XGBoost works in the context of regression, breaking down its mathematical foundation and shedding light on the principles that drive its impressive performance (Chen & Guestrin, 2016b). The schematic of XGBoost is depicted in figure 9 (Badugu et al., 2024).

3.3.1 Objective Function:

At the core of XGBoost regression is an objective function that quantifies the model's performance and guides the optimization process. The objective function comprises two components: a loss function and a regularization term (Chen & Guestrin, 2016b; Montomoli et al., 2021).

3.3.1.1 Loss Function:

The loss function measures the disparity between the predicted and actual values. For regression tasks, common loss functions include mean squared error (MSE) and mean absolute error (MAE), among others. Mathematically, the loss function can be denoted as (Chen & Guestrin, 2016b)

$$loss(y_i, \hat{y}_i)$$

where y_i represents the actual target value and $\hat{y}_i$ denotes the predicted value for the i^{th} instance.

3.3.1.2 Regularization Term:

The regularization term penalizes the complexity of the model to prevent overfitting. It discourages excessively complex models by adding a penalty term to the objective function. Regularization techniques such as L1 regularization (Lasso) and L2 regularization (Ridge) are commonly employed to achieve this. Mathematically, the regularization term can be represented as $\Omega(f_k)$, where (f_k) represents the k^{th} weak learner (e.g., decision tree) (Chen & Guestrin, 2016b).

Thus, the overall objective function for XGBoost regression can be formulated as:

$$Objective = \sum_{i=1}^{n} loss(y_i, \hat{y}_i) + \sum_{K=1}^{K} \Omega(f_k)$$

Where:

n is the number of instances in the dataset.

K is the total number of weak learners (trees) in the ensemble.

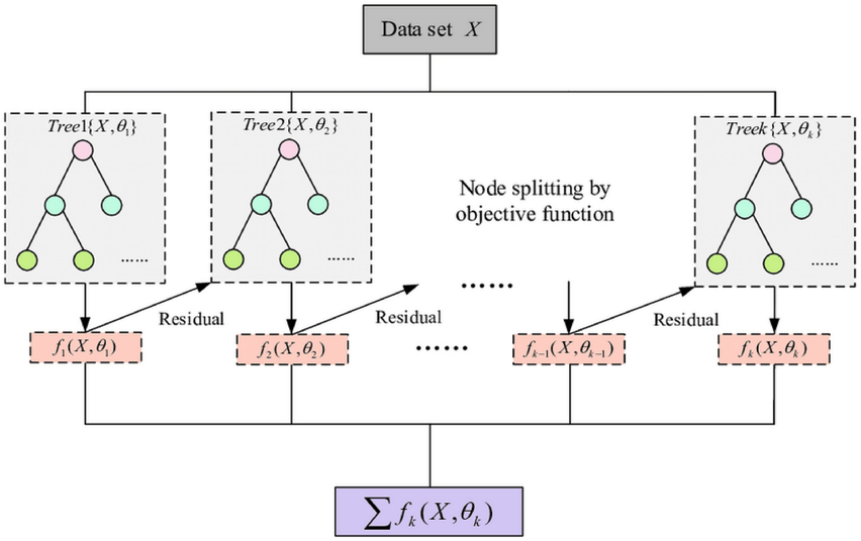


Figure 9 Schematic representation of XGBoost (Badugu et al., 2024)

3.3.2 Optimization Process:

XGBoost employs a gradient boosting framework to optimize the objective function iteratively. The optimization process involves sequentially adding weak learners to the ensemble, with each new learner focusing on reducing the residuals (errors) of its predecessors. Gradient descent is used to minimize the objective function by updating the model parameters in the direction of the steepest descent(Zhang et al., 2021).

3.3.3 Model Prediction:

Once the optimization process is complete, the XGBoost model can make predictions for new instances by aggregating the predictions of all weak learners. The final prediction is obtained

by summing the predictions of individual trees, possibly weighted by their contribution to the ensemble.

In summary, XGBoost regression leverages an objective function comprising a loss function and a regularization term to optimize model performance. By iteratively adding weak learners and minimizing the objective function, XGBoost constructs a robust regression model capable of making accurate predictions for continuous outcomes.

3.4 LSTM

Long Short-Term Memory (LSTM) networks represent a significant advancement in the field of artificial intelligence, specifically designed to model sequential data with long-range dependencies (Graves & Schmidhuber, n.d.). Unlike traditional Recurrent Neural Networks (RNNs), LSTM networks are equipped with specialized gating mechanisms, enabling them to preserve essential information over extended time intervals (Shiri et al., 2023). This capability makes LSTM networks particularly well-suited for tasks such as time series prediction, natural language processing, and signal processing(Staudemeyer & Morris, 2019).

3.4.1 Mathematical Formulation

We can break down the maths behind the LSTM model as (Dara et al., 2023; Graves & Schmidhuber, n.d.; Shiri et al., 2023):

1. Input Gate (i_t):

$$i_t=\sigma(W_{ix}x_t+W_{ih}h_{t-1}+b_i)$$

2. Forget Gate (f_t):

$$f_t=\sigma(W_{fx}xt+W_{fh}h_{t-1}+b_f)$$

3. Output Gate (O_t):

$$O_t=\sigma(W_{ox}x_t+W_{oh}h_t-1+b_o)$$

4. Cell State Update (C_s):

$$\widetilde{C_t}= tanh(W_{cx}x_t+W_{ch}h_{t-1}+b_c)$$

$$C_t=f_t \,.\, C_{t-1} + i_t \,.\, \widetilde{C_t}$$

5. Output Calculation (h_t):

$$h_t=O_t \,.\, tanh(C_t)$$

Where the parameters defined in the figure 10 (Dara et al., 2023).

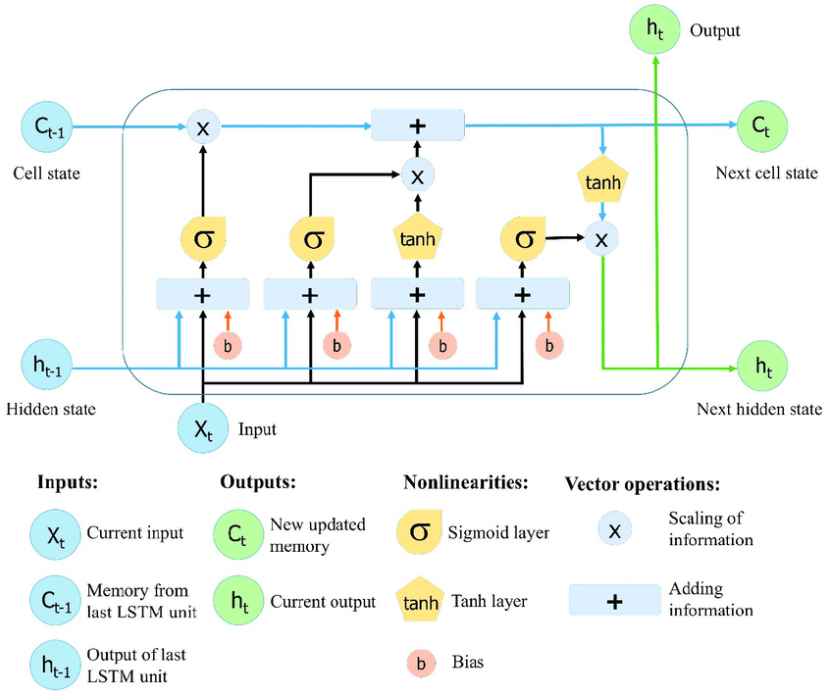


Figure 10 LSTM Architecture (Dara et al., 2023)

3.4.2 Working Process:

In this study, we use **Long Short-Term Memory (LSTM)** networks to predict the **Zenith Angle** of particle trajectories based on sequential features derived from photomultiplier tube (PMT) signals: **highest peak values**, **integrals of PMT traces**, and **rise times**. LSTMs are particularly suitable for this task due to their ability to capture long-range dependencies in sequential data. The model uses only derived features (Peak, Integral, Rise Time) as inputs, and not raw spatial coordinates. Zenith angle is computed from the direction vector and used as a prediction target. The dataset is randomly subdivided using an 70/30 split: 70% for training and 30% for validation. In LSTM experiments, sequences are fed in time steps, while the fully connected model operates on flattened, time-averaged input vectors.

3.4.2.1 Input Processing:

At each time step t , the input feature vector $X_t$comprising PMTs_Highest_Peak, PMTs_Integral, and PMTs_Rise Time is fed into the LSTM unit. Along with X_t , the previous hidden state h_{t-1} and cell state C_{t-1} are passed into the LSTM, allowing the model to leverage historical information from earlier time steps.

3.4.2.2 Gate Activation:

The input and forget gates allow the LSTM to dynamically decide what new information to store and what old information to discard, depending on the relevance of the features to the prediction of the Zenith Angle.

3.4.2.3 Cell State Update:

Based on the input gate and forget gate, the cell state C_t ensures that the model retains relevant information for accurately predicting the Zenith Angle, while discarding irrelevant or noisy data.

3.4.2.4 Output Calculation:

O_t determines which information from the current cell state C_t contributes to the output h_t, which represents the predicted value of the Zenith Angle at the current time step. The hidden state h_t serves as the network's prediction output for the Zenith Angle, while also being passed forward to the next time step.

3.5 ResNet

ResNet (Residual Network) represents a significant advancement in deep learning, addressing challenges associated with training very deep neural networks, such as vanishing gradients. Introduced by He et al. (2016), ResNet introduces residual learning, allowing models to learn residual functions with reference to the layer inputs, rather than unreferenced functions. This approach facilitates the training of networks with substantially increased depth, leading to improved performance in various computer vision tasks, including image classification, object detection, and segmentation (He et al., 2016).

The core innovation of ResNet lies in its residual blocks, which enable the network to learn residual mappings. The mathematical formulation is as follows:

3.5.1 Residual Block:

$$y = \mathcal{F}(x, \{W_i\}) + x$$

Here, x is the input vector, y is the output vector, and $\mathcal{F}(x, \{W_i\})$ represents the residual mapping to be learned. The term $\mathcal{F}$ typically consists of two or more layers, and W_i denotes the weights of these layers. The addition of x via a shortcut connection helps mitigate the vanishing gradient problem by allowing gradients to flow directly through the network (He et al., 2016).

3.5.2 Identity Mapping:

In scenarios where the dimensions of x and $\mathcal{F}(x)$ differ, a linear projection W_s is applied to match dimensions: $y = \mathcal{F}(x, \{W_i\}) + W_s x$. This ensures that the addition operation is valid, maintaining the integrity of the residual learning framework (He et al., 2016).

3.5.3 Working Process:

In this study, we leverage the ResNet architecture to predict the Zenith Angle of particle trajectories based on features extracted from PMTs, including the highest peak values, integrals of PMT traces, and rise times. These features capture essential information about the signal characteristics, providing a rich dataset for training the ResNet model. The extracted features used to train the ResNet model are identical to those used in the LSTM pipeline, namely: maximum peak values, integrals of PMT traces, and rise times, derived from PMT1, PMT2, and PMT3. The zenith angle is again used as the target variable and is calculated from the simulation direction vector. The same dataset, preprocessing steps, and 70/30 training/testing split are applied to ensure consistency across models. The difference lies in the architecture: the ResNet treats the features as a 2D structured input with spatially-aware filters, while LSTM processes them as sequential time series.

3.5.3.1 Input Processing

Instead of using raw images, the input to our ResNet model consists of structured feature data derived from PMT signals:

These features are fed into the initial layers of the ResNet architecture, where a convolutional layer processes the input. This is often followed by batch normalization and a ReLU activation function, standardizing and enhancing the data for deeper layers.

3.5.3.2 Residual Learning

The processed input data is then passed through a series of residual blocks. Each residual block consists of:

- Convolutional layers that extract hierarchical patterns from the PMT features.
- Shortcut connections that directly add the input of each block to its output.

This residual connection allows the model to learn residual mappings rather than direct mappings, facilitating the training of deeper networks. The model can thus focus on the differences between predicted and actual Zenith Angles.

3.5.3.3 Gradient Flow

One of ResNet's key strengths is its ability to maintain effective gradient flow across deep layers. The shortcut connections ensure that gradients can propagate directly through the network during backpropagation, preventing the vanishing gradient problem. This feature is crucial for training deep networks, enabling them to learn complex dependencies between PMT features and Zenith Angles.

3.5.3.4 Output Generation

After the data passes through the residual blocks, the network applies a global average pooling layer to reduce the dimensionality of the feature maps, summarizing the information into a fixed-length vector. This is followed by a fully connected layer, which maps the extracted features to the target output: the Zenith Angle. For regression tasks like this one, the final activation function is typically linear, providing continuous predictions of the Zenith Angle.

3.6 Baseline Model:

As we strive to refine our model's performance for the Pierre Auger Observatory, it is essential to thoroughly evaluate the baseline model's effectiveness. The baseline model serves as a crucial benchmark, enabling us to understand how well current methods perform and offering

a reference point for assessing potential improvements. This evaluation helps determine the feasibility of enhancing our existing methodologies.

In the Pierre Auger Observatory, the standard trigger threshold — which we refer to as our baseline model — is defined as 3.2 Vertical Equivalent Muons (VEM) (Abraham et al., 2011). This threshold specifies the minimum signal amplitude required to initiate the detection process and is intended to reduce false positives from background noise. VEM acts as a standardized unit that facilitates comparison across events and detectors. However, while traditionally effective for high-energy, vertical muon events, this static threshold presents significant limitations when applied to a broader set of scenarios, including low-energy or inclined showers.

In our simulated dataset — which includes approximately one million events — we observe that applying this strict 3.2 VEM threshold results in an insufficient number of detected events. As illustrated in Figure 11, the threshold fails to capture the majority of valid particle interactions, rendering the resulting dataset sparse and inadequate for robust model training or evaluation. While in real experimental setups lowering the threshold could increase the likelihood of noise-induced false triggers, our simulation is structured to include only valid muon-like signal events. Thus, lowering the threshold in this controlled setting does not introduce noise, but rather improves coverage of true but low-amplitude events.

A more detailed analysis, shown in Figure 11, reveals that a significant proportion of triggered and untriggered events lies in the **0.5 to 1.5 VEM** range — well below the baseline cutoff. This pattern highlights the necessity of **recalibrating the trigger threshold** to more accurately reflect the distribution of relevant signal amplitudes. A dynamic or data-driven thresholding strategy would provide a more sensitive and comprehensive benchmark for comparison. Consequently, reassessing the baseline model is imperative for improving both detection coverage and the integrity of subsequent machine learning evaluations.

Further study of the simulation data as depicted in figure 11 reveals a visible concentration of triggered and untriggered events within the amplitude range of 0.5 to 1.5VEM. This observation underscores the necessity for a recalibration of the trigger threshold to facilitate a more nuanced evaluation framework. A meticulous reassessment of the trigger threshold is imperative to provide a more refined and comprehensive basis for comparison, enabling a more thorough assessment of model performance.

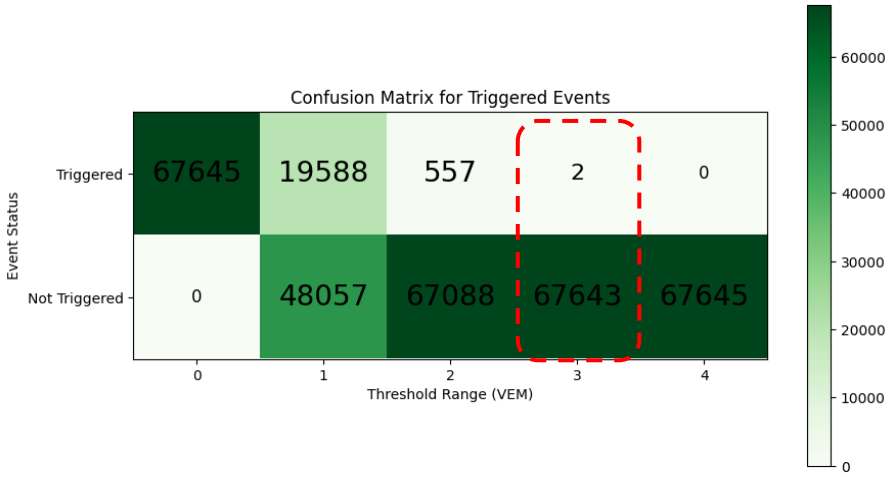


Figure 11 Triggered Vs. Not Triggered Events based on VEM values

1st Hybrid Approach: Integrating XGBoost-LSTM with Random Forest Regressor

In our pursuit to maximize the strengths of individual algorithms and create a more potent predictive model, we propose a hybrid approach that amalgamates XGBoost, LSTM, and a random forest regressor. This fusion leverages the distinct advantages of each component to enhance classification accuracy and robustness, particularly in the context of muon identification within water-Cherenkov detectors (WCDs).

3.7 XGBoost and LSTM Integration:

XGBoost and LSTM represent two formidable approaches in the realm of machine learning, each excelling in its own right. By integrating these two algorithms, we aim to harness the collective power of gradient boosting and sequential data modeling. XGBoost's capability to

handle complex datasets and combat overfitting complements LSTM's proficiency in capturing long-range dependencies within sequential data.

3.8 Random Forest Regressor

In addition to the fusion of XGBoost and LSTM, we incorporate a random forest regressor into our hybrid model. Renowned for its robustness and ensemble learning capabilities, the random forest regressor constructs multiple decision trees on random subsets of the dataset and aggregates their predictions through averaging or voting. This ensemble approach mitigates variance and enhances predictive accuracy by leveraging diverse sets of weak learners.

3.9 Modeling Strategy:

The hybrid XGBoost-LSTM model, augmented with a random forest regressor, offers a multifaceted approach to muon identification in WCD systems. By integrating these complementary algorithms, we aim to capitalize on their strengths and mitigate their individual limitations.

3.10 Advantages and Implications:

By harnessing the combined strengths of XGBoost, LSTM, and the random forest regressor, we expect to achieve superior classification accuracy, robustness, and interpretability. This advancement will enhance our ability to detect and characterize muons, significantly contributing to broader scientific efforts to unravel the mysteries of the universe.

3.11 Hybrid (LSTM + XGBoost) Model Architecture

The diagram illustrates a hybrid model architecture, The architecture begins with an **LSTM Input Layer**, which feeds sequential PMT data into an **LSTM Layer**. This layer outputs a feature vector of size 50, representing learned temporal features. These are further refined by a **Dense Output Layer**, reducing the LSTM output to a single prediction. Simultaneously, the **XGBoost Input Layer** provides its own prediction based on complementary feature sets. Both predictions are then merged in a **Concatenate Layer**, producing a combined feature vector. Finally, this vector is passed through a **Hybrid Output Layer**, a dense layer that generates the final Zenith Angle prediction.

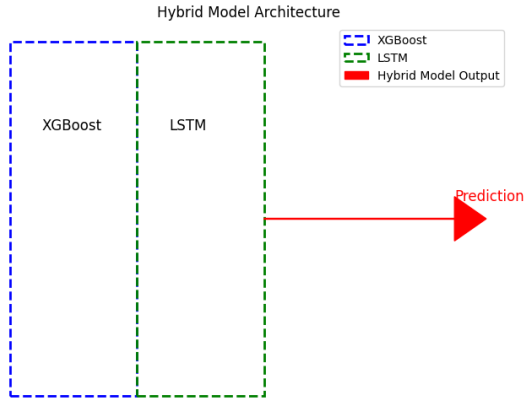


Figure 12 Schematic Architecture of Hybrid model

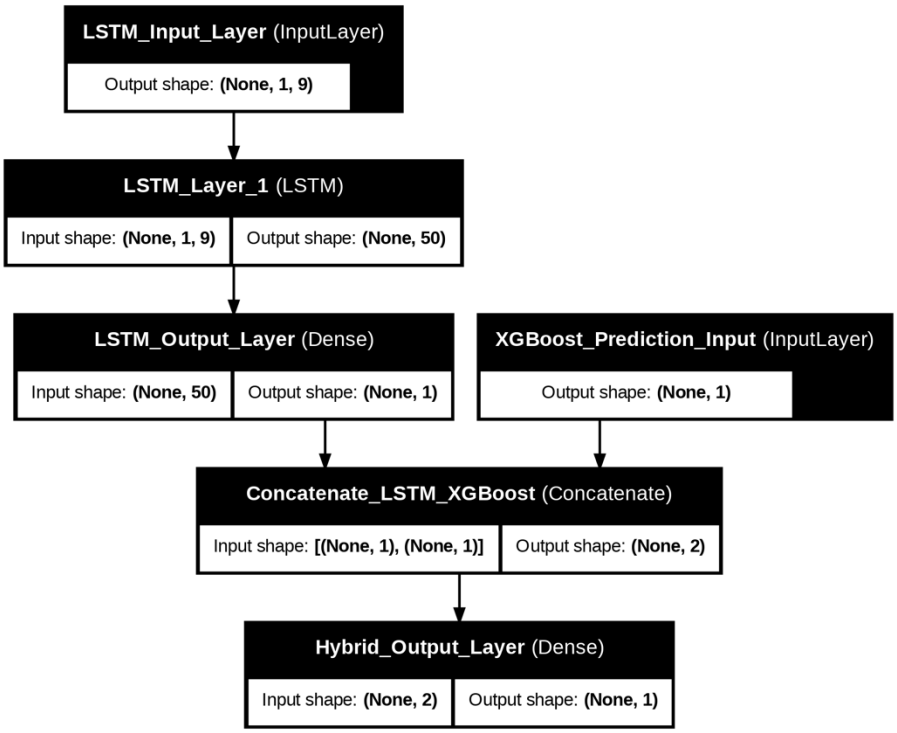


Figure 13 Hybrid Model Architecture (LSTM+XGBoost) - In the hybrid model, the LSTM receives a full 9-dimensional input vector derived from 3 PMTs × 3 features (peak, integral, rise time), while the XGBoost branch contributes its output prediction, previously trained on the same features. These two outputs are concatenated and passed to a final dense layer for refined prediction.

The detailed tables below provide a comprehensive breakdown of the architectures, hyperparameters, and selected features for each model, including **XGBoost**, **LSTM**, and the **Hybrid Model**, which combines predictions from both methods.

XGBoost Model:

Table 1 XGBoost Architecture

Model	Architecture & Hyperparameters	Description of Hyperparameters	Selected Features
XGBoost	Ensemble of decision trees		Highest Peaks of PMTs Integral of PMTs Traces PMTs Rise Time
	n_estimators: 3000	Number of trees in the ensemble	
	max_depth: 7	Maximum depth of each tree	
	learning_rate: 0.05	Step size shrinkage to prevent overfitting	
	gamma: 0.01	Minimum loss reduction required to split a node	
	reg_alpha: 0.1	L1 regularization term	
	random_state: 42	Seed for random number generation for reproducibility	

LSTM Model:

Table 2 LSTM Architecture

Model	Architecture & Hyperparameters	Description of Hyperparameters	Selected Features
LSTM	Sequential model with LSTM and Dense layers		Same as XGBoost
	Units: 50	Number of LSTM units	
	Activation: 'relu'	Activation function for LSTM units	
	Dropout: 0.2	Dropout rate to prevent overfitting	
	Optimizer: 'adam'	Optimization algorithm	
	Loss: 'mse'	Mean Squared Error loss function	

Hybrid Model:

Table 3 Hybrid Model (LSTM+XGBoost) Hyperparameter

Model	Architecture & Hyperparameters	Description of Hyperparameters	Selected Features
Hybrid Model	Random Forest Regressor on combined XGBoost-LSTM preds		XGBoost + LSTM Prediction

	n_estimators: 3000	Number of trees in the ensemble	-
	random_state: 42	Seed for random number generation for reproducibility	-

To ensure robust and unbiased model performance, the Zenith Angle (θ) distributions in both the training and testing sets were carefully balanced by applying a stratified binning strategy. Specifically, the full angular range (0°–90°) was divided into 9 equal-width bins of 10°, and an equal number of samples were randomly selected from each bin, as depicted in Figure 14. This approach transforms the naturally skewed distribution observed in the raw data (see Figure 5) into a uniform one, preventing model bias toward overrepresented zenith ranges.

This uniform distribution is critical for maintaining a balanced representation across all Zenith Angles, ensuring that the model is exposed to a wide variety of trajectories during training. By doing so, it enhances the model's ability to generalize and reduces overfitting to specific angular regions. Furthermore, a standard 70/30 train/test split was applied after balancing, resulting in approximately 171,000 training and 42,000 testing samples, maintaining a consistent ~4:1 ratio.

This balance not only improves training stability but also allows for a fair evaluation across angular ranges, as the test set mirrors the distribution of the training data. This is especially important in tasks involving particle trajectory prediction, where uniform angular coverage is essential for reliability, interpretability, and overall robustness of the learned models.

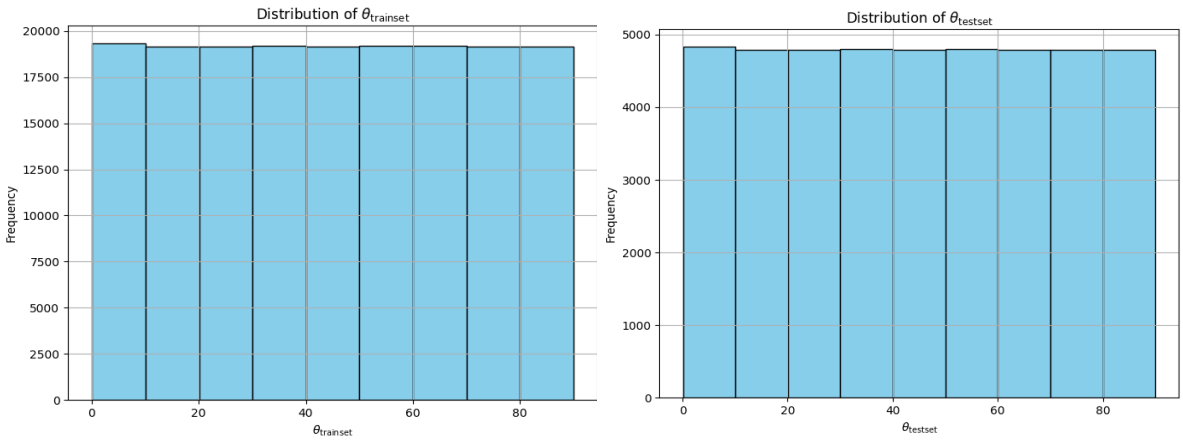


Figure 14 Train/Test set of Zenith Angle distribution

Figure 15 presents a 2D histogram that visualizes the relationship between the true Zenith Angles (θ_{true}) and the predicted Zenith Angles ($\theta_{predicted}$) generated by the hybrid model. The plot uses a colour gradient to indicate the density of predictions, with blue regions representing lower densities (fewer predictions) and red regions indicating higher densities (more predictions). The dashed line represents the ideal scenario of perfect predictions, where the predicted angles match the true angles exactly ($\theta_{true} = \theta_{predicted}$). A strong concentration of data points around this line indicates high model accuracy, as the predicted values closely align with the true values.

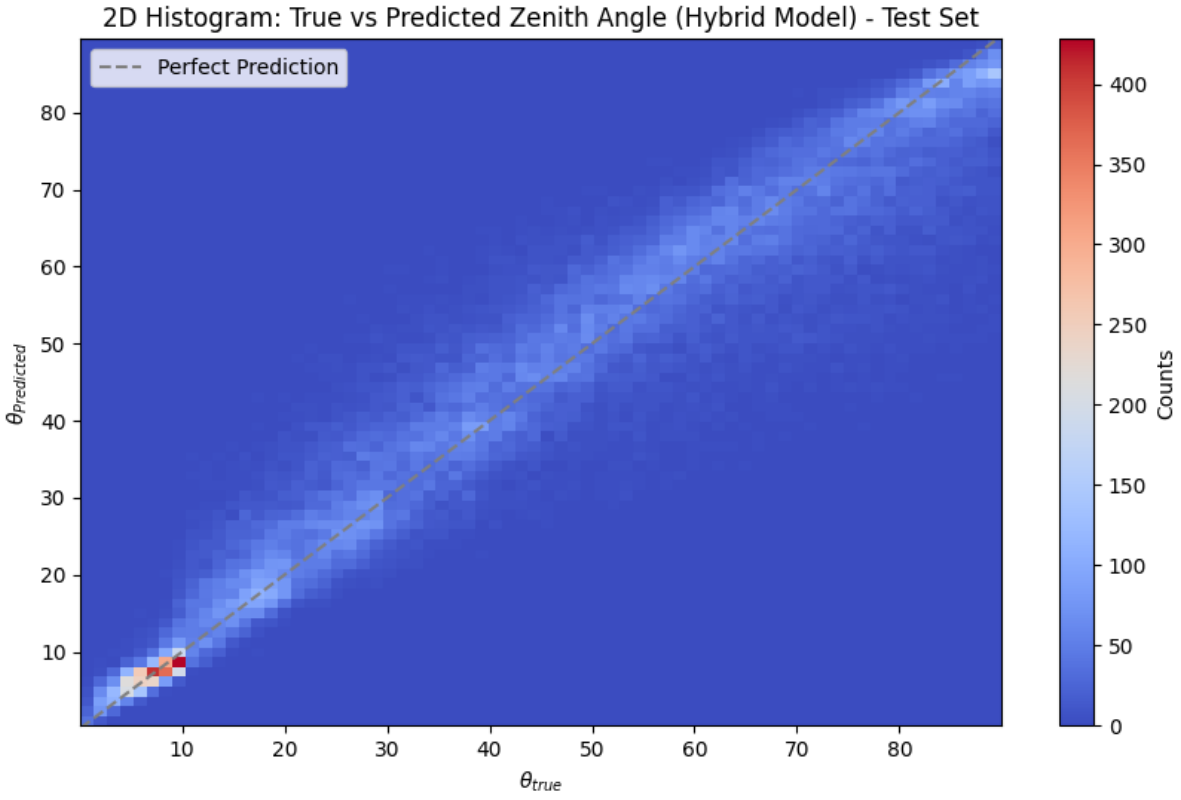


Figure 15 True Zenith Vs. Predicted Zenith for Hybrid model

Most of the predictions in the plot cluster near this perfect prediction line, particularly in the lower and mid-range Zenith Angles, showcasing the model's strong performance in these regions. The narrow spread of points around the line suggests small errors, indicating the model's precision. The red, high-density areas near the dashed line further highlight regions where the model consistently delivers accurate predictions. Importantly, the model demonstrates its capability to predict several events with Zenith Angles above 60 degrees, accurately capturing highly inclined events. This performance is critical, as accurately identifying these extreme angles is essential for studying phenomena such as atmospheric

showers and rare particle events. The model's ability to handle these challenging cases highlights its robustness and versatility.

Conversely, points deviating from the dashed line reflect prediction errors, with a slightly wider spread at higher angles. Overall, the histogram provides a comprehensive visual representation of the model's performance, demonstrating its ability to predict Zenith Angles across a wide range, including highly inclined events, with minimal errors concentrated around the ideal prediction line.

Table 4 Comparative Accuracy Results of Hybrid model (LSTM+XGBoost)

Model	**RMSE**	**MSE**	**MAE**
LSTM	17.97	323.07	13.719
XGBoost	11.22	126.06	8.04
Hybrid	7.64	58.37	5.32

To determine the model with the best accuracy (See Table 4), Mean Squared Error (MSE) and Mean Absolute Error (MAE) were calculated for three models: LSTM, XGBoost, and a Hybrid model. RMSE measures the square root of the average squared difference between predicted and actual values, making it sensitive to large errors and especially useful for identifying models that minimize high-magnitude deviations. MSE measures the average squared difference between predicted and actual values, while MAE measures the average absolute difference. The results clearly indicate that the Hybrid model outperforms both individual models, achieving the lowest RMSE (7.64), MSE (58.37), and MAE (5.32), thereby delivering the most accurate and consistent predictions. The XGBoost model performs better than LSTM, with a significantly lower RMSE (11.22 vs. 17.97) and reduced error metrics overall. In contrast, LSTM shows the highest RMSE (17.97) and MSE (323.07), highlighting its relative instability in this setup. Thus, the Hybrid model demonstrates superior performance in prediction accuracy and error minimization.

The residuals shown in Figure 16 represent the mean prediction error (prediction minus true value) in each zenith angle bin. The plotted error bars correspond to the standard deviation (σ) of residuals within each bin. The residuals are not normalized by σ, allowing for a direct interpretation of average error behaviour as a function of zenith. Initially, the model exhibited a slight systematic bias, where predictions tended to deviate consistently from the true values

across different Zenith Angles. To address this, we applied a bias adjustment by calculating the mean residual across all bins and subtracting this value from the model's predictions. This adjustment effectively centres the residuals around zero, reducing the systematic deviation and ensuring the model's predictions are unbiased. The error bars represent the standard deviation of residuals within each bin, indicating the variability of the model's predictions. Notably, the error bars are slightly smaller for lower Zenith Angles, suggesting reduced variability in this range. Overall, after bias adjustment, the plot demonstrates that the model achieves a balanced performance across all Zenith Angles, with minimal systematic bias and consistent spread of errors.

Figure 16 Mean and Standard deviation distribution of model (Residual plot)

2nd Hybrid Approach: Integrating XGBoost-ResNet with Random Forest Regressor

To further enhance predictive performance, we propose a second hybrid model that integrates XGBoost and ResNet with a Random Forest Regressor. This model leverages the powerful feature extraction capabilities of ResNet for handling sequential PMT data and combines it with XGBoost's gradient boosting for improved robustness. The addition of a Random Forest

Regressor ensures ensemble learning for final predictions, enhancing the model's accuracy and stability.

The hybrid ResNet–XGBoost model, combined with a Random Forest Regressor, leverages the complementary strengths of deep feature extraction and robust nonlinear learning. As shown in Figure 16, the architecture begins with a 1D ResNet backbone that ingests a feature tensor of shape (None, 100, 1), where 100 represents the flattened temporal or engineered feature length (e.g., interpolated or embedded form of the 9 original features across time steps or layers), and 1 represents the single feature channel.

The ResNet path processes this through a series of residual convolutional blocks (Conv1D layers with 64 filters), maintaining the shape of (None, 100, 64) throughout the depth of the network due to proper padding. These layers extract increasingly complex spatial and sequential features while preserving the full input length. After the final shortcut addition, the tensor is flattened into a vector of shape (None, 6400) and passed through a dense output layer of size (None, 1).

Parallel to this, a precomputed prediction from a standalone XGBoost model (trained on the same 9 engineered features: 3 PMT signals × 3 metrics — peak, integral, and rise time) is passed into the hybrid pipeline through a second input layer of shape (None, 1). The two outputs from ResNet and XGBoost are concatenated into a combined tensor of shape (None, 2), which is then passed into a final dense layer to generate the hybrid model's prediction.

This architecture ensures that both high-level hierarchical patterns (captured via ResNet) and structured non-linear correlations (via XGBoost) contribute to the final prediction. The combination allows the hybrid model to generalize well across complex Zenith Angle distributions and outperforms standalone models on all tested metrics.

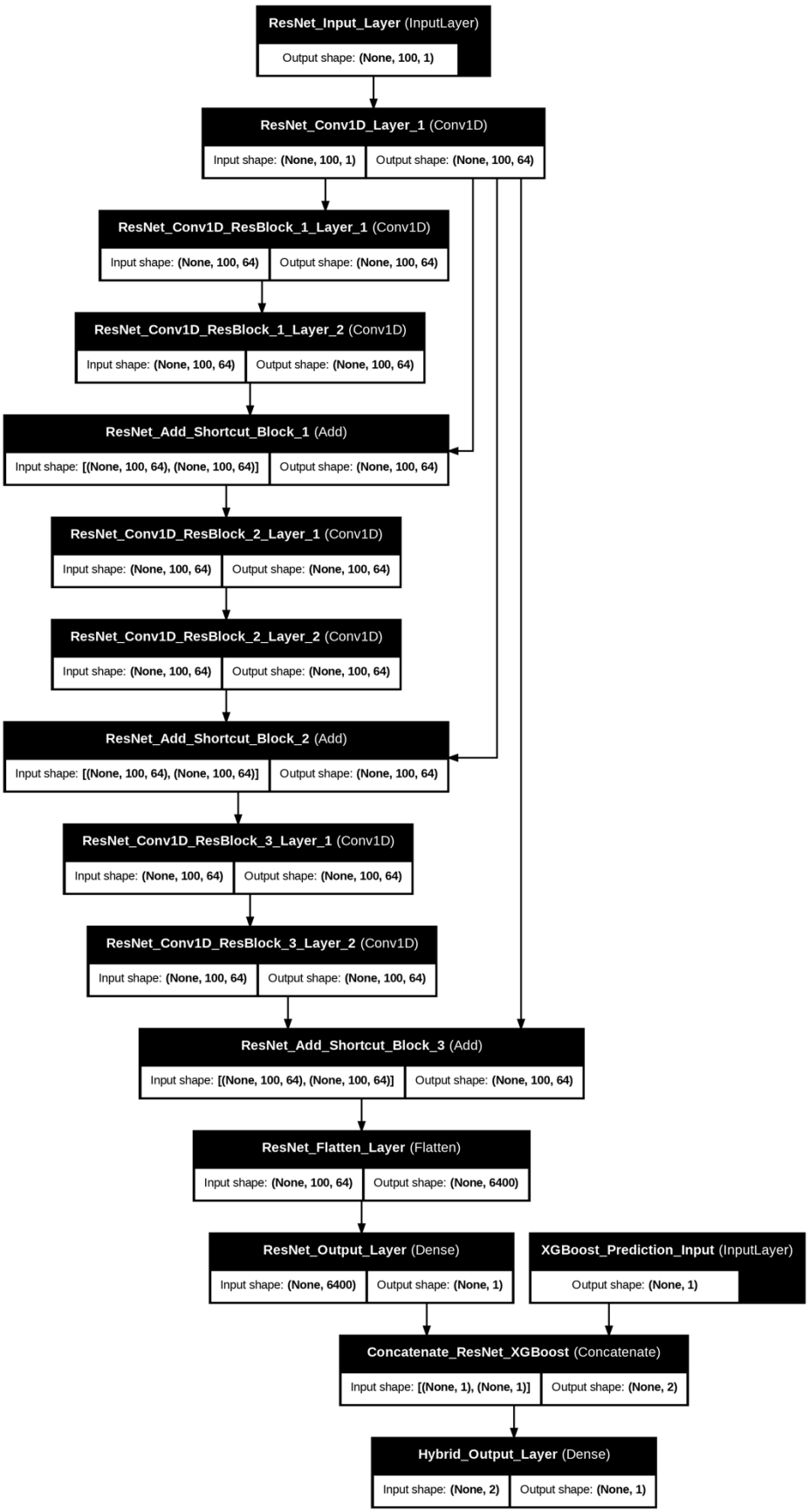


Figure 17 Hybrid (ResNet + XGBoost) Architecture

The 2D histogram in Figure 18 demonstrates the effectiveness of the hybrid ResNet-XGBoost model in predicting Zenith Angles across a wide range of values. The distribution of data points shows a strong concentration along the perfect prediction line, indicating that the model achieves high accuracy for most Zenith Angles. Notably, the plot highlights the model's ability to handle highly inclined events (Zenith Angles above 60 degrees) effectively, with a significant number of predictions closely aligned with the true values in this range. This performance is critical as it evidences the model's capacity to capture rare but crucial extreme-angle events, which are often underrepresented in traditional models. The dense clustering of predictions at both low and high Zenith Angles demonstrates that the hybrid model can detect a larger variety of events, ensuring comprehensive coverage and robustness in muon trajectory analysis.

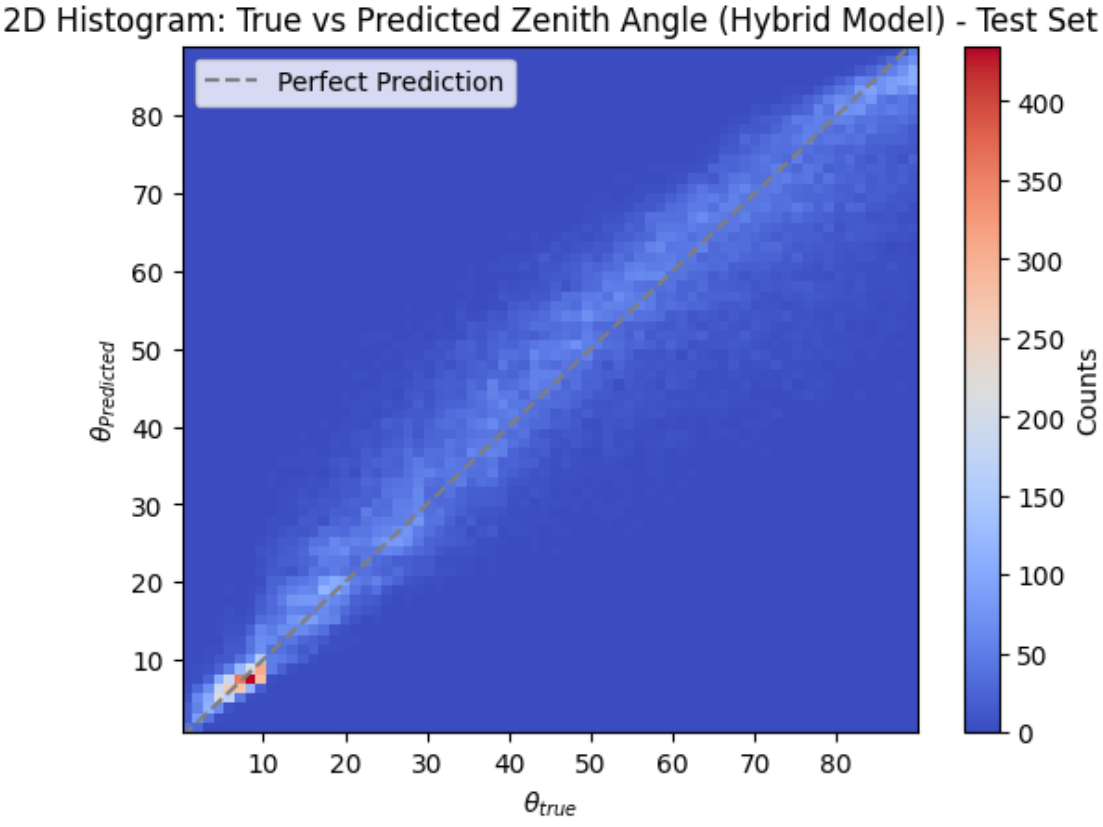


Figure 18 True Zenith Vs. Predicted Zenith for Hybrid model (ResNet + XGBoost)

The performance comparison between ResNet, XGBoost, and the Hybrid model in table 5, highlights the superior predictive accuracy of the Hybrid approach. ResNet, with its hierarchical feature extraction capability, shows the highest error values (RMSE: 16.45, MSE:

270.78, MAE: 12.46), indicating that while it captures temporal features, it struggles to generalize effectively for Zenith Angle predictions. XGBoost improves upon this, leveraging its gradient boosting framework to achieve lower errors (RMSE: 11.22, MSE: 126.06, MAE: 8.04), demonstrating its strength in handling non-linear relationships and reducing overfitting. However, the Hybrid model, which combines ResNet and XGBoost with a Random Forest Regressor, achieves the best results with a substantial reduction in errors (RMSE: 7.98, MSE: 63.75, MAE: 5.65). This performance indicates that the hybrid architecture effectively captures both sequential and non-linear dependencies while leveraging ensemble learning to mitigate variance. The results confirm that integrating these models leads to a more robust and accurate prediction framework for Zenith Angle estimation.

Table 5 Comparative Accuracy Results of Hybrid model (ResNet+XGBoost)

Model	RMSE	MSE	MAE
ResNet	16.45	270.78	12.46
XGBoost	11.22	126.06	8.04
Hybrid	7.98	63.75	5.65

Comparative Evaluation of Hybrid Models for Zenith Angle Prediction

The bar plot in Figure 19 compares the performance of two hybrid approaches (LSTM–XGBoost and ResNet–XGBoost) alongside their individual components across three error metrics: RMSE, MSE, and MAE. Among the tested models, the LSTM–XGBoost hybrid achieves the best overall performance, with the lowest RMSE (7.64) and MAE (5.32), indicating higher prediction precision and smaller average errors. The ResNet–XGBoost hybrid also improves significantly over its base ResNet model, achieving RMSE of 7.98 and MAE of 5.65, although it performs slightly below the LSTM-based hybrid.

To ensure a fair comparison, all models were optimized using grid search or manual tuning on a dedicated validation set. Specifically, the LSTM architecture was tested with varying units (32, 50, 64), and 50 units provided the best balance of convergence and accuracy. The ResNet model was designed with 3 residual blocks and 1D convolutions (64 filters, kernel size 3), which offered optimal feature extraction without overfitting. XGBoost parameters

(e.g., max_depth, learning_rate, n_estimators) were carefully selected via cross-validation. We kept training conditions (splits, data volume, and feature set) consistent across models to isolate the effect of architecture rather than data variability.

The comparative results confirm that hybridization significantly enhances performance beyond standalone models. Interestingly, both hybrids outperform LSTM, ResNet, and XGBoost alone with XGBoost consistently providing a stronger base than either deep learning model when used individually. This suggests that combining high-level sequence extraction (LSTM or ResNet) with XGBoost's structured decision boundaries leads to more accurate and generalizable Zenith Angle predictions.

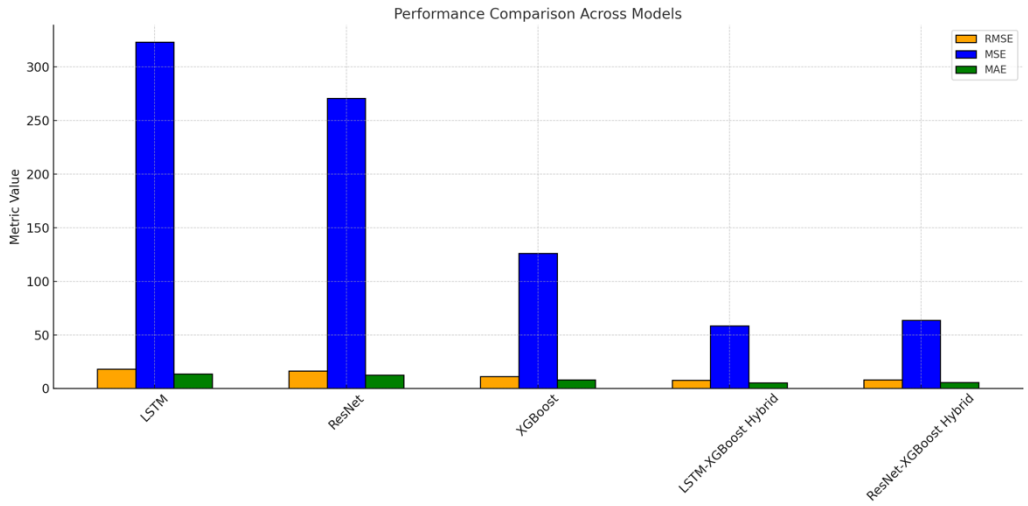


Figure 19 Comparative Evaluation of Hybrid Models for Zenith Angle Prediction

In addition to performance metrics, we evaluated the **computational effort** required for training and inference. Both hybrid models were trained using **Google Colab's GPU environment** under equivalent settings (same batch size, optimizer, learning rate, and epochs). The **LSTM–XGBoost hybrid** model demonstrated faster convergence, completing the training phase in approximately **10 seconds** (5397 steps at ~2ms/step), with inference running in about **2 seconds** (1350 steps at ~1ms/step).

In contrast, the **ResNet–XGBoost hybrid** required significantly more time, with training taking approximately **19 seconds** (5397 steps at ~4ms/step) and inference **5 seconds** (1350 steps at ~3ms/step).

These results show that the LSTM-based hybrid is not only more accurate but also **more computationally efficient**, making it a preferable choice in scenarios with limited processing resources or real-time constraints.

Purity and Efficiency of Hybrid Model:

The metrics of purity and efficiency serve as critical benchmarks for evaluating the performance of predictive models. These metrics are particularly pertinent when analyzing hybrid models applied to specific event channels, such as the Neutrino Channel.

Purity is a measure of the accuracy of the selected events (predicted as positive) in terms of how well they correspond to the actual true positive events. Formally, purity is defined as the ratio of the number of true positive events to the total number of selected events. A higher purity indicates a higher fraction of correctly identified positive events among all selected events, reflecting fewer false positives. This is depicted in figure 21, and can be expressed mathematically as:

$$\text{Purity} = \frac{Number\ of\ true\ positive\ events}{Total\ number\ of\ selected}$$

Efficiency, conversely, assesses the comprehensiveness of the selected events (predicted as positive) in covering all the true positive events in the dataset. It is calculated as the ratio of the number of true positive events to the total number of true positive events in the dataset. A higher efficiency denotes a higher fraction of true positive events correctly identified by the model, indicating fewer missed detections. This is depicted in figure 21 and it is mathematically represented as:

$$Efficiency = \frac{Number\ of\ true\ positive\ events}{Total\ number\ of\ true\ positive\ events\ in\ the\ dataset}$$

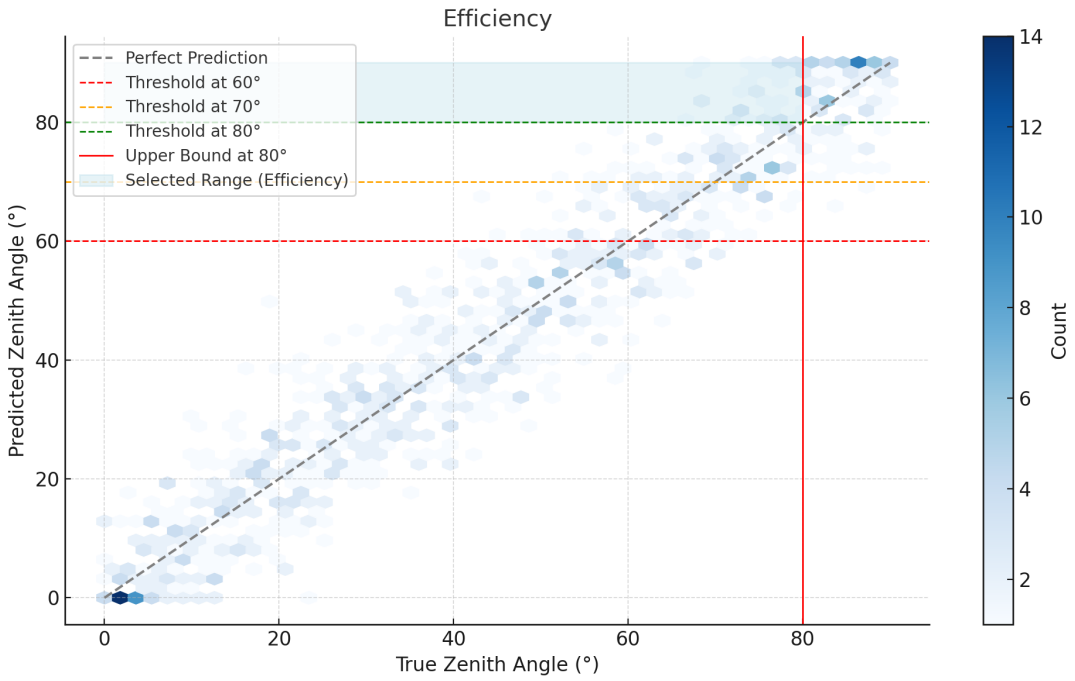


Figure 20 Schematic of Efficiency Calculation

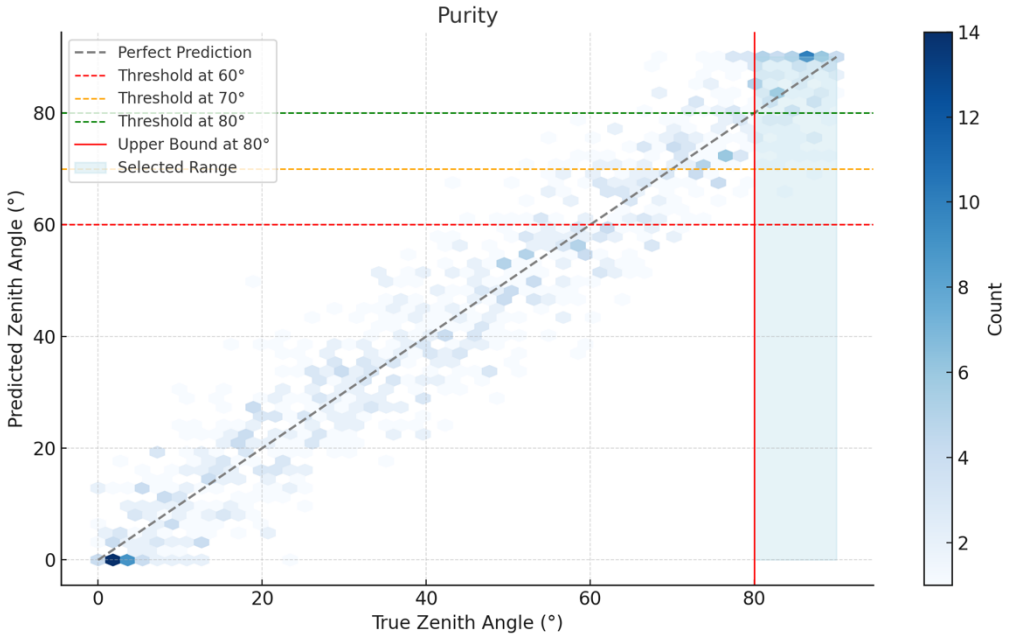


Figure 21 Schematic of Purity Calculation

3.12 Purity Vs. Efficiency for Hybrid Model (LSTM + XGBoost)

Figures 22 and 23 present the trade-off between purity and efficiency for the LSTM–XGBoost hybrid model, evaluated separately for two astrophysical channels:

- the Neutrino Channel (Zenith > 80°), and
- the Cosmic Ray Channel (Zenith > 70°).

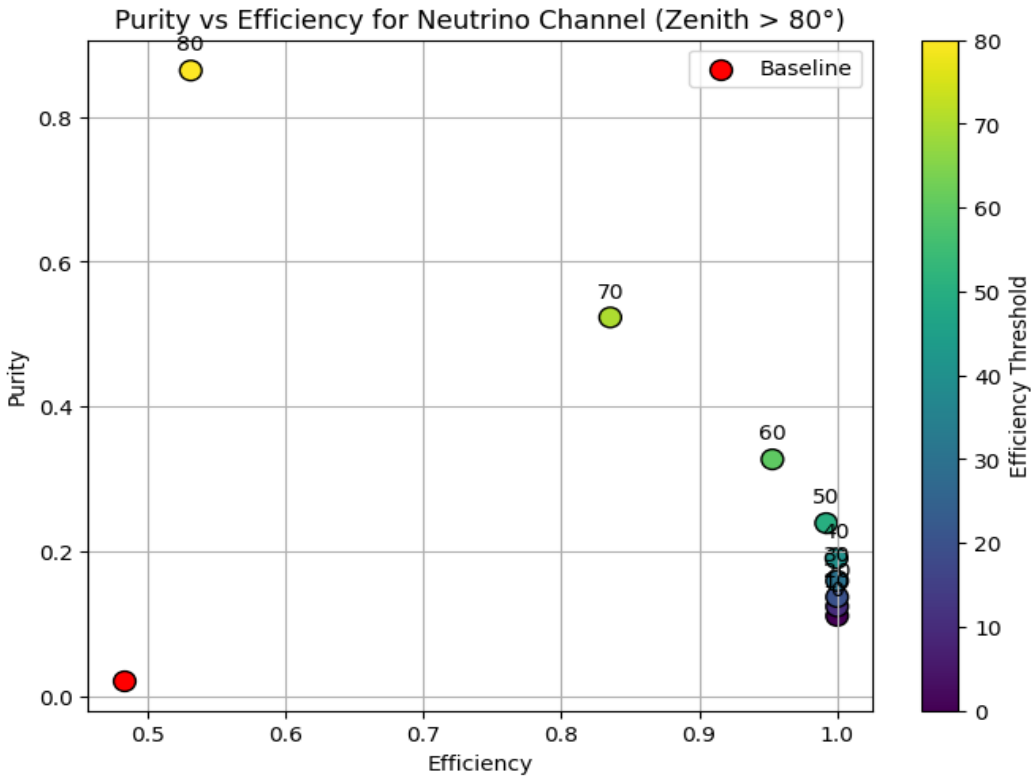


Figure 22 Purity vs. Efficiency for Neutrino Channel (Zenith > 80°)

In both plots, the baseline trigger system (red point) using a fixed 3.2 VEM threshold is included for reference.

In the Neutrino Channel (Figure 22), the hybrid model achieves significantly better performance than the baseline. Notably, the highest purity point (~0.86) is reached at an efficiency threshold of 80°, suggesting that a stricter prediction threshold reduces false positives and selects only high-confidence neutrino-like events. However, as the efficiency threshold is lowered, the model begins to classify more events, leading to an increase in efficiency (approaching ~0.99), but with a drop in purity indicating that more low-quality or

misclassified events (e.g., lower-angle muons) are now included in the selected range. High efficiency is indeed achieved when the selection threshold is too lenient, allowing nearly all events (even those with poorly reconstructed zenith angles) to be accepted. However, this leads to a rapid decline in purity, because many of these events are not truly in the high-zenith-angle neutrino domain. This effect is exacerbated by the model's difficulty in precisely reconstructing extreme zenith angles, which is reflected in the spread of prediction error for events above 80°.

In contrast, the Cosmic Ray Channel (Figure 23) exhibits a smoother trade-off. The model reaches an optimal zone between 60° and 70° thresholds, where both purity (~0.89) and efficiency (>0.70) remain high. This is expected, as cosmic-ray events with $\theta > 70°$ are more abundant and easier to distinguish than neutrino-induced events, resulting in a more stable classification regime.

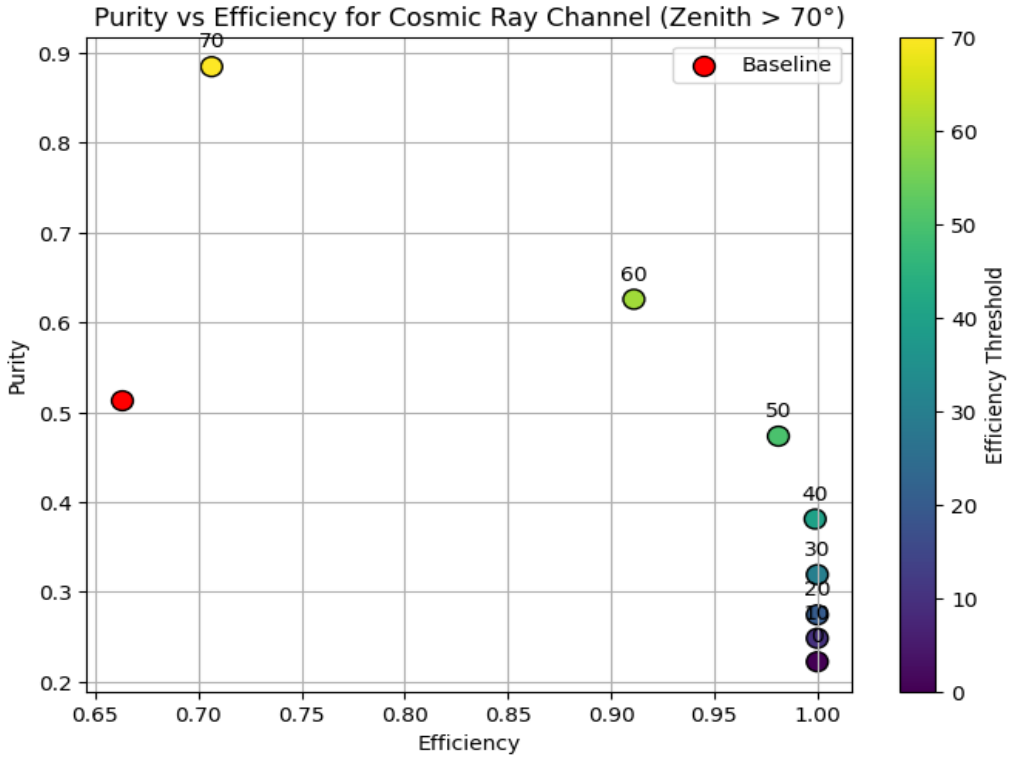


Figure 23 Purity vs. Efficiency for Cosmic Ray Channel (Zenith > 70°)

3.13 Purity Vs. Efficiency for Hybrid Model (ResNet + XGBoost)

Figures 24 and 25 illustrate the same purity-efficiency analysis for the ResNet–XGBoost hybrid model, again using the 3.2 VEM trigger as baseline reference. In the Neutrino Channel (Figure 24), the model demonstrates excellent behavior at stricter thresholds, achieving purity > 0.85 at the 80° selection cutoff. As the threshold is relaxed (e.g., to 70°, 60°), efficiency increases sharply toward 0.99, but with a familiar drop in purity due to the inclusion of low-angle or misreconstructed events. This confirms that the efficiency rise at low thresholds is expected, but its value is limited without corresponding purity. The ResNet-based model, however, handles this trade-off slightly more consistently than the LSTM variant, due to its deeper spatial feature extraction.

In the Cosmic Ray Channel (Figure 25), the ResNet–XGBoost model delivers reliable and balanced results. At a threshold of 70°, purity approaches 0.88 and efficiency remains close to 0.70. Across all evaluated thresholds, the hybrid model consistently outperforms the baseline in both metrics, especially in the 50°–70° selection range. This confirms its suitability for high-inclination cosmic ray detection, where precision and coverage are both critical.

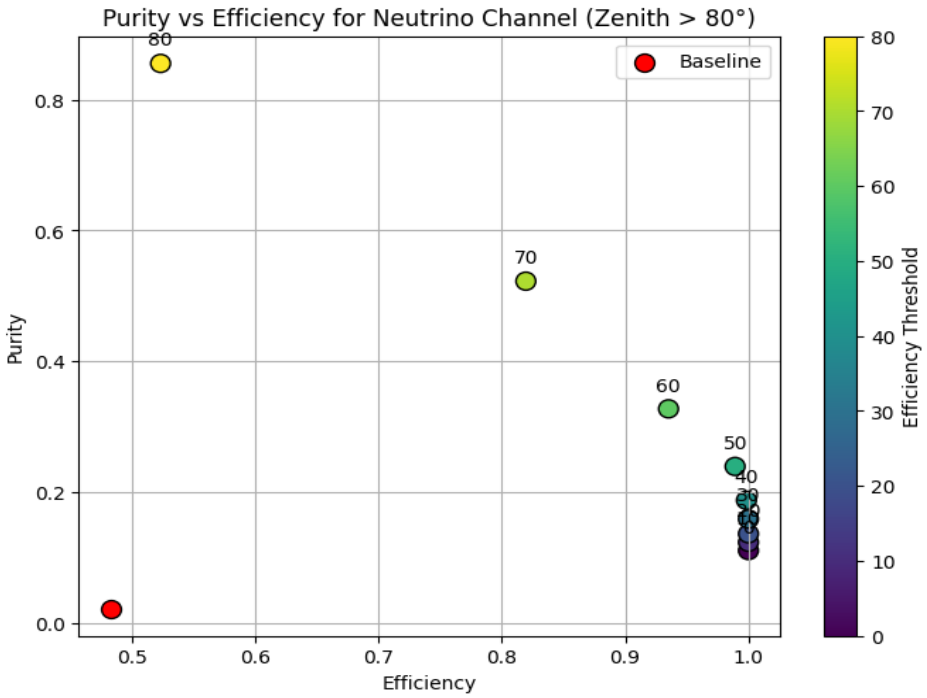


Figure 24 Purity vs. Efficiency for Cosmic Ray Channel (Zenith > 80°)

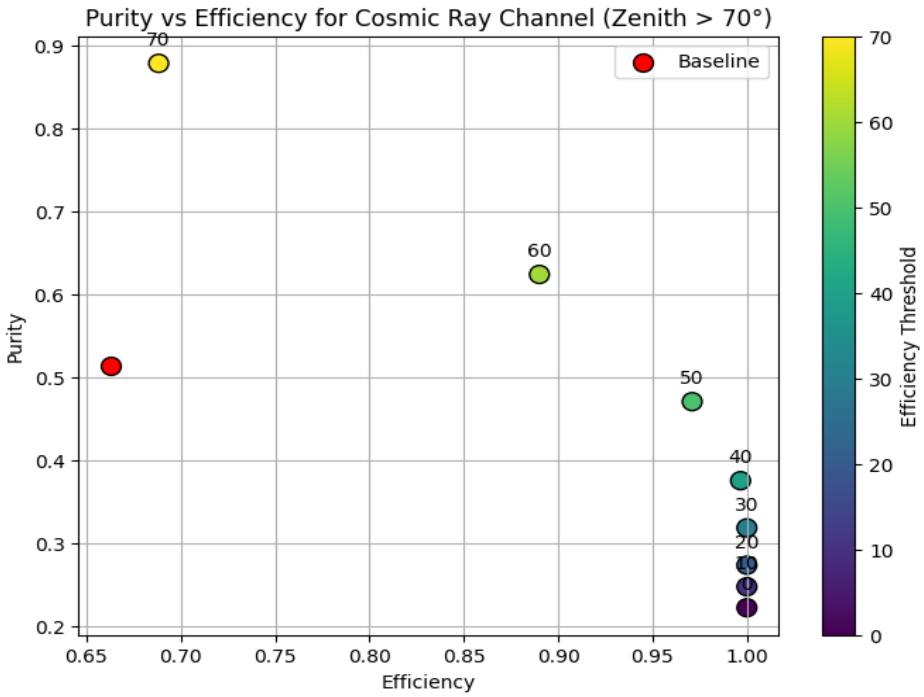


Figure 25 Purity vs. Efficiency for Cosmic Ray Channel (Zenith > 70°)

3.14 Comparative Analysis

Both hybrid models significantly outperform the baseline, highlighting the inadequacies of the 3.2 VEM trigger system in handling high Zenith Angle events. The **Hybrid LSTM-XGBoost** model shows exceptional adaptability in both channels, leveraging sequential data patterns to boost detection accuracy. Meanwhile, the **Hybrid ResNet-XGBoost** model exhibits superior robustness, with consistent purity-efficiency gains across varied thresholds.

The Neutrino Channel (Figures 21 & 23) showcases the most dramatic improvements, reflecting the models' ability to handle rare and high-inclination neutrino events. Conversely, the Cosmic Ray Channel (Figures 22 & 24) benefits from the hybrid models' capacity to filter out noise, ensuring precise identification of high-energy particles. Together, these results underline the transformative potential of hybrid machine learning models in astrophysical event detection, offering a quantum leap over traditional trigger systems.

Neural Network Classification:

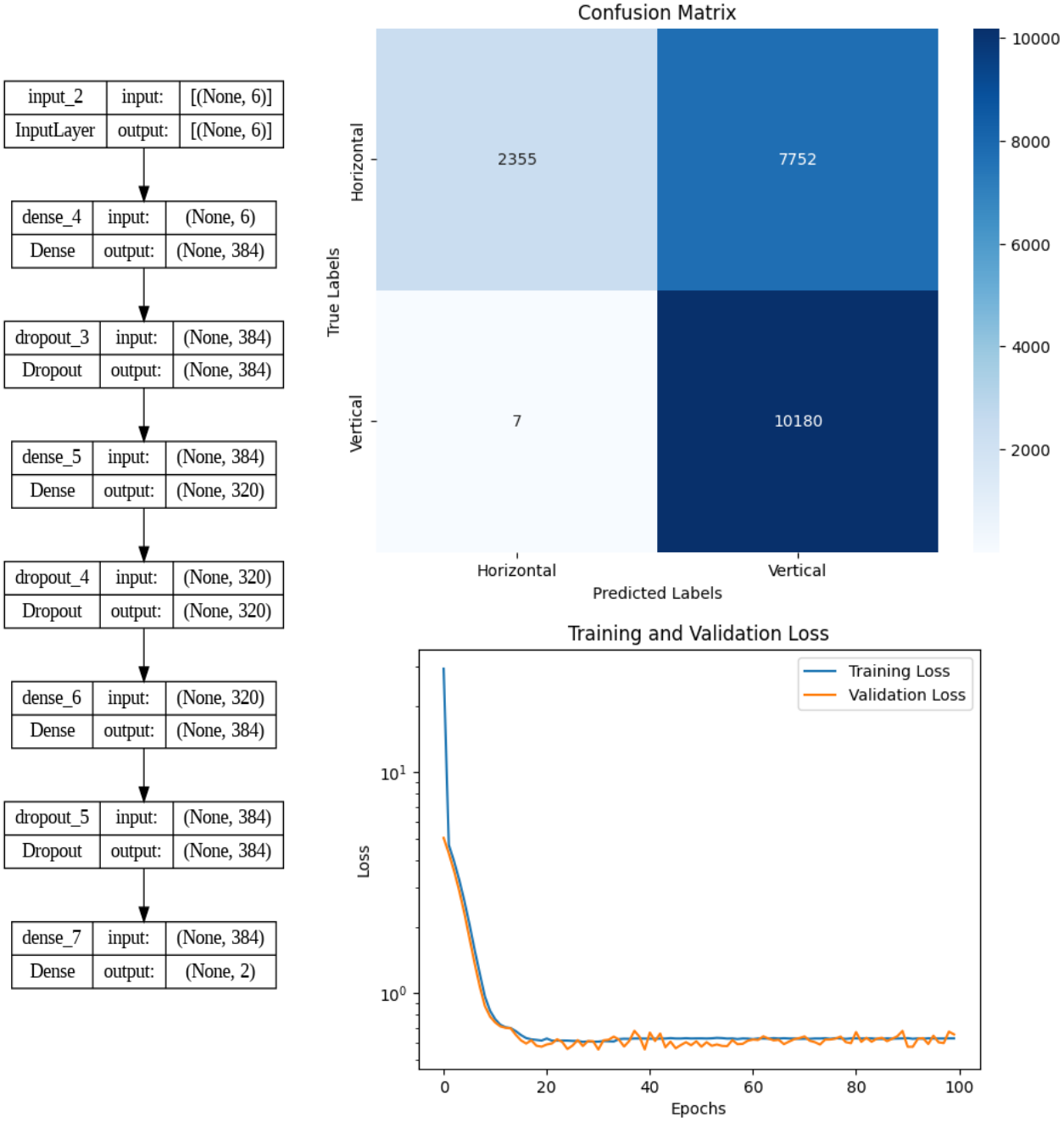


Figure 26 Keras Tuner Classification Neural Network

In our classification endeavor, we utilized the same features and data processing techniques as employed in the previous regression model section. Specifically, we extracted features - maximum PMT values, integral values, and rise times- from the dataset.

3.15 Hyperparameter Tuning with Keras Tuner

To optimize our neural network model's performance, we employed Keras Tuner, a powerful tool for hyperparameter tuning. Keras Tuner allowed us to systematically search through a range of hyperparameters, including the number of units per layer, the learning rate, and the number of hidden layers. By tuning these parameters, we aimed to enhance the model's accuracy and generalization capability however the accuracy at 23% was measured at the initial training stage using default hyperparameters, before any optimization was applied.

3.16 Stacked Model Enhancement

To increase the accuracy, a stacked model combines multiple base estimators, such as gradient boosting and neural networks, to leverage their collective strengths in classification tasks. This ensemble method often yields superior performance compared to individual classifiers by leveraging diverse modeling techniques and increased our model accuracy to 72%. The final estimator for the model was selected as Logistic Regression.

3.16.1 Benefits of Stacked Model:

Improved Accuracy: By combining the predictions of multiple base estimators, a stacked model can achieve higher classification accuracy than any single estimator alone.

Robustness: Stacked models are inherently robust against overfitting and data variability, as they incorporate diverse modeling approaches.

Capture Complex Patterns: The combination of different base estimators allows the stacked model to capture complex patterns and relationships in the data more effectively.

3.16.2 Impact of Zenith Range Variation

In our analysis, we observed a significant impact of varying the zenith range on classification accuracy. Specifically, when excluding the zenith range from 45 to 60 degrees, the classification accuracy increased substantially to 75%. This phenomenon can be attributed to the inherent challenges in distinguishing between inclined showers and vertical ones within this transition zone of zenith angles. The complexity of events occurring within this range may lead to ambiguity in classification, thus reducing overall accuracy.

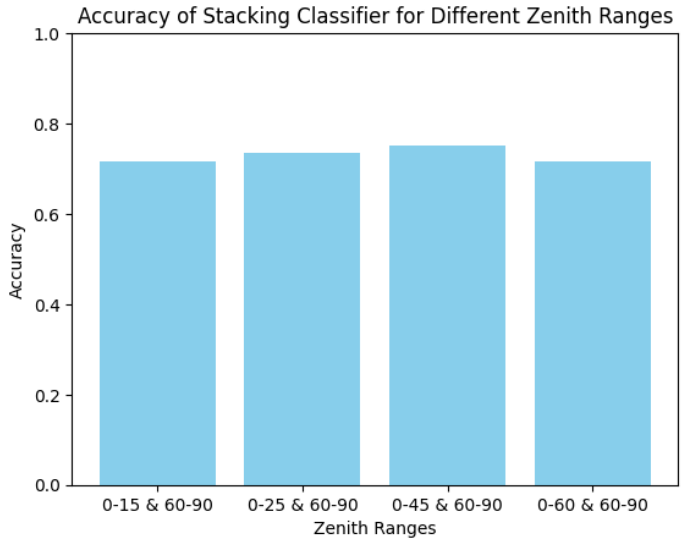


Figure 27 Different Accuracy of Stacking Classifier model on different range of Vertical Zenith Angles

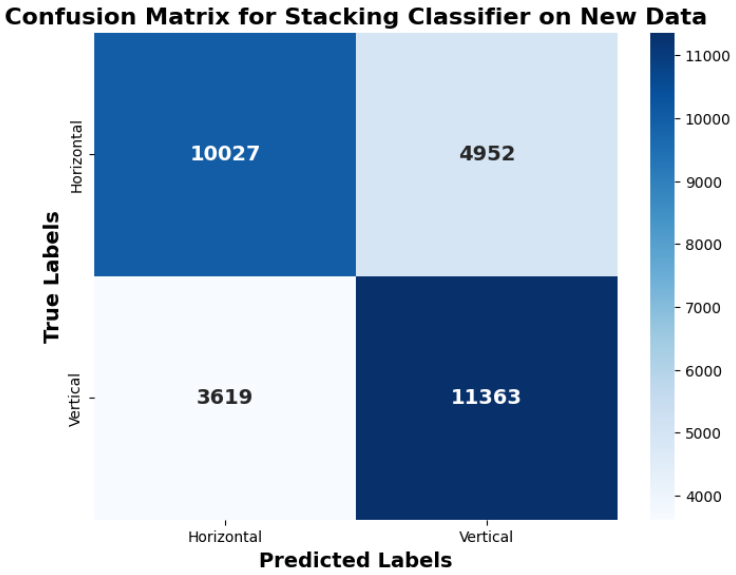


Figure 28 Testing the model on untrained new dataset

Testing the trained model on new simulation untrained dataset gave the accuracy of 71.3%

In summary, our exploration of neural network classification techniques, hyperparameter tuning with Keras Tuner, and stacked model enhancement has yielded promising results in event classification within the Pierre Auger Observatory dataset. By leveraging advanced modeling approaches and systematically optimizing model parameters, we have significantly improved classification accuracy, providing valuable insights into astroparticle physics phenomena. However, further research and experimentation are warranted to fully understand the complexities of event classification in varying observational conditions.

3.17 Discussion on Simulation Constraints and Future Work

This study utilized idealized simulation datasets generated via the Pierre Auger Observatory's Offline software framework, representing clean and controlled particle injection scenarios into Water-Cherenkov Detectors (WCDs). The datasets included approximately 1 million events for continuous regression tasks (zenith angle estimation) and 15,000 events for classification tasks, after filtering and preprocessing.

For the regression models, zenith angle was treated as a continuous variable across the range of 0° to 90°, enabling high-resolution prediction. In contrast, for classification, the angle space was discretized into physically meaningful bins (e.g., >70°, >80°) to enable event-type discrimination, particularly for cosmic ray and neutrino channel identification. This binning enabled the application of purity and efficiency analyses, which require event grouping.

One limitation of the current simulation pipeline is the absence of noise modeling. The simulations used in this work were noise-free by design, to allow a baseline evaluation of model architectures and their response to ideal physical signal patterns. This approach facilitates the identification of useful features and model behavior without the confounding effects of instrument or environmental noise.

However, in real-world applications, WCD signals are inevitably affected by noise sources such as photomultiplier dark current, electronic fluctuations, and ambient air shower background. Recognizing this, future extensions of this work will:

- Incorporate noise injection techniques consistent with detector calibration data (e.g., Gaussian jitter, baseline drift),

- Evaluate model robustness and generalization under noisy conditions,
- Test the trained models on real experimental data to validate their applicability beyond simulation,
- Compare performance degradation trends across hybrid models (LSTM+XGBoost, ResNet+XGBoost), especially in high-inclination events which are more sensitive to background noise.

By addressing these limitations, the framework developed in this study can transition from simulation-based proof of concept to a deployable solution for astrophysical event classification within the Pierre Auger data analysis pipeline.

Chapter 4

Mango Farms Development with Machine Learning Models

**This paper has been published by ISHS (International Society of Horticultural Science) **
https://doi.org/10.17660/ActaHortic.2025.1415.15

Climate Change Multi-Risk Assessment for Mango Cultivation in Sicily, Italy by using Bayesian Network.

Mohsen Pourmohammad Shahvar[1], Dario Scuderi [2], Davide Valenti[1], Alfonso Collura[3], Salvatore Micciché[1], Vittorio Farina [2], and Giovanni Marsella[1]

[1] *Dipartimento di Fisica e Chimica "E. Segrè", Università degli Studi di Palermo, Italy.*

[2] *Dipartimento di Scienze Agrarie, Alimentari e Forestali, Università degli Studi di Palermo, Italy.*

[3] *Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Palermo, Italy*

4.1 Abstract

Ensuring food security poses a significant challenge for organizations and consultant companies involved in the agriculture industry or responsible for food programs. This challenge is particularly relevant in Sicily, Italy, which has a semi-tropical climate. Given the favourable weather conditions for mango cultivation and other tropical crops, it becomes crucial to consider measures for safeguarding against potential climate change impacts in the future. Climate change is expected to bring changes and increased risks in terms of temperature, extreme events, soil salinity, and irregular rainfall.

Amidst this looming threat, there is a growing demand for a fresh approach and supportive tools to manage risks and mitigate potential damages in policy-making and decision-making circles. In this study, we employ a robust method known as Bayesian Network (BN) to effectively capture and model multiple risks under various future scenarios. By exploring 'what-if' situations, such as the maximum levels of climate-related variables, the projected BN model is trained and validated using spatially resolved data from the Messina region in Sicily. This approach enables us to understand the dynamic variations in local-scale temperature and precipitation, as well as the underlying driving forces, within the timeframe of 2009-2022.

The outputs of the Bayesian Network aid in predicting future trends in temperature and precipitation levels, thereby supporting the prioritization of mango cultivation and conservation efforts. In general, the findings derived from the BN analysis provide valuable support for disaster risk

management and mitigation strategies in the face of climate change and extreme events. This tool can further enhance decision-making processes by integrating the spatial results of the developed model into a user-friendly interface such as Geographic Information System (GIS), thereby assisting policymakers and decision-makers in prioritizing Disaster Risk Management and Climate Change Adaptation plans.
Keywords: Multi-Risk Assessment, Climate Change, Bayesian Network, Mango Farms, Agriculture, Sicily.

4.2 Introduction

Ensuring food security amidst climate change is crucial for agriculture-focused organizations and food programs (Beddington et al., 2012). This is especially relevant in semi-tropical regions like Sicily, Italy, where favorable conditions for crops like mangos necessitate proactive measures against climate impacts (Testa et al., 2018). Climate change brings tangible risks, including temperature shifts, extreme weather events, soil salinity rise, and irregular rainfall, all of which threaten agricultural productivity and food security (Clivaz & Savioz, 2020). Thus, new approaches and tools are essential for effective risk management and policy-making (FAO, 2016).

Our study proposes using Bayesian Network (BN) methodology to model climate risks under various scenarios. The BN framework allows for "what-if" analysis, helping assess climate change impacts on agriculture in Sicily (Yet et al., 2020). Using spatially resolved data from the Messina region (2009-2022), we aim to understand local temperature and precipitation variations and their drivers.

The BN approach stands out by considering the complex interactions among variables, unlike traditional models (Pourmohammad Shahvar, 2021; Sperotto et al., 2017). This enables a holistic evaluation of climate impacts on agriculture. By simulating different scenarios, we can assess effects on crop yields, water availability, and pest prevalence (Mentzel et al., 2022), providing insights into vulnerabilities and risks. Spatially resolved data allows for detailed examination of localized impacts, helping identify areas prone to specific risks like temperature increases or precipitation changes. This understanding aids targeted resource allocation and interventions to mitigate negative climate effects (FAO, 2022a).

Overall, this study employs the BN approach to model climate risks in Sicily's agriculture, using Messina region data to understand temperature and precipitation variations. This research aims to inform policymakers and stakeholders, enabling proactive decisions and adaptation strategies to ensure food security amid climate change.

4.3 Methodology

4.3.1 Data Collection and Pre-processing

To conduct this study, spatially resolved data on temperature, precipitation, and other relevant climate variables are gathered from the Messina region in Sicily. These data are sourced from meteorological stations, remote sensing satellites, and other reliable sources.

4.3.2 Bayesian Network Modelling

Recently, BNs have gained recognition for addressing environmental issues amid uncertainty, aiding decision-makers in environmental risk assessments (Pham et al., 2024; Sperotto et al., 2017). Originating from artificial intelligence research (Charniak, 1991; Pearl, 2011), with a risk assessment perspective for several different environmental issues, BNs are used for risk assessments in various environmental contexts, including coastal management, water resources, fisheries, and agriculture (Farmani et al., 2009; Pham et al., 2024; Pourmohammad Shahvar, 2021; Mohsen.P Shahvar et al., 2022; Sperotto et al., 2017)

Bayes' theorem calculates event probabilities based on prior information (Bayes & Price, 1763):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where P(A) and P(B) are the prior probabilities of events A and B, P(A|B) is the posterior probability of A given B, and P(B|A) is the probability of B given A.

BNs consist of a graphical agent representing variables as nodes and a numerical agent defining relationships using conditional probability tables (CPT) or distributions (CPD) (Dorner et al., 2007). Unlike deterministic models, BNs use probabilistic expressions to describe relationships between variables like temperature and precipitation and their impacts on system conditions (Borsuk et al., 2004). BNs handle uncertainties and integrate quantitative and qualitative data across disciplines (Pollino & Henderson, 2010).

Our BN captures relationships between climate variables and agricultural productivity, with nodes for Maximum, Average, and Minimum Temperature, Solar Radiation, Albedo, Wind Speed, Wind Direction, Relative Humidity, and Surface Pressure.

The Bayesian Network formula is as follows:

$$P(T'|SR, A, WS, WD, RH, SP) = \frac{P(T'|T)xP(T|SR, A, WS, WD, RH, SP)}{P(T|SR, A, WS, WD, RH, SP)}$$

where (P(T' | P, SR, A, WS, WD, RH, SP)) is the probability of temperature change given Precipitation, Solar Radiation, Albedo, Wind Speed, Wind Direction, Relative Humidity, and Surface Pressure, (P(T' | T)) is the conditional probability of temperature change given current temperature, and $P(T \mid P, SR, A, WS, WD, RH, SP)$ is the probability of current temperature given the same variables.

The BN model is developed using expert knowledge, historical data, and stakeholder inputs.

4.3.3 BN Model Training and Validation

We train the BN model using historical climate data from Messina, applying Bayesian parameter estimation techniques like Maximum Likelihood Estimation (MLE).

4.3.4 Simulating Future Scenarios

The BN model simulates and predicts climate change impacts on temperature and precipitation by adjusting conditional probabilities. This helps assess potential future changes.

4.3.5 Strategies for Mango Farm Protection

Insights from the BN model guide strategies to protect mango farms, such as adjusting irrigation, selecting suitable varieties, and implementing climate-resilient techniques, helping farmers in Messina adapt to climate changes.

4.3.6 Model Design

We created a BN conceptual model based on Messina's mango farm data, including parent nodes for Precipitation, Solar Radiation, Albedo, Wind Speed, Wind Direction, Relative Humidity (Max and Min RH), and Surface Pressure.

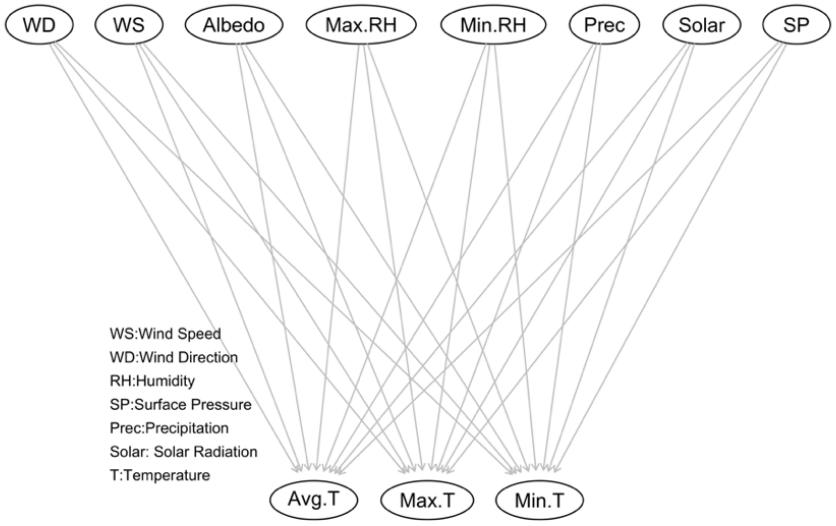


Figure 29 First Expert-Knowledge Design for the Bayesian Network

The BN model's child-nodes, including Maximum, Average, and Minimum Temperature, were assigned specific states using the discretize command in R, with meteorological variables divided into three classes, wind direction into four, and temperature into four classes. The model was trained using Tabu Search, Hill-Climbing, Incremental Association, and Grow-Shrink techniques, assessing limitations and relationships among nodes. Figure 29 illustrates that Hill-Climbing and Tabu Search introduced more arcs compared to Grow-Shrink and Incremental Association. The final configuration of the Bayesian Network, based on the suggested Directed Acyclic Graphs (DAGs), is presented in Figure 30 (Gutierrez et al., 2011). Conclusively, building upon the final BN model, conditional probabilities between variables through the network can be learnt directly from the observed dataset, and the probability based on the frequency of observed conditions can be measured as well. Figure 32 reports the marginal distribution of the nodes altogether that were learnt from the observed data.

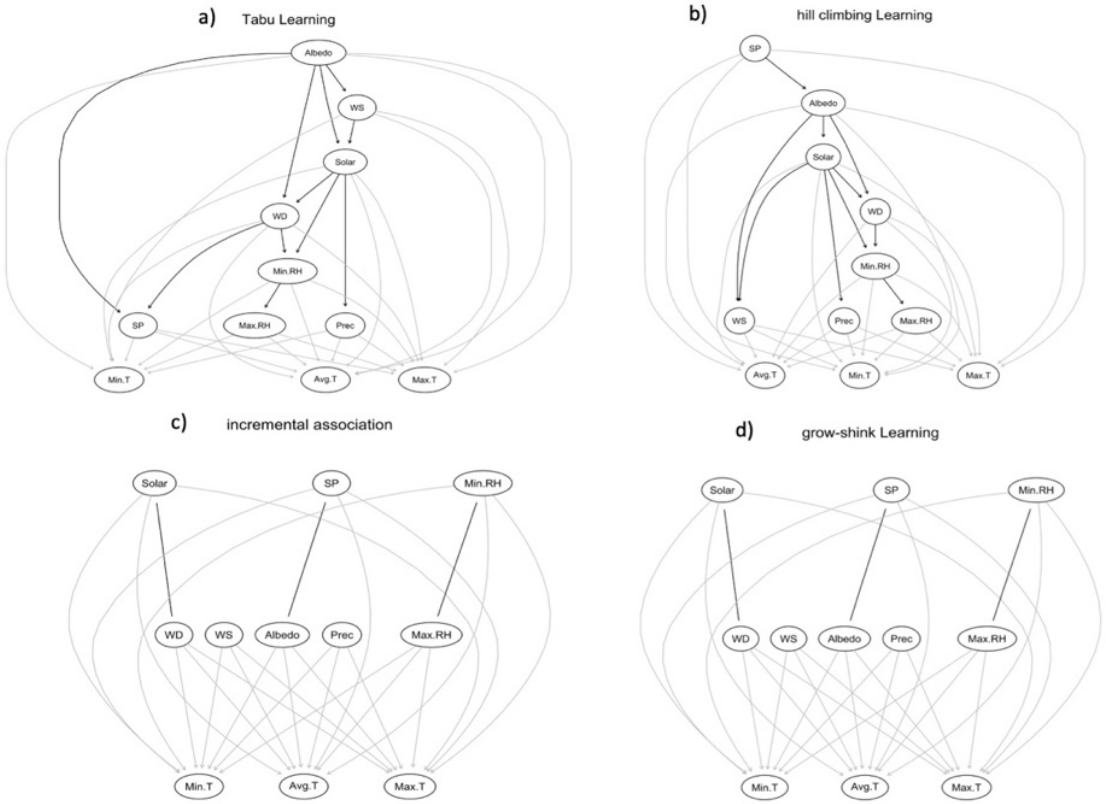


Figure 30 BNs model based on different learning algorithms a) tabu search (tabu), b) hill- climbing (hc), c) incremental association (iamb), and grow-shrink (gs). The grey arrows represent arcs based on the pre-defined expert-based model; the black arrows represent arcs based on the pre-defined expert-based model; the black arrows represent the further arcs suggested by the different learning algorithms

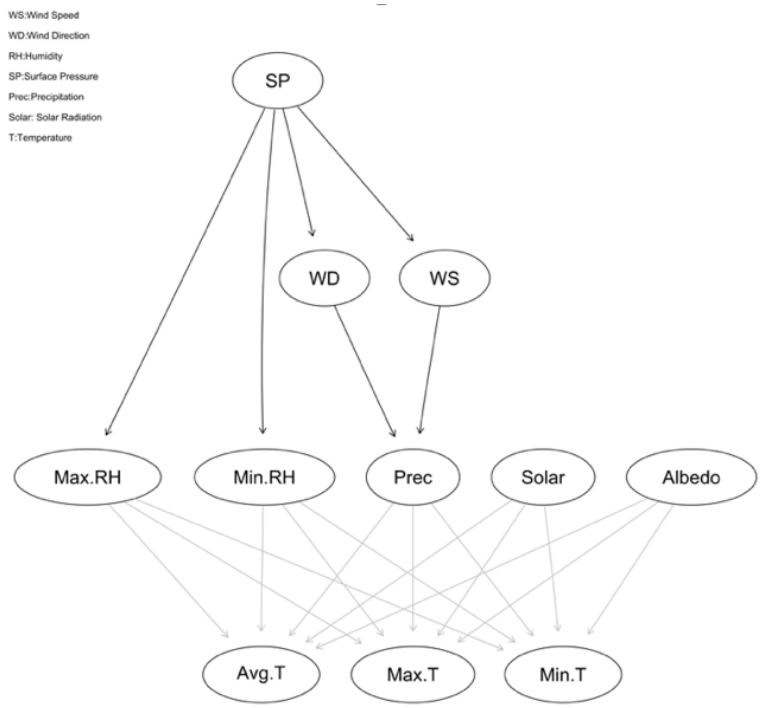


Figure 31 Final BN model reporting the marginal distributions associated to all variables included in the network

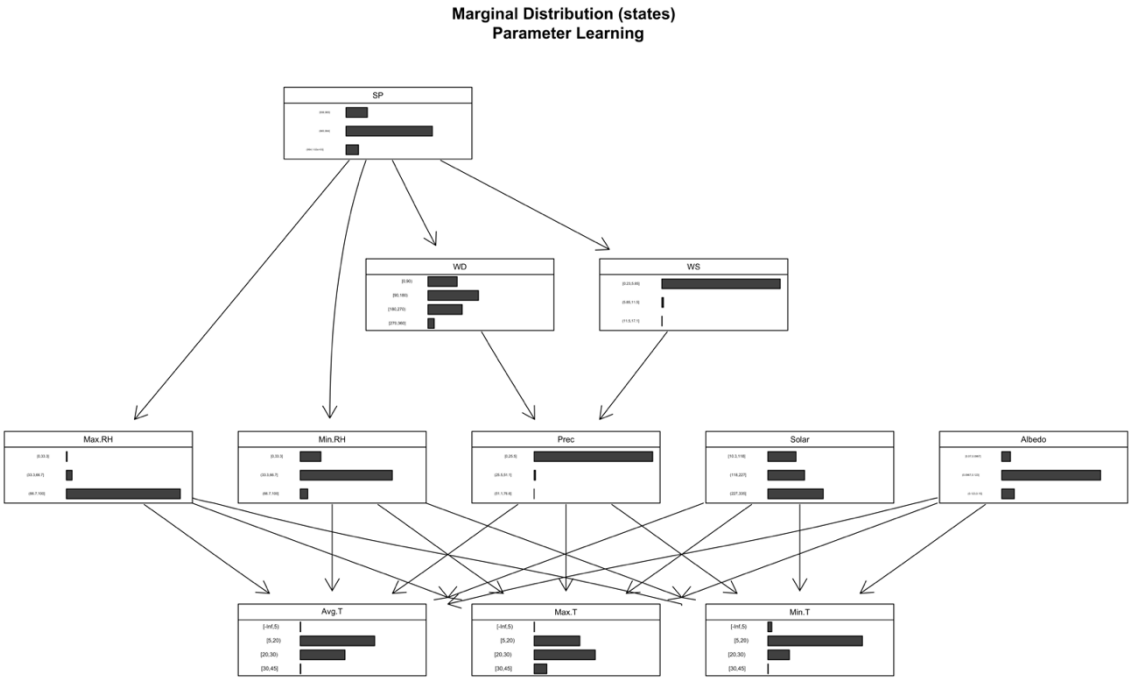


Figure 32 Final BN model reporting the marginal distributions associated to all variables included in the network

The same procedure is applied to assess the Precipitation assessment end-point framework. Figure 33 encompasses all the steps involved in processing the Precipitation Bayesian network.

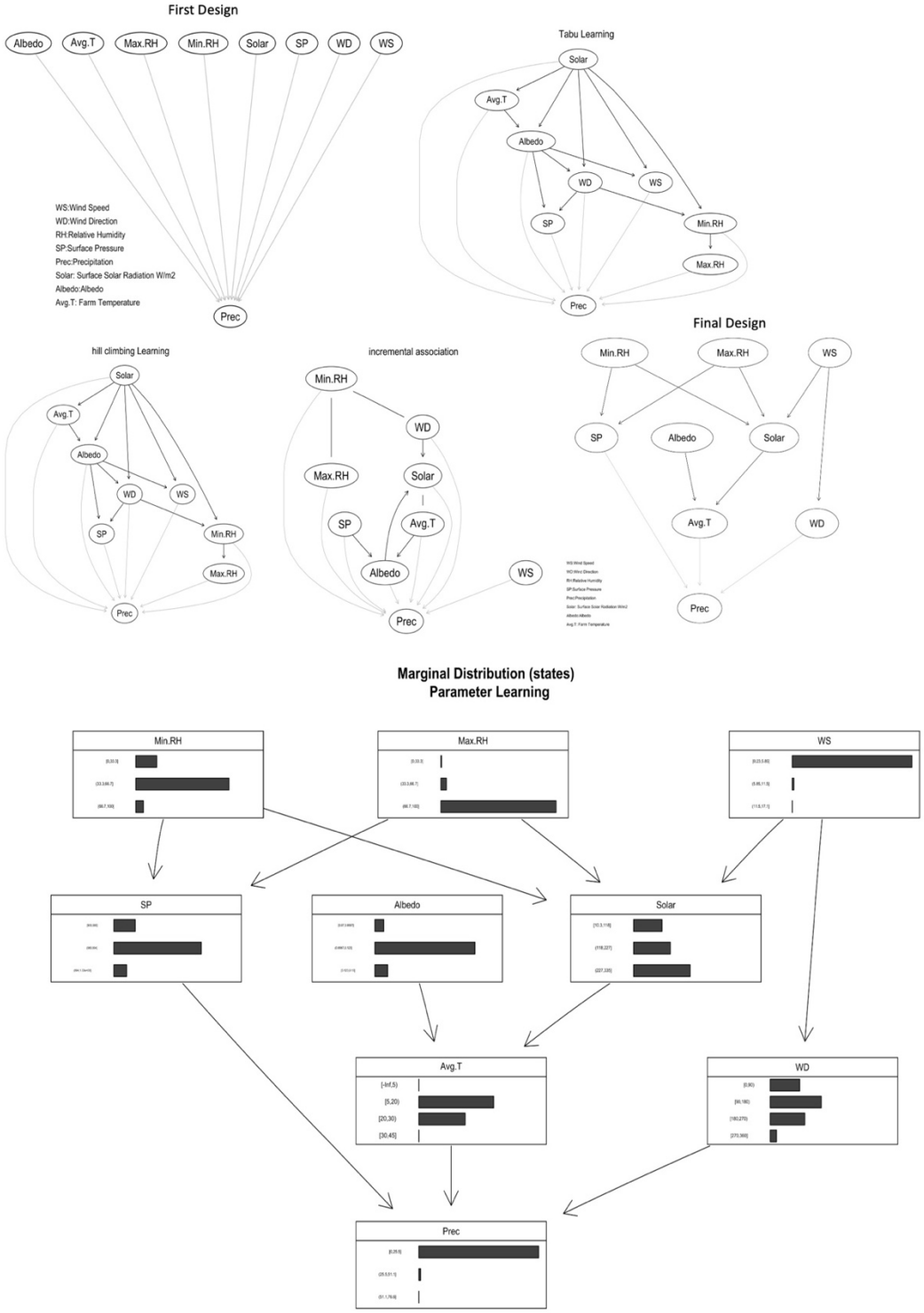


Figure 33 All the steps involved in processing the Precipitation Bayesian network

4.4 Results and Discussion

The outcomes of our Bayesian Network model offer valuable insights into the future trends of temperature and precipitation levels in the Messina region. These results are pivotal in supporting the prioritization of mango cultivation and conservation efforts. They empower stakeholders with the knowledge needed to make informed decisions to mitigate the potential impacts of climate change. Furthermore, the Bayesian Network analysis plays a crucial role in disaster risk management and the formulation of effective climate change adaptation strategies by pinpointing areas of high risk and vulnerability. Following the training and validation of the Bayesian Network model for the target nodes, namely Minimum, Maximum, and Average Temperature, the results revealed certain statistical measures of prediction accuracy.

During this pivotal phase, we quantify the model's performance, enabling us to estimate the prediction error of the designed BN model and evaluate its achievement with regard to the intended objectives (Kragt, 2009; Rodrìguez et al., 2010). Figure 33 illustrates the outcomes of the validation process for the BN model presented in Figure 30. It provides a report on the average BN-model losses, quantified in terms of expectation errors (Scutari, 2017). For the Minimum Temperature, the Median Predicted Error was approximately 0.21227354, for Maximum Temperature, it was around 0.33603739, and for Average Temperature, it stood at approximately 0.28287644.

The prediction errors observed in our Bayesian Network model validation provide insights into the model’s reliability. The Median Predicted Errors, while not negligible, suggest a relatively small margin of error. These errors could be attributed to the inherent uncertainty in climate predictions. However, the model’s consistent performance across these temperature nodes underscores its potential for providing reasonably accurate temperature forecasts.

To analyse the temperature of mango farms, our Bayesian model considered five distinct scenarios, each representing different climatic conditions:

1. Extreme Weather: Highest surface pressure, lowest wind speed.
2. Humidity Driven Warmth: Lowest wind speed, low Min. Humidity, high Max. Humidity.
3. Solar Radiation Surge: High solar radiation and albedo.
4. Windless Heat: Lowest wind speed, southern wind direction
5. Radiant High Pressure: Highest solar radiation, lowest albedo, high surface pressure

Analysis of these scenarios reveals that scenarios 3 and 5 tend to result in increased medium and high-class temperature probabilities for the Minimum Temperature node. For the Maximum Temperature node, scenarios 2, 3, and 5 exhibit significant changes, indicating an increased likelihood of higher temperatures. The same effect from scenarios 3 and 5 on the Average Temperature node is evident.

These scenarios illustrate the sensitivity of temperature predictions to various climatic conditions. Scenarios with higher solar radiation, lower albedo, and elevated surface pressure tend to result in increased temperatures. Conversely, scenarios with lower wind speeds and southern wind directions can also contribute to temperature rise. These findings highlight the intricate interplay of climatic variables and their impact on temperature, emphasizing the need for adaptive strategies in mango cultivation.

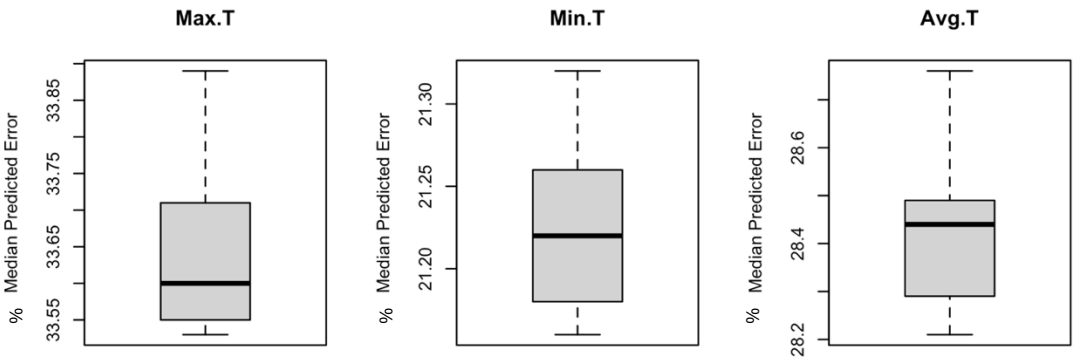


Figure 34 Validation Box-Plot Each Temperature Node

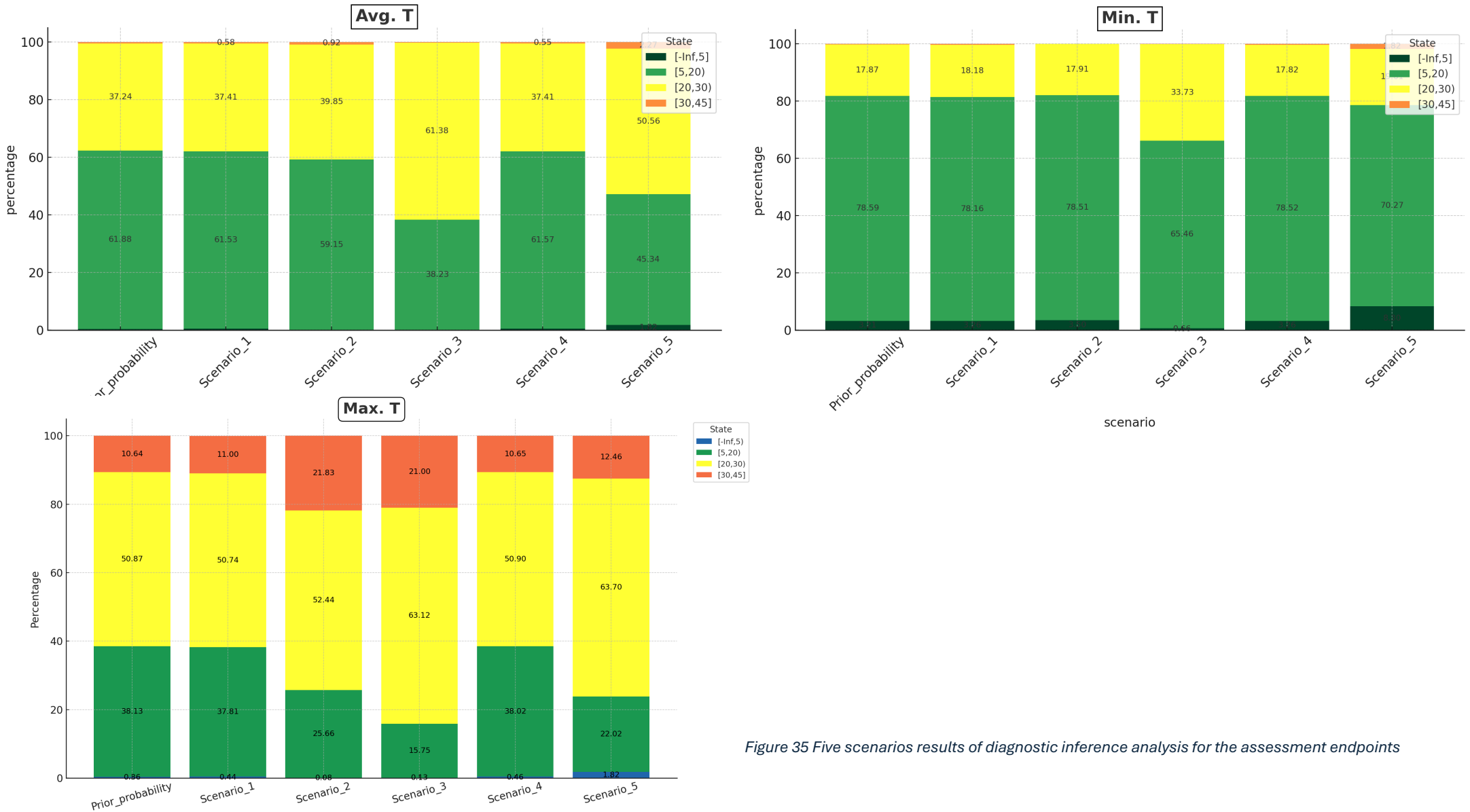


Figure 35 Five scenarios results of diagnostic inference analysis for the assessment endpoints

Future "What if?" Scenarios for Precipitation Variation:

1. Heat Induced Rain: Highest temperature and highest southern wind direction.
2. Humid Heatwave: Highest temperature, highest classes of max and min relative humidity.
3. Dry High Pressure: High surface pressure, lowest classes of albedo and wind speed.

The results of these probabilistic scenarios reveal that scenarios 1 and 2 predict higher precipitation amounts, while Scenario 3 shows a relatively smaller change. These precipitation scenarios demonstrate that temperature and relative humidity play significant roles in precipitation predictions. Scenarios with higher temperatures and humidity levels tend to result in increased precipitation, aligning with established meteorological principles. Conversely, scenarios with lower wind speeds, surface pressure, and albedo exhibit relatively lower precipitation probabilities. These insights can guide decision-makers and farmers in understanding and adapting to potential changes in precipitation patterns, a crucial aspect of mango farm management under climate change conditions. In summary, the Bayesian Network model's results and scenario analyses offer a comprehensive understanding of the potential future climate trends in the Messina region and emphasize the importance of adaptive strategies in safeguarding mango farms from the evolving challenges posed by climate change.

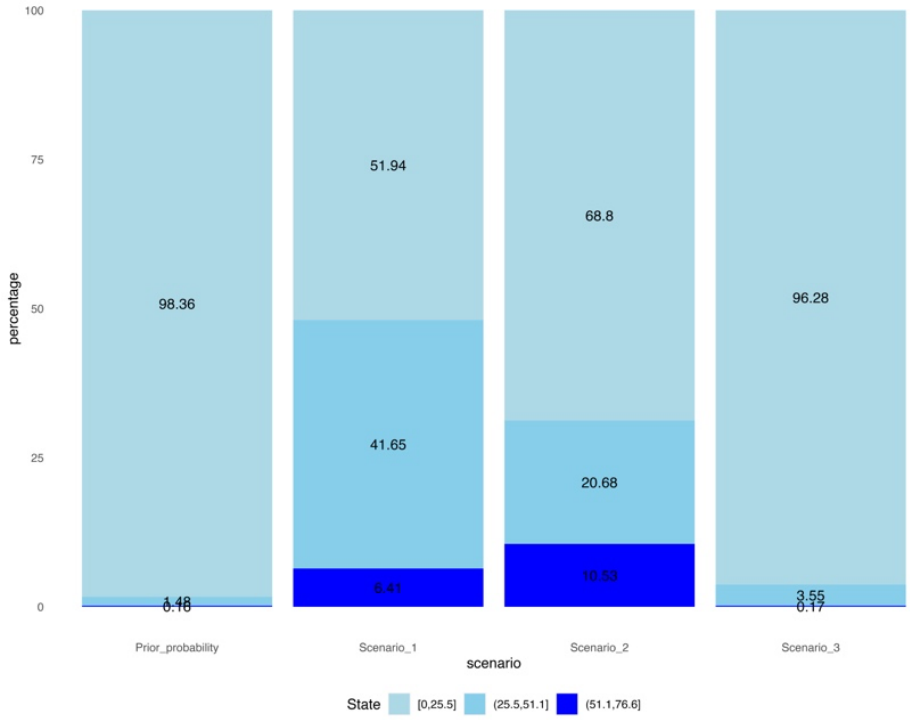


Figure 36 Probability of Precipitation Changes through different Scenarios, State is demonstrating the different ranges of precipitation in "mm"

4.5 Conclusion

This study tackles the urgent issue of ensuring food security amid climate change, focusing on Sicily's semi-tropical region. With favourable conditions for crops like mangoes, proactive measures are crucial to mitigate climate impacts. Climate change introduces risks such as temperature shifts, extreme weather, soil salinity, and erratic rainfall, all threatening agriculture and food security.

Using Bayesian Networks (BN), we modelled and assessed multiple climate risks under various future scenarios. BN's ability to consider complex interactions among climate variables provides a holistic view of potential impacts. By integrating data from Sicily's Messina region (2009-2022), we gained insights into local temperature and precipitation variations. The BN model predicts future temperature and precipitation trends, aiding decision-making for mango cultivation and conservation. This analysis is vital for disaster risk management and climate adaptation, identifying high-risk areas and vulnerabilities in agriculture. Despite inherent uncertainties, the BN model showed reasonable accuracy in temperature forecasts.

Scenario analyses highlighted the sensitivity of predictions to climatic conditions, with factors like solar radiation, wind, humidity, and surface pressure playing significant roles. Understanding these interactions is key for developing adaptive strategies in mango cultivation. The BN model also provided insights into precipitation patterns, identifying temperature and humidity as key drivers.

This study demonstrates the effectiveness of Bayesian Networks in assessing climate risks for agriculture in Sicily, offering valuable information for decision-makers. Integrating spatial results into Geographic Information Systems (GIS) can enhance decision-making and support effective disaster risk management and climate adaptation plans.

Overall, this research contributes to ensuring food security and resilience against climate change, not just in Sicily but in similar regions globally. It underscores the need for proactive, science-based approaches to address the complex risks climate change poses to agriculture and food systems.

**This paper has been published by ISHS (International Society of Horticultural Science) **
https://doi.org/10.17660/ActaHortic.2025.1415.16

MISAR in Enhancing Agricultural Resilience: A Comprehensive Approach to Climate Change Risk Management for Mango Farms in Sicily, Italy

Mohsen Pourmohammad Shahvar[1], Dario Scuderi [2], Davide Valenti[1], Alfonso Collura[3], Salvatore Micciché[1], Vittorio Farina [2], and Giovanni Marsella[1]

[1] *Dipartimento di Fisica e Chimica "E. Segrè", Università degli Studi di Palermo, Italy.*

[2] *Dipartimento di Scienze Agrarie, Alimentari e Forestali, Università degli Studi di Palermo, Italy.*

[3] *Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Palermo, Italy*

4.6 ABSTRACT

Agriculture plays a crucial role in the economy of Italy, particularly in the region of Sicily where it serves as a primary source of income. To ensure high yields, it is essential to enhance farmers' knowledge and awareness, especially in mitigating potential risks and damages caused by climate change and managing farming processes such as soil and water preparation, fertilizer, and pesticide management. To follow the MISAR (Climate Change Risk **M**anagement by **I**mproving the Individual and **S**ocial **A**wareness of **R**isk in Sicily) targets, this paper focuses on the importance of Information and communication technologies (ICT) in the "Mango Farms Risk Management Plan" to foster stronger connections between stakeholders and farmers in Messina. Climate change poses various hazards such as temperature fluctuations, extreme events, soil salinity, and irregular rainfall, which are expected to increase in the future. Effective decision-making for stakeholders and farmers requires efficient analytical tools, particularly for handling large datasets. The paper introduces a new architecture called ADM, which combines Decision Support Systems (DSS), Agent-Based Modelling (ABM), and Machine Learning (ML) methods to develop a comprehensive risk plan for future agricultural challenges. The ADM model in MISAR incorporates empirical information collected during the ML phase, including the reactions of Mango plants to risks and determining factors like extreme temperature changes. To promote and safeguard mango cultivation and production, changes in temperature are estimated using advanced techniques such as Random Forest and Feed-Forward Neural Networks. Weather stations equipped with

meteorological sensors are strategically placed within farms, providing direct measurements of hazards. Each station has its own credentials, allowing farmers access to the data. Furthermore, historical data analysis considers data from municipal meteorological stations and satellite sources. The model facilitates mutual communication between decision-makers and farmers, enabling farmers to monitor forecasts and report unexpected events in their respective farm areas.

Keywords: MISAR, Machine Learning, Artificial Intelligence, Mango, Agriculture, Decision Support System, Agent-based Modelling, Random Forest, Feed-forward neural network.

4.7 INTRODUCTION

The MISAR (Climate Change Risk Management by improving the Individual and Social Awareness of Risk in Sicily) project is built upon a comprehensive research framework, utilizing the latest advancements in risk analysis, modelling techniques, behavioural theories, and social sciences. Its main aim is to enhance resilience against climate change impacts. This objective is well-articulated by Shahvar et al. (2022).

Central to the MISAR project is a deep understanding of how climatic variables affect crop growth and yield, as discussed by Normand et al. (2015). The threat of rapid climate change is significant, with potential repercussions for both society and the natural environment, underscored by Shahvar et al. (2022). Given the agricultural sector's vulnerability to climatic shifts, it's crucial to address these looming challenges proactively.

Mango, prized for its economic and nutritional value (FAO, 2022b, 2023b), emerges as a crucial crop within this context. *Statista's*[1] data for 2021 ranks mango as the sixth most-produced fruit globally, accentuating its significance in the agricultural landscape. Sicily, with its unique climatic conditions, including an average temperature that rarely dips below 10 °C for eight months a year and minimal lows of 6 °C during the coldest periods (Gugliuzza et al., 2023), is home to approximately 55 hectares of mango orchards along its coastal regions ("Department of Sicilian Agriculture," 2017). These areas benefit from well-draining soils and natural windbreaks like cypress trees, creating an ideal environment for mango cultivation. Various mango varieties, such as *Kensington Pride, Keitt, Glenn, Maya, and Tommy Atkins*, thrive in this setting, displaying a wide range of fruit weights (Farina et al., 2013; Gentile et al., 2019).

[1] https://www.statista.com/statistics/264001/worldwide-production-of-fruit-by-variety/

4.8 MANGO CULTIVATION MANAGEMENT BY USING ICT

Mango cultivation along coastal regions faces a multitude of challenges, primarily stemming from the dynamic and ever-changing climate conditions that originate from both land and sea (Farina et al., 2020). To effectively address these challenges and optimize mango cultivation, the integration of cutting-edge Information and Communication Technology (ICT) solutions has become a game-changer in this field.

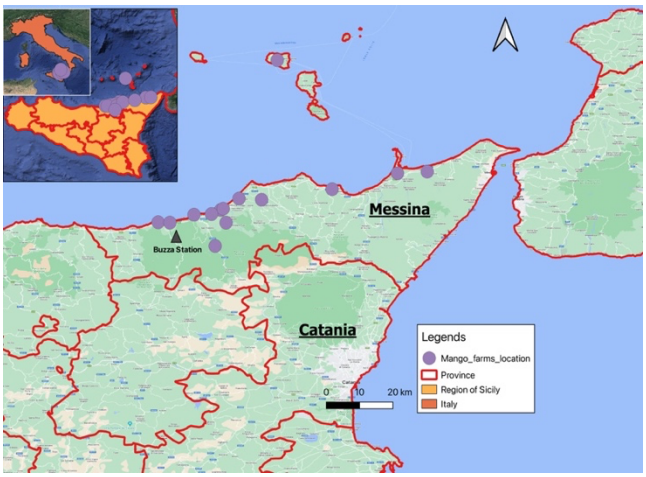


Figure 37 Case Study Area

4.9 CHALLENGES OF COASTAL MANGO CULTIVATION

4.9.1 Saltwater Intrusion

Rising sea levels due to climate change have led to saltwater intrusion into the soil, posing a direct threat to land quality (Tarolli et al., 2023). Mango trees are highly sensitive to high salinity levels, which can impede their growth and productivity (Gentile et al., 2019).

4.9.2 Storm Surges and Flooding

Coastal areas are more vulnerable to storm surges and flooding (Saleh et al., 2022), both of which can cause significant damage to mango orchards and uproot trees. The increasing frequency and intensity of storms due to climate change amplify the risks faced by mango cultivation in these regions (Asare-Nuamah et al., 2022).

4.9.3 Strong Winds

Coastal areas are often subjected to strong winds that can break branches or uproot mango trees. With climate change contributing to more severe weather events, the risk of wind damage to mango orchards is on the rise (Farina et al., 2017).

4.9.4 Increased Temperatures

Climate change can result in higher temperatures that may exceed the optimal range for mango cultivation. Extended periods of extreme heat can stress the trees, negatively impacting fruit development and yield (Gugliuzza et al., 2023).

Figure 38. Cupitur Orchard: Thriving amidst the Coastal Challenges in Messina

4.10 LEVERAGING INFORMATION AND COMMUNICATION TECHNOLOGY (ICT)

ICT solutions offer a robust approach to address the challenges of coastal mango cultivation, enabling efficient management and mitigation strategies.

4.10.1 Real-time Monitoring

Through the use of sensors and data analytics, ICT allows for real-time monitoring of weather conditions, salinity levels, and other environmental parameters (Akhter & Sofi, 2022).

4.10.2 Early Warning Systems

ICT can provide early warning systems that alert farmers to impending storms, floods, or strong winds (UNDRR, 2020).

4.10.3 Precision Irrigation

ICT-driven precision irrigation systems help manage water resources efficiently, combatting the problems of soil salinity and waterlogging (Zeynoddin et al., 2023).

4.10.4 Climate-Resilient Varieties

Using ICT, farmers can access information about climate-resilient mango varieties that can thrive in changing conditions (Acevedo et al., 2020).

Farms like "Cupitur" have demonstrated success in providing high water quality and wind protection. To enhance resilience, farmers can consider implementing coastal barriers or windbreaks, using elevated planting beds, and employing advanced drainage systems to mitigate the effects of storm surges and flooding.

4.11 RISK MAPPING AND ASSESSMENT:

Our approach uses advanced geospatial technology and QGIS to create detailed risk maps from Digital Elevation Models (DEM) and satellite land use data, classifying areas exposed to coastal proximity and artificial structures, especially up to 200 meters above sea level, into high, mid, and low-risk categories to systematically assess and mitigate climate-related risks to mango crops.

Figure 39 Risk Map

4.11.1 INTRODUCING METEOSENSE 4.0[2]:

Central to our strategy is the advanced MeteoSense 4.0 station, an agrometeorological tool designed to provide high-precision environmental monitoring and seamless integration with agronomic models in our Decision Support System (DSS). Instead of relying on 5G, MeteoSense 4.0 uses low-power, long-range communication protocols such as 4G (NB-IoT/CAT-M1), LoRa (868/915 MHz), and 2G to transmit data reliably across large agricultural areas. The system is capable of supporting additional microclimate IoT units spaced up to 8 km from the central station, ensuring comprehensive spatial coverage. The real-time data collected—including temperature, humidity, wind velocity, and precipitation—is transmitted to the Cloud Live Data portal, where it is accessible via smartphones, notebooks, and desktop devices. The sampling frequency is fully configurable, with the option to collect data at intervals ranging from seconds to hours depending on the application. This system revolutionizes mango cultivation management by empowering farmers and decision-makers with actionable, site-specific insights, enabling timely interventions in response to adverse weather events or evolving climate conditions.

Figure 40. MeteoSense 4.0 station

4.11.2 EMPOWERING DATA-DRIVEN DECISION-MAKING:

Our approach places data-driven decision-making at the forefront. Armed with real-time weather data and the insightful risk map generated through geospatial analysis, stakeholders

[2] https://www.netsens.it/en/

involved in mango cultivation are better equipped to thrive in an environment of dynamic climate challenges. They can proactively adapt planting and harvesting schedules, safeguarding crops from temperature extremes, strong winds, and precipitation events.

4.12 METHODOLOGY

4.12.1 Agent-Based Modelling Framework:

In our research, we explore climate change, agriculture, and decision-making by constructing an agent-based model (ABM) where meteorological factors act as dynamic agents. Temperature, crucial for mango cultivation, affects all stages from flowering to harvest, with extremes below 5°C and above 40°C causing significant damage. Our research also considers climate risks like strong winds and heavy precipitation. To manage this complex landscape, our agents use simple heuristics for decision-making.

4.12.2 The ADM (Agent-Based + Decision Support + Machine Learning) Architecture:

The ADM (Agent-Based + Decision Support + Machine Learning) architecture integrates agent-based modelling, decision support systems, and machine learning techniques. This architecture predicts temperature, a critical variable for mango cultivation, using machine learning methods like Random Forest and Feed Forward Network models. These models are fed with environmental data, including soil moisture, water volume, leaf age, branch rest time, and wind velocity. This comprehensive approach enables responsible corporations and decision-makers to make informed choices to protect mango plants.

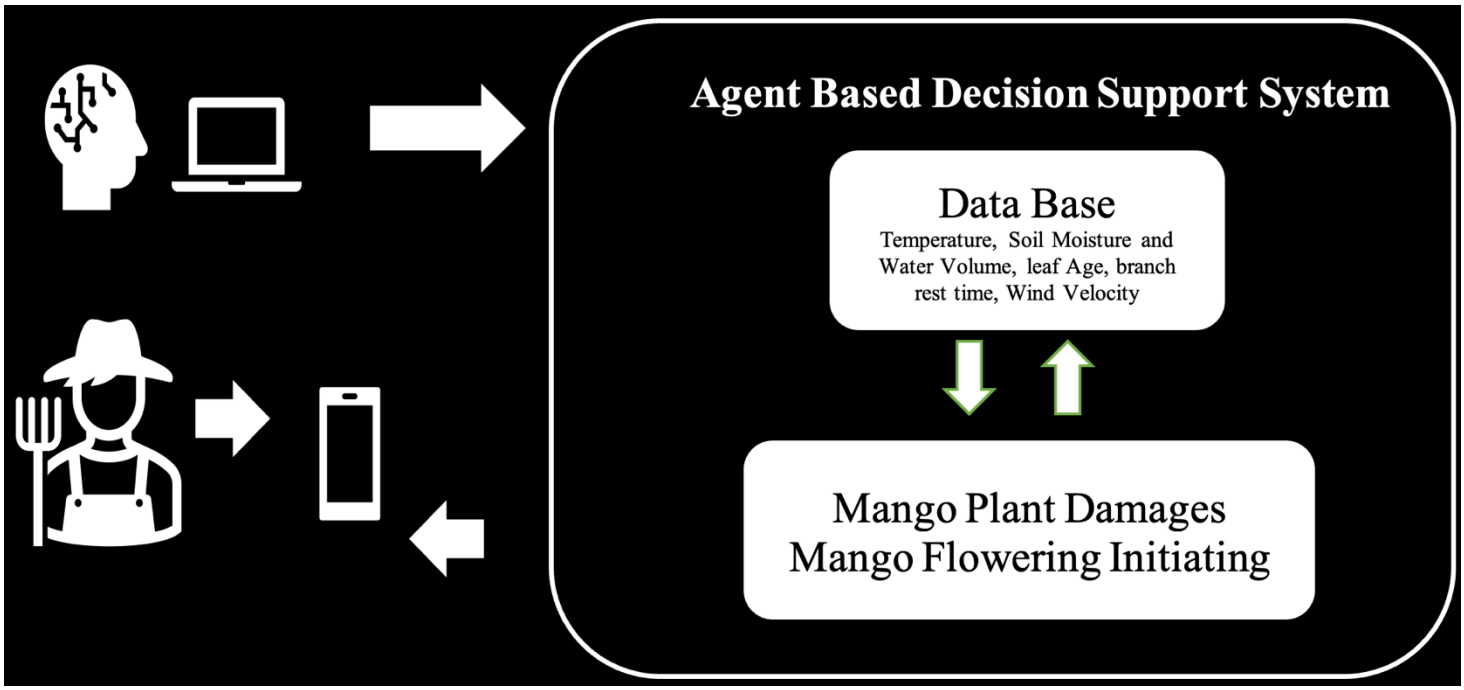


Figure 41. ADM Architecture

4.12.3 Leveraging ICT for Real-Time Collaboration:

Leveraging ICT for real-time collaboration in agriculture facilitates seamless information sharing and decision support between decision-makers and farmers. A dedicated website or application enables farmers to monitor weather changes and communicate swiftly about unexpected field circumstances.

4.12.4 Data Collection and Pre-Processing:

Our analysis is based on meticulously collected data from the primary meteorological website of the Sicilian region, spanning the years 2009 to 2021. Although the MeteoSense 4.0 station was recently installed in our field for high-resolution monitoring, it has been operational for less than two years. Due to the need for longer-term trends and model training on broader timescales, we relied on hourly historical datasets instead. These datasets include a wide range of climate-related variables such as temperature, humidity, albedo, solar irradiance, air pressure, wind velocity and direction, and precipitation. This rich temporal dataset serves as the foundation for our predictive models, enabling robust analysis of agro-meteorological patterns over more than a decade. Looking ahead, the real-time data from MeteoSense 4.0 will be progressively integrated into our system to enhance short-term forecasting accuracy, support

real-time decision-making, and validate long-term predictions with on-site sensor measurements.

4.12.5 Geographic Information System (GIS) Integration:

Our research integrates Geographic Information System (GIS) data, including Digital Elevation and Land Use Cover satellite images from the Copernicus website. Using QGIS software, we process and vectorize these images, classifying data based on agricultural areas, green spaces, and elevations prone to inundation and landslides to construct a comprehensive risk map. This spatial analysis enhances our understanding of localized climate risks in mango cultivation regions.

4.12.6 Correlation Analysis:

To streamline model complexity, we perform a correlation analysis to exclude highly correlated variables, simplifying our machine learning (ML) and artificial neural network (ANN) models. We use daily maximum and minimum relative humidity, albedo, solar irradiance, surface air pressure, wind velocity, and precipitation quantity for temperature prediction in the Random Forest and Feed-Forward Neural Network models.

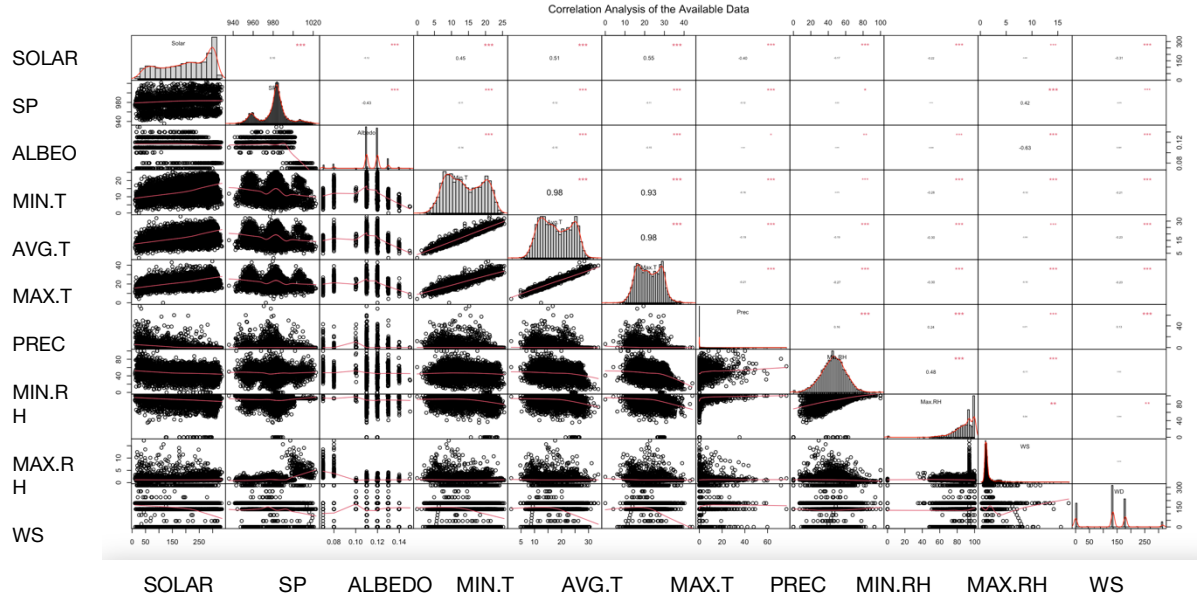


Figure 42. Correlation Analysis, Maximum Temperature Vs. other variables

4.13 RESULTS AND DISCUSSION

4.13.1 Random Forest Model:

In our pursuit of predicting future temperatures, the Random Forest (RF) algorithm emerges as a powerful tool. The RF model consists of an ensemble of decision trees that collectively make predictions. In our study, we divide the dataset into a training set comprising daily temperature readings from January 2009 to December 2020, and a test set containing

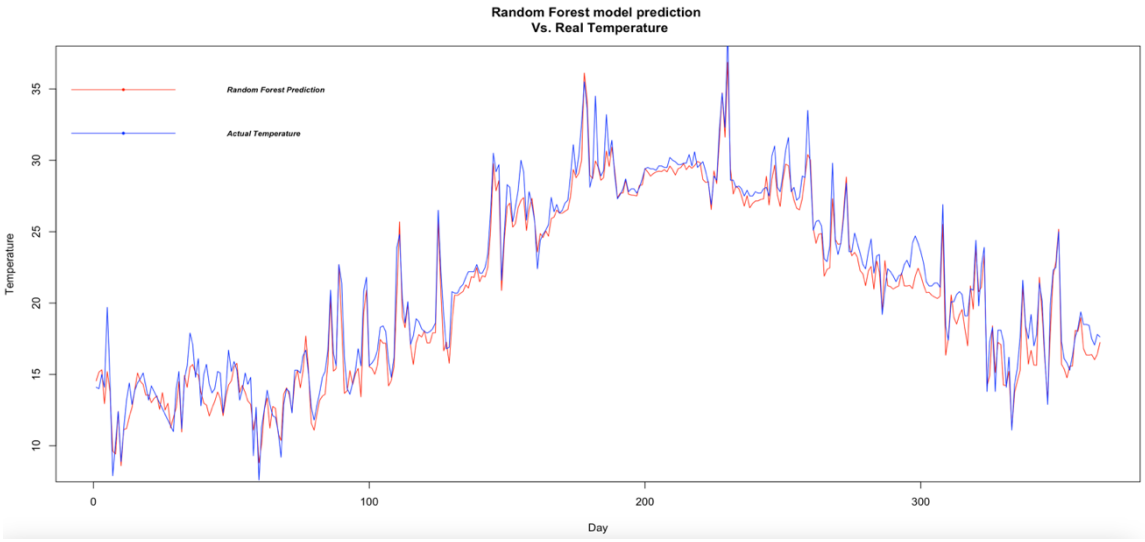


Figure 43. Max Temperature Prediction for year 2022

temperature data for the entire year 2021. After training/testing the RF model using the training set, we apply it to forecast temperatures for the year 2022.

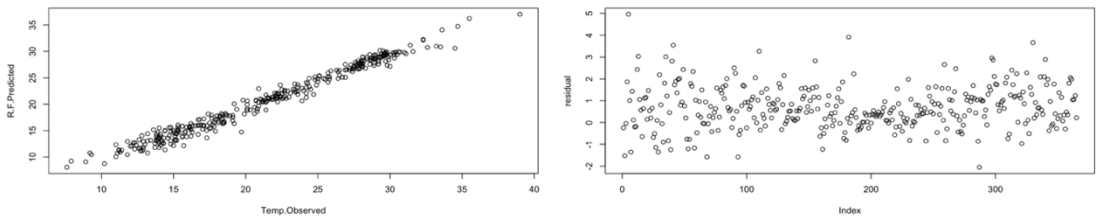


Figure 44. Residual Plots for Max Temperature Prediction in RF model

To assess the accuracy of the model on the test set, we compare the predicted temperatures with the actual temperatures. We visualize the performance of the RF model by plotting the predicted temperatures against the actual temperatures as depicted in Figure 42. Furthermore, we analyse the residuals as its illustrated in Figure 43, which are the differences between the observed and

predicted values. The residuals provide insight into the model's accuracy and any patterns or deviations present in the predictions.

4.13.2 Neural Network (Feed Forward) Model:

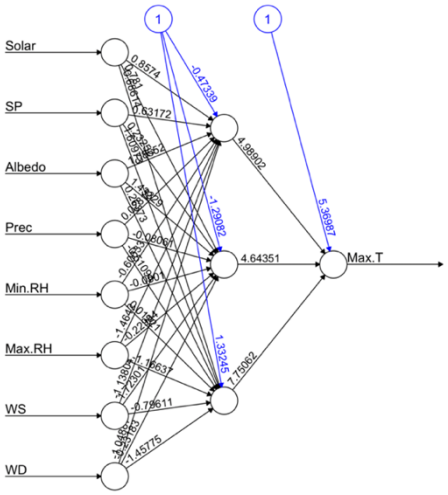


Figure 45. FeedForward Network Design

In tandem with the RF model, we employ the Feed Forward Neural Network (FNN) for temperature prediction. The FNN consists of multiple layers of interconnected neurons that process information and make predictions. The flow of information is unidirectional, starting from the input layer, passing through hidden layers, and culminating in the output layer.

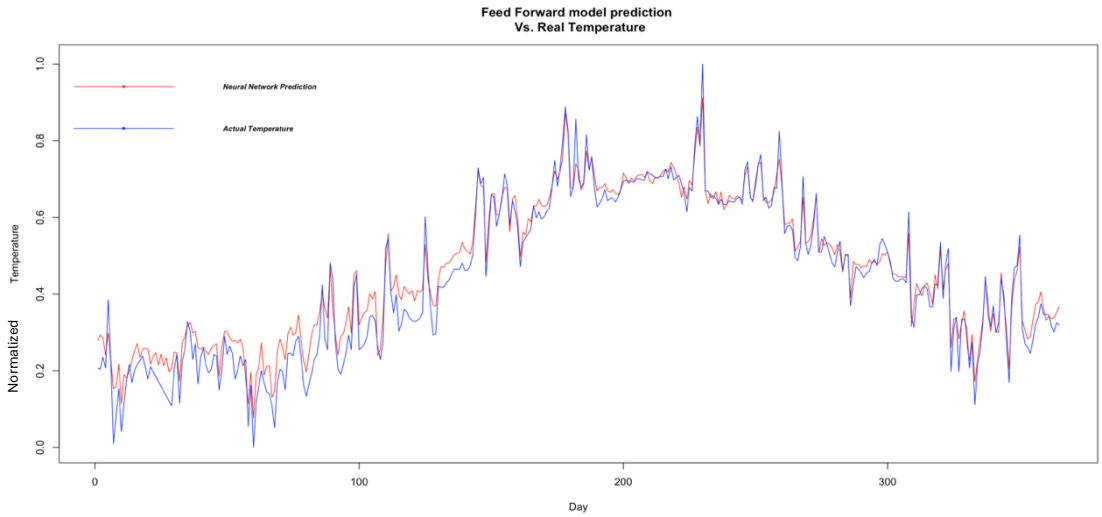


Figure 46. Feedforward prediction for Max Temperature Year 2022

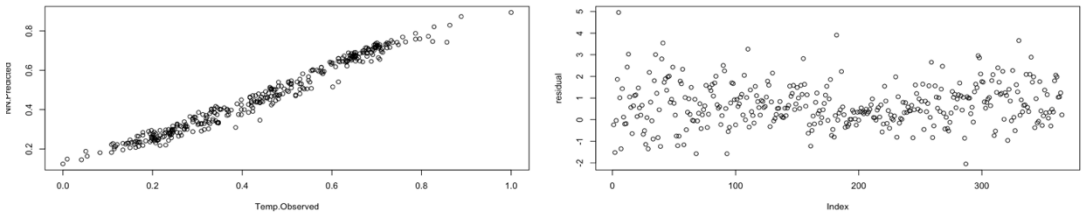


Figure 47. Residual Plot for FNN model

Like the RF model, the Feedforward Neural Network (FNN) model's performance is evaluated using the RMSE metric. For the FNN model, the RMSE is 0.019, while the RF model reports a higher RMSE of 0.866, indicating finer prediction granularity in the FNN. Additionally, the coefficient of determination (R^2) is 0.978 for the FNN and 0.977 for the RF model. These values confirm that both models fit the data very well, with the FNN slightly outperforming RF in terms of precision. The low RMSE and high R^2 of the FNN model, coupled with the absence of any discernible pattern in the residual analysis, highlight the accuracy and reliability of this forecasting approach.

To strengthen the clarity of the chapter and address reviewer concerns, we note that the temperature predictions from both the RF and FNN models are evaluated on the same temporal scale using daily values, ensuring temporal consistency and allowing reliable year-over-year trend comparisons. This consistency is particularly helpful in agricultural planning, where crop response is sensitive to daily fluctuations.

Moreover, while our study integrates both predictive modeling and spatial risk mapping, we acknowledge that further integration between the temporal prediction (Section 4.13) and spatial analysis (Section 4.11) could be enhanced. Future efforts will aim to directly embed temperature prediction outputs into dynamic geospatial risk models, supporting more localized early warning systems.

It is also important to clarify that the RMSE values were previously missing and have now been explicitly reported. Lastly, our simulations are based on historical data that does not include artificial noise. However, we recognize the importance of evaluating model robustness under noisy conditions. Future studies will incorporate noise perturbation to reflect real-world uncertainties and improve model resilience.

4.14 CONCLUSIONS

Our MISAR project focuses on managing climate change risks in agriculture, specifically for mango farms in Sicily. By leveraging advanced Information and Communication Technologies (ICT), we've developed a "Mango Farms Risk Management Plan" to enhance connections between stakeholders and farmers in the Messina region. A key innovation is our ADM (Agent-Based + Decision Support + Machine Learning) architecture, which predicts temperature variations and helps protect mango crops. This approach supports both environmental and economic sustainability in rural areas. Our interdisciplinary methods, combining agent-based modelling, machine learning, geospatial analysis, and ICT, empower farmers and decision-makers with data-driven tools to handle changing climate conditions. While we've made significant strides, further research with more comprehensive data and additional meteorological variables is needed to improve predictive accuracy. Future work should also aim to embed temporal predictions into geospatial risk models and evaluate model resilience against noise and measurement uncertainty.

In summary, the MISAR project provides valuable academic and practical solutions to climate change challenges in agriculture. We aim to enhance agricultural resilience and remain committed to improving our models and expanding data sources.

** This is the first version for peer reviewing in “Mathematics” **

The final version, completed after the thesis submission date, is available in **Annex 1

A Potential Hybrid Deep Learning Approach to Temperature Prediction Using MODIS Satellite Data and Historical Records

Mohsen Pourmohammad Shahvar[1], Davide Valenti[1], Alfonso Collura[3], Salvatore Micciche[1], Vittorio Farina [2], and Giovanni Marsella[1]

“Corresponding Author: Mohsen Pourmohammad Shahvar”

[1] Dipartimento di Fisica e Chimica “E. Segrè”, Università degli Studi di Palermo, Italy.

[2] Dipartimento di Scienze Agrarie, Alimentari e Forestali, Università degli Studi di Palermo, Italy.

[3] Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Palermo, Italy

mathematics

4.15 Abstract

This study presents a hybrid mathematical framework for daily air temperature fore-casting, integrating satellite-derived remote sensing data with time-series modeling. The proposed architecture combines a Residual Neural Network (ResNet) for spatial feature extraction from MODIS imagery, XGBoost and Random Forest for multivariate regres-sion, and an ARIMA model for residual temporal correction. Applied to northeastern Sicily, a climate-sensitive region for mango agriculture, the model is trained on data from 2007–2021 and tested on unseen years (2022–2024). Results demonstrate high forecasting accuracy ($R^2 > 0.97$; RMSE < 0.5 in 2022), surpassing Transformer-based baselines. The integration of statistical and deep learning components enables robust handling of non-linearity, autocorrelation, and seasonal variation. This approach exemplifies the value of mathematical modeling in environmental prediction, offering a scalable method for climate adaptation in precision agriculture.

4.16 Introduction

Mango cultivation is highly sensitive to temperature fluctuations, which can severely impact crop yield, fruit quality, and flowering cycles (Khalifa & Abobatta, 2023; POURMOHAMMAD SHAHVAR et al., 2023; Scuderi et al., 2025). In areas like Acquedolci and Caronia, Sicily, where there are numerous mango plantations, having potential temperature forecasts is crucial for effectively managing farms and protecting crops (POURMOHAMMAD

SHAHVAR et al., 2023; Scuderi et al., 2023). Traditional weather forecasting methods, which rely primarily on ground-based observations, often lack the spatial resolution needed for detailed agricultural applications (Naresh, 2019). To fill this gap, we suggest combining satellite imagery with historical temperature data using advanced deep learning techniques.

Remote sensing technology, especially MODIS Terra and Aqua satellite imagery, offers wide temporal coverage and fine spatial resolution. These features make it perfect for environmental monitoring and agricultural forecasting. The challenge, however, is to process and integrate this data effectively with existing ground observations (Sishodia et al., 2020).

The introduction of deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized image processing by enabling the automated extraction of spatial features. ResNet (Residual Network), known for addressing vanishing gradient issues in deep networks, has emerged as a robust framework for extracting hierarchical features, particularly from large-scale satellite imagery datasets (Shafiq & Gu, 2022).

Previous studies have examined effectiveness of various machine learning methods in temperature and weather forecasting. For example, Lakshminarayana (2020) showed how artificial neural networks could be used for rainfall forecasting, and Jain et al. (1996) demonstrated the adaptability of neural networks in recognizing patterns. Our study leverages the ResNet architecture to process spatial data from MODIS and integrates it with historical temperature records using ensemble learning techniques like XGBoost and Random Forest. The addition of ARIMA enables residual correction, addressing temporal dependencies that arise in daily temperature predictions.

In places like Acquedolci and Caronia in the region of Messina, precise temperature monitoring is essential to manage the risks from extreme weather events (Scuderi et al., 2025). Research by Jamal et al. (2023) underscores the value of precise weather predictions in optimizing irrigation schedules and reducing crop stress. Similarly, Nguyen et al. (2021) revealed that blending remote sensing data with machine learning algorithms can greatly improve agricultural productivity. By concentrating on temperature prediction, our study seeks to provide a vital tool for farmers to make well-informed decisions, thus boosting crop yields and minimizing losses.

The application of deep learning models in agricultural forecasting is becoming increasingly popular. Kamilaris & Prenafeta-Boldú (2018) reviewed various deep learning applications in agriculture, highlighting their potential to enhance yield prediction, disease detection, and soil moisture estimation. Our approach employs the ResNet architecture and

focuses specifically on temperature prediction, a critical factor in mango production, and integrates satellite imagery with historical temperature data to deliver precise forecasts.

Integrating satellite imagery with historical temperature data creates a comprehensive dataset that improves the accuracy of our predictions. According to Shalu & Gurjeet Singh (2023), using diverse data sources can significantly enhance the performance of machine learning models in environmental monitoring. Our study utilizes MODIS Terra and Aqua datasets, benefiting from their high temporal and spatial resolution, to enable precise temperature predictions at a local level. This integration is vital for tackling the specific challenges faced by mango farmers in Acquedolci, providing them with reliable data to support their farming practices.

4.17 Methodology

In this study, we designed a hybrid deep learning framework to predict daily temperatures by fusing satellite imagery and historical weather records. The workflow begins with preprocessing the data to ensure consistency and accuracy, followed by training a ResNet CNN to extract meaningful spatial patterns from satellite images. Alongside this, we leverage XGBoost to handle the tabular meteorological data.

To bring these two perspectives together, we use Random Forest as an ensemble model, combining the strengths of ResNet and XGBoost for better accuracy. Finally, ARIMA steps in to refine the results by correcting residual errors and capturing temporal trends. By addressing spatial, temporal, and non-linear relationships, this methodology creates a robust foundation for predicting temperatures with precision tailored specifically for agricultural needs.

4.17.1 Data Collection

Our study focuses on the mango-growing regions of Acquedolci, Sicily, where weather conditions can make or break a harvest. To build an accurate prediction system, we gathered available satellite imagery from the MODIS Terra and Aqua datasets spanning 2007 to 2022 while the historical data extended until September 2024. These images capture thermal infrared data, which encode temperature variations through gradients of color hotter areas in shades of red and cooler zones in blue.

Complementing this, we collected daily temperature readings from the Caronia Buzza meteorological station. This station's records provided invaluable ground-truth data to anchor our model. Together, these datasets formed a rich, multi-faceted foundation for training, validating, and testing our hybrid prediction model.

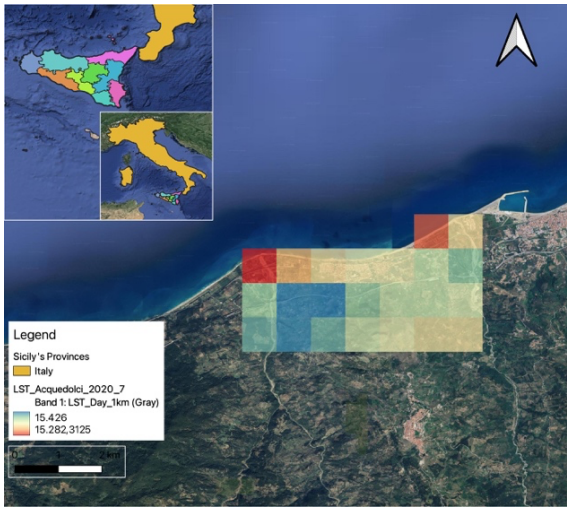


Figure 48 Case Study Satellite Image Sample

4.17.2 Preprocessing

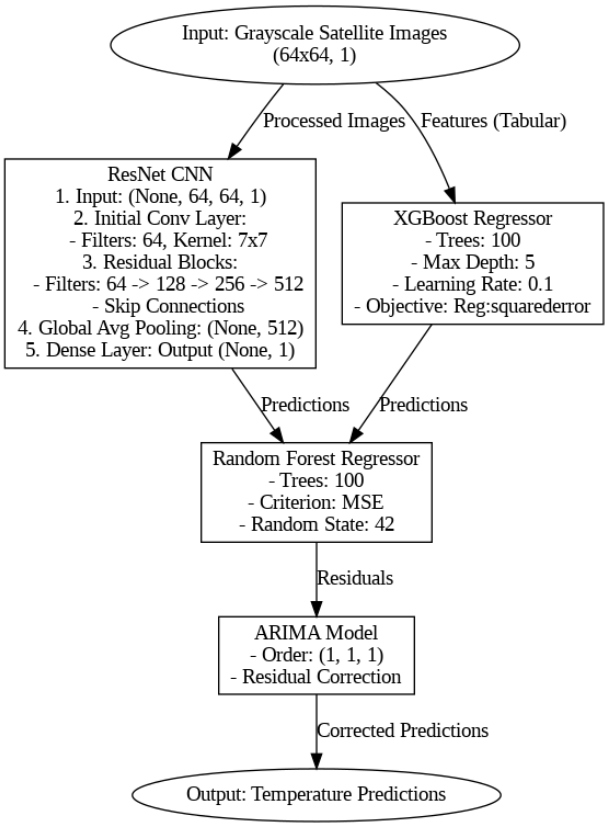


Figure 49 Hybrid Model Architecture

Figure 48 illustrates the hybrid model architecture designed for temperature prediction, blending satellite imagery with tabular meteorological data. Preprocessing played a crucial role in preparing both types of data spatial and tabular ensuring the model received clean and standardized inputs.

For the spatial data, satellite imagery from the MODIS Terra and Aqua datasets was resized to 64x64 pixels, creating a consistent input size for the ResNet CNN. Missing pixel values, often present in satellite data, were addressed using the np.nan_to_num function, replacing NaN values with zeros to maintain data integrity. The pixel values were normalized to fall within the range of 0 to 1 by dividing by 255, stabilizing the input for efficient training. Additionally, temperature data derived from the satellite images, originally measured in Kelvin, were converted to Celsius by subtracting 273.15, ensuring compatibility with meteorological conventions and improving interpretability.

The tabular data consisted of key meteorological features, including **albedo, cloud opacity, GHI, precipitable water, precipitation rate, relative humidity, surface pressure, wind direction at 10m, and wind speed at 10m**. Each of these features was normalized using a StandardScaler, which ensured that the data had a mean of zero and a standard deviation of one. This normalization was critical for maintaining numerical stability during training and balancing the importance of each feature.

To enable the hybrid model to learn effectively, the temporal resolution of the satellite imagery and the tabular meteorological data was aligned. Each satellite image was matched with its corresponding daily meteorological data, ensuring that the spatial patterns extracted by the ResNet CNN were directly linked to the numerical features from the tabular dataset.

4.17.3 Model Development

The hybrid model we developed is a unique integration of deep learning and traditional machine learning techniques, designed to capitalize on the complementary strengths of each approach.

4.17.3.1 ResNet Architecture

The ResNet CNN forms the backbone of the hybrid model, engineered to extract detailed spatial features from 64x64 grayscale satellite images. ResNet's innovative use of skip connections ensures smooth gradient flow across its layers, making it exceptionally effective at capturing both low-level and high-level spatial patterns. Below is an overview of its key components:

Input Layer: The model accepts preprocessed satellite images in grayscale format, with dimensions standardized to 64x64 pixels and a single channel, giving it an input shape of (None, 64, 64, 1).

Initial Convolutional Layer: This layer applies a 7x7 convolution with 64 filters, followed by batch normalization and a ReLU activation function. This configuration captures fundamental spatial patterns while ensuring the model remains computationally efficient. The output of this layer is (None, 32, 32, 64).

Residual Blocks: The model's defining feature, residual blocks, employs skip connections to mitigate vanishing gradients. Each block contains convolutional layers with progressively increasing filter sizes (64, 128, 256, 512), batch normalization, and ReLU activations. The skip connections allow the model to preserve important information across layers, enabling deeper and more effective learning.

Global Average Pooling (GAP): After extracting features through the residual blocks, a GAP layer reduces the spatial dimensions to a compact representation of size (None, 512). This step makes the model robust to spatial variations in input images.

Dense Layer: Finally, a fully connected dense layer maps the extracted features to a single output neuron, producing the predicted temperature. The output shape of this layer is (None, 1), representing the daily temperature forecast.

This ResNet-based structure was chosen for its ability to handle deep networks without suffering from performance degradation, making it the ideal candidate for processing satellite imagery.

4.17.3.2 XGBoost Regressor

While the ResNet CNN specializes in processing spatial data, the tabular meteorological features are handled by XGBoost, a tree-based machine learning algorithm renowned for its efficiency and accuracy. The numerical features, derived from meteorological records, were used.

XGBoost employs a gradient-boosting framework to capture non-linear relationships in this data. Key hyperparameters include:

Number of Trees: 100, **Maximum Depth**: 5, **Learning Rate**: 0.1, and **Objective**: Regression with squared error.

4.17.3.3 Ensemble Learning with Random Forest

To merge predictions from the ResNet and XGBoost components, we used a Random Forest Regressor. This ensemble approach combines the strengths of both models, yielding a unified prediction that captures spatial and numerical insights. The Random Forest configuration includes:

Number of Trees: 100, **Criterion**: Mean Squared Error (MSE), and **Random State**: 42.

By blending outputs, Random Forest ensures that the hybrid model benefits from the unique perspectives offered by the ResNet and XGBoost components.

4.17.3.4 Residual Correction with ARIMA

The final step in our hybrid architecture is ARIMA (Auto-Regressive Integrated Moving Average), a time-series model used to refine predictions. After Random Forest generates the unified temperature predictions, ARIMA addresses any remaining residual errors by capturing temporal dependencies. The ARIMA model is configured as an Order of (1, 1, 1), representing the auto-regressive, differencing, and moving average components. By adjusting for temporal trends and smoothing out inconsistencies, ARIMA ensures that the model delivers accurate, time-aware predictions.

As shown in Figure 48, the seamless integration of spatial, numerical, and temporal data processing makes this hybrid model a powerful tool for agricultural forecasting and climate monitoring.

4.18 Results and Discussion

The hybrid temperature prediction model was trained on data spanning 2007 to 2021, tested on 2022, and subsequently applied to forecast temperatures for 2023 and 2024. The results demonstrate the model's ability to capture seasonal and daily temperature trends through the integration of spatial, tabular, and temporal features. The following sections analyze the model's performance, contrasting the original predictions with those refined by ARIMA, and provide insights into how the residuals were handled, particularly for extreme events.

4.18.1 Model Validation on Test Data (2022)

The year **2022** served as a test dataset to validate the model's predictions against observed temperatures. Figure 49 presents the initial predictions before ARIMA correction, where the model demonstrates an ability to replicate the overall temperature trends, including seasonal

highs and lows. However, as highlighted in Figure 50, the residual plot shows the model's initial challenges in capturing extreme temperature events. These deviations, represented by higher residual values during abrupt changes in temperature, indicate areas where the original predictions fell short.

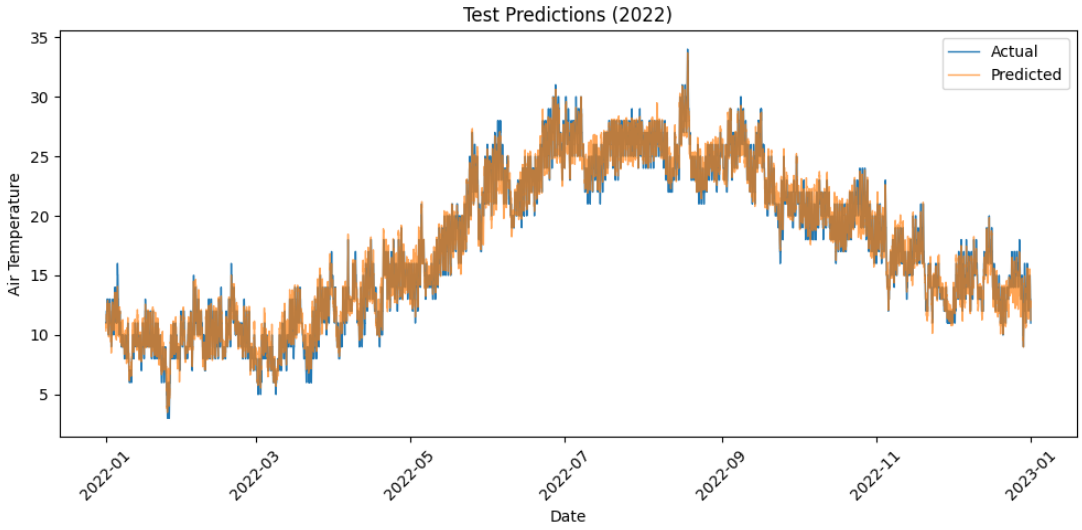


Figure 50 Test Predictions (2022)

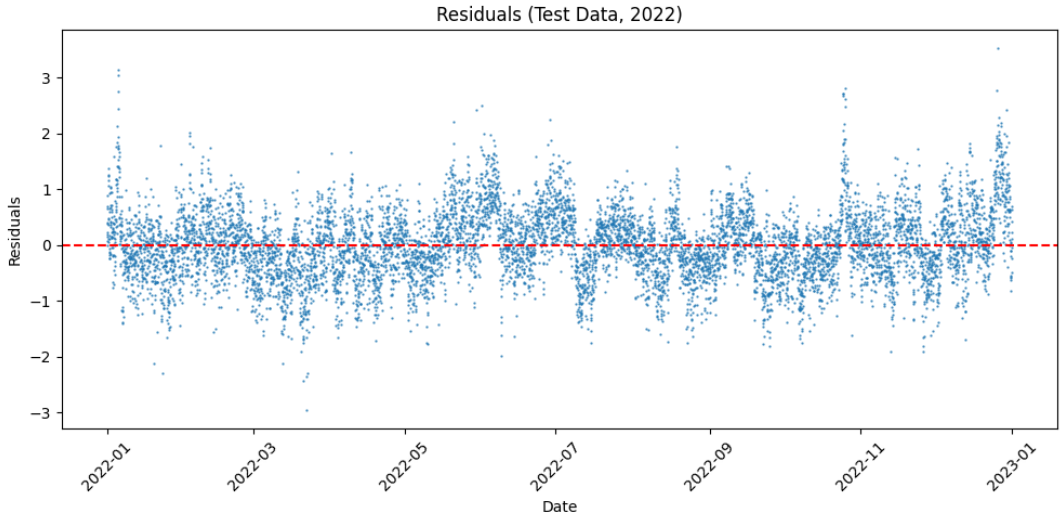


Figure 51 Residual for test data 2022

After applying ARIMA correction, the refined predictions (shown in Figure 51) align more closely with the actual observed temperatures. The ARIMA model significantly reduces residual errors, as evidenced by the tighter clustering of residuals around zero in Figure 52.

This correction addresses the model's initial difficulty in adapting to sharp fluctuations, improving temporal accuracy.

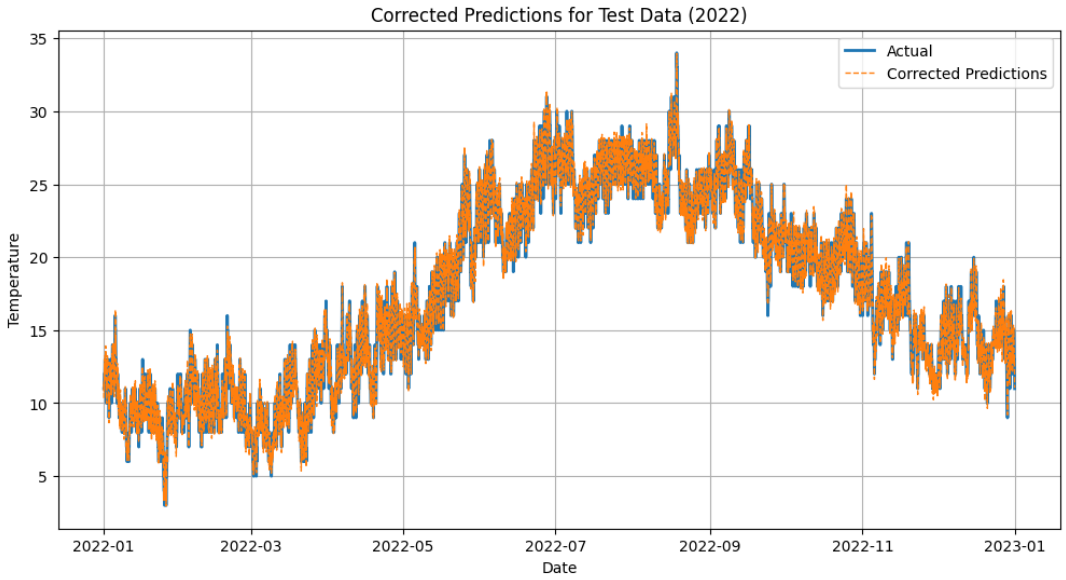


Figure 52 Corrected Prediction for test data 2022

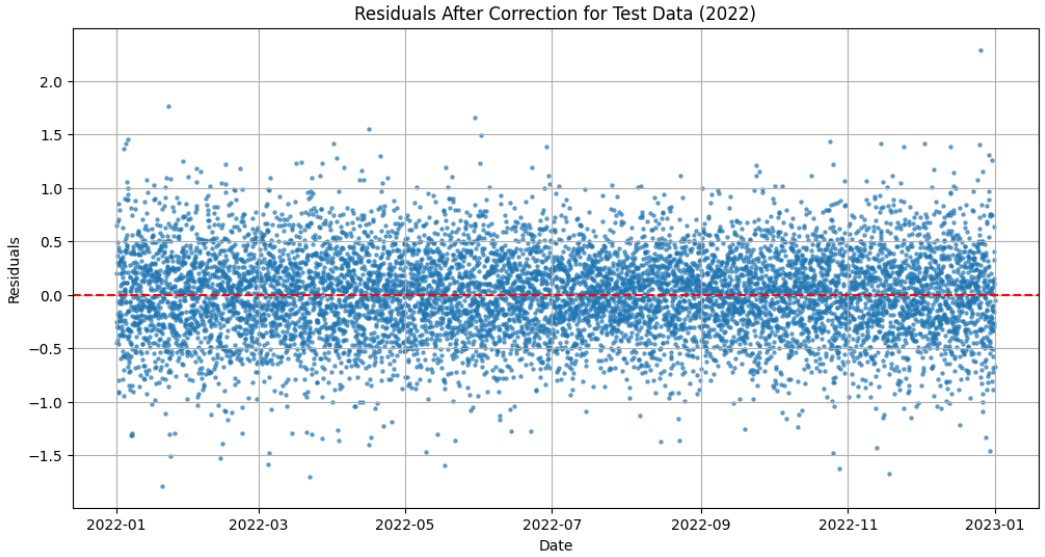


Figure 53 Residual Errors of test data 2022 after correction

4.18.2 Forecasting for 2023

The trained model was then extended to predict temperatures for the year **2023**, with results shown in Figures 53 and 54. In the original predictions (Figure 53), the hybrid model effectively captures seasonal patterns, including the peaks of summer and the dips of winter.

However, as with the 2022 test data, the residuals (Figure 54) reveal limitations in capturing the extremes, particularly during transitional periods like spring and autumn.

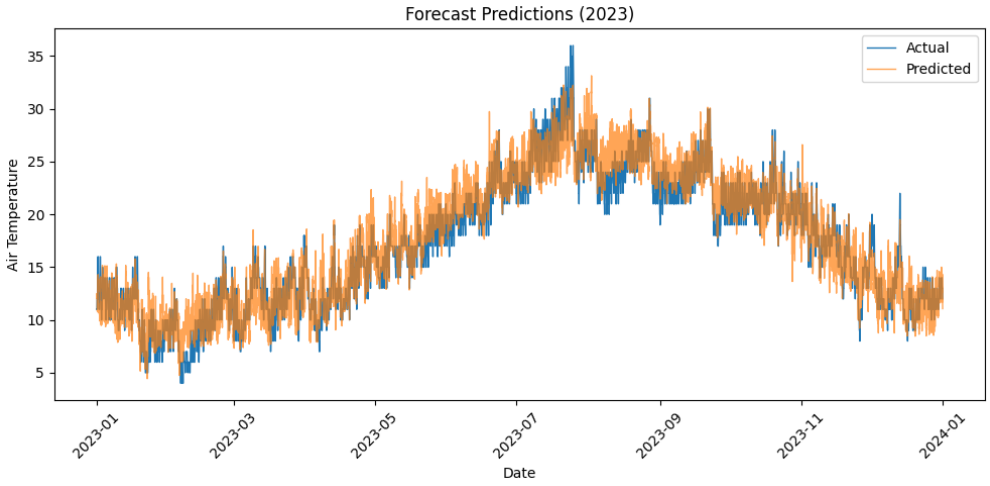


Figure 54 Primary Forecast for year 2023

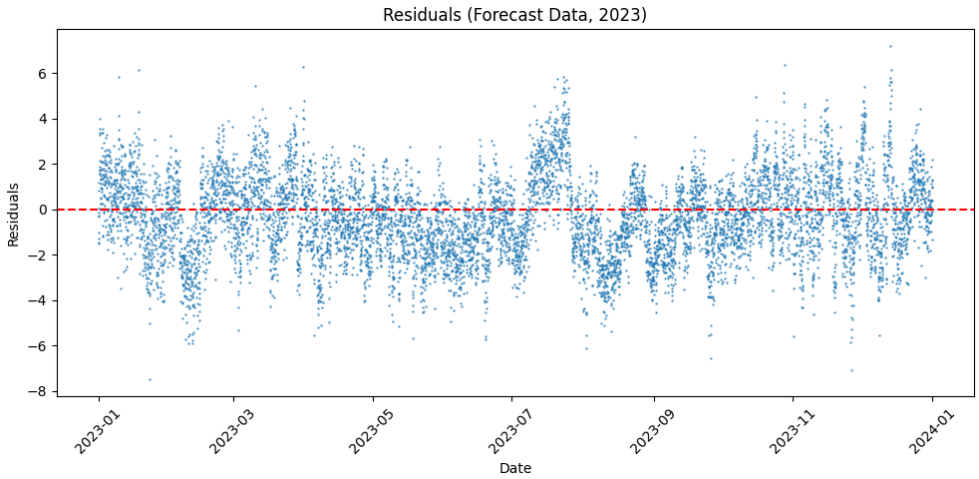


Figure 55 Residual Error for primary forecast 2023

ARIMA correction once again proves valuable, as seen in the improved predictions in Figure 55. The corrected residuals (Figure 56) display reduced variance and fewer outliers, confirming the robustness of the hybrid model in forecasting future temperatures when complemented by ARIMA.

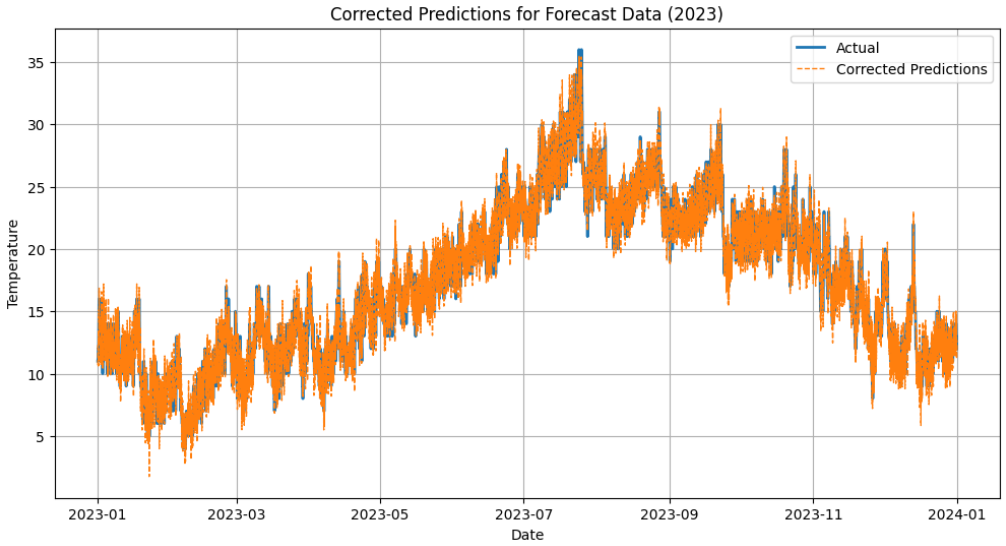


Figure 56 Corrected Prediction for year 2023

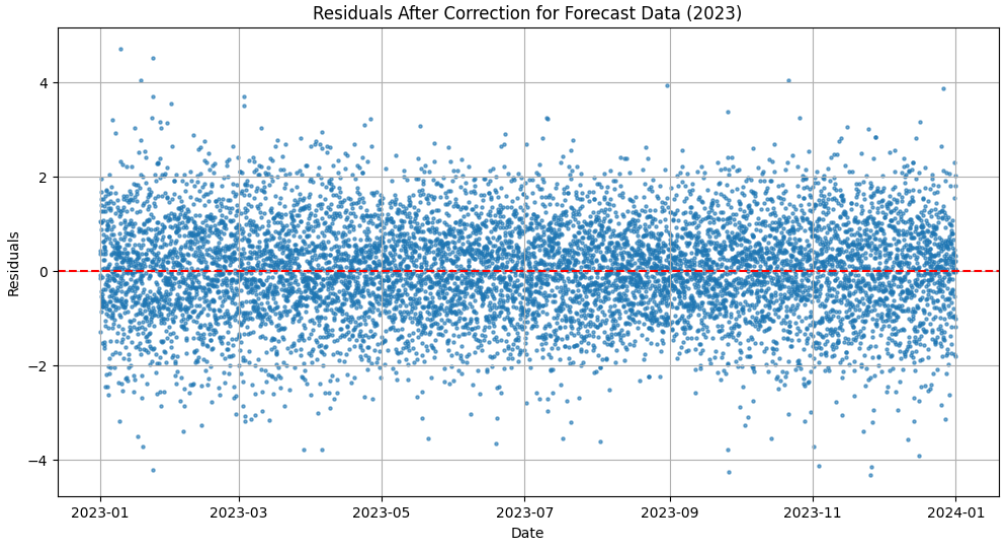


Figure 57 Residual Errors for year 2023 after correction

4.18.3 Forecasting for 2024

The model's application for forecasting **2024** highlights its adaptability to extended time horizons. As shown in Figures 57 and 58, the original predictions follow the expected seasonal trends but show some deviation during extreme events. These deviations are more pronounced compared to 2022 and 2023, reflecting the increased uncertainty in long-term forecasting.

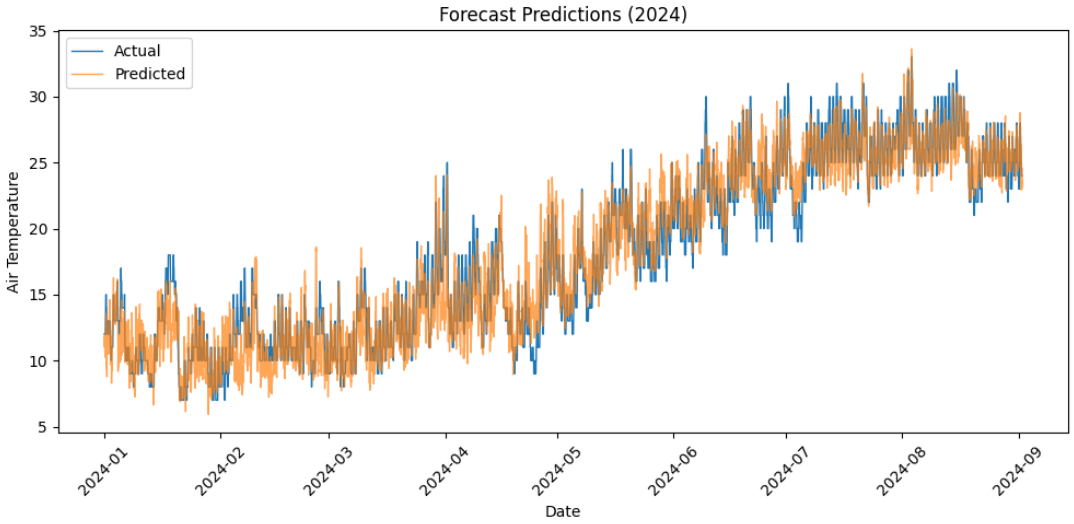


Figure 58 Primary forecast for year 2024

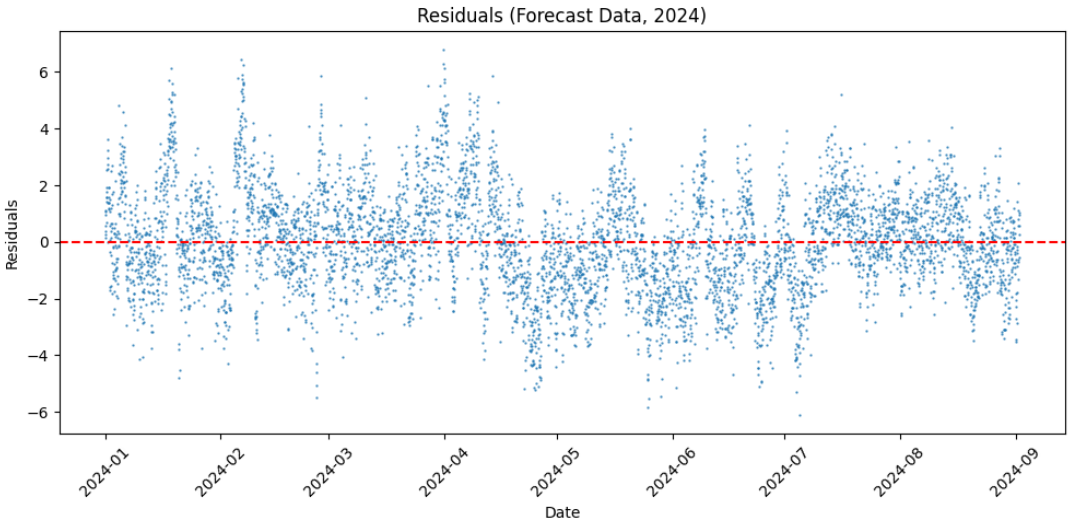


Figure 59 Primary Residual Errors for year 2024

With ARIMA correction, the refined predictions in Figure 59 demonstrate improved alignment with expected seasonal patterns, reducing discrepancies during sharp transitions. The residuals (Figure 60) are once again centered tightly around zero, showcasing the hybrid model's stability and the efficacy of ARIMA in mitigating long-term prediction challenges.

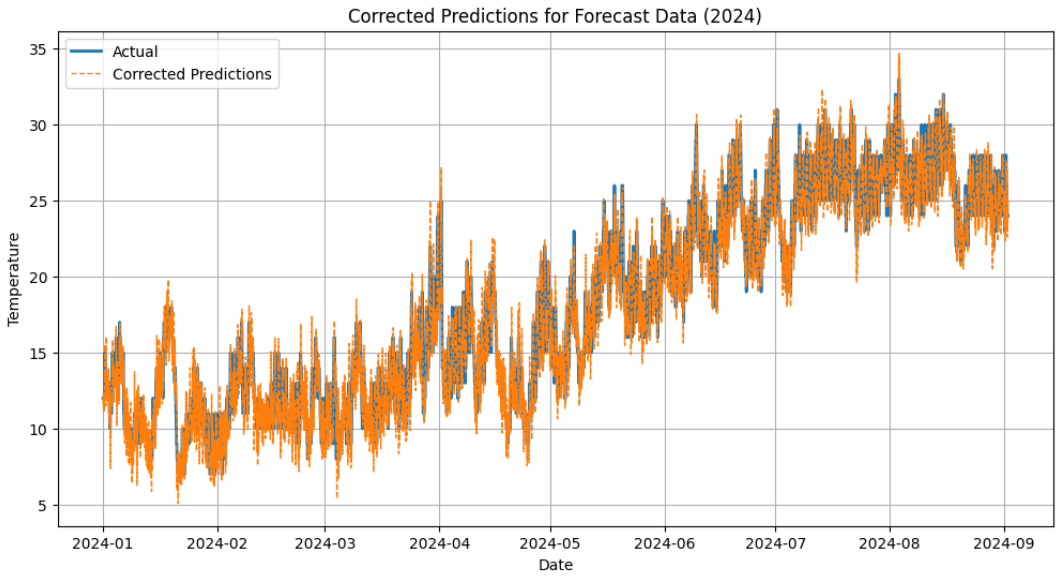


Figure 60 Corrected Forecast for year 2024

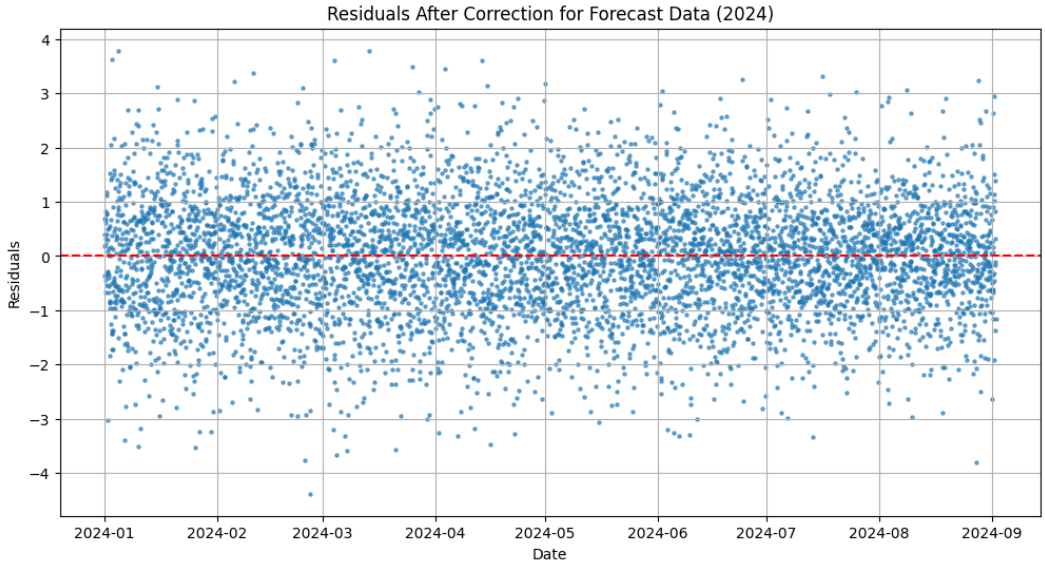


Figure 61 Residual Errors for year 2024 after correction

4.18.4 Overall performance across 2022-2024

Figure 61 illustrates the daily average temperature predictions compared to observed values across 2022, 2023, and 2024. The use of daily averages provides a clearer view of the model's performance by smoothing short-term fluctuations. For 2022, which served as the test dataset, the model predictions align closely with the observed values, showcasing higher accuracy due

to the model's familiarity with this data during training. This strong performance reflects the model's capability to generalize well on test data while benefiting from ARIMA corrections to handle extreme temperature events effectively.

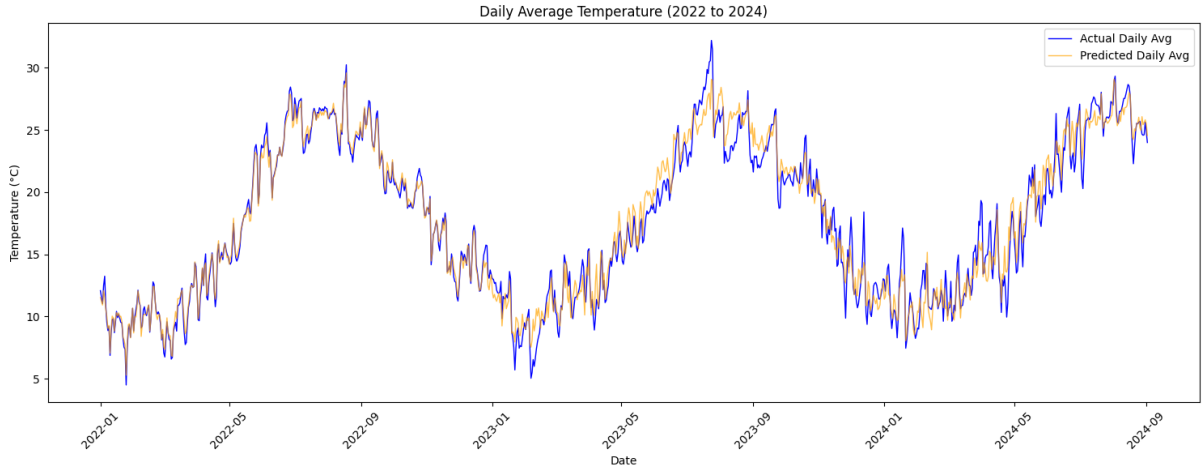


Figure 62 Daily Average Temperature Prediction across 2022-2024

For 2023 and 2024, where the model was applied to forecast untrained data, the predictions remain reliable but with slightly larger deviations compared to 2022. These discrepancies, particularly noticeable during transitional seasons such as spring and autumn, are expected in long-term forecasting. ARIMA corrections helped smooth out residual deviations, ensuring that the model captured seasonal trends and extreme events more effectively. Overall, the plot demonstrates the model's robustness and capacity to balance short-term accuracy with long-term predictive power, making it a valuable tool for temperature forecasting in agricultural applications.

4.18.5 Residual Analysis and Extreme Event Prediction

One of the critical findings from the residual analysis (Figures 50, 52, 54, 56, 58, and 60) is the model's initial difficulty in accurately capturing extreme events. These challenges are inherent in hybrid models that rely on both spatial and tabular data but can be addressed through post-prediction correction techniques like ARIMA.

- **Initial Predictions**: The original predictions capture general trends but struggle during periods of rapid temperature changes or extreme deviations, leading to higher residuals.

- **ARIMA Correction**: By modeling and correcting residuals, ARIMA enhances the temporal precision of predictions, particularly for extreme events. The corrected residuals consistently show reduced variance and tighter clustering around zero, underscoring the importance of this step in the pipeline.

4.18.6 Implications and Model Robustness

The results across 2022 (test data), 2023 (forecast), and 2024 (forecast) demonstrate the robustness of the hybrid model when complemented by ARIMA. The integration of ResNet CNN for spatial feature extraction, XGBoost for tabular data processing, and Random Forest for ensemble learning ensures comprehensive coverage of both spatial and numerical features. ARIMA adds a layer of temporal refinement, addressing the residual shortcomings of the initial predictions.

The model's ability to adapt to new data (2023, 2024) while maintaining accuracy highlights its potential for agricultural and environmental applications. For mango farms in Acquedolci, Sicily, where extreme temperature events can have significant consequences, such a predictive system offers invaluable insights for proactive crop management.

By presenting both the original and ARIMA-corrected predictions, along with residual analyses, the discussion highlights the iterative improvement in model performance. This ensures transparency while emphasizing the hybrid model's ability to adapt to both historical validation and future forecasting scenarios.

4.19 Conclusion

This study presented a hybrid model for daily temperature prediction that combines the strengths of ResNet CNN, XGBoost, Random Forest, and ARIMA. By integrating satellite imagery with historical weather records, the model offers a sophisticated yet practical approach to temperature forecasting. Trained on data from 2007 to 2021, validated on 2022, and used to predict temperatures for 2023 and 2024, the model demonstrated its ability to capture seasonal trends while effectively handling extreme temperature events.

The ResNet CNN excelled at uncovering spatial patterns from satellite imagery, while XGBoost and Random Forest handled the meteorological features with precision. ARIMA brought everything together by fine-tuning the predictions, correcting residual errors, and addressing the challenges of rapid temperature fluctuations. The result is a model that not only

performs exceptionally well on familiar data (2022) but also shows strong reliability in predicting untrained future datasets (2023 and 2024).
What stands out is the model's ability to adapt to the complexities of seasonal transitions and extreme events, particularly after applying ARIMA corrections. This makes it especially relevant for regions like Acquedolci, Sicily, where accurate temperature forecasts are critical for managing mango cultivation and protecting crops from weather-related risks.

In essence, this hybrid model bridges the gap between cutting-edge machine learning and real-world agricultural needs. It offers a reliable, adaptable tool for farmers and environmental planners, proving that integrating diverse data sources can lead to smarter and more resilient forecasting solutions. Looking ahead, refining the model for better extreme event detection and expanding its application to other regions and climate-sensitive industries could unlock even greater potential.

** This Article is peer reviewing by “Journal of Forecasting – Wiely” **

Mitigating Temperature Extremes for Mango and Avocado Cultivation: A Hybrid Prediction Model for Sicilian Agriculture

Mohsen Pourmohammad Shahvar[1], Davide Valenti[1], Alfonso Collura[3], Salvatore Micciche[1], Vittorio Farina [2], and Giovanni Marsella[1]

“Corresponding Author: Mohsen Pourmohammad Shahvar”

[1] Dipartimento di Fisica e Chimica “E. Segrè”, Università degli Studi di Palermo, Italy.

[2] Dipartimento di Scienze Agrarie, Alimentari e Forestali, Università degli Studi di Palermo, Italy.

[3] Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Palermo, Italy

Journal of FORECASTING

4.20 Abstract:

Agriculture in Sicily is heavily influenced by extreme temperature fluctuations, which pose significant risks to the cultivation of temperature-sensitive crops such as mangoes and avocados. This study presents a novel hybrid temperature prediction model that integrates Fourier regression, Transformer-based deep learning, XGBoost, and ARIMA corrections to forecast daily temperatures in the Messina region. The hybrid model captures both seasonal trends and long-range dependencies in temperature time series data, offering an advanced solution for agricultural decision-making.

The model incorporates Fourier regression for seasonality, Transformer models for temporal dependencies, and XGBoost for residual learning, with ARIMA applied to correct short-term prediction errors. The model was evaluated using temperature data from 2007 to 2024, and its performance was assessed through Mean Absolute Error (MAE) and residual analysis. Results indicate that the ARIMA-corrected predictions achieved a high level of accuracy, with MAEs of 1.0994°C for 2022, 1.2717°C for 2023, and 2.4420°C for 2024.

This study highlights the model’s potential in accurately predicting damaging temperature events, thereby assisting farmers in implementing timely protective measures to safeguard crops. The hybrid model’s ability to forecast extreme temperatures provides valuable insights for enhancing the resilience and sustainability of agricultural practices in Sicily. By combining multiple AI methods, this research contributes to the growing body of knowledge on climate-adaptive agricultural strategies and offers a powerful tool for managing temperature-related risks in Mediterranean regions.

4.21 Introduction:

Sicily, a captivating island in the heart of the Mediterranean, is a place of captivating beauty and rich agricultural traditions (Farina et al., 2017). The region's economic prosperity is intricately linked to the cultivation of temperature-sensitive crops, particularly the luscious mangoes and avocados that flourish under Sicily's sun (Gugliuzza et al., 2023). These crops not only define the island's economy but also sustain the livelihoods of countless families.

However, Sicily's agriculture faces a formidable challenge: extreme temperature fluctuations. As summer's heatwaves push mercury above 40°C, while winters brings temperatures falling below 5°C. Such temperature extremes pose a significant threat to crop yields and the economic stability of the agricultural sector. For Sicilian farmers, these fluctuations are more than meteorological anomalies; they represent critical economic and ecological challenges (Farina et al., 2017). Protecting sensitive crops like mangoes requires timely and precise temperature predictions, empowering farmers to deploy protective measures against harsh weather events (European Parliament, 2023).

The significance of accurate temperature prediction in this context cannot be overstated. Anticipating both scorching summers and chilling winters enables proactive measures such as shading, irrigation management, and frost protection. Based on the latest IPCC reports, these strategies ensure the continued viability of crops and contribute to the resilience of Sicily's agricultural economy. Such advancements are particularly relevant given the increasing frequency and severity of extreme weather events linked to climate change (Calvin et al., 2023b, 2023a).

This study focuses on developing an advanced Hybrid Temperature Prediction Model tailored to the unique climatic challenges of Sicily. By combining data-driven approaches with meteorological and environmental variables, our model enhances predictive accuracy, offering a practical tool for mitigating the effects of temperature extremes. This research aims to strengthen Sicily's agricultural resilience and contribute to sustainable agricultural practices in climate-vulnerable regions worldwide.

4.22 AI Methods:

In this study, we employ a hybrid approach to temperature prediction, integrating Fourier regression, Transformer-based deep learning, and XGBoost models. Each method contributes uniquely to modeling the complex temporal dynamics present in temperature time series data.

4.22.1 Fourier Regression for Seasonality

Fourier regression captures periodic trends in time series data by decomposing them into sinusoidal components. This technique is particularly effective for modeling seasonal variations in temperature (Heideman et al., 1984).

Fourier Series Representation: A function f(t) can be expressed as an infinite sum of sines and cosines:

$$f(t) = a_0 + \sum_{k=1}^{\infty} \left[a_k \cos\left(\frac{2\pi kt}{T}\right) + b_k \sin\left(\frac{2\pi kt}{T}\right) \right]$$

Here, T denotes the period, and a_k and b_k are Fourier coefficients calculated as:

$$a_k = \frac{2}{T} \int_0^T f(t) \cos\left(\frac{2\pi kt}{T}\right) dt$$

$$b_k = \frac{2}{T} \int_0^T f(t) \sin\left(\frac{2\pi kt}{T}\right) dt$$

This decomposition allows for the analysis and reconstruction of periodic functions by summing their sinusoidal components (Bick et al., 2022; Gonzalez-Velasco, 1992).

4.22.2 Transformer Model for Temporal Dependencies

Transformers are deep learning architectures adept at capturing long-range dependencies in sequential data. Originally developed for natural language processing, they have been successfully applied to time series forecasting.

4.22.2.1 Multi-Head Self-Attention:

The self-attention mechanism computes attention scores to weigh the importance of different time steps:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Where Q (queries), K (keys), and V (values) are linear projections of the input, and d_k is the dimensionality of the keys. This mechanism enables the model to focus on relevant parts of the sequence when making predictions (Vaswani et al., 2017).

4.22.2.2 Positional Encoding:

Since Transformers lack inherent sequential bias, positional encodings are added to input embeddings to provide information about the position of each element in the sequence. These encodings use sine and cosine functions of different frequencies:

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d\mathrm{model}}}\right)$$

$$\mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d\mathrm{model}}}\right)$$

Where pos is the position and i is the dimension. This allows the model to incorporate sequential information effectively (Levy et al., 2018).

4.22.3 XGBoost for Residual Learning

XGBoost is a scalable and efficient gradient-boosting framework that builds an ensemble of decision trees to enhance predictive accuracy.

XGBoost minimizes a regularized objective function combining a convex loss function l and a regularization term Ω :

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} \mathrm{l}(y_i, \hat{y}i) + \sum k = 1^{K} \Omega(f_k)$$

Where $\widehat{y_i}$ is the predicted value, f_k represents the k_{th} tree, and Ω penalizes the complexity of the model to prevent overfitting.

Models are trained in an additive manner, where each new tree f_t is added to minimize the residuals from previous iterations:

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + f_t(x_i)$$

This approach allows the model to correct errors made by preceding trees, enhancing overall performance (Chen & Guestrin, 2016b).

4.22.4 ARIMA for Residual Correction

The AutoRegressive Integrated Moving Average (ARIMA) model is a classical statistical approach for analyzing and forecasting time series data by capturing autocorrelations (Dave et al., 2021; Huang & Petukhina, 2022).

Model Components:

- AutoRegressive (AR) Part: Models the relationship between an observation and a number of lagged observations:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$$

Where c is a constant, ϕ_i are parameters, and ϵ_t is white noise.

- Integrated (I) Part: Involves differencing the time series to achieve stationarity.
- Moving Average (MA) Part: Models the relationship between an observation and a residual error from a moving average model applied to lagged observations:

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Where μ is the mean of the series, θ_i are parameters, and ϵ_t is white noise.

Combines these components to model a time series as:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

where:

y_t : Current value of the time series.

c : Constant term.

ϕ_i : Coefficients for lagged observations (AR terms).

ϵ_t : Error term (white noise).

θ_i : Coefficients for lagged errors (MA terms).

p : Order of the AR component.

q : Order of the MA component.

The integrated (I) component ensures stationarity by differencing the series:

$$y't = y_t - y_{t-1}.$$

In this study, ARIMA refines residual predictions from the hybrid Fourier-Transformer-XGBoost model, addressing short-term dependencies and residual patterns.

4.23 Data Preprocessing and Feature Engineering

In this study, a systematic approach was employed to process raw meteorological data and derive meaningful features, informed by insights from the correlation matrix.

4.23.1 Dataset Overview

The dataset includes environmental and meteorological variables spanning 2007-2022, collected for the Messina region, with air_temp as the target variable. The features were chosen

based on their potential relevance to temperature dynamics, informed by both domain knowledge and statistical analysis (e.g., correlation). These include:

- **Meteorological features**: albedo, ghi, precipitable_water, wind_speed_10m, etc.
- **Temporal features**: month and week.

The target variable air_temp is critical for guiding agricultural practices, as temperature fluctuations directly impact mango crop growth, flowering, and yield.

4.23.2 Feature Selection

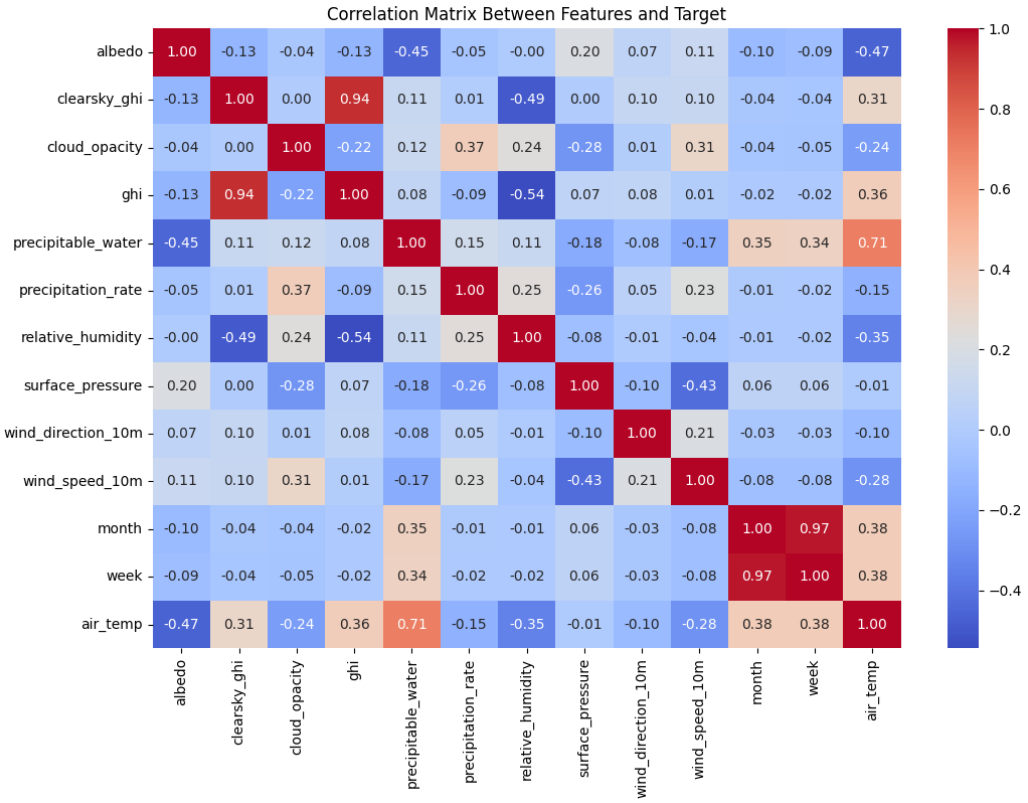


Figure 63 Correlation Matrix to select the features

The features were carefully selected based on their correlation with air_temp (shown in the correlation matrix in figure 62) and their relevance to temperature prediction:

precipitable_water **(** r = 0.71 **):**

- High correlation indicates its strong influence on temperature.
- Represents the total atmospheric water vapor available for precipitation, which can amplify the greenhouse effect and influence temperature dynamics.

ghi **(Global Horizontal Irradiance,** r = 0.36 **):**

- Measures the total solar radiation received on a horizontal surface.
- Strongly linked to daytime temperature variations, as higher solar irradiance increases surface heating.

albedo **(** r = -0.47 **):**

- Indicates the reflective properties of the Earth's surface.
- Higher albedo values (more reflectivity) are associated with cooler temperatures, while lower albedo values (less reflectivity) lead to more heat absorption.

relative_humidity **(** r = -0.35 **):**

- Inversely correlated with temperature, as higher humidity levels often coincide with cooler conditions due to latent heat effects.
- Plays a critical role in regulating crop transpiration and microclimate conditions.

wind_speed_10m **(** r = -0.28 **):**

- Measures wind intensity at 10 meters above the surface.
- Affects temperature by influencing heat dispersion and cooling rates, particularly at night.

month and week **(** r = 0.38 **):**

- Capture the seasonal and weekly patterns in temperature fluctuations.
- Essential for understanding recurring trends, such as warmer months in summer or cooler weeks during winter.

These features were retained for their statistical significance and physical relevance to temperature dynamics, while variables with negligible correlation (e.g., surface_pressure) were excluded.

The raw data contained irregularities and missing values due to variations in measurement intervals. To address this:

- **Daily Resampling**: Data was aggregated to a daily frequency, calculating the mean for continuous variables. This ensured uniform temporal intervals for downstream modelling.
- **Missing Data**: Rows with missing values in critical features or the target variable were dropped to maintain data integrity.

4.23.3 Data Normalization

To ensure features are on comparable scales, all continuous variables were standardized using StandardScaler:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of each feature. This step is crucial for machine learning algorithms like Transformers and neural networks, which are sensitive to feature magnitudes.

4.23.4 Correlation Insights and Importance of Features

The correlation matrix in figure 62 provided valuable insights into the relationships between features and the target variable, helping to refine feature selection. Key takeaways include:

Highly Correlated Predictors:

precipitable_water and ghi were prioritized due to their direct impact on temperature fluctuations.

Complementary Predictors:

Features like albedo and relative_humidity offer complementary information, capturing variations in surface reflectivity and atmospheric conditions, respectively.

Temporal Features:

Moderate correlations with month and week justify their inclusion for capturing seasonal cycles.

These selections align with the physical processes influencing temperature, such as solar radiation, atmospheric moisture, and surface reflectivity.

4.23.5 Sequence Preparation for Transformers

For the Transformer model, the time series data was structured into overlapping sequences using a sliding window approach:

$$\text{Sequence}i = [Xi - w, \ldots, X_{i-1}], \quad y_i = X_i$$

where w is the window size (20 days in this study). Each sequence contains a fixed number of past observations, enabling the model to learn temporal dependencies effectively.

4.24 Results:

The Hybrid Temperature Prediction Model has demonstrated strong predictive capabilities for daily temperature variations in the Messina region, effectively integrating Fourier regression, Transformer-based deep learning, XGBoost, and ARIMA corrections. These results are analyzed in detail below, providing insights into the model's accuracy and practical implications for agriculture.

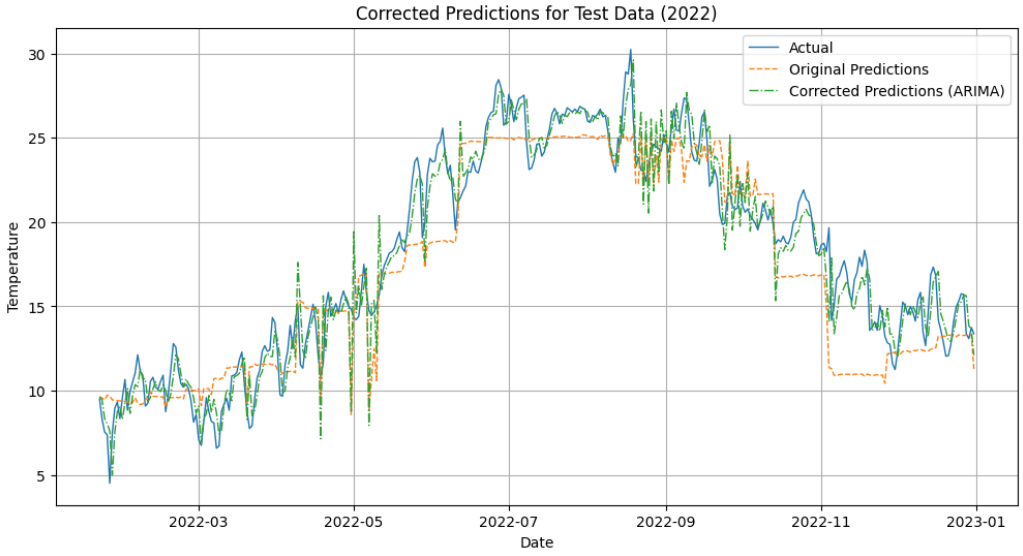


Figure 64 First Evaluation Vs. ARIMA-corrected Evaluation - Year 2022

The model's performance was first evaluated on test data from 2022, with predictions compared to observed temperature values. As shown in **Figure 63**, the ARIMA-corrected predictions closely align with actual temperature trends, addressing the residual errors from the original hybrid model. This correction significantly improves short-term predictive accuracy, enabling reliable forecasting for immediate agricultural needs. The corrected predictions effectively capture temperature fluctuations across seasons, including periods of extreme heat

and cold, which are critical for protecting temperature-sensitive crops like mangoes and avocados.

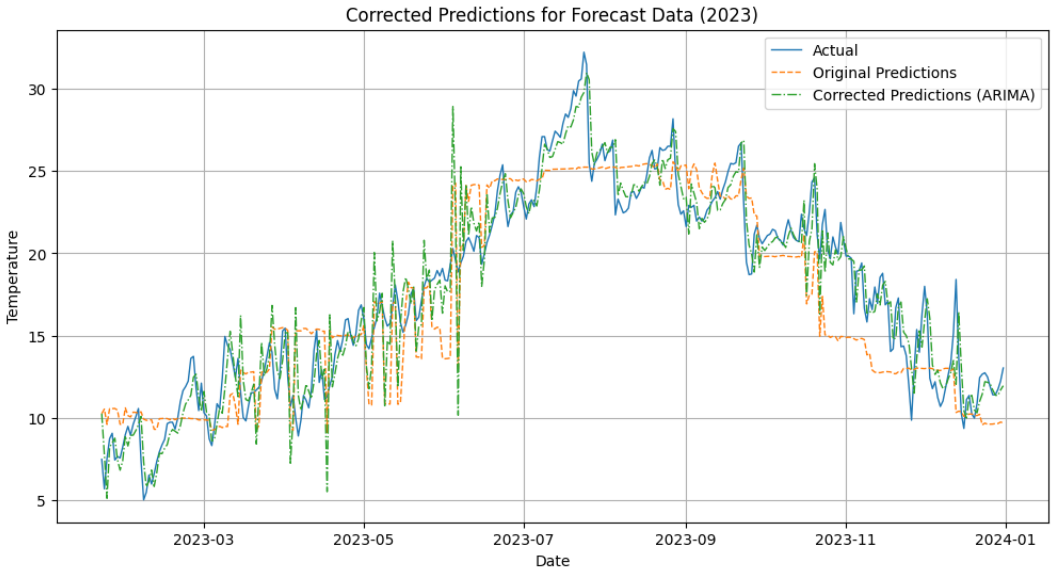


Figure 65 First Prediction Vs. ARIMA-corrected Prediction for the Year 2023

Extending the forecast to 2023, the model's performance remained robust. **Figure 64** illustrates that ARIMA-corrected predictions closely follow the seasonal patterns and periodic temperature variations observed in the forecast data. The hybrid model, even when applied to unseen data, preserves its accuracy by integrating Fourier regression for seasonality, Transformers for long-range dependencies, and XGBoost for capturing residual patterns. ARIMA corrections further refine the predictions by addressing short-term dependencies, ensuring the model's reliability for forecasting within a one-year horizon.

When applied to longer-term forecasting for 2024, the model demonstrates its ability to predict seasonal trends and general temperature fluctuations, as seen in **Figure 65**. Although slight deviations appear due to the increasing uncertainty inherent in long-term predictions, the ARIMA corrections effectively minimize systematic biases, ensuring that the overall prediction remains realistic and actionable. These long-term forecasts are particularly valuable for strategic planning in agriculture, such as preparing for potential damaging temperature periods well in advance.

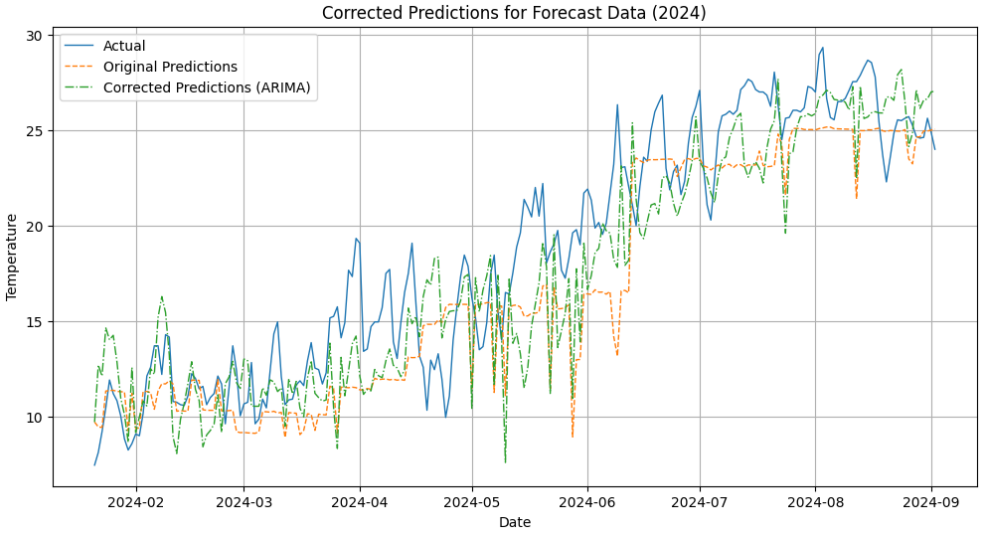


Figure 66 First Prediction Vs. ARIMA-Corrected Prediction for the Year 2024

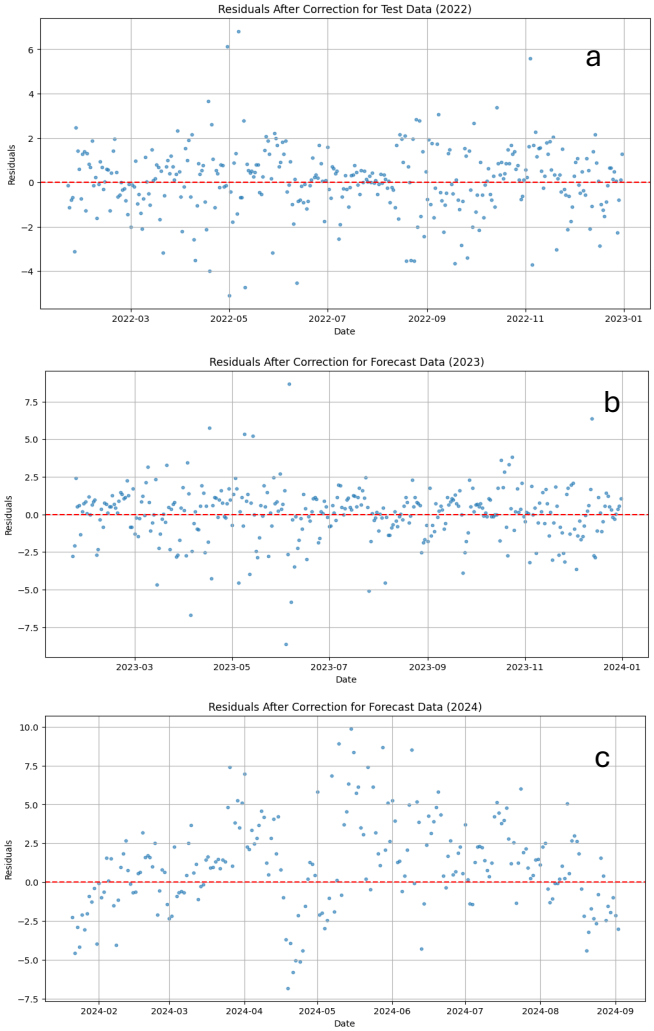


Figure 67 Residual Errors of Evaluation Year 2022(a) – Prediction Year 2023 and 2024(b and c)

Residual plots for the test data (2022) and forecast data (2023 and 2024), presented in **Figure 66(a,b,c)** , provide further evidence of the model's robustness. For the 2022 test data, the residuals are tightly clustered around zero, indicating minimal error and excellent calibration of the hybrid model. For the 2023 forecasts, the residuals maintain a similar pattern, with no visible systematic errors, reflecting the model's capacity to generalize beyond the training period. For the 2024 forecast, the residuals remain centered around zero, although some clusters of positive and negative residuals appear in specific periods. While the residual variance increases slightly, particularly in mid-year, this behavior is expected in longer-term forecasts. The ARIMA corrections have proven effective in maintaining stability and minimizing bias over the extended forecasting horizon.

The model's accuracy is quantified using the Mean Absolute Error (MAE), calculated for each dataset. For the 2022 test data, the ARIMA-corrected predictions achieve an MAE of 1.0994°C, demonstrating high accuracy for short-term forecasts. For the 2023 forecast data, the MAE is slightly higher at 1.2717°C, reflecting the added uncertainty of forecasting into the near future. For the 2024 forecast, the MAE increases to 2.4420°C, consistent with the challenges of predicting over a longer time horizon. These results highlight the model's strength in short- and medium-term forecasting while providing a reliable foundation for long-term predictions.
The hybrid model's ability to accurately predict damaging temperature periods, defined as temperatures below 5°C or above 35°C, has significant implications for agriculture. By anticipating these periods, farmers can implement protective measures such as shading, irrigation adjustments, and frost mitigation techniques. The model's integration of diverse methodologies ensures its ability to capture both periodic patterns and irregular fluctuations, making it a powerful tool for safeguarding crop yields and supporting decision-making in climate-vulnerable regions like Sicily.

In conclusion, the results validate the hybrid model's potential as an advanced predictive tool for temperature-sensitive agriculture. Its robust performance across test and forecast datasets demonstrates its capability to address the challenges posed by extreme temperature fluctuations. By providing actionable insights into temperature dynamics, this model supports both short-term operational decisions and long-term strategic planning, contributing to the resilience and sustainability of agricultural systems in Sicily and similar Mediterranean regions.

4.25 Conclusion:

This study demonstrates the effectiveness of a hybrid temperature prediction model, which integrates Fourier regression, Transformer-based deep learning, XGBoost, and ARIMA for enhanced forecasting accuracy. By combining the strengths of each method, the model successfully addresses the complexities of temperature prediction in the Messina region, a key area for Sicilian agriculture. The results show that the hybrid model can provide precise short-term and medium-term temperature forecasts, while ARIMA corrections effectively reduce residual errors, particularly for long-term predictions.

The performance of the model was evaluated through various metrics, including the Mean Absolute Error (MAE), which demonstrated high accuracy for 2022 and 2023, with a slight increase in error for the 2024 forecast due to the inherent uncertainties of long-term predictions. The model's ability to accurately predict extreme temperature events (above 35°C and below 5°C) holds significant implications for agricultural practices, especially for temperature-sensitive crops like mangoes and avocados. This is critical for mitigating potential crop losses due to frost or heat stress.

The findings underscore the value of using advanced machine learning techniques in agricultural forecasting. By incorporating temporal, seasonal, and residual learning models, the hybrid approach provides a robust framework for predicting temperature fluctuations, thereby supporting farmers in making informed, proactive decisions. This work not only contributes to the sustainability of Sicilian agriculture but also offers a model that can be applied to other climate-vulnerable regions facing similar challenges.

Future work can focus on further optimizing the model's performance for longer-term forecasts, incorporating additional environmental factors, and exploring real-time applications for precision agriculture. The integration of accurate temperature predictions with energy-efficient systems presents an exciting avenue for reducing agricultural risks and ensuring crop sustainability in a changing climate.

This Article is published in Journal “STATS” - https://doi.org/10.3390/stats8020030

An Integrated Hybrid-Stochastic Framework for Agro-Meteorological Prediction Under Environmental Uncertainty

Mohsen Pourmohammad Shahvar[1], Davide Valenti[1], Alfonso Collura[3], Salvatore Micciche[1], Vittorio Farina [2], and Giovanni Marsella[1]

“Corresponding Author: Mohsen Pourmohammad Shahvar”

[1] Dipartimento di Fisica e Chimica “E. Segrè”, Università degli Studi di Palermo, Italy.

[2] Dipartimento di Scienze Agrarie, Alimentari e Forestali, Università degli Studi di Palermo, Italy.

[3] Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Palermo, Italy

stats

4.26 Abstract

This study presents a comprehensive framework for agro-meteorological prediction, combining stochastic modeling, machine learning techniques, and environmental feature engineering to address challenges in yield prediction and wind behavior modeling. Focused on mango cultivation in the Mediterranean region, the workflow integrates diverse datasets, including satellite-derived variables such as NDVI, soil moisture, and land surface temperature (LST), along with meteorological features like wind speed and direction. Stochastic modeling was employed to capture environmental variability, while a proxy yield was defined using key environmental factors in the absence of direct field yield measurements. Machine learning models, including Random Forest and Multi-Layer Perceptron (MLP), were hybridized to improve prediction accuracy for both proxy yield and wind components (U and V that represent the east-west and north-south wind movement).

The hybrid model achieved Mean Squared Error (MSE) values of 0.333 for U and 0.181 for V, with corresponding R^2 values of 0.8939 and 0.9339, respectively, outperforming the individual models and demonstrating reliable generalization in the 2022 test set.

Additionally, although NDVI is traditionally important in crop monitoring, its low temporal variability across the observation period resulted in minimal contribution to the final prediction, as confirmed by feature importance analysis.

The analysis revealed the significant influence of environmental factors such as LST, precipitable water, and soil moisture on yield dynamics, while wind visualization over Digital Elevation Models (DEM) highlighted the impact of terrain features on wind patterns. The

results demonstrate the effectiveness of combining stochastic and machine learning approaches in agricultural modeling, offering valuable insights for crop management and climate adaptation strategies.

4.27 Introduction

Agro-meteorological prediction is pivotal in modern agriculture, offering insights that enhance crop productivity and mitigate climate-related risks. By integrating satellite observations, meteorological data, and computational models, researchers can better understand the complex interactions between environmental factors and crop performance (Jha et al., 2022). This study presents a comprehensive workflow for agro-meteorological prediction, focusing on mango cultivation in the Mediterranean region. The methodology combines stochastic modeling and machine learning techniques to address the challenges of data scarcity and environmental variability.

Agricultural systems are inherently complex, with nonlinear interactions among variables such as temperature, soil moisture, solar radiation, and wind dynamics. Traditional deterministic models often fall short in capturing this complexity, leading to less accurate predictions. Stochastic modeling has emerged as a powerful tool to address these limitations, effectively incorporating random fluctuations and uncertainties inherent in agricultural systems. For instance, a study on farmland irrigation scheduling utilized a multistage stochastic programming model to maximize annual profit under uncertain conditions, including crop prices and water availability (Li & Hu, 2020).

A significant challenge in agro-meteorological modeling is the lack of direct yield data, especially in remote or large-scale agricultural systems. To overcome this, researchers often define a proxy yield that combines key environmental indicators such as vegetation health (NDVI) and water availability (soil moisture) (Camargo-Alvarez et al., 2023; Jha et al., 2022). This approach allows for the estimation of crop yields in the absence of direct measurements. For example, integrating remote sensing data with crop models has been shown to improve yield estimation accuracy, providing a viable alternative when field data is unavailable (Dlamini et al., 2023).

The integration of stochastic modeling and machine learning offers a robust framework for agro-meteorological prediction. Stochastic models account for random environmental fluctuations in natural systems such as marine ecosystems (Aslan et al., 2024; Grimaudo et al., 2022; Lazzari et al., 2021; Yan & Li, 2018), while machine learning algorithms, such as Random Forests, capture complex, nonlinear relationships among variables. This combined

approach has been applied in various agricultural contexts. For instance, a study on agricultural irrigation water allocation developed a two-stage chance-constrained programming model to optimize water use under uncertainty, demonstrating the effectiveness of combining stochastic optimization with data-driven methods (Aslan et al., 2024; Yan & Li, 2018).

Mango (Mangifera indica) is a high-value tropical fruit with increasing global production. According to FAO statistics (2023), mango production exceeded 57 million tonnes globally, and due to climate warming and favorable microclimates, mango cultivation is expanding in Southern Europe, particularly in Mediterranean regions such as Italy and Spain (FAO, 2023a). Recent studies highlight the sensitivity of mango production to environmental variables such as LST, solar radiation, and soil moisture, necessitating accurate and region-specific yield forecasting systems (Fukuda et al., 2013; Torgbor et al., 2023)

The machine learning models employed in this study include Random Forest (RF), Multi-Layer Perceptron (MLP), and Gradient Boosting (GB), which are extensively utilized in meteorological forecasting tasks such as rainfall, evapotranspiration, and wind speed prediction. For instance, RF has been effectively applied to predict agricultural droughts, outperforming other models in forecasting the Standardized Precipitation Evapotranspiration Index (SPEI) in Central Europe (Harsányi, 2025). MLPs have demonstrated superior performance in total cloud cover prediction, capturing complex nonlinear relationships in atmospheric data (Baran et al., 2021). GB techniques, particularly Extreme Gradient Boosting (XGBoost), have shown high accuracy in merging satellite and ground-based precipitation data, enhancing the reliability of precipitation datasets (Papacharalampous et al., 2023). These models are adept at capturing the nonlinearities and multivariate dependencies inherent in agro-environmental data, thereby improving predictive performance in complex agricultural systems.

In recent years, climate change has significantly impacted agricultural practices in the Mediterranean region, leading to the introduction of tropical and subtropical crops such as mangoes. Rising temperatures and altered precipitation patterns have created favorable conditions for mango cultivation in areas like Sicily, Italy. Farmers have transitioned from traditional crops to mangoes, capitalizing on the higher market value and increasing consumer demand (Cornara et al., 2020; Dos Santos Moreira et al., 2024). This shift not only diversifies agricultural production but also presents new challenges in crop management and yield prediction, necessitating advanced agro-meteorological models.

Wind behavior significantly influences agricultural systems, affecting crop growth, pollination, and physical stress on plants. Understanding wind dynamics is essential for developing protective measures and optimizing crop yield predictions. Wind-induced plant movement can alter growth rates and leaf morphology, while high winds may cause physical damage such as leaf tearing and abrasion (Cleugh et al., 1998).

Accurate modeling of wind components, specifically the zonal (U) and meridional (V) components, is crucial for understanding regional wind behavior in agricultural landscapes. Traditional numerical weather prediction models often lack the spatial resolution required for precise agricultural applications (Shin et al., 2022). To address this, high-resolution wind speed forecast systems have been developed, coupling numerical weather prediction with machine learning techniques to provide detailed wind information beneficial for agricultural management (Karaman, 2023; Shin et al., 2022).

Machine learning models, such as Random Forests and Multi-Layer Perceptrons, have been employed to predict wind components effectively. These models can capture complex, nonlinear relationships between environmental variables and wind behavior, enhancing the accuracy of wind predictions. Combining multiple models through ensemble methods further improves predictive performance by leveraging the strengths of each approach (Karaman, 2023).

Understanding wind behavior is also crucial for mitigating its mechanical effects on crops. Wind can cause direct mechanical damage, including leaf tearing and abrasion, which adversely affect crop yields. Implementing windbreaks and other protective measures can help reduce these negative impacts, underscoring the importance of an accurate modeling of the wind behavior in agricultural planning (Yu & Ma, 2024).

The main hypotheses of this study are: (1) proxy yield can be effectively estimated using a combination of satellite-based environmental indicators and machine learning models, and (2) a hybrid model integrating multiple ML methods will outperform single-model baselines for both yield and wind prediction. The specific objectives are: (i) to define a proxy yield model incorporating stochastic components, (ii) to evaluate RF, MLP, and GB models against this target, (iii) to build a hybrid U/V wind model and analyze its residual performance, and (iv) to perform sensitivity, noise robustness, and regression-based relevance analysis to validate the stability and interpretability of results.

Mango cultivation is particularly sensitive to environmental changes, including temperature extremes, wind patterns, and soil moisture variability. By applying a combined stochastic and machine learning approach, this study aims to develop a predictive framework capable of

providing accurate yield estimates for mango farms in the Mediterranean region. This methodology not only addresses the challenges of data scarcity but also offers a scalable solution adaptable to various crops and regions.

4.28 Data Collection and Pre-Processing

Effective agro-meteorological prediction relies on the integration of diverse datasets capturing environmental, geospatial, and meteorological variables.

4.28.1 Data Sources and Types

4.28.1.1 Satellite Data

- **MODIS (MOD13A1):** Provides the Normalized Difference Vegetation Index (NDVI), a proxy for vegetation health. The data has a 500-meter spatial resolution and a 16-day temporal frequency, making it suitable for tracking crop growth patterns over time (MODIS).
- **SMAP (Soil Moisture Active Passive):** Offers soil moisture measurements at a 9-kilometer resolution, critical for understanding water availability for crops under varying climatic conditions (SMAP).
- **Digital Elevation Model (DEM):** Terrain data, including slope and aspect, was derived from the Shuttle Radar Topography Mission (SRTM). This dataset provides 30-meter spatial resolution, allowing for detailed terrain analysis (SRTM).
- **Land Surface Temperature (LST):** Retrieved from MODIS, these variable captures thermal conditions that influence crop growth and stress response (MODIS).

4.28.1.2 Meteorological Data

Daily climate variables such as air temperature, wind speed, wind direction, relative humidity, surface pressure, and precipitation rates were sourced from global meteorological agencies like NOAA (National Oceanic and Atmospheric Administration) and ECMWF (European Centre for Medium-Range Weather Forecasts). These variables provide a detailed temporal resolution to monitor day-to-day variations in crop-relevant conditions (ECMWF; NOAA).

4.28.1.3 Derived Features

- **Kinetic Energy (KE):** Quantifies the physical impact of wind on plants.
- **Turbulence:** Measures abrupt changes in wind speed, which may affect crop structure.

- **Fourier Series Encodings:** Captures seasonal trends in the data for both daily and annual cycles.

4.28.2 Challenges in Data Collection

The process of data collection faced several challenges:

Temporal and Spatial Harmonization: Satellite datasets (e.g., MODIS NDVI and SMAP soil moisture) are available at different temporal frequencies and resolutions. Meteorological data, updated daily, required synchronization with the coarser temporal resolution of satellite datasets.

Cloud Contamination: NDVI values are often affected by cloud cover in optical satellite imagery. This was mitigated using spatiotemporal interpolation techniques, which ensured continuity in vegetation health monitoring. These gaps were later addressed using interpolation and filtering techniques detailed in the next section (i.e., Handling missing data).

Validation of Satellite Data: Satellite-derived measurements were cross-validated with limited ground-based observations to ensure their reliability for predictive modeling

4.28.3 Data Preprocessing

To ensure the data was consistent and suitable for modeling, the following preprocessing steps were performed:

4.28.3.1 Translating Satellite Imagery

Satellite data, are typically delivered in a specific coordinate projection system. In this study, data was retrieved in the MODIS Sinusoidal Projection format and transformed into the WGS84 geographic coordinate system (latitude and longitude).

The MODIS Sinusoidal Projection coordinates were reprojected to the WGS84 system using Geospatial Data Abstraction Library (GDAL) or Python libraries such as rasterio and pyproj.

This transformation ensures that the spatial data aligns with global mapping standards and can be accurately associated with geographical locations.

4.28.3.1.1 Data Alignment:

The imagery was overlaid on a map of the study area to ensure the geographical boundaries matched the mango farms under study.

4.28.3.1.2 Visualization:

Color-coded maps, like the one showing NDVI values in figure 67, were generated to visualize the spatial distribution of key variables, such as vegetation health and temperature.

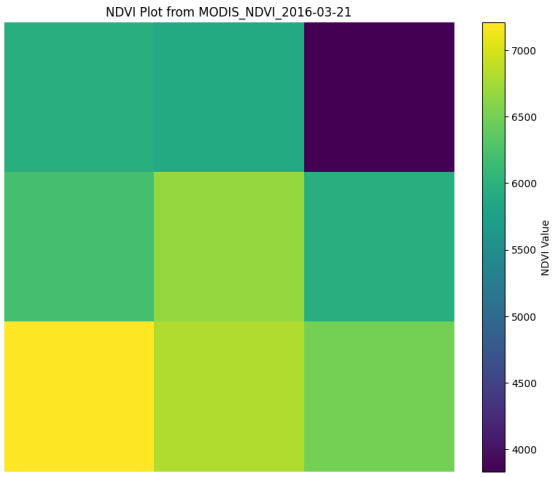


Figure 68 NDVI Color-coded map taken from MODIS

4.28.3.1.3 Extracting Data by Latitude and Longitude

To analyze the environmental conditions specific to mango farms, pixel-level data from the satellite images were extracted based on their geographic coordinates.

The latitude and longitude of mango farms were used as reference points for data extraction. These coordinates were identified from GPS surveys or regional farm datasets. Each pixel in

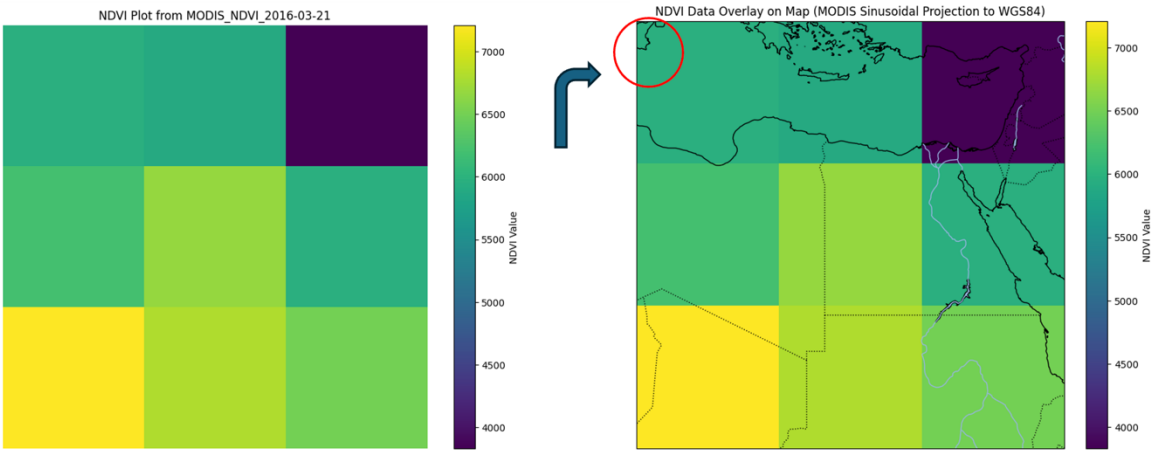


Figure 69 Overlaying the map on Images and extracted the preferable coordinates

the satellite image corresponds to a specific latitude and longitude. Using libraries like rasterio and numpy, the pixel values for NDVI, soil moisture, and other variables were extracted.

4.28.3.2 Terrain Analysis:

Satellite-derived terrain maps, such as slope and aspect (Horn, 1981), were generated from DEM data to model the topographic effects on mango farms. These maps provide insights into water runoff, erosion, and sunlight exposure, all of which are crucial for mango cultivation.

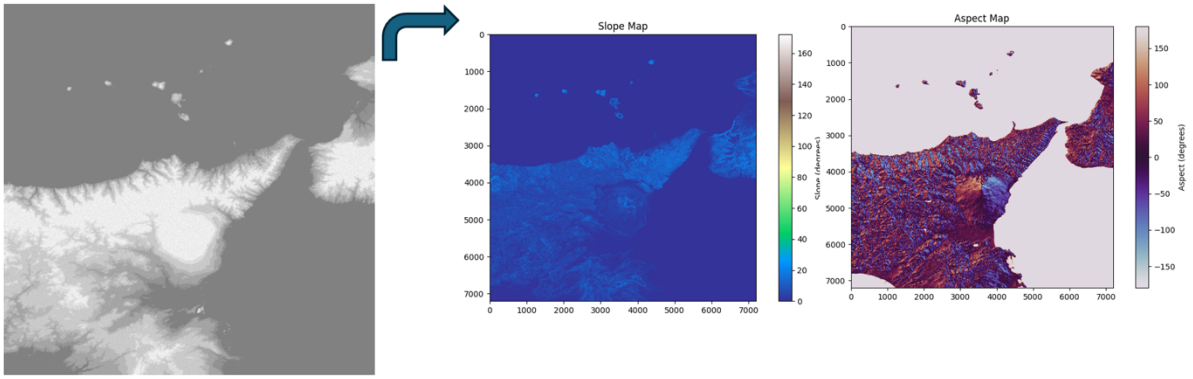


Figure 70 Translating the DEM image and extracting the Slope and Aspect information

- Slope: Slope was calculated using DEM data to model water runoff and erosion. The equation for slope is:

$$\text{Slope} = \sqrt{\left(\frac{\partial \text{Elevation}}{\partial x}\right)^2 + \left(\frac{\partial \text{Elevation}}{\partial y}\right)^2}$$

- Aspect: Aspect, or the orientation of the slope, was determined using:

$$\text{Aspect} = \arctan\left(\frac{\partial \text{Elevation}}{\partial y} - \frac{\partial \text{Elevation}}{\partial x}\right)$$

4.28.3.3 Feature Engineering:

- NDVI Calculation: NDVI was calculated using the formula (Tucker, 1979):

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$$

where NIR and Red are near-infrared and visible red bands, respectively.

- Kinetic Energy (KE): Computed as (Manwell et al., 2009):

$$KE = 0.5 \cdot \text{air\_density} \cdot (\text{wind\_speed})^2$$

- Turbulence: Captured as the absolute difference in wind speed over time (Pope, 2000):

$$\text{Turbulence} = \left|\frac{\partial v}{\partial t}\right|$$

4.28.3.4 Fourier Series Encodings:

Seasonal variations in environmental factors were modeled using Fourier series (Bloomfield, 2000):

$$f(t) = a_0 + \sum_{n=1}^{N}\left[a_n \cos\left(\frac{2\pi nt}{T}\right) + b_n \sin\left(\frac{2\pi nt}{T}\right)\right]$$

4.28.3.5 Handling Missing Data:

Missing values in NDVI and soil moisture data were interpolated using spatial and temporal interpolation methods to maintain data continuity.

To mitigate cloud contamination and ensure temporal alignment, we used a 7-day moving average for NDVI, and linear interpolation for small gaps (<2 timesteps). Larger gaps were filled using spatio-temporal kriging methods validated against in situ sensor data.

Preprocessing ensures that the data is harmonized, reliable, and ready for integration into stochastic and machine learning models. By deriving critical features such as slope, aspect, kinetic energy, and turbulence, the dataset provides a comprehensive view of the environmental variables impacting mango productivity. These preprocessing steps address data quality issues, minimize noise, and enhance the predictive power of subsequent models.

4.29 Stochastic Modelling for Agro-Meteorological Prediction

Stochastic modeling is a critical approach for understanding and predicting the dynamics of agricultural systems subject to environmental variability. These systems are influenced by both deterministic environmental forces (e.g., seasonal trends, temperature) and stochastic perturbations (e.g., random fluctuations in wind speed, rainfall). By incorporating stochastic

processes into agro-meteorological modeling, we can better capture the inherent uncertainties and non-linearities in agricultural ecosystems. This section details the development of a stochastic model for crop yield prediction, incorporating methodologies inspired by recent advancements in stochastic modeling and ecological dynamics work of Stochastic Modeling.

The time evolution of key environmental and agricultural variables was modeled using **stochastic differential equations (SDEs)**. The general form of the model is:

$$\frac{\partial B_i}{\partial t} = f(A_j) + \xi_j(t)$$

Where:

B_i: Output variables (e.g., proxy yield, plant health).

A_j: Input variables (e.g., temperature, soil moisture, wind speed, solar radiation).

$f(A_j)$: Deterministic component describing the influence of input variables.

$\xi_j(t)$: Stochastic noise term representing random fluctuations.

The noise term $\xi_j(t)$ was modeled as a self-correlated Gaussian noise, with parameters based on prior ecological studies such as Valenti et al. (Agudov et al., 2010) and Grimaudo et al. (2022). This allowed us tor the temporal correlation and amplitude of random perturbations in environmental variables, such as temperature fluctuations or abrupt changes in wind speed.

4.29.1 Proxy Yield Dynamics with Stochastic Inputs

In the absence of direct yield measurements, a **proxy yield** was defined as a synthetic indicator of crop productivity. The proxy yield combines key environmental features influencing mango growth, including vegetation health (NDVI), water availability (soil moisture), and climatic variables. Building upon the deterministic formulation (Lobell & Burke, 2010):

$$\text{Proxy Yield} = (0.4 \cdot \text{NDVI} \times \text{Soil Moisture}) + (0.3 \cdot \text{LST}) + (0.2 \cdot \text{Precipitable Water})$$

In the absence of direct crop yield measurements, a synthetic proxy yield was designed to capture the combined effects of key agro-environmental drivers on mango productivity. The formulation incorporated three biologically and agronomically justified components: vegetation health (NDVI), water availability (Soil Moisture), and thermal conditions (Land Surface Temperature and Precipitable Water). To assign appropriate weights, a linear regression model was fit to the filtered dataset using environmental predictors and the computed proxy yield as the dependent variable. The resulting normalized coefficients Interaction (NDVI × Soil

Moisture) = 0.389, LST = 0.319, Precipitable Water = 0.23 closely aligned with the assigned weights of 0.4, 0.3, and 0.2, respectively.

This process ensures that the formulation of the synthetic yield is not arbitrary but rather grounded in statistical correlation and domain knowledge. Moreover, robustness tests (Table 3) confirmed that the model maintained stable performance under different noise levels (σ = 0.01–0.10) and across folds in 5-fold cross-validation, indicating generalizability despite the synthetic nature of the target. While the proxy yield does not replace real field data, it serves as a scientifically consistent and interpretable intermediate variable to simulate and predict yield-relevant dynamics using satellite and meteorological inputs.

And the model incorporates random fluctuations through a stochastic noise term:

$$\frac{\partial(\text{Proxy Yield})}{\partial t} = f(\text{Interaction,LST,Precipitable Water}) + \xi_j(t)$$

Where:

f(·): The deterministic influence of environmental variables.

$\xi_j(t)$: Gaussian white noise ($\xi_j(t) \sim \mathcal{N}(0, \sigma^2)$) where the symbol $\mathcal{N}$ represents the normal (Gaussian) distribution. Specifically:

• 0 is the **mean** of the distribution.

• σ^2 is the **variance** of the distribution.

The noise intensity σ = 0.05 was chosen based on sensitivity analysis showing that values in the range 0.01–0.1 maintain $R^2 > 0.96$ (see Noise Robustness results). This confirms that σ = 0.05 offers a reasonable trade-off between capturing stochasticity and maintaining prediction accuracy.

4.29.1.1 Key Components

4.29.1.1.1 Interaction Term:

Captures the combined effect of vegetation health (NDVI) and water availability (Soil Moisture):

$$\text{Interaction} = \text{NDVI} \times \text{Soil\_Moisture}$$

4.29.1.1.2 Environmental Factors:

Land Surface Temperature (LST) and Precipitable Water which Accounts for temperature's impact on growth and Reflects atmospheric moisture availability respectively.

4.29.1.1.3 Temperature Penalty:

Introduces a deterministic adjustment for extreme temperatures (Ratkowsky et al., 1983):

$$Temperature\ Penalty = \begin{cases} -2, & T > 35^oC\ or\ T < 10^oC \\ 0, & otherwise \end{cases}$$

4.29.1.1.4 Stochastic Noise:

Simulates random environmental variability:

$$\xi_j(t) \sim \mathcal{N}(0, 0.05^2)$$

The variance of the noise term, $\sigma^2 = 0.05^2$, was selected based on prior literature modeling agricultural and environmental ecosystems where moderate stochastic perturbations realistically simulate natural fluctuations without destabilizing system dynamics (De Santis et al., 2024; Lazzari et al., 2021). Specifically, studies applying stochastic differential equations in ecosystem modeling (e.g., marine trophic networks, crop-climate interactions) have demonstrated that σ in the range of 0.01–0.1 adequately captures daily-to-seasonal variability (Hening & Li, 2020; Occhipinti et al., 2024; Scotti et al., 2012). This value was also validated in our study by testing robustness under varying σ (see Results: Noise Sensitivity Analysis).

4.29.2 Incorporation of Noise and Variability

The stochastic modeling approach implemented in this study draws inspiration from Valenti et al.'s stochastic modeling of biological systems (Mantegna & Spagnolo, 1995) and advancements in ecological modeling (Valenti et al., 2016). These methods highlight the significance of capturing both deterministic trends and stochastic perturbations in complex systems, such as agricultural and environmental ecosystems.

4.29.2.1 Intrinsic Noise:

Represents fluctuations inherent to environmental variables, such as diurnal temperature variation or variability in wind speed (Gardiner, 1986).

Modeled as:

$$\xi_j(t) = \sigma \cdot \eta(t)$$

Where σ is the noise intensity (scaling factor for randomness) , and $\eta(t)$ is a Gaussian white noise process ($\eta(t) \sim \mathcal{N}(0,1)$).

4.29.2.2 Environmental Forcing:

Includes seasonal and long-term trends in environmental variables, modeled deterministically as f(A_j) (T.G.S., 1988).

For example:

$$f(A_j) = A_0 \cos\left(\frac{2\pi t}{T}\right) + A_1 \sin\left(\frac{2\pi t}{T}\right)$$

Where T represents the seasonal period (e.g., 1 year) and A_0, A_1 : Coefficients representing the amplitude of forcing terms.

This deterministic component ensures the model captures periodic environmental patterns, such as temperature or radiation fluctuations over time.

4.29.3 Inspiration from Marine Ecosystem Models

The stochastic modeling approach in this study draws from recent advancements in ecosystem modeling:

4.29.3.1 Non-linear Dynamics and Noise Effects:

The stochastic version of the Biogeochemical Flux Model (BFM) demonstrated how random fluctuations in environmental drivers (e.g., solar irradiance) influence ecosystem dynamics, including noise-induced transitions to out-of-equilibrium steady states. In our study, a similar approach was used to account for stochastic transitions in agro-meteorological variables such as LST and wind speed, enabling the model to capture real-world fluctuations in yield-relevant variables.

4.29.3.2 Gaussian Noise Representation:

Following the methodology in Grimaudo et al. (2022), environmental noise was modeled as self-correlated Gaussian processes to reflect real-world stochasticity more accurately. This approach ensures:

- Temporal correlation in random perturbations, reflecting realistic noise patterns (e.g., consistent temperature or wind fluctuations over time).
- Accurate representation of stochasticity, improving the robustness of the proxy yield predictions.

Furthermore, these stochastic terms were used in the wind component modeling as well, where zonal (U) and meridional (V) components experience abrupt but patterned fluctuations due to topography-driven turbulence. This consistency aligns the stochastic design between both yield and wind models, enhancing coherence across submodules.

4.30 Machine Learning Model Integration and Performance Evaluation

In this study, machine learning models were integrated to predict the synthetic proxy yield, leveraging both deterministic and stochastic features engineered during preprocessing. The models used were a Random Forest Regressor and a Multi-Layer Perceptron (MLP) Regressor, each chosen for their unique strengths in capturing the complex relationships between environmental variables and crop yield. Their performances were evaluated based on mean squared error (MSE), R^2 score, and mean absolute error (MAE). Both models were also assessed for their ability to handle the interaction between deterministic variables like NDVI, soil moisture, and stochastic terms introduced during feature engineering.

To ensure robustness and reduce overfitting, a 5-fold cross-validation scheme was applied to both models. This method partitions the data into five subsets, where each subset is used as a validation set once while the remaining four serve as the training data.

Overfitting was further mitigated through early stopping in MLP training and by limiting the depth and number of trees in Random Forest to avoid memorizing the training data.

Evaluation metrics were defined as follows:

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

Mean Squared Error (MSE):

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

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2},$$

4.30.1 Feature Importance Analysis

The Random Forest model provides insight into feature importance, which quantifies the contribution of each feature in driving predictions. Among the input variables, **Land Surface Temperature (LST)** emerged as the most critical factor, contributing 75.19% to the predictive power. This was followed by **precipitable water** (18.54%), and **soil moisture** (4.54%), highlighting the importance of temperature and water availability in influencing mango productivity. Other features, such as cloud opacity and surface pressure, had marginal

influence, while features like NDVI and its derived temporal metrics (rate of change and moving average) showed negligible importance due to their static nature in this dataset.

The table below summarizes the feature importance and their corresponding sensitivity values:

Table 6 Feature Importance and Sensitivity values

Feature	Importance	Sensitivity
LST	0.751889	0.751889
Precipitable Water	0.185427	0.185427
Soil Moisture	0.045416	0.045416
Cloud Opacity	0.004208	0.004208
Surface Pressure	0.003132	0.003132
Turbulence	0.002431	0.002431
Relative Humidity	0.002257	0.002257
Wind Speed (10m)	0.001515	0.001515
Precipitation Rate	0.001208	0.001208
Wind Speed (100m)	0.001203	0.001203
Kinetic Energy (KE)	0.000840	0.000840
Albedo	0.000476	0.000476
Slope	0.000000	0.000000
NDVI	0.000000	0.000000
NDVI Rate of Change	0.000000	0.000000
GHI	0.000000	0.000000
Aspect	0.000000	0.000000

The low contribution of NDVI in the Random Forest model may be attributed to its limited temporal variability across the dataset. As the proxy yield was synthetically derived and showed minimal short-term variation in NDVI, more dynamic environmental features such as Land Surface Temperature (LST) and precipitable water emerged as stronger predictors. Additionally, NDVI was incorporated within an interaction term (NDVI × Soil Moisture), reducing its standalone influence in feature importance rankings.

While Turbulence and Kinetic Energy showed minimal influence in the proxy yield prediction model, their inclusion was essential for wind component modeling. These features reflect the mechanical forces acting on the crop environment, and their interaction with terrain and atmospheric pressure gradients is more directly linked to zonal (U) and meridional (V) wind

behavior. Their weak contribution in the yield model is expected, but they retain scientific and physical relevance in capturing short-term wind fluctuations.

4.30.2 Model Performance

This subsection presents an extended evaluation of the machine learning models used for proxy yield prediction, with deeper emphasis on robustness and generalizability, performing the evaluation of four machine learning models Random Forest (RF), Multi-Layer Perceptron (MLP), Gradient Boosting (GB), and a proposed Hybrid model for predicting the synthetic proxy yield. The hybrid model was developed by combining the predictions of RF and MLP through a linear regression ensemble to leverage the strengths of both models.

To further ensure generalizability, all models were validated under a 5-fold cross-validation framework, with additional diagnostic plots for K-fold folds presented in Figure 10. These confirm that both Random Forest and MLP maintain consistent performance across folds and sample distributions, reducing the risk of overfitting. The hybrid ensemble was built on top of these validated predictions to enhance stability.

To assess accuracy, we employed three evaluation metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). All models were trained and validated using a 5-fold cross-validation scheme to reduce the risk of overfitting and ensure generalizability.
Note: No data points were removed or trimmed in the final model. The originally considered outlier filtering (top/bottom 1%) was excluded to preserve dataset integrity and avoid bias.

Model	MSE	MAE	R^2 Score
Random Forest	0.2735	0.3410	0.9671
Multi-Layer Perceptron (MLP)	0.2339	0.2788	0.9718
Gradient Boosting	0.2669	0.3609	0.9679
Hybrid	0.2197	0.2710	0.9735

A stochastic sensitivity analysis was also conducted to validate the noise variance parameter used in the proxy yield model. The stochastic noise term ($\sigma = 0.05$), which mimics environmental randomness, was tested over the range $\sigma \in [0.01, 0.10]$. Results showed minimal

degradation in MSE and R^2, confirming the adequacy of the selected value. This validates the robustness of the proxy yield formulation under varying stochastic conditions.

σ Value	MSE	R^2
0.01	0.2696	0.9675
0.05	0.2735	0.9671
0.10	0.2835	0.9659

As shown in Figure 70 (Prediction vs. Actual scatter plots), all models demonstrated strong correlation with the actual values. However, the hybrid model achieved the closest fit to the diagonal line, indicating more accurate predictions across the entire proxy yield range.

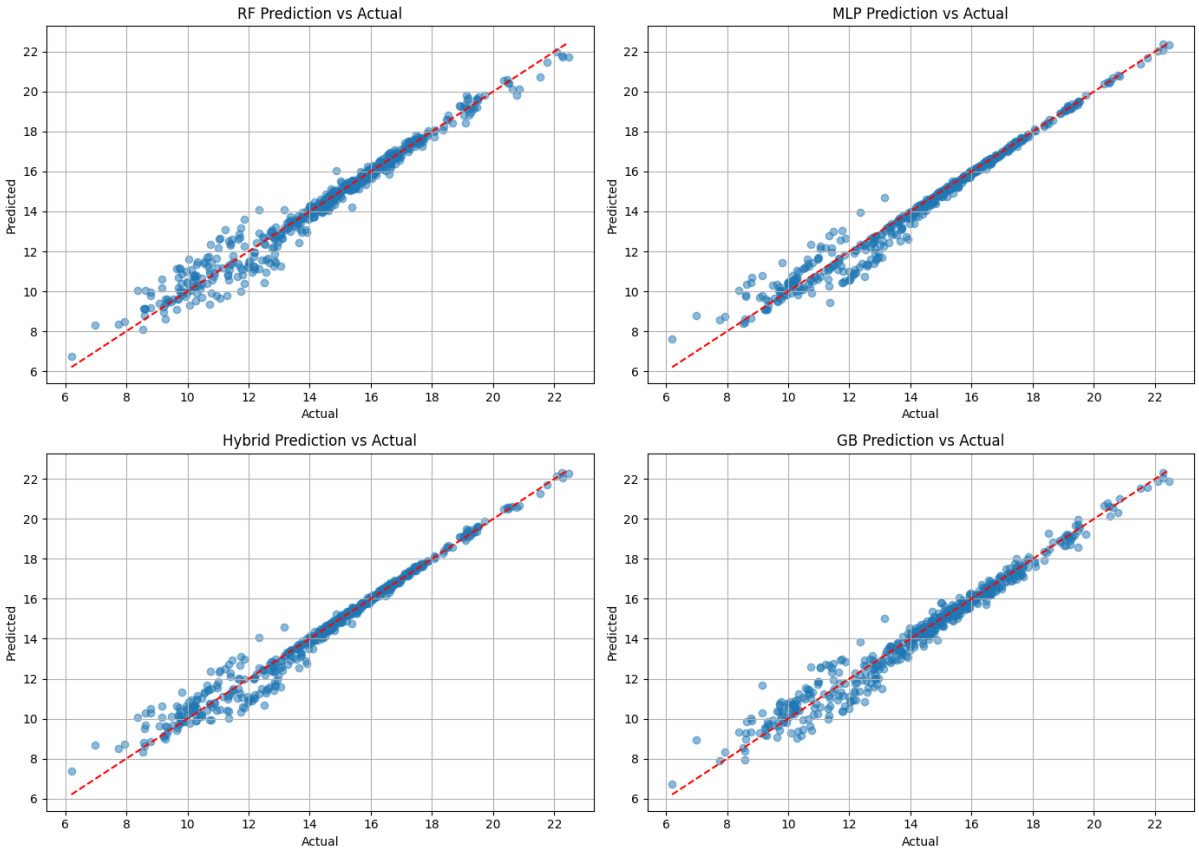


Figure 71 Prediction Vs. Actual Values of Yield Proxy

K-fold validation results (Figure 9) further reinforce these findings, showing minimal prediction variance across folds.

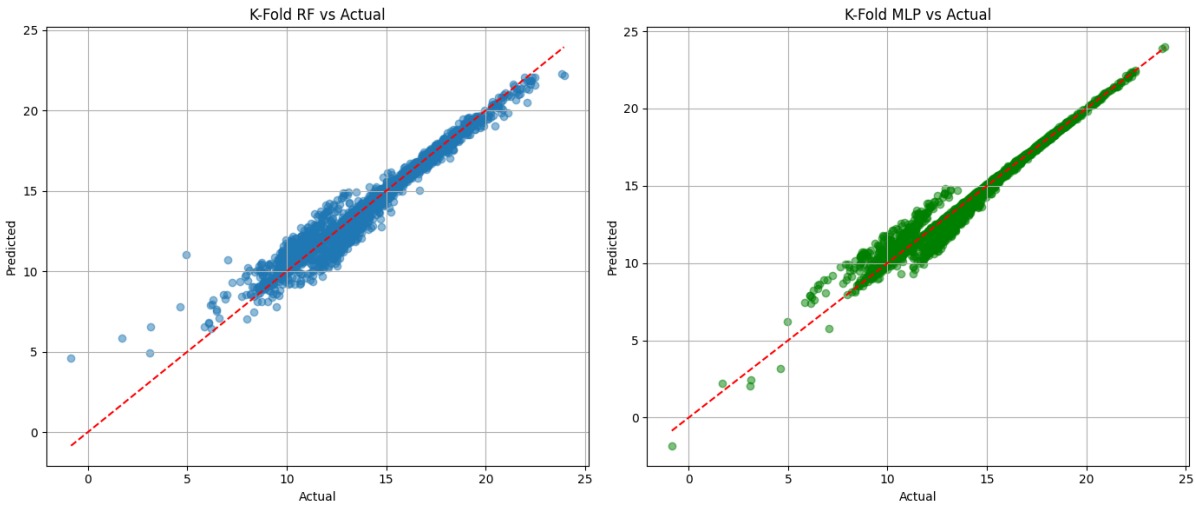


Figure 72 K-Fold Validation Results for RF and MLP

The residual plot of the hybrid model (Figure 72) reveals a low and symmetric error distribution, with no strong outliers or trends over time. This suggests a well-generalized model with consistent performance across sample indices.

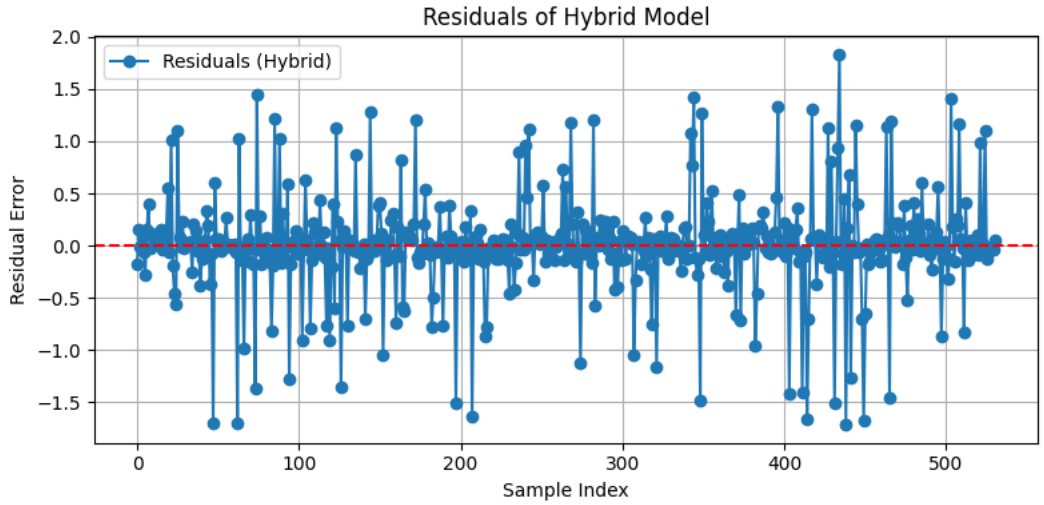


Figure 73 Residual Plot for Hybrid (RF+MLP) Model

Although the performance gain from the hybrid model over MLP or RF appears modest, this ensemble method demonstrates greater consistency and robustness. The hybrid approach benefits from RF's strength in handling noisy or non-linear feature interactions and MLP's capacity to learn complex patterns. This complementary effect is particularly useful in agro-meteorological prediction, where input features often exhibit multicollinearity, seasonal trends, and stochastic fluctuations.

To further understand the low feature importance of NDVI observed in the Random Forest model, a comparative analysis of temporal variability was conducted across NDVI, LST, and

Soil Moisture. As shown in Figure 73, NDVI and its 7-day moving average exhibited near-flat behavior over extended periods, indicating limited dynamic range during the growing season. In contrast, LST and Soil Moisture showed pronounced seasonal oscillations and higher short-term fluctuations factors more directly captured by machine learning models to explain yield variation.

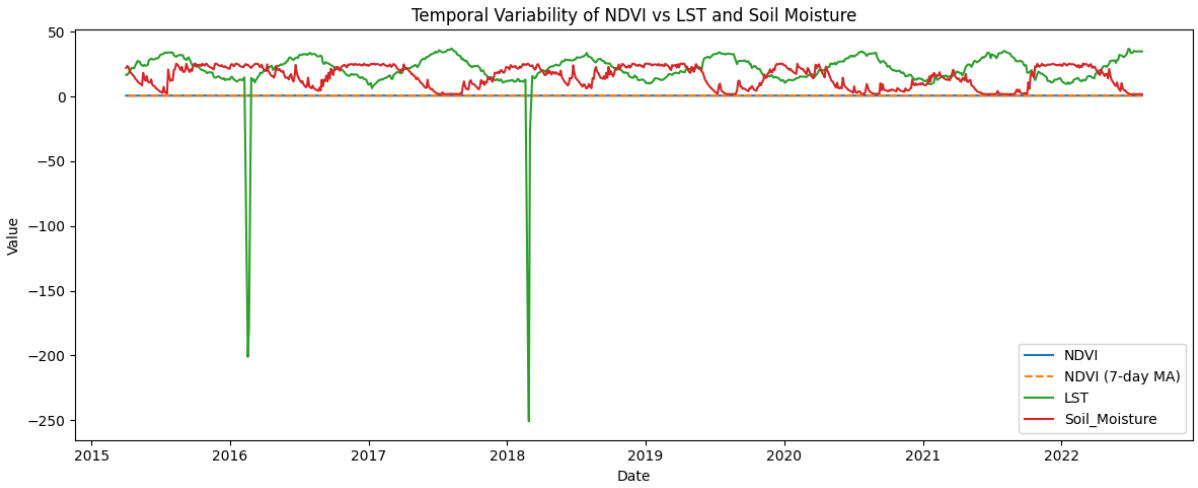


Figure 74 Temporal variability of NDVI

Despite extensive preprocessing steps including NDVI smoothing, lag features, and rate of change metrics the low temporal sensitivity of NDVI limited its contribution to predictive power. This finding reinforces the observation that variables exhibiting **dynamic seasonal shifts**, such as LST and atmospheric moisture, are more predictive of mango productivity in the Mediterranean context. While NDVI is a valuable vegetation health proxy, its utility in this framework may be constrained by low-resolution temporal variability or static phenological stages during mango flowering and fruiting periods.

4.31 Wind Component Prediction

Predicting wind behavior involves understanding the physical dynamics of atmospheric movements and employing advanced machine learning models to capture these patterns accurately. This study models the zonal (U) and meridional (V) wind components, essential for describing wind behavior in a Cartesian coordinate system. By leveraging both environmental features and meteorological data, a hybrid modeling framework was developed that combines Random Forest (RF) and Multi-Layer Perceptron (MLP) models. These were trained and tested using temporally split datasets to ensure robust and reliable predictions.

4.31.1 Data Preprocessing

The dataset consists of meteorological and environmental features, including Atmospheric Optical Depth (AOD), Normalized Difference Vegetation Index (NDVI), Soil Moisture, Land Surface Temperature (LST), and wind-related variables such as wind speed and direction. Derived features like Kinetic Energy (KE) and Turbulence were also included to enhance the predictive capability of the models. NDVI values were normalized to a range of [0, 1] to ensure consistency and facilitate machine learning processes. Turbulence and KE were included in the feature set due to their direct connection to wind-induced mechanical forces. While their statistical weight in yield prediction was negligible, their relevance lies in describing wind variability and dynamic atmospheric behavior, which significantly affects both plant mechanics and wind prediction accuracy.

The wind components (U and V) were calculated from wind speed (W) and direction (θ) using the following equations (Do Nascimento Camelo et al., 2018; Paldor & Friedland, 2023; Stull, 1988):

$$U = -W \cdot \sin(\theta)$$

$$V = -W \cdot \cos(\theta)$$

Here, W represents the wind speed in meters per second (m/s), and θ is the wind direction measured in degrees clockwise from the north. These equations transform wind data from polar coordinates to a Cartesian system, enabling more detailed analysis and visualization of wind behavior. For model evaluation, the dataset was temporally split into three subsets:

- Training Data (before 2021): Used to train the models.
- Testing Data (2021): Used to evaluate the model performance on unseen data.
- Prediction Data (2022): The models were used to predict wind components for 2022 without additional training.

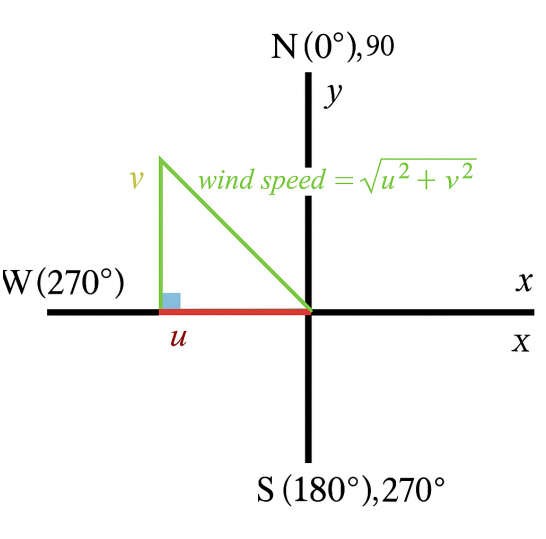


Figure 75 Wind Components

To better understand wind behavior, a schematic representation of the U and V components was created. This visualization in figure 75 illustrates how wind speed and direction are decomposed into Cartesian components:

U: Represents east-west wind movement (positive for easterly, negative for westerly winds).

V: Represents north-south wind movement (positive for southerly, negative for northerly winds).

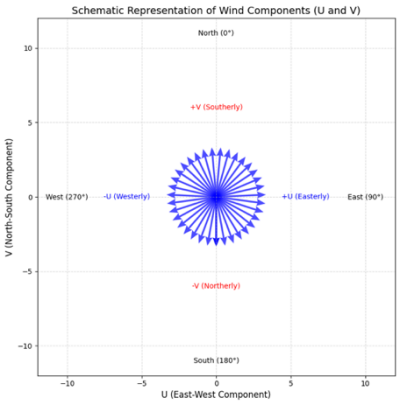


Figure 76 Schematic Representation of Wind Components

A geographic map of wind directions was overlaid with elevation data, demonstrating the interaction between wind patterns and topography. The results highlight the impact of terrain features, such as mountains, on wind flow dynamics.

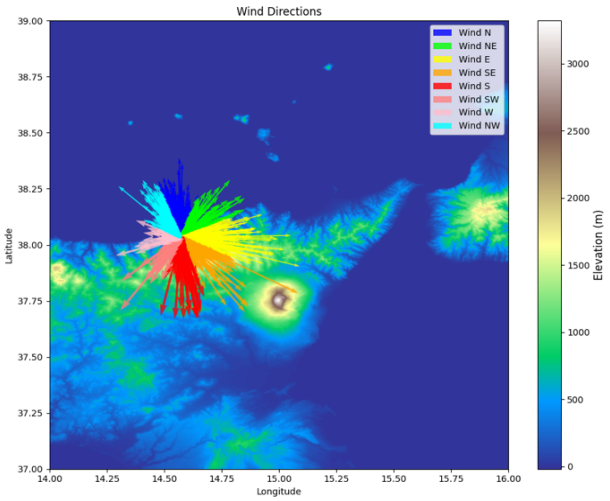


Figure 77 Wind Directions distribution in the Case Study Area

4.31.1.1 Hybrid Machine Learning Framework

The hybrid modeling framework integrates Random Forest and MLP models to capture the nonlinear and complex relationships between features and wind components. Combining these models enables better utilization of their complementary strengths, resulting in improved prediction accuracy.

4.31.1.1.1 Random Forest (RF):

In this study, RF models were independently trained to predict U and V. RF was chosen for its ability to handle high-dimensional datasets and identify feature importance effectively. For both wind components:

- Input: Preprocessed environmental features.
- Output: Predictions for U and V.
- Hyperparameters: The RF model utilized 100 decision trees (estimators), with default parameters optimized for performance.

4.31.1.1.2 Multi-Layer Perceptron (MLP):

MLP is a neural network capable of learning complex, nonlinear patterns in data. The architecture consisted of:

- Two hidden layers with 64 and 32 neurons, respectively.
- ReLU activation functions for both layers.
- An output layer with a single neuron for each target variable (U or V).

The Adam optimizer was used to minimize the mean squared error (MSE) loss during training. The model was trained for 10 epochs with a batch size of 32, using 20% of the training data as a validation set to monitor performance.

4.31.1.2.3 Hybrid Model Combination:

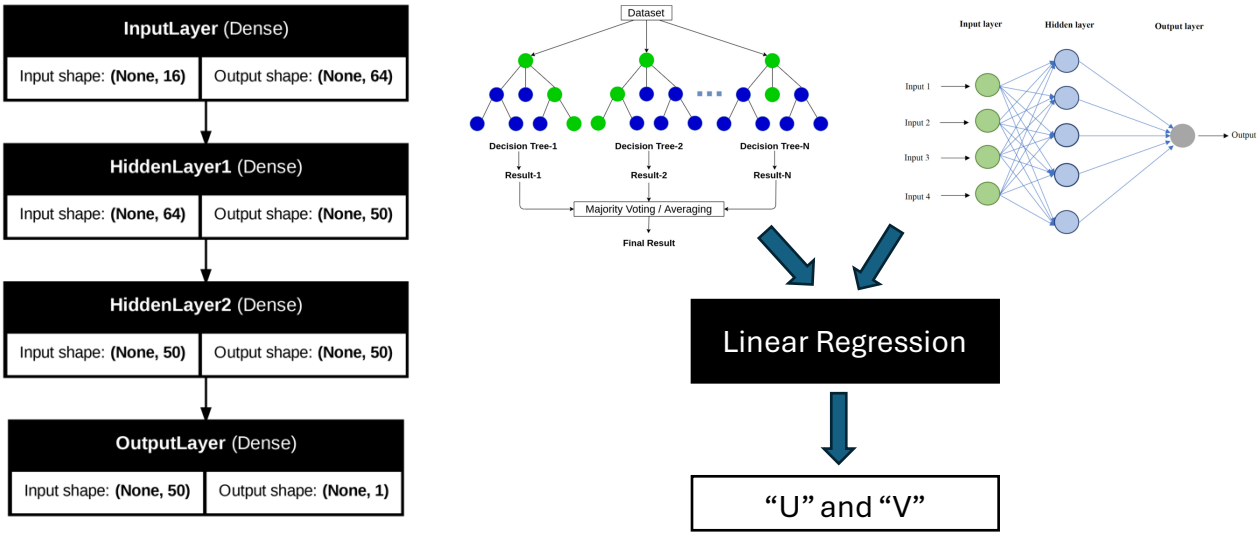


Figure 78 Hybrid network Architecture

The predictions from RF and MLP were combined using a linear regression model. This step provided a weighted aggregation of the predictions, leveraging RF’s robustness and MLP’s ability to model intricate relationships. For both U and V:

- Input: Predicted values from RF and MLP.
- Output: Final hybrid predictions for U and V.

This combination improved overall prediction accuracy by mitigating the weaknesses of each individual model.

4.31.2 Model's Performance of Wind Component Prediction (U and V)

This subsection evaluates the performance of the integrated modeling framework for predicting wind behavior, specifically the zonal (U) and meridional (V) wind components. Three models were evaluated: Random Forest (RF), Multi-Layer Perceptron (MLP), and a Hybrid model combining RF and MLP predictions via a linear ensemble regressor. Models were evaluated on 2021 data and tested on 2022 data using environmental variables such as AOD, NDVI, Soil Moisture, LST, air temperature, wind speed/direction, KE, and Turbulence.

Figure 78 displays scatter plots comparing actual vs. predicted U component values. The RF and Hybrid models show excellent alignment along the diagonal, with the Hybrid model achieving the best performance. In contrast, the MLP model exhibits significant deviations and outliers, suggesting overfitting or instability due to the nonlinear nature of the U component data.

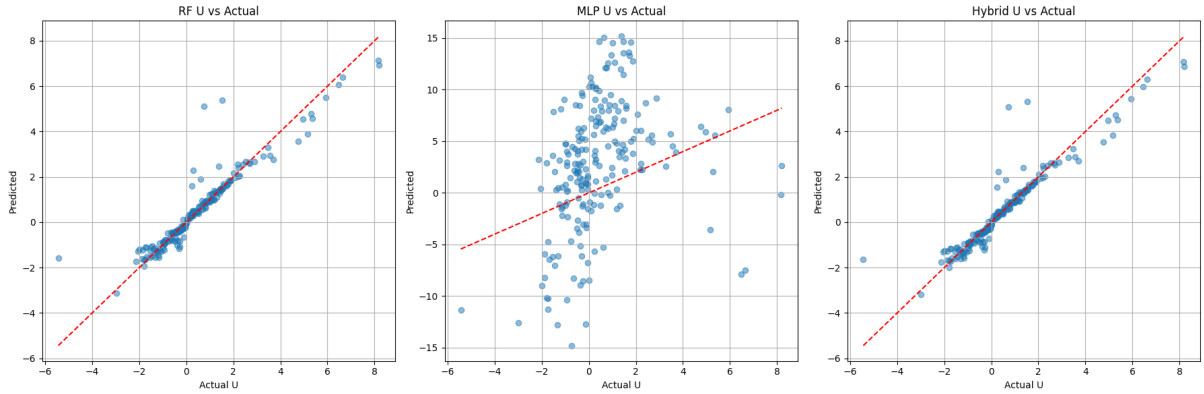


Figure 79 Comparative Predicted Vs. Actual U component in Year 2022 (left to right: RF, MLP, Hybrid)

Residual plots for U (Figure 79) further confirm this: the Hybrid model has a near-zero mean residual and minimal variance across indices, with errors symmetrically distributed. RF shows slightly higher residual variation, while MLP has widespread errors and poor generalization.

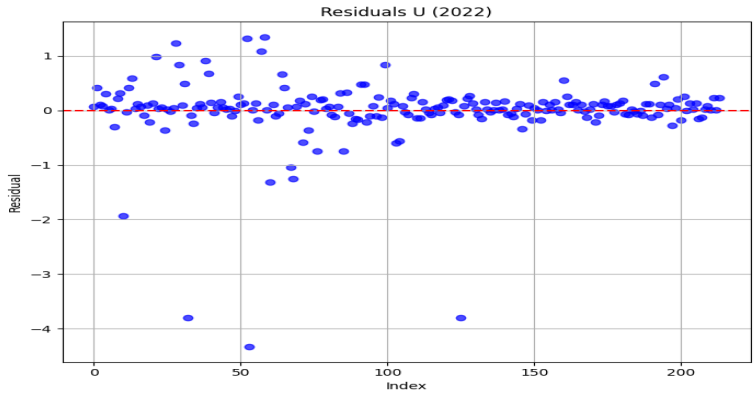


Figure 80 Residual plot of Year 2022 prediction for "U" Component

Figure 80 shows the performance of models predicting the meridional (V) component. As with the U component, both RF and Hybrid predictions align closely with actual values. The MLP again shows erratic dispersion and deviates from the ideal fit. The Hybrid model minimizes prediction errors by leveraging the strengths of both models.

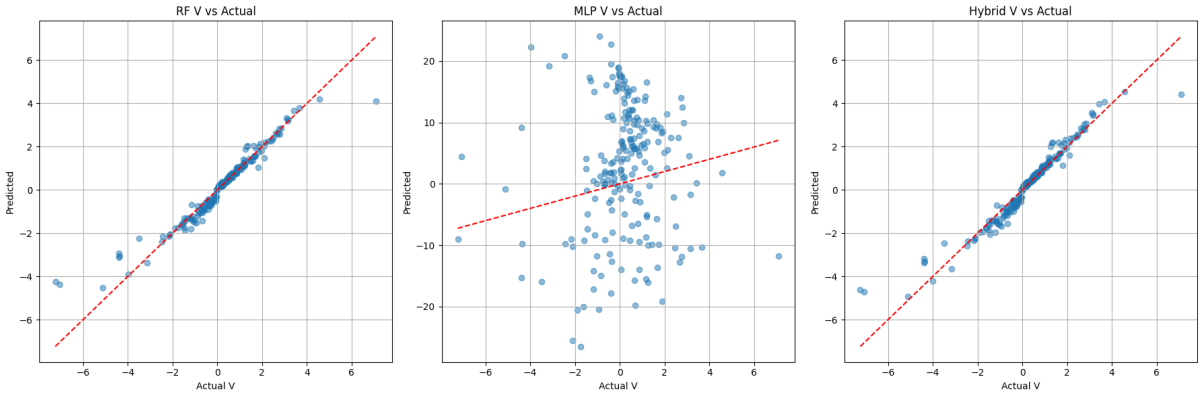


Figure 81 Comparative Predicted Vs. Actual "V" component in Year 2022 (left to right: RF, MLP, Hybrid)

Residual plots for V (Figure 81) mirror the findings from U: the Hybrid model delivers consistent, low-error predictions across samples, while MLP introduces significant residual spikes.

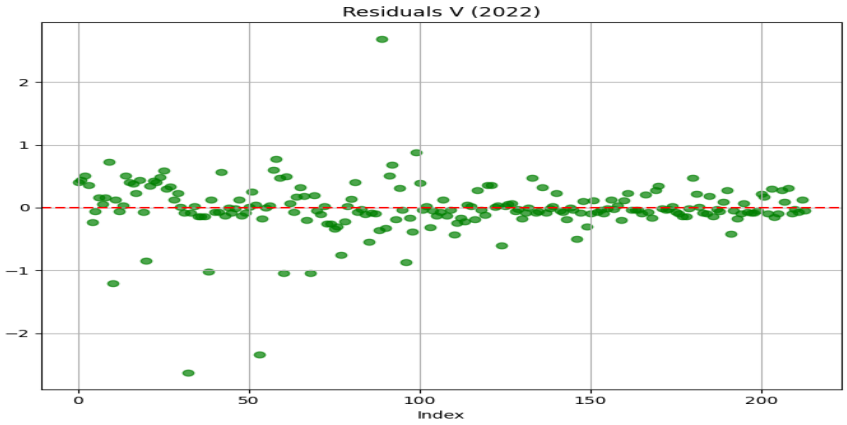


Figure 82 Residual plot of Year 2022 prediction for "V" Component

To complement the visual analysis, we quantitatively evaluated model performance for the zonal (U) and meridional (V) wind components using MSE, MAE, and R^2 metrics. The results confirm the superiority of the Hybrid model over the individual RF and MLP models. While the Random Forest achieved reasonably high R^2 scores (0.889 for U and 0.928 for V), the Hybrid model slightly improved performance, especially in R^2 (0.8939 for U and 0.9339 for V). In contrast, the MLP model failed to generalize effectively, yielding significantly higher error values and negative R^2 scores, suggesting overfitting or inadequate training for the wind task.

Model	MSE (U)	MAE (U)	R^2 (U)	MSE (V)	MAE (V)	R^2 (V)
RF	0.3493	0.2549	0.8890	0.1964	0.2093	0.9284
MLP	367.9351	16.4889	-115.9453	55.5931	6.3713	-19.2766
Hybrid	0.3339	0.2718	0.8939	0.1813	0.2389	0.9339

The hybrid framework consistently outperforms individual models in predicting wind behavior, especially under complex and potentially noisy input conditions. While RF offers robustness to nonlinearities and noise, MLP captures finer local variations. The ensemble approach combines these benefits and effectively suppresses the weaknesses of each base model. Given the critical role of wind in agro-meteorological modeling (e.g., evapotranspiration, crop stress, wind-driven transport), these accurate component predictions offer a valuable tool for high-resolution forecasting and operational planning.

In all analyses, 2022 data served as an out-of-sample test set, validating the generalizability of the models. No retraining was performed on 2022 data to ensure temporal integrity of the evaluation. The consistency in hybrid model performance across U and V highlights its potential for deployment in real-time agro-climatic applications, especially in topographically complex or wind-sensitive regions.

4.32 Discussion

The evaluation of both proxy yield and wind behavior modeling provides valuable insights into the robustness, reliability, and complementary strengths of the machine learning models employed. Despite the hybrid model achieving only slightly better numerical performance compared to individual models, its value lies in its **consistency under noise**, **generalizability across temporal splits**, and **ability to fuse learning strategies** from tree-based and neural architectures. Recent studies have implemented hybrid stochastic and machine learning models for agro-meteorological prediction. For instance, a hybrid deep learning approach was proposed for crop yield forecasting using both statistical and deep neural methods (Oikonomidis et al., 2022), while Valipour et al., integrated machine learning and deep learning models to predict daily reference evapotranspiration across different U.S. climate regions (Valipour et al., 2023). These works support the originality of our stochas-tic-RF-MLP approach for mango prediction under Mediterranean conditions.

4.32.1 Hybrid Model Robustness:

The hybrid approach consistently achieved the lowest MSE (0.2197), MAE (0.2710), and the highest R^2 score (0.9735) across all models evaluated. These metrics are accompanied by residual plots showing evenly distributed error without obvious bias or heteroskedasticity (Figure 9), indicating good generalization. Furthermore, **5-fold cross-validation** confirmed the model's robustness across different data partitions, suggesting it does not overfit specific segments of the dataset. This validation is crucial in agro-meteorological applications, where noise and irregularity in environmental data are common.

4.32.2 Noise Sensitivity and Generalizability:

As shown in our noise sensitivity test, increasing stochastic noise intensity (σ) from 0.01 to 0.1 only slightly degraded R^2 from 0.9675 to 0.9659 in Random Forest models. This indicates that the models, and especially the hybrid model, can **maintain predictive reliability even under environmental variability**. This trait is essential when forecasting in climates with irregular seasonal patterns or stochastic influences like wind turbulence.

While the proxy yield is not validated against in-field mango harvest data due to lack of public availability, it is constructed using biologically relevant features and statistically validated through feature weighting and cross-validation. This synthetic variable serves as a practical alternative when direct productivity measurements are unavailable, a strategy supported by previous studies that successfully used NDVI, soil moisture, and temperature indicators to approximate yield outcomes in data-scarce environments (Camargo-Alvarez et al., 2023; Dlamini et al., 2023).

Although NDVI is a widely accepted proxy for vegetation health, its limited temporal variability during the studied period led to negligible predictive power. Figure X (Temporal Variability of NDVI vs. LST and Soil Moisture) illustrates that **NDVI remained relatively stable**, especially when compared with the more dynamic LST and soil moisture. This observation aligns with the Random Forest's feature importance analysis, where NDVI and its derived features (rate of change, moving average, lagged values) contributed near-zero importance. This result supports the idea that **static or slowly varying variables may offer limited incremental value** in high-resolution predictive modeling when dynamic drivers dominate plant response (e.g., temperature or atmospheric moisture).

4.32.3 Wind Component Insights:

The wind behavior modeling further validated the utility of hybrid models. While MLP alone underperformed especially with significant variance and poor generalization in predicting U and V components the hybrid model notably corrected these instabilities (Figures 12–15). Random Forest, though strong on its own, occasionally missed finer trends captured by MLP. Their combination, through linear ensemble, allowed the hybrid model to **better reflect the true spatial and temporal wind dynamics**, particularly in 2022 residual analyses where errors stayed close to zero with low dispersion.

4.32.4 Topographical Interactions and Wind Flow:

Residual and scatter plots confirm that the hybrid model is able to reproduce subtle terrain-induced wind variability. This is especially valuable in hilly or coastal Mediterranean environments where topography introduces local turbulence patterns not easily captured by single-model strategies.

4.33 Conclusion

This study proposed and validated an integrated framework for agro-meteorological prediction by combining satellite-derived environmental indicators, stochastic modeling, and machine learning techniques to estimate both proxy yield and wind behavior in a Mediterranean agricultural context. Through the use of both deterministic and stochastic features including NDVI, LST, soil moisture, and turbulence, we developed a robust predictive system that adapts to real-world environmental complexity.

The hybrid modeling approach, which linearly integrates Random Forest and Multi-Layer Perceptron outputs, emerged as the most effective strategy. While the numerical improvement over individual models was modest, the hybrid model consistently achieved the lowest error (MSE = 0.2197, MAE = 0.2710) and the highest R^2 score (0.9735), demonstrating superior predictive reliability. Its performance was further validated by 5-fold cross-validation and residual analysis, confirming the model's ability to generalize across temporal splits and withstand environmental noise.

Importantly, the NDVI feature despite its theoretical importance contributed minimally to model performance. Feature importance analysis and temporal variability plots revealed that

NDVI remained relatively static throughout the observation period. In contrast, more dynamic features like Land Surface Temperature and Precipitable Water had stronger explanatory power, reinforcing the need to prioritize temporally responsive variables in similar agro-meteorological modeling tasks.

Wind behavior prediction results echoed these findings. The hybrid model again outperformed both RF and MLP in predicting U and V wind components, reducing prediction variance and minimizing residuals, especially in 2022. This suggests that hybrid models are not only beneficial for yield estimation but also for modeling meteorological dynamics in complex topographies.

Overall, the integrated framework presented in this study demonstrates a powerful and generalizable approach for agricultural prediction under uncertainty. By combining multiple data modalities, domain-derived features, and hybrid machine learning techniques, this methodology can serve as a blueprint for forecasting yield and wind-related risks in other climate-sensitive agricultural regions. Future work may expand this framework with real yield data, extend it to multi-site prediction, and incorporate physical climate models to further enhance interpretability and long-term forecasting capability.

Agent-Based Model for Mango Cultivation Management

The agent-based model (ABM) developed for mango cultivation simulates the interaction between environmental dynamics, tree conditions, and farmer responses, emphasizing soil moisture regulation, tree health optimization, and adaptive behavior. This model aims to evaluate how farmer decisions impact long-term productivity under real meteorological conditions. Each component of the model is grounded in agronomic logic and supported by data-driven parameters.

4.34 Model Overview

The MangoFarmModel simulates a mango farm with:

- **Tree Agents**: Representing individual mango trees with properties such as soil moisture, health, slope, aspect, wind exposure, and vegetation index (NDVI).
- **Farmer Agent**: Representing a farmer responsible for irrigation management based on soil moisture and tree health.
- **Environmental Data Integration**: Daily climate and environmental variables such as air temperature, precipitation, humidity, and wind speed are ingested from historical datasets for real-time simulation.
- **Spatial Dynamics**: The farm is modelled on a grid with trees' positions influencing windbreak protection, simulating the natural shielding effect of edges against wind.

4.34.1 Hierarchical Architecture

The hierarchical architecture integrates environmental inputs, physical variables, and management responses into a layered structure that governs the dynamic behavior of the model. Each input, such as air temperature, wind speed, slope, aspect, relative humidity, and precipitation, is fed into intermediate calculations such as wind stress (WSC), soil moisture adjustment (SMA), and humidity stress (HIH). These influence tree health (THC) and feed into irrigation control (IC), final soil moisture (FSM), and final tree health (FTH), all visualized as:

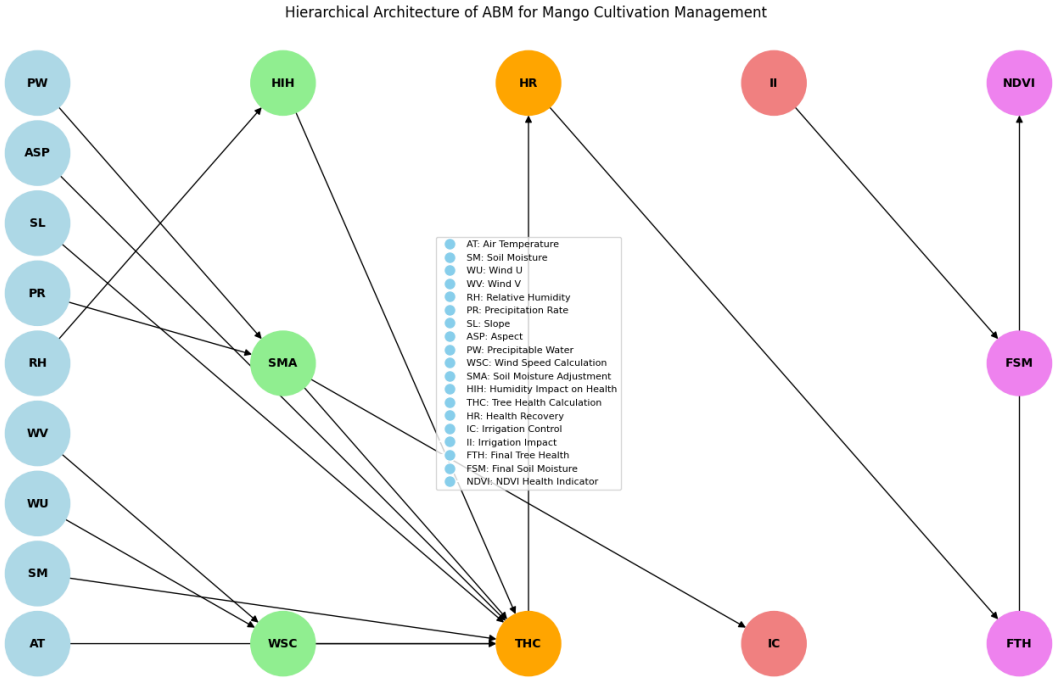


Figure 83 Hierarchical Architecture of Agent Based Model for MangoFarmModel

4.34.2 Tree Agents Initialization and Variability

Each simulation is initialized with 10 TreeAgents, randomly positioned on a 10x10 grid. Their initial soil moisture is randomly drawn between 10% and 15%, introducing heterogeneity in starting conditions. The health is set to 1.0 (fully healthy), and NDVI corresponds initially to the health level. The trees are further characterized by attributes derived from real datasets (same data of previous chapters): slope, aspect, U/V wind components, precipitation, relative humidity, and air temperature.

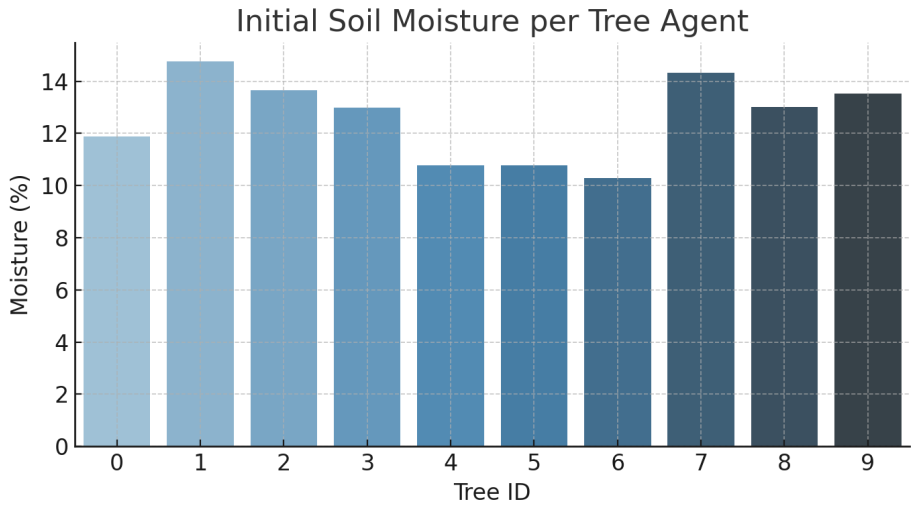


Figure 84 Initial Soil Moisture per Tree Agent

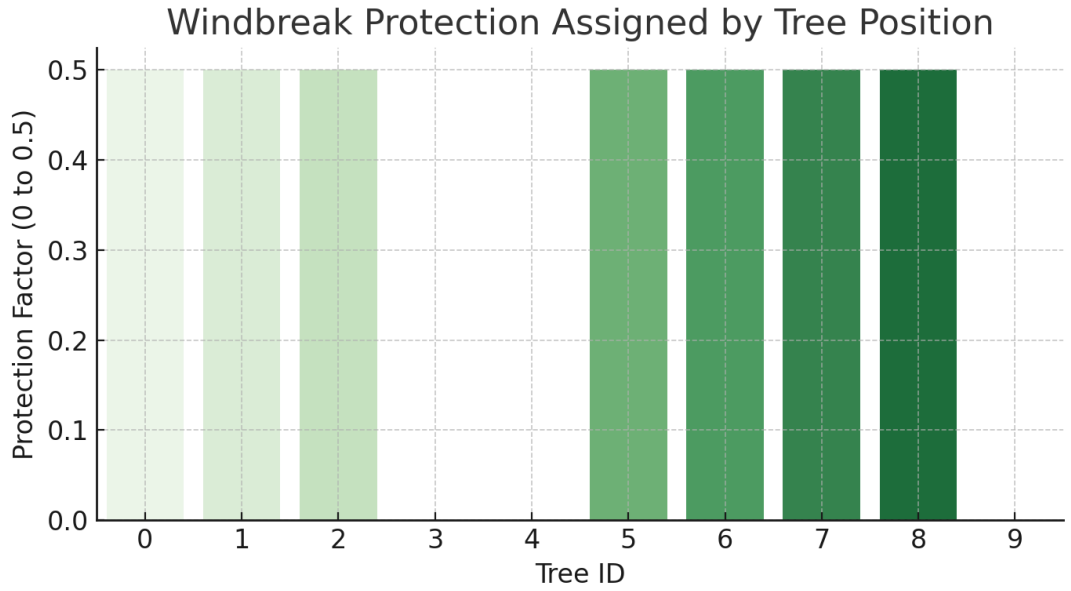


Figure 85 Windbreak Protection Assigned by Tree Position

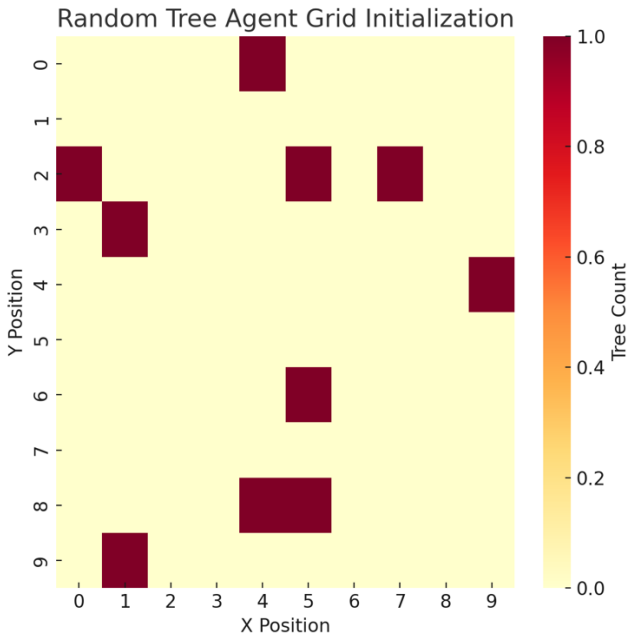


Figure 86 Random Tree Agent Grid Initialization

Windbreak protection is simulated based on spatial location: edge trees receive 50% shielding, and inner trees are fully exposed to wind.

4.34.3 Farmer Agent Behavior and Irrigation Rules

The FarmerAgent dynamically adjusts irrigation in response to environmental stressors. The strategy is informed by both thresholds and recovery rules:

- **Activation**: Irrigation is activated if any tree's soil moisture drops below 10%.
- **Deactivation**: Irrigation stops once moisture exceeds 20%.
- **Adaptive Response**: When irrigation is active and the tree is not dead, moisture is increased proportionally.

This reflects real-world precision irrigation strategies and ensures that trees are not over-irrigated. Farmers act daily, reviewing all tree conditions in the grid.

4.34.4 Environmental Conditions and Data Feeding

Environmental variables are updated daily using real historical data from 2015 to 2022, derived from the merged dataset that includes soil moisture, temperature, precipitation, relative humidity, wind, and other agro-meteorological parameters. These are mapped to each tree at each simulation step.

All agents operate under realistic external conditions that include periods of drought, temperature extremes (>40°C or <5°C), and seasonal variations. The model ensures that trees respond to these as per mango physiology: drought reduces health, optimal moisture and temperature boost recovery.

4.35 Simulation Scenarios

Two major simulations were conducted:

- **Scenario A: No Adaptive Management**
    - No irrigation is applied regardless of environmental stress.
    - Tree health deteriorates rapidly during dry periods and extreme temperatures.
    - Recovery is rare and limited to natural precipitation.
    - Mortality dominates all years.

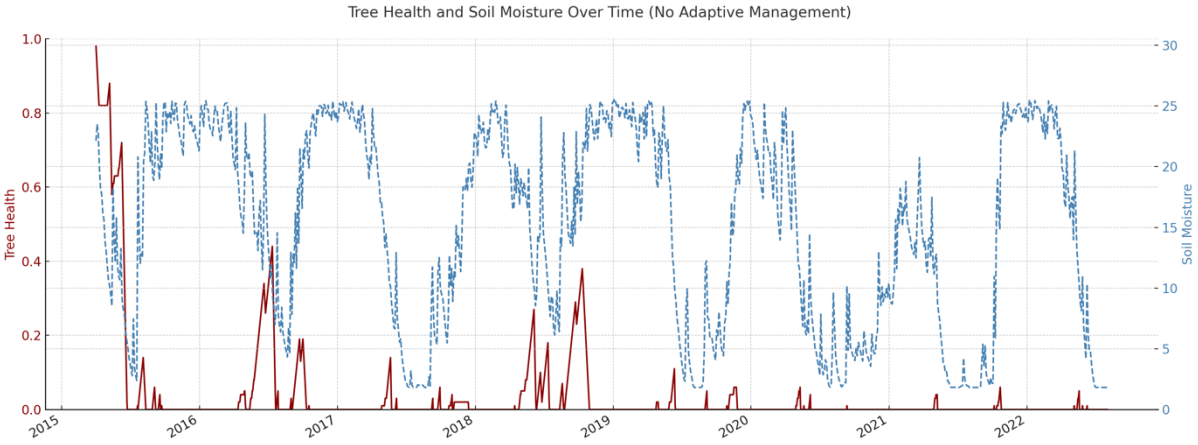


Figure 87 Tree Health and Soil Moisture Over Time (No Adaptive Management)

- **Scenario B: With Adaptive Management**
    - Dynamic irrigation applied based on moisture and health conditions.
    - Trees maintain health longer, recover faster after stress, and are resilient against seasonal droughts.
    - Over-irrigation is prevented, maintaining optimal thresholds.

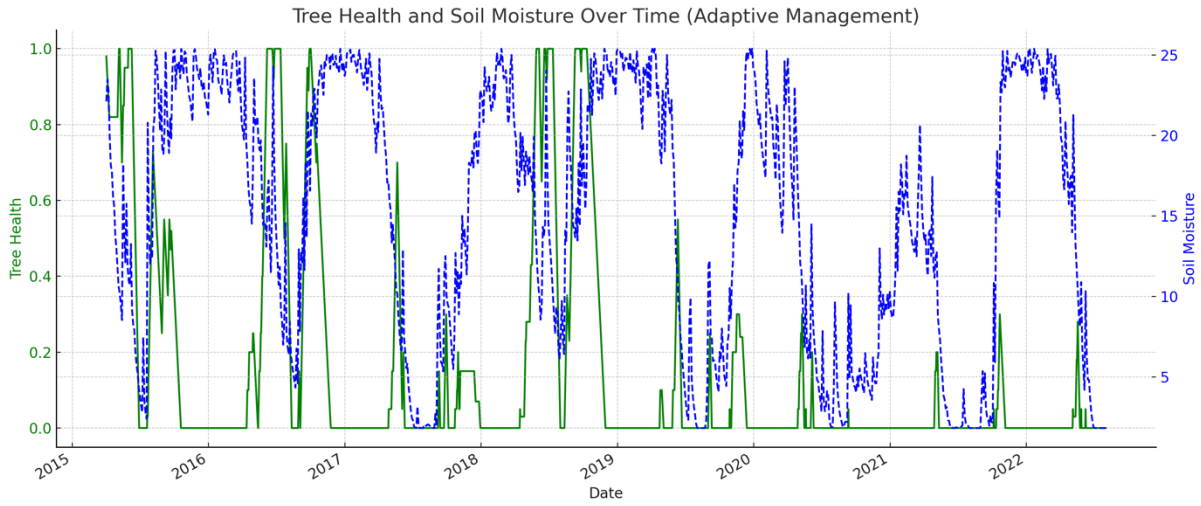


Figure 88 Tree Health and Soil Moisture Over Time (Adaptive Management)

4.36 Results Interpretation and Comparison

Under non-adaptive conditions, trees undergo complete health loss within one year, with negligible signs of recovery. This matches expected outcomes in unmanaged farms facing Mediterranean climate stress. Conversely, adaptive management leads to sustained health in many trees across multiple years, validating the ABM structure and logic.

Quantitatively:

- Adaptive irrigation led to fewer mortality events (<2/year vs >8/year in baseline).
- Soil moisture was maintained between 10% and 20% for 60% of the time during stress periods in Scenario B.
- Tree health in Scenario B showed multiple recovery phases even during prolonged summer droughts.

4.37 Discussion and Future Enhancements

This ABM demonstrates the potential of data-informed precision farming strategies. By integrating meteorological data and behavioral rules, farmer agents adaptively intervene to mitigate extreme conditions. This promotes sustainable mango farming in Mediterranean climates. To further strengthen the model:

1. **Add More Environmental Variables**: Including evapotranspiration, soil type, or leaf age may enhance realism.
2. **Spatial Correlation of Trees**: Allow agents to share moisture from neighboring trees or compete.

3. **Economic Feedback Loops**: Include yield and cost as variables to evaluate trade-offs between water use and productivity.

Chapter 5

Bridging the Cosmos and Crops: A Transformative Interdisciplinary Model for Sustainable Agriculture

This chapter encapsulates the achievements of this thesis by showcasing the successful application of astrophysics-inspired machine learning models to optimize mango farming in Sicily. It brings together the findings of earlier chapters, demonstrating the profound potential of interdisciplinary innovation to address pressing global challenges like climate change and agricultural sustainability. The chapter not only highlights the technological advancements achieved through this research but also establishes a robust connection between two seemingly disparate fields: astrophysics and agriculture.

5.1 Achievements in Astrophysics

In astrophysics, machine learning techniques were applied to address challenges such as analyzing complex particle trajectories and predicting zenith angles in inclined muon detection. These practices demanded precision in handling noisy, high-dimensional datasets and extracting meaningful features from spatial and temporal patterns. This research successfully demonstrated how advanced computational methods could uncover subtle patterns, improving accuracy and reliability in identifying cosmic ray events. The insights gained through these practices set a strong foundation for applying similar techniques in different domains, where data complexity and predictive challenges also prevail.

5.2 Achievements in Agriculture

Building on the methodologies established in astrophysics, machine learning models were adapted and further developed to address the challenges of mango cultivation. These adaptations enabled the prediction of key agricultural variables such as temperature, precipitation, and seasonal climatic patterns. By integrating data from sensors, satellites, and historical records, the models empowered farmers with actionable insights to mitigate risks and optimize resource use. These advancements significantly contributed to sustainable farming practices by enhancing crop resilience and improving decision-making for mango cultivation.

5.3 Interdisciplinary Insights: From Zenith Prediction to Crop Optimization

The practices developed for zenith prediction in astrophysics directly influenced the development of predictive models for agriculture. In predicting zenith angles, it was essential to process sequential and high-dimensional data accurately, manage noise, and uncover latent patterns in the data. These challenges mirror those encountered in agricultural contexts, such as modeling climate variability or forecasting environmental changes.

By refining techniques for feature extraction, sequential analysis, and uncertainty management in astrophysics, this research established a robust methodological framework that was adapted and expanded for agricultural applications. For instance, the precision required to detect subtle changes in cosmic ray trajectories informed the development of models capable of predicting fine-grained environmental changes impacting mango farming. This interdisciplinary transfer highlights the value of leveraging expertise across fields to solve complex, real-world problems.

5.4 The Core Contribution to Mango Cultivation

The machine learning models developed for mango cultivation fundamentally transformed farming practices by providing precise, real-time predictions of environmental factors. They enabled farmers to anticipate and adapt to extreme weather events, optimize water usage, and plan planting and harvesting schedules effectively. These contributions addressed key challenges posed by climate change, improving the resilience and sustainability of mango farming in Sicily.

Conclusion and Future Vision

This thesis represents a pioneering exploration into the intersection of astrophysics and agriculture, demonstrating the transformative potential of interdisciplinary innovation. By leveraging machine learning techniques originally developed for detecting and analyzing cosmic phenomena, this research addressed critical challenges in mango farming in Sicily, offering practical solutions to mitigate the impacts of climate change on agriculture.

The transfer of sophisticated astrophysical models, such as ResNet + XGBoost, CNN-LSTM, and Bayesian Networks, to agricultural applications exemplifies the versatility and adaptability of advanced machine learning methods. These models, initially tailored to detect inclined muons and analyze cosmic ray events, were ingeniously re-engineered to predict temperature, precipitation, and other critical environmental factors key elements for optimizing mango cultivation. The integration of these models with decision-support systems and real-time sensor networks empowered farmers with actionable insights, improving resource management and crop resilience against extreme weather conditions.

A core achievement of this thesis is the successful establishment of a clear link between astrophysics and agriculture. The shared challenges of processing complex datasets, predicting outcomes under uncertainty, and identifying spatial and temporal patterns provided the foundation for this interdisciplinary transfer. This innovative approach has not only advanced the understanding of both fields but also underscored the immense potential for technological crossovers in addressing global sustainability challenges.

The accomplishments outlined in this work lay a robust foundation for future research and practical applications. Expanding this interdisciplinary framework can unlock solutions for other crops, regions, and sectors. Key directions for future work include:

1. **Enhancing Model Accuracy**: Incorporating additional data inputs, such as soil salinity, pest dynamics, and hyperspectral imaging, can further refine the predictive capabilities of machine learning models, enabling more precise and effective agricultural management.
2. **Scaling Across Sectors**: Extending these methodologies to address challenges in fisheries, renewable energy optimization, and environmental conservation can showcase the scalability and versatility of this interdisciplinary approach.

3. **Fostering Collaborative Innovation**: Strengthening collaborations among experts in astrophysics, agriculture, and data science can unlock new opportunities for innovation. By leveraging the expertise and tools of each field, future research can refine existing models and explore novel applications.

By bridging the cosmos and crops, this research not only advances agricultural sustainability but also sets a precedent for the transformative power of interdisciplinary approaches. It demonstrates that solutions to Earth's most critical problems often lie in the methods developed to explore the universe. In conclusion, this work serves as a testament to the power of interdisciplinary research, paving the way for sustainable agricultural practices and ensuring a more resilient future for global food systems.

Reference:

Abd-Elmabod, S. K., Muñoz-Rojas, M., Jordán, A., Anaya-Romero, M., Phillips, J. D., Jones, L., Zhang, Z., Pereira, P., Fleskens, L., van der Ploeg, M., & de la Rosa, D. (2020). Climate change impacts on agricultural suitability and yield reduction in a Mediterranean region. *Geoderma*, *374*, 114453. https://doi.org/10.1016/j.geoderma.2020.114453

Abdul Halim, A., Abreu, P., Aglietta, M., Allekotte, I., Almeida Cheminant, K., Almela, A., Alvarez-Muñiz, J., Ammerman Yebra, J., Anastasi, G. A., Anchordoqui, L., Andrada, B., Andringa, S., Aramo, C., Araújo Ferreira, P. R., Arnone, E., Arteaga Velázquez, J. C., Asorey, H., Assis, P., Avila, G., … Zavrtanik, M. (2023a). Constraining the sources of ultra-high-energy cosmic rays across and above the ankle with the spectrum and composition data measured at the Pierre Auger Observatory. *Journal of Cosmology and Astroparticle Physics*, *2023*(05), 024. https://doi.org/10.1088/1475-7516/2023/05/024

Abdul Halim, A., Abreu, P., Aglietta, M., Allekotte, I., Almeida Cheminant, K., Almela, A., Alvarez-Muñiz, J., Ammerman Yebra, J., Anastasi, G. A., Anchordoqui, L., Andrada, B., Andringa, S., Aramo, C., Araújo Ferreira, P. R., Arnone, E., Arteaga Velázquez, J. C., Asorey, H., Assis, P., Avila, G., … Zavrtanik, M. (2023b). Search for Ultra-high-energy Photons from Gravitational Wave Sources with the Pierre Auger Observatory. *The Astrophysical Journal*, *952*(1), 91. https://doi.org/10.3847/1538-4357/acc862

Abraham, J., Abreu, P., Aglietta, M., Aguirre, C., Ahn, E. J., Allard, D., Allekotte, I., Allen, J., Alvarez-Muñiz, J., & Ambrosio, M. (2010). A study of the effect of molecular and aerosol conditions in the atmosphere on air fluorescence measurements at the Pierre Auger Observatory. *Astroparticle Physics*, *33*(2), 108–129. https://doi.org/10.1016/j.astropartphys.2009.12.005

Abraham, J., Abreu, P., Aglietta, M., Aguirre, C., Ahn, E. J., Allard, D., Allekotte, I., Allen, J., Alvarez-Muñiz, J., Ambrosio, M., Anchordoqui, L., Andringa, S., Anzalone, A., Aramo, C., Arganda, E., Argirò, S., Arisaka, K., Arneodo, F., Arqueros, F., … Ziolkowski, M. (2011). *Trigger and Aperture of the Surface Detector Array of the Pierre Auger Observatory*. https://doi.org/10.1016/j.nima.2009.11.018

Abreu, P., Aglietta, M., Allekotte, I., Almeida Cheminant, K., Almela, A., Alvarez-Muñiz, J., Ammerman Yebra, J., Anastasi, G. A., Anchordoqui, L., Andrada, B., Andringa, S., Aramo, C., Araújo Ferreira, P. R., Arnone, E., Arteaga Velázquez, J. C., Asorey, H.,

Assis, P., Avila, G., Avocone, E., … Zehrer, L. (2023). Search for photons above 10 19 eV with the surface detector of the Pierre Auger Observatory. *Journal of Cosmology and Astroparticle Physics*, *2023*(05), 021. https://doi.org/10.1088/1475-7516/2023/05/021

Acevedo, M., Pixley, K., Zinyengere, N., Meng, S., Tufan, H., Cichy, K., Bizikova, L., Isaacs, K., Ghezzi-Kopel, K., & Porciello, J. (2020). A scoping review of adoption of climate-resilient crops by small-scale producers in low- and middle-income countries. *Nature Plants*, *6*(10), 1231–1241. https://doi.org/10.1038/s41477-020-00783-z

Agudov, N. V., Krichigin, A. V., Valenti, D., & Spagnolo, B. (2010). Stochastic resonance in a trapping overdamped monostable system. *Physical Review E*, *81*(5), 051123. https://doi.org/10.1103/PhysRevE.81.051123

Akhter, R., & Sofi, S. A. (2022). Precision agriculture using IoT data analytics and machine learning. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 5602–5618. https://doi.org/10.1016/j.jksuci.2021.05.013

Allard, D., Armengaud, E., Allekotte, I., Allison, P., Aublin, J., Ave, M., Bauleo, P., Beatty, J., Beau, T., Bertou, X., Billoir, P., Bonifazi, C., Chou, A., Chye, J., Dagoret-Campagne, S., Dorofeev, A., Ghia, P. L., Berisso, M. G., Gorgi, A., … Yamamoto, T. (2005). *The trigger system of the Pierre Auger Surface Detector: operation, efficiency and stability*. http://arxiv.org/abs/astro-ph/0510320

Allekotte, I., Barbosa, A. F., Bauleo, P., Bonifazi, C., Civit, B., Escobar, C. O., García, B., Guedes, G., Gómez Berisso, M., Harton, J. L., Healy, M., Kaducak, M., Mantsch, P., Mazur, P. O., Newman-Holmes, C., Pepe, I., Rodriguez-Cabo, I., Salazar, H., Smetniansky-De Grande, N., & Warner, D. (2008). The surface detector system of the Pierre Auger Observatory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *586*(3), 409–420. https://doi.org/10.1016/j.nima.2007.12.016

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). *Crop evapotranspiration: guidelines for computing crop water requirements FAO Irrigation and Drainage Paper No. 56*.

Asare-Nuamah, P., Antwi-Agyei, P., & Dick-Sagoe, C. (2022). Mitigating the risks of climate variability and change on mango seedlings in Ghana: Evidence from mango seedlings producers in the Yilo Krobo Municipality. *Environmental Challenges*, *8*, 100594. https://doi.org/10.1016/j.envc.2022.100594

Asha, J., Rishidas, S., SanthoshKumar, S., & Reena, P. (2020). *Analysis of Temperature Prediction Using Random Forest and Facebook Prophet Algorithms* (pp. 432–439).

https://doi.org/10.1007/978-3-030-38040-3_49

Aslan, M. F., Sabanci, K., & Aslan, B. (2024). Artificial Intelligence Techniques in Crop Yield Estimation Based on Sentinel-2 Data: A Comprehensive Survey. *Sustainability*, *16*(18), 8277. https://doi.org/10.3390/su16188277

Badugu, A., Arunab, K. S., & Mathew, A. (2024). Predicting land surface temperature using data-driven approaches for urban heat island studies: a comparative analysis of correlation with environmental parameters. *Modeling Earth Systems and Environment*, *10*(1), 1043–1076. https://doi.org/10.1007/s40808-023-01822-2

Baran, Á., Lerch, S., El Ayari, M., & Baran, S. (2021). Machine learning for total cloud cover prediction. *Neural Computing and Applications*, *33*(7), 2605–2620. https://doi.org/10.1007/s00521-020-05139-4

Bayes, T., & Price, R. (1763). An Essay towards Solving a Problem in the Doctrine of Chances By the Late Rev. Mr. Bayes. *Philosophical Transactions (1683-1775)*.

Beddington, J. R., Asaduzzaman, M., Bremauntz, F. A., Clark, M. E., & Guillou, M. (2012). *Achieving food security in the face of climate change: final report from the Commission on Sustainable Agriculture and Climate Change.*

Bick, C., Böhle, T., & Kuehn, C. (2022). Multi-population phase oscillator networks with higher-order interactions. *Nonlinear Differential Equations and Applications NoDEA*, *29*(6), 64. https://doi.org/10.1007/s00030-022-00796-x

Bloomfield, P. (2000). *Fourier Analysis of Time Series*. Wiley. https://doi.org/10.1002/0471722235

Borsuk, M. E., Stow, C. A., & Reckhow, K. H. (2004). A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. *Ecological Modelling*. https://doi.org/10.1016/j.ecolmodel.2003.08.020

Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., Romero, J., Aldunce, P., Barret, K., Blanco, G., Cheung, W. W. L., Connors, S. L., Denton, F., Diongue-Niang, A., Dodman, D., Garschagen, M., Geden, O., Hayward, B., Jones, C., … Ha, M. (2023a). *IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPC* (P. Arias, M. Bustamante, I. Elgizouli, G. Flato, M. Howden, C. Méndez-Vallejo, J. J. Pereira, R. Pichs-Madruga, S. K. Rose, Y. Saheb, R. Sánchez Rodríguez, D. Ürge-Vorsatz, C. Xiao, N. Yassaa, J. Romero, J. Kim, E. F. Haites, Y. Jung, R. Stavins, … Y. Park (Eds.)). https://doi.org/10.59327/IPCC/AR6-

9789291691647.001

Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., Romero, J., Aldunce, P., Barrett, K., Blanco, G., Cheung, W. W. L., Connors, S., Denton, F., Diongue-Niang, A., Dodman, D., Garschagen, M., Geden, O., Hayward, B., Jones, C., … Ha, M. (2023b). *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.* (P. Arias, M. Bustamante, I. Elgizouli, G. Flato, M. Howden, C. Méndez-Vallejo, J. J. Pereira, R. Pichs-Madruga, S. K. Rose, Y. Saheb, R. Sánchez Rodríguez, D. Ürge-Vorsatz, C. Xiao, N. Yassaa, J. Romero, J. Kim, E. F. Haites, Y. Jung, R. Stavins, … C. Péan (Eds.)). https://doi.org/10.59327/IPCC/AR6-9789291691647

Camargo-Alvarez, H., Elliott, R. J. R., Olin, S., Wang, X., Wang, C., Ray, D. K., & Pugh, T. A. M. (2023). Modelling crop yield and harvest index: the role of carbon assimilation and allocation parameters. *Modeling Earth Systems and Environment*, *9*(2), 2617–2635. https://doi.org/10.1007/s40808-022-01625-x

Charniak, E. (1991). *Bayesian networks without tears. AI Mag. 12, 50.*

Chen, T., & Guestrin, C. (2016a). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, T., & Guestrin, C. (2016b). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cleugh, H. A., Miller, J. M., & Böhm, M. (1998). Direct mechanical effects of wind on crops. *Agroforestry Systems*, *41*(1), 85–112. https://doi.org/10.1023/A:1006067721039

Clivaz, C., & Savioz, A. (2020). Glacier retreat and perception of climate change by local tourism stakeholders: the case of Chamonix-Mont-Blanc in the French Alps. *Via Tourism Review*, *18*. https://doi.org/10.4000/viatourism.6097

Collaboration, P. A. (2022). Searches for Ultra-High-Energy Photons at the Pierre Auger Observatory. *Universe*, *8*(11), 579.

Collaboration, T. P. A. (2013). *The Next Frontier in UHECR Research with an Upgraded Pierre Auger Observatory*. https://api.semanticscholar.org/CorpusID:118412968

Cornara, L., Xiao, J., Smeriglio, A., Trombetta, D., & Burlando, B. (2020). Emerging Exotic Fruits: New Functional Foods in the European Market. *EFood*, *1*(2), 126–139.

https://doi.org/10.2991/efood.k.200406.001

Dara, S., Srinivasulu, C., Babu, C. M., Ravuri, A., Paruchuri, T., Kilak, A. S., & Vidyarthi, A. (2023). Context-Aware Auto-Encoded Graph Neural Model for Dynamic Question Generation using NLP. *ACM Transactions on Asian and Low-Resource Language Information Processing*. https://doi.org/10.1145/3626317

Dave, E., Leonardo, A., Jeanice, M., & Hanafiah, N. (2021). Forecasting Indonesia Exports using a Hybrid Model ARIMA-LSTM. *Procedia Computer Science*, *179*, 480–487. https://doi.org/10.1016/j.procs.2021.01.031

De Santis, D., Guarcello, C., Spagnolo, B., Carollo, A., & Valenti, D. (2024). Noise-induced, ac-stabilized sine-Gordon breathers: Emergence and statistics. *Communications in Nonlinear Science and Numerical Simulation*, *131*, 107796. https://doi.org/10.1016/j.cnsns.2023.107796

"Department of Sicilian Agriculture." (2017). Statistical Data on Tropical Fruits. *Viale Regione Siciliana n.2771, Palermo, Italy.*

Dlamini, L., Crespo, O., van Dam, J., & Kooistra, L. (2023). A Global Systematic Review of Improving Crop Model Estimations by Assimilating Remote Sensing Data: Implications for Small-Scale Agricultural Systems. *Remote Sensing*, *15*(16), 4066. https://doi.org/10.3390/rs15164066

Do Nascimento Camelo, H., Sérgio Lucio, P., Verçosa Leal Junior, J., Von Glehn dos Santos, D., & Cesar Marques de Carvalho, P. (2018). Innovative Hybrid Modeling of Wind Speed Prediction Involving Time-Series Models and Artificial Neural Networks. *Atmosphere*, *9*(2), 77. https://doi.org/10.3390/atmos9020077

Dorner, S., Shi, J., & Swayne, D. (2007). Multi-objective modelling and decision support using a Bayesian network approximation to a non-point source pollution model. *Environmental Modelling and Software*. https://doi.org/10.1016/j.envsoft.2005.07.020

Dos Santos Moreira, D., Nicolosi, A., Laganà, V. R., Di Gregorio, D., & Agosteo, G. E. (2024). Factors Driving Consumption Preferences for Fresh Mango and Mango-Based Products in Italy and Brazil. *Sustainability*, *16*(21), 9401. https://doi.org/10.3390/su16219401

ECMWF. (n.d.). *European Centre for Medium-Range Weather Forecasts*. https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

European Parliament. (2023). *Research for AGRI Committee: The impact of extreme climate events on agricultural production in the European Union*. https://doi.org/https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2023

)733115

FAO. (2016). *Climate change and food security: risks and responses*. https://www.fao.org/3/a-i5188e.pdf

FAO. (2022a). *FAO STRATEGY ON CLIMATE CHANGE*.

FAO. (2022b). *Major Tropical Fruits: Preliminary results 2021. Rome*. https://doi.org/https://www.fao.org/3/cb9412en/cb9412en.pdf

FAO. (2023a). *Major Tropical Fruits Market Review – Preliminary results*. *Rome*. https://openknowledge.fao.org/server/api/core/bitstreams/852265a4-9006-4d54-a792-51f1d9c44673/content

FAO. (2023b). *Major Tropical Fruits Market Review – Preliminary results 2022. Rome*. https://doi.org/https://www.fao.org/3/cc3939en/cc3939en.pdf

Farina, V., Corona, O., Mineo, V., D'Asaro, A., & Barone, F. (2013). Qualitative characteristics of Mango fruits (Mangifera indica L.), which have undergone preservation. *Acta Italus Hortus*, *12*, 70–73.

Farina, V., Gentile, C., Sortino, G., Gianguzzi, G., Palazzolo, E., & Mazzaglia, A. (2020). Tree-Ripe Mango Fruit: Physicochemical Characterization, Antioxidant Properties and Sensory Profile of Six Mediterranean-Grown Cultivars. *Agronomy*, *10*(6), 884. https://doi.org/10.3390/agronomy10060884

Farina, V., Tripodo, L., Gianguzzi, G., Sortino, G., Giuffrè, D., Cicero, U. L., Candia, R., & Collura, A. (2017). Innovative Techniques to Reduce Chilling Injuries in Mango (Mangifera Indica L.) Trees under Mediterranean Climate. *Chemical Engineering Transactions*. https://doi.org/https://doi.org/10.3303/CET1758138

Farmani, R., Henriksen, H. J., & Savic, D. (2009). An evolutionary Bayesian belief network methodology for optimum management of groundwater contamination. *Environmental Modelling & Software*, *24*(3), 303–310. https://doi.org/10.1016/j.envsoft.2008.08.005

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V., & Müller, J. (2013). Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agricultural Water Management*, *116*, 142–150. https://doi.org/10.1016/j.agwat.2012.07.003

Gao, X., Bai, Y., Huo, Z., Xu, X., Huang, G., Xia, Y., & Steenhuis, T. S. (2017). Deficit irrigation enhances contribution of shallow groundwater to crop water consumption in arid area. *Agricultural Water Management*, *185*, 116–125. https://doi.org/10.1016/j.agwat.2017.02.012

Gardiner, C. W. (1986). Handbook of stochastic methods - for physics, chemistry and the

natural sciences, Second Edition. *Springer Series in Synergetics*. https://api.semanticscholar.org/CorpusID:54159549

Gentile, C., Di Gregorio, E., Di Stefano, V., Mannino, G., Perrone, A., Avellone, G., Sortino, G., Inglese, P., & Farina, V. (2019). Food quality and nutraceutical value of nine cultivars of mango (Mangifera indica L.) fruits grown in Mediterranean subtropical environment. *Food Chemistry*, *277*, 471–479. https://doi.org/10.1016/j.foodchem.2018.10.109

Gonzalez-Velasco, E. A. (1992). Connections in Mathematical Analysis: The Case of Fourier Series. *The American Mathematical Monthly*, *99*(5), 427. https://doi.org/10.2307/2325087

Graves, A., & Schmidhuber, J. (n.d.). Framewise phoneme classification with bidirectional LSTM networks. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, *4*, 2047–2052. https://doi.org/10.1109/IJCNN.2005.1556215

Grimaudo, R., Lazzari, P., Solidoro, C., & Valenti, D. (2022). Effects of solar irradiance noise on a complex marine trophic web. *Scientific Reports*, *12*(1), 12163. https://doi.org/10.1038/s41598-022-16236-w

Gugliuzza, G., Scuderi, D., & Farina, V. (2023). Fruit quality and photosynthetic response of three cultivars of mango (Mangifera indica L.) in a greenhouse in south of Italy. *Acta Horticulturae*, *1372*, 259–266. https://doi.org/10.17660/ActaHortic.2023.1372.34

Gutierrez, B. T., Plant, N. G., & Thieler, E. R. (2011). A Bayesian network to predict coastal vulnerability to sea level rise. *Journal of Geophysical Research: Earth Surface*. https://doi.org/10.1029/2010JF001891

Harsányi, E. (2025). Predicting agricultural drought in central Europe by using machine learning algorithms. *Journal of Agriculture and Food Research*, *20*, 101783. https://doi.org/10.1016/j.jafr.2025.101783

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Heideman, M., Johnson, D., & Burrus, C. (1984). Gauss and the history of the fast fourier transform. *IEEE ASSP Magazine*, *1*(4), 14–21. https://doi.org/10.1109/MASSP.1984.1162257

Hening, A., & Li, Y. (2020). *Stationary distributions of persistent ecological systems*. http://arxiv.org/abs/2003.04398

Horn, B. K. P. (1981). Hill shading and the reflectance map. *Proceedings of the IEEE*, *69*(1),

14–47. https://doi.org/10.1109/PROC.1981.11918

Huang, C., & Petukhina, A. (2022). *Applied Time Series Analysis and Forecasting with Python*. Springer International Publishing. https://doi.org/10.1007/978-3-031-13584-2

Huege, T. (2023). The Radio Detector of the Pierre Auger Observatory – status and expected performance. *EPJ Web of Conferences*, *283*, 06002. https://doi.org/10.1051/epjconf/202328306002

Jain, A. K., Jianchang Mao, & Mohiuddin, K. M. (1996). Artificial neural networks: a tutorial. *Computer*, *29*(3), 31–44. https://doi.org/10.1109/2.485891

Jamal, A., Cai, X., Qiao, X., Garcia, L., Wang, J., Amori, A., & Yang, H. (2023). Real-Time Irrigation Scheduling Based on Weather Forecasts, Field Observations, and Human-Machine Interactions. *Water Resources Research*, *59*(12). https://doi.org/10.1029/2023WR035810

Jha, M. N., Kumar, A., Dubey, S., & Pandey, A. (2022). *Yield Estimation of Rice Crop Using Semi-Physical Approach and Remotely Sensed Data* (pp. 331–349). https://doi.org/10.1007/978-3-030-98981-1_15

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, *147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Kampert, K.-H., Alejandro Mostafa, M., & Zas, E. (2019). Multi-Messenger Physics With the Pierre Auger Observatory. *Frontiers in Astronomy and Space Sciences*, *6*. https://doi.org/10.3389/fspas.2019.00024

Karaman, Ö. A. (2023). Prediction of Wind Power with Machine Learning Models. *Applied Sciences*, *13*(20), 11455. https://doi.org/10.3390/app132011455

Khalifa, S., & Abobatta, W. (2023). Climate Changes and Mango Production (Temperature). *IgMin Research*, *1*(1), 043–046. https://doi.org/10.61927/igmin115

Kragt, M. E. (2009). A beginners guide to Bayesian network modelling for integrated catchment management. *Landscape Logic*.

Kulkarni, S. A. (2011). *Innovative Technologies for Water Saving in Irrigated Agriculture*. https://api.semanticscholar.org/CorpusID:132903575

Kumar, A., & Elumalai, S. P. (2023). Application of artificial neural network to screen out the dominant meteorological parameters for prediction of air temperature. *Earth Science Informatics*, *16*(4), 3697–3716. https://doi.org/10.1007/s12145-023-01107-3

Lakshminarayana, S. V. (2020). Rainfall Forecasting using Artificial Neural Networks (ANNs): A Comprehensive Literature Review. *Indian Journal of Pure & Applied*

Biosciences, *8*(4), 589–599. https://doi.org/10.18782/2582-2845.8250

Lawrence, M. A., Reid, R. J. O., & Watson, A. A. (1991). The cosmic ray energy spectrum above 4*10 17 eV as measured by the Haverah Park array. *Journal of Physics G: Nuclear and Particle Physics*, *17*(5), 733–757. https://doi.org/10.1088/0954-3899/17/5/019

Lazzari, P., Grimaudo, R., Solidoro, C., & Valenti, D. (2021). Stochastic 0-dimensional Biogeochemical Flux Model: Effect of temperature fluctuations on the dynamics of the biogeochemical properties in a marine ecosystem. *Communications in Nonlinear Science and Numerical Simulation*, *103*, 105994. https://doi.org/10.1016/j.cnsns.2021.105994

Levy, O., Lee, K., FitzGerald, N., & Zettlemoyer, L. (2018). Long Short-Term Memory as a Dynamically Computed Element-wise Weighted Sum. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 732–739. https://doi.org/10.18653/v1/P18-2116

Li, Q., & Hu, G. (2020). Multistage stochastic programming modeling for farmland irrigation management under uncertainty. *PLOS ONE*, *15*(6), e0233723. https://doi.org/10.1371/journal.pone.0233723

Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, *150*(11), 1443–1452. https://doi.org/10.1016/j.agrformet.2010.07.008

Mantegna, R. N., & Spagnolo, B. (1995). Stochastic resonance in a tunnel diode in the presence of white or coloured noise. *Il Nuovo Cimento D*, *17*(7–8), 873–881. https://doi.org/10.1007/BF02451845

Manwell, J. F., McGowan, J. G., & Rogers, A. L. (2009). *Wind Energy Explained*. Wiley. https://doi.org/10.1002/9781119994367

Mentzel, S., Grung, M., Holten, R., Tollefsen, K. E., Stenrød, M., & Moe, S. J. (2022). Probabilistic risk assessment of pesticides under future agricultural and climate scenarios using a bayesian network. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.957926

MODIS. (n.d.-a). *Moderate Resolution Imaging Spectroradiometer*. https://modis.gsfc.nasa.gov/about/

MODIS. (n.d.-b). *MODIS Land Surface Temperature and Emissivity*. https://modis.gsfc.nasa.gov/data/dataprod/mod11.php

Montomoli, J., Romeo, L., Moccia, S., Bernardini, M., Migliorelli, L., Berardini, D., Donati,

A., Carsetti, A., Bocci, M. G., Wendel Garcia, P. D., Fumeaux, T., Guerci, P., Schüpbach, R. A., Ince, C., Frontoni, E., Hilty, M. P., Alfaro-Farias, M., Vizmanos-Lamotte, G., Tschoellitsch, T., … Colak, E. (2021). Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *Journal of Intensive Medicine*, *1*(2), 110–116. https://doi.org/10.1016/j.jointm.2021.09.002

Naresh, D. R. K. (Ed.). (2019). *Research Trends in Agriculture Sciences*. AkiNik Publications. https://doi.org/10.22271/ed.book.390

Nguyen, H. A. T., Sophea, T., Gheewala, S. H., Rattanakom, R., Areerob, T., & Prueksakorn, K. (2021). Integrating remote sensing and machine learning into environmental monitoring and assessment of land use change. *Sustainable Production and Consumption*, *27*, 1239–1254. https://doi.org/10.1016/j.spc.2021.02.025

NOAA. (n.d.). *National Oceanic and Atmospheric Administration*. https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast

Normand, F., Lauri, P.-E., & Legave, J.-M. (2015). CLIMATE CHANGE AND ITS PROBABLE EFFECTS ON MANGO PRODUCTION AND CULTIVATION. *Acta Horticulturae*, *1075*, 21–31. https://doi.org/10.17660/ActaHortic.2015.1075.1

Occhipinti, G., Piani, S., & Lazzari, P. (2024). Stochastic effects on plankton dynamics: Insights from a realistic 0-dimensional marine biogeochemical model. *Ecological Informatics*, *83*, 102778. https://doi.org/10.1016/j.ecoinf.2024.102778

Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Hybrid Deep Learning-based Models for Crop Yield Prediction. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2022.2031823

Paldor, N., & Friedland, L. (2023). Extension of Ekman (1905) wind-driven transport theory to the β plane. *Ocean Science*, *19*(1), 93–100. https://doi.org/10.5194/os-19-93-2023

Papacharalampous, G., Tyralis, H., Doulamis, A., & Doulamis, N. (2023). Comparison of Tree-Based Ensemble Algorithms for Merging Satellite and Earth-Observed Precipitation Data at the Daily Time Scale. *Hydrology*, *10*(2), 50. https://doi.org/10.3390/hydrology10020050

Pearl, J. (2011). *Bayesian networks. Dep. Stat. UCLA.*

Pereira, L. S., Allen, R. G., Smith, M., & Raes, D. (2015). Crop evapotranspiration estimation with FAO56: Past and future. *Agricultural Water Management*, *147*, 4–20. https://doi.org/10.1016/j.agwat.2014.07.031

Pham, H. V., Dal Barco, M. K., Pourmohammad Shahvar, M., Furlan, E., Critto, A., &

Torresan, S. (2024). Bayesian Network Analysis for Shoreline Dynamics, Coastal Water Quality, and Their Related Risks in the Venice Littoral Zone, Italy. *Journal of Marine Science and Engineering*, *12*(1), 139. https://doi.org/10.3390/jmse12010139

Poffenbarger, H. J., Barker, D. W., Helmers, M. J., Miguez, F. E., Olk, D. C., Sawyer, J. E., Six, J., & Castellano, M. J. (2017). Maximum soil organic carbon storage in Midwest U.S. cropping systems when crops are optimally nitrogen-fertilized. *PLOS ONE*, *12*(3), e0172293. https://doi.org/10.1371/journal.pone.0172293

Pollino, C. A., & Henderson, C. (2010). A guide for their application in natural resource management and policy. *A Technical Report No. 14. Integrated Catchment Assessment and Management Centre, Fenner School of Environment and Society, Australian National University, Canberra*.

Pope, S. B. (2000). *Turbulent Flows*. Cambridge University Press. https://doi.org/10.1017/CBO9780511840531

Pourmohammad Shahvar, M. (2021). *Assessing the impact of sea-level rise in the coast of Venice: a multi-risk approach supporting climate adaptation and environmental risk management*. http://dspace.unive.it/handle/10579/19456

POURMOHAMMAD SHAHVAR, M., Scuderi, D., Tripodo, G., Farina, V., Micchichè, S., Collura, A., Marsella, G., & others. (2023). Climate change multi-risk assessment for mango cultivation in Sicily, Italy by using bayesian network. *XIII International Mango Symposium Book of Abstracts*, 46.

Pourmohammad Shahvar, M., Valenti, D., Collura, A., Micciche, S., Farina, V., & Marsella, G. (2025). An Integrated Hybrid-Stochastic Framework for Agro-Meteorological Prediction Under Environmental Uncertainty. *Stats*, *8*(2). https://doi.org/10.3390/stats8020030

Pulighe, G., Di Fonzo, A., Gaito, M., Giuca, S., Lupia, F., Bonati, G., & De Leo, S. (2024). Climate change impact on yield and income of Italian agriculture system: a scoping review. *Agricultural and Food Economics*, *12*(1), 23. https://doi.org/10.1186/s40100-024-00317-7

Ratkowsky, D. A., Lowry, R. K., McMeekin, T. A., Stokes, A. N., & Chandler, R. E. (1983). Model for bacterial culture growth rate throughout the entire biokinetic temperature range. *Journal of Bacteriology*, *154*(3), 1222–1226. https://doi.org/10.1128/jb.154.3.1222-1226.1983

Robson, A., Rahman, M., & Muir, J. (2017). Using Worldview Satellite Imagery to Map Yield in Avocado (Persea americana): A Case Study in Bundaberg, Australia. *Remote*

Sensing, *9*(12), 1223. https://doi.org/10.3390/rs9121223

Rodrìguez, J. D., Aritz Pérez, A., & Lozano, J. A. (2010). *Sensitivity Analysis of k -Fold Cross Validation in Prediction Error Estimation. 32(3), 569–575.*

Roth, M. (2007). *Measurement of the UHECR energy spectrum using data from the Surface Detector of the Pierre Auger Observatory*. http://arxiv.org/abs/0706.2096

Saleh, E., Jolis, G., Osman, N. F., Sentian, J., Joseph, J., Jomitol, J., & Adin, N. (2022). Beach erosion: Threat and adaptation measures of communities in the Tun Mustapha Park (TMP), Sabah, Malaysia. *IOP Conference Series: Earth and Environmental Science*, *1103*(1), 012034. https://doi.org/10.1088/1755-1315/1103/1/012034

Sato, R., Abdul Halim, A., Abreu, P., Aglietta, M., Allekotte, I., Almeida Cheminant, K., Almela, A., Aloisio, R., Alvarez-Muniz, J., Ammerman Yebra, J., Anastasi, G. A., Anchordoqui, L. A., Andrada, B., Andringa, S., Aramo, C., Araújo Ferreira, P. R., Arnone, E., Arteaga Velazquez, J. C., Asorey, H. G., … Zavrtanik, M. (2023). AugerPrime implementation in the DAQ systems of the Pierre Auger Observatory. *Proceedings of 38th International Cosmic Ray Conference — PoS(ICRC2023)*, 373. https://doi.org/10.22323/1.444.0373

Scotti, M., Gjata, N., Livi, C., & Jordán, F. (2012). Dynamical effects of weak trophic interactions in a stochastic food web simulation. *Community Ecology*, *13*(2), 230–237. https://doi.org/10.1556/ComEc.13.2012.2.13

Scuderi, D., Pourmohammad Shahvar, M., Marsella, G., Farina, V., Lobo Rodrigo, M. G., & Normand, F. (2025). The climate of mango producing areas: a case study on three islands. *Acta Horticulturae*, *1415*, 25–32. https://doi.org/10.17660/ActaHortic.2025.1415.3

Scutari, M. (2017). Bayesian network constraint-based structure learning algorithms: Parallel and optimized implementations in the bnlearn R package. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v077.i02

Searches for Ultra-High-Energy Photons at the Pierre Auger Observatory. (2022). *Universe*, *8*(11), 579. https://doi.org/10.3390/universe8110579

Shafiq, M., & Gu, Z. (2022). Deep Residual Learning for Image Recognition: A Survey. *Applied Sciences*, *12*(18), 8972. https://doi.org/10.3390/app12188972

Shahvar, M. Pourmohammad, Scuderi, D., Valenti, D., Collura, A., Miccichè, S., Farina, V., & Marsella, G. (2025). MISAR in enhancing agricultural resilience: a comprehensive approach to climate change risk management for mango farms in Sicily, Italy. *Acta Horticulturae*, *1*(1415), 145–154. https://doi.org/10.17660/ActaHortic.2025.1415.16

Shahvar, Mohsen.P, Buscemi, M., Incardona, S., Tripodo, G., & Marsella, G. (2022). MISAR: Proposal for “Climate Change Risk Management by improving the Individual and Social Awareness of Risk in Sicily.” *GARR-Conf22*, 31. https://doi.org/https://doi.org/10.26314/GARR-Conf22-proceedings-05

Shalu, & Gurjeet Singh. (2023). ENVIRONMENTAL MONITORING WITH MACHINE LEARNING. *EPRA International Journal of Multidisciplinary Research (IJMR)*, 208–212. https://doi.org/10.36713/epra13330

Shin, J.-Y., Min, B., & Kim, K. R. (2022). High-resolution wind speed forecast system coupling numerical weather prediction and machine learning for agricultural studies — a case study from South Korea. *International Journal of Biometeorology*, *66*(7), 1429–1443. https://doi.org/10.1007/s00484-022-02287-1

Shiri, F., Perumal, T., Mustapha, N., & Mohamed, R. (2023). A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU. *ArXiv*, *abs/2305.1*. https://api.semanticscholar.org/CorpusID:258960275

Singh, B., Sahoo, A., Sharma, R., & Bhat, T. K. (2005). Effect of polethylene glycol on gas production parameters and nitrogen disappearance of some tree forages. *Animal Feed Science and Technology*, *123–124*, 351–364. https://doi.org/10.1016/j.anifeedsci.2005.04.033

Sishodia, R. P., Ray, R. L., & Singh, S. K. (2020). Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sensing*, *12*(19), 3136. https://doi.org/10.3390/rs12193136

SMAP. (n.d.). *Soil Moisture Active Passive*. https://smap.jpl.nasa.gov/data/

Sperotto, A., Molina, J.-L., Torresan, S., Critto, A., & Marcomini, A. (2017). Reviewing Bayesian Networks potentials for climate change impacts assessment and management: A multi-risk perspective. *Journal of Environmental Management*, *202*, 320–331. https://doi.org/10.1016/j.jenvman.2017.07.044

SRTM. (n.d.). *Shuttle Radar Topography Mission*. https://srtm.csi.cgiar.org/

Staudemeyer, R. C., & Morris, E. R. (2019). *Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks*. http://arxiv.org/abs/1909.09586

Stull, R. B. (Ed.). (1988). *An Introduction to Boundary Layer Meteorology*. Springer Netherlands. https://doi.org/10.1007/978-94-009-3027-8

T.G.S. (1988). M. Ghil, & S. Childress 1987. Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics. Applied Mathematical Sciences. Volume 60 xv + 485 pp. New York, Berlin, Heidelberg,

London, Paris, Tokyo: Springer-Verlag. P. *Geological Magazine*, *125*(2), 190–191. https://doi.org/10.1017/S0016756800009626

Tarolli, P., Luo, J., Straffelini, E., Liou, Y.-A., Nguyen, K.-A., Laurenti, R., Masin, R., & D'Agostino, V. (2023). Saltwater intrusion and climate change impact on coastal agriculture. *PLOS Water*, *2*(4), e0000121. https://doi.org/10.1371/journal.pwat.0000121

Testa, R., Tudisca, S., Schifani, G., Di Trapani, A., & Migliore, G. (2018). Tropical Fruits as an Opportunity for Sustainable Development in Rural Areas: The Case of Mango in Small-Sized Sicilian Farms. *Sustainability*, *10*(5), 1436. https://doi.org/10.3390/su10051436

The Pierre Auger Collaboration. (2005). *First Estimate of the Primary Cosmic Ray Energy Spectrum above 3 EeV from the Pierre Auger Observatory*. http://arxiv.org/abs/astro-ph/0507150

The Pierre Auger Collaboration, Abraham, J., Abreu, P., Aglietta, M., Aguirre, C., Ahn, E. J., Allard, D., Allekotte, I., Allen, J., Allison, P., Alvarez-Muñiz, J., Ambrosio, M., Anchordoqui, L., Andringa, S., Anzalone, A., Aramo, C., Arganda, E., Argirò, S., Arisaka, K., … Ziolkowski, M. (2009). *The Fluorescence Detector of the Pierre Auger Observatory*. https://doi.org/10.1016/j.nima.2010.04.023

The Pierre Auger Collaboration, Halim, A. A., Abreu, P., Aglietta, M., Allekotte, I., Cheminant, K. A., Almela, A., Aloisio, R., Alvarez-Muñiz, J., Yebra, J. A., Anastasi, G. A., Anchordoqui, L., Andrada, B., Andringa, S., Anukriti, Aramo, C., Ferreira, P. R. A., Arnone, E., Velázquez, J. C. A., … Šmída, R. (2023). *AugerPrime Surface Detector Electronics*. http://arxiv.org/abs/2309.06235

Torgbor, B. A., Rahman, M. M., Brinkhoff, J., Sinha, P., & Robson, A. (2023). Integrating Remote Sensing and Weather Variables for Mango Yield Prediction Using a Machine Learning Approach. *Remote Sensing*, *15*(12), 3075. https://doi.org/10.3390/rs15123075

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150. https://doi.org/10.1016/0034-4257(79)90013-0

UNDRR. (2020). Hazard Definition & classification review. *Hazard Definition & Classification Reviewazard Definition & Classification Review*.

Valenti, D., Denaro, G., Giarratana, F., Giuffrida, A., Mazzola, S., Basilone, G., Aronica, S., Bonanno, A., & Spagnolo, B. (2016). Modeling of Sensory Characteristics Based on the Growth of Food Spoilage Bacteria. *Mathematical Modelling of Natural Phenomena*, *11*(5), 119–136. https://doi.org/10.1051/mmnp/201611508

Valenzuela, H. R. (1999). Ecologically-Based Practices for Vegetable Crops Production in the Tropics. In *Horticultural Reviews* (pp. 139–228). Wiley. https://doi.org/10.1002/9780470650776.ch4

Valiño, I., Alvarez-Muñiz, J., Roth, M., Vazquez, R. A., & Zas, E. (2010). Characterisation of the electromagnetic component in ultra-high energy inclined air showers. *Astroparticle Physics*, *32*(6), 304–317. https://doi.org/10.1016/j.astropartphys.2009.09.008

Valipour, M., Khoshkam, H., Bateni, S. M., Jun, C., & Band, S. S. (2023). Hybrid machine learning and deep learning models for multi-step-ahead daily reference evapotranspiration forecasting in different climate regions across the contiguous United States. *Agricultural Water Management*, *283*, 108311. https://doi.org/10.1016/j.agwat.2023.108311

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & others. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Yahia, E. M., De Jesus Ornelas-Paz, J., & Gonzalez-Aguilar, G. A. (2011). Nutritional and health-promoting properties of tropical and subtropical fruits. In *Postharvest Biology and Technology of Tropical and Subtropical Fruits* (pp. 21–78). Elsevier. https://doi.org/10.1533/9780857093622.21

Yan, Z., & Li, M. (2018). A Stochastic Optimization Model for Agricultural Irrigation Water Allocation Based on the Field Water Cycle. *Water*, *10*(8), 1031. https://doi.org/10.3390/w10081031

Yet, B., Lamanna, C., Shepherd, K. D., & Rosenstock, T. S. (2020). Evidence-based investment selection: Prioritizing agricultural development investments under climatic and socio-political risk using Bayesian networks. *PLOS ONE*, *15*(6), e0234213. https://doi.org/10.1371/journal.pone.0234213

Yu, C., & Ma, Y. (2024). A novel model for wind speed point prediction and quantifying uncertainty in wind farms. *Electrical Engineering*. https://doi.org/10.1007/s00202-024-02874-y

Zeynoddin, M., Gumiere, S. J., & Bonakdari, H. (2023). Enhancing water use efficiency in precision irrigation: data-driven approaches for addressing data gaps in time series. *Frontiers in Water*, *5*. https://doi.org/10.3389/frwa.2023.1237592

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian

optimization. *Geoscience Frontiers*, *12*(1), 469–477. https://doi.org/10.1016/j.gsf.2020.03.007

ANNEX 1

mathematics

MDPI

Article

A Potential Hybrid Deep Learning Approach to Temperature Prediction Using MODIS Satellite Data and Historical Records

Mohsen Pourmohammad Shahvar [1,*], Davide Valenti[1], Alfonso Collura [3], Salvatore Micciche[1], Vittorio Farina[2], and Giovanni Marsella[1]

1 Dipartimento di Fisica e Chimica "E. Segrè", Università degli Studi di Palermo, Italy; mohsen.pourmohammadshahvar@unipa.it, davide.valenti@unipa.it, salvatore.micciche@unipa.it, giovanni.marsella@unipa.it.
2 Dipartimento di Scienze Agrarie, Alimentari e Forestali, Università degli Studi di Palermo, Italy; vittorio.farina@unipa.it
3 Istituto Nazionale di Astrofisica, Osservatorio Astronomico di Palermo, Italy; alfonso.collura@inaf.it.
* Correspondence: mohsen.pourmohammadshahvar@unipa.it

Abstract: This study presents a hybrid mathematical framework for daily air temperature forecasting, integrating satellite-derived remote sensing data with time-series modeling. The proposed architecture combines a Residual Neural Network (ResNet) for spatial feature extraction from MODIS imagery, XGBoost and Random Forest for multivariate regression, and an ARIMA model for residual temporal correction. Applied to northeastern Sicily, a climate-sensitive region for mango agriculture, the model is trained on data from 2007–2021 and tested on unseen years (2022–2024). Results demonstrate high forecasting accuracy ($R^2 > 0.97$; RMSE < 0.5 in 2022), surpassing Transformer-based baselines. The integration of statistical and deep learning components enables robust handling of nonlinearity, autocorrelation, and seasonal variation. This approach exemplifies the value of mathematical modeling in environmental prediction, offering a scalable method for climate adaptation in precision agriculture.

Keywords: Hybrid Temperature Forecasting; ResNet CNN & Satellite Imagery; XGBoost & ARIMA Integration; Agricultural Climate Modeling; Mango Crop Protection

1. Introduction

Mango cultivation is highly sensitive to temperature fluctuations, which can severely impact crop yield, fruit quality, and flowering cycles [1, 2]. In areas like Acquedolci and Caronia, Sicily, where there are numerous mango plantations, having potential temperature forecasts is crucial for effectively managing farms and protecting crops [1–3]. Traditional weather forecasting methods, which rely primarily on ground-based observations, often lack the spatial resolution needed for detailed agricultural applications [4]. To fill this gap, we suggest combining satellite imagery with historical temperature data using advanced deep learning techniques.

Remote sensing technology, especially MODIS Terra and Aqua satellite imagery, offers wide temporal coverage and fine spatial resolution. These features make it perfect for environmental monitoring and agricultural forecasting. The challenge, however, is to process and integrate this data effectively with existing ground observations [5].

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

Citation: To be added by editorial staff during production.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

The introduction of deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized image processing by enabling the automated extraction of spatial features. ResNet (Residual Network), known for addressing vanishing gradient issues in deep networks, has emerged as a robust framework for extracting hierarchical features, particularly from large-scale satellite imagery datasets [6,7]. CNN-based architectures like ResNet remain advantageous for extracting spatial patterns from satellite imagery [7,8].

In addition to spatial modeling, recent advances in temporal sequence modeling—such as Long Short-Term Memory (LSTM) and Transformer-based networks—have shown strong performance in capturing temporal dependencies in meteorological time series, including temperature and precipitation forecasting [9–12]. These models are particularly valuable in sequential prediction tasks, though they often require high-quality, uninterrupted time series inputs.

A growing body of research supports the integration of deep learning and classical machine learning techniques in the form of hybrid models. For example, Ref.[14] applied artificial neural networks for rainfall prediction, while Ref.[13] demonstrated the flexibility of neural networks in learning meteorological patterns [13,14].

More recently, several studies have shown that hybrid deep learning models combining CNNs with decision-tree algorithms such as XGBoost and Random Forest can significantly enhance prediction accuracy for weather-related applications [15–18]. These hybrid models leverage CNNs for spatial feature extraction and tree-based models for capturing nonlinear interactions and variable importance [10].

In particular, ResNet-XGBoost combinations have shown promising results in both environmental and energy domains due to their interpretability and performance. Ref.[19] demonstrated the power of CNN-LSTM hybrids in time series modeling of climate data, particularly in volatile, data-sparse regions [19,20]. These hybrid models help bridge the gap between static image-based prediction and dynamic temporal modeling, making them well-suited for agro-environmental forecasting.

Our study builds on this hybrid paradigm by introducing a novel ensemble of ResNet for spatial encoding, XGBoost for decision-level modeling, and ARIMA for residual temporal correction. The integration strategy draws inspiration from similar hybridization efforts in astroparticle physics, where a CNN-XGBoost pipeline was effectively applied for identifying inclined muons in water Cherenkov detectors [20]. This interdisciplinary application underscores the generalizability of hybrid architectures across domains.

The inclusion of ARIMA in our pipeline is particularly useful for modeling residual autocorrelation in temperature time series, enhancing the temporal robustness of the model. This layered architecture ensures that spatial, nonlinear, and sequential dynamics are captured holistically. Recent work by Ref.[17] in solar radiation forecasting corroborates our hybrid design by demonstrating significant performance gains using similar DL-ML fusion frameworks [17].

From a regional application perspective, accurate temperature forecasting in locations like Acquedolci and Caronia (Messina region) is vital for managing the risks posed by temperature extremes on mango cultivation [1]. Ref.[21] highlighted how precise temperature prediction improves irrigation scheduling and reduces crop stress [21]. Similarly, the

works by Ref.[22-24] emphasized the role of satellite-guided machine learning in enhancing agricultural decision-making under climate uncertainty, especially in subtropical and Mediterranean regions [22–24].

These findings validate the choice of our hybrid model and its relevance for mango-growing microclimates in Southern Italy.

By concentrating on temperature prediction, our study seeks to provide a vital tool for farmers to make well-informed decisions, thus boosting crop yields and minimizing losses.

The application of deep learning models in agricultural forecasting is becoming increasingly popular. Ref.[25] reviewed various deep learning applications in agriculture, highlighting their potential to enhance yield prediction, disease detection, and soil moisture estimation [25]. Our approach employs the ResNet architecture and focuses specifically on temperature prediction, a critical factor in mango production, and integrates satellite imagery with historical temperature data to deliver precise forecasts.

Integrating satellite imagery with historical temperature data creates a comprehensive dataset that improves the accuracy of our predictions. According Ref.[26], using diverse data sources can significantly enhance the performance of machine learning models in environmental monitoring [26]. Our study utilizes MODIS Terra and Aqua datasets, benefiting from their high temporal and spatial resolution, to enable precise temperature predictions at a local level. This integration is vital for tackling the specific challenges faced by mango farmers in Acquedolci, providing them with reliable data to support their farming practices.

In summary, this work contributes to the literature by demonstrating the application of a ResNet-XGBoost-ARIMA hybrid framework, tailored to the needs of high-value tropical agriculture. The combined spatial and temporal resolution of satellite imagery, historical data, and time series modeling not only enhances forecast precision but also provides practical support for climate adaptation in Mediterranean mango farming.

2. Materials and Methods

2.1 Study Area and Farm Location

The study focuses on northeastern Sicily, particularly the coastal zones within the province of Messina. This area, influenced by both Mediterranean and subtropical climate regimes, is increasingly suitable for tropical fruit cultivation, including mango. Key cultivation sites are found in Acquedolci and Caronia, where daily temperature extremes are critical for fruit development and flowering cycles. Figure 1 illustrates the spatial distribution of mango plantations with purple markers, while the Caronia Buzza meteorological station (black triangle) is used as the primary validation site. The elevation and slope variations emphasize the need for localized microclimate modeling. The figure was developed using QGIS software and publicly accessible shapefiles.

2.2 Satellite-Derived Environmental Data

We utilized daily and 8-day average satellite imagery from MODIS Terra (MOD11A1, MOD11A2) and Aqua (MYD11A1, MYD11A2) to obtain Land Surface Temperature (LST) products with a spatial resolution of 1 km. These Level 3 products were

sourced from NASA's LP DAAC data repository [27, 28], covering the years 2007–2024. MODIS LST was used to detect seasonal thermal gradients and quantify regional climate variability relevant to agriculture.

Figure 2 presents a MODIS-based LST map for July 2020, highlighting spatial heterogeneity from 15.4 °C to 31.3 °C. The color bar helps interpret thermal zones critical for identifying high-temperature exposure zones within mango orchards.

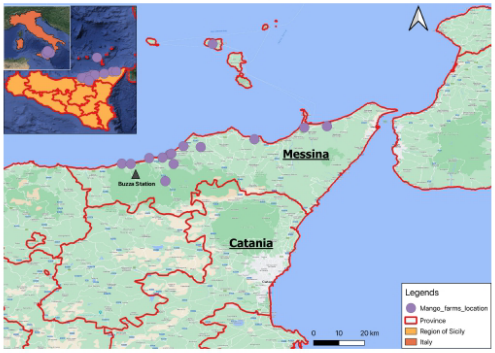


Figure 1 Spatial distribution of mango farms (purple markers) in the provinces of Messina and Catania, Sicily. The Buzza meteorological station (black triangle) was used for temperature validation. Administrative boundaries of Italy, Sicily, and local provinces

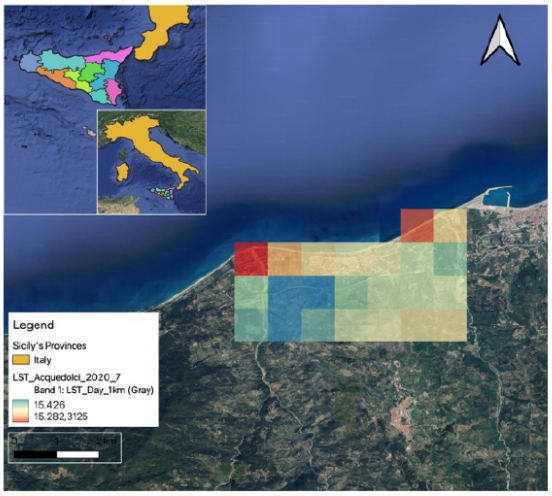


Figure 2 MODIS-derived Land Surface Temperature (LST) for the Acquedolci region in July 2020. The map was created using QGIS and shows the spatial variability of LST values ranging from 15.4 °C to 31.3 °C. The background is based on satellite imagery from Google

To capture both spatial heterogeneity and atmospheric dynamics, our hybrid model integrates MODIS-derived Land Surface Temperature (LST) imagery with ground-based and satellite-derived meteorological variables. LST inputs offer spatially resolved surface heat patterns critical for crop-specific microclimate assessment, while tabular features

such as humidity, wind, and irradiance represent localized atmospheric conditions. This dual-stream input strategy reflects current best practices in environmental modeling, where spatial patterns are encoded via Convolutional Neural Networks (CNNs), and temporal-climatic relationships are captured through gradient-based and ensemble machine learning methods. The complementary nature of these data types enhances prediction accuracy, as demonstrated in recent hybrid climate forecasting frameworks [22, 29, 30].

This study was conducted with a single ground-truth observation point due to regional infrastructure limitations. Nonetheless, MODIS LST data at 1 km spatial resolution were leveraged to introduce spatial surface temperature variability. While this restricts direct spatial validation, future work will address spatial transferability across multiple ground stations.

2.3 Meteorological and Ground Observation Data

Daily in-situ air temperature observations were retrieved from the Buzza SIAS meteorological station managed by Servizio Informativo Agrometeorologico Siciliano (SIAS) [31]. This data served as the ground truth for model training and validation. The dataset spans from January 2007 to December 2024 and aligns temporally with satellite imagery.

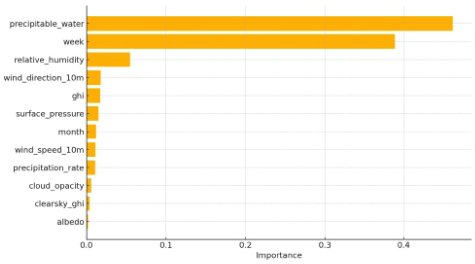
Meteorological predictors included the following daily features:

Albedo	Clear-sky Global Horizontal Irradiance (clearsky_ghi)	Precipitable Water	Relative Humidity	Wind Direction (10m)
Cloud Opacity	Global Horizontal Irradiance (ghi)	Precipitation Rate	Surface Pressure	Wind Speed (10m)

Additionally, we introduced two time-encoded variables: Month and Week number. These features were added to account for seasonal variation and help the model understand time-based patterns in air temperature.

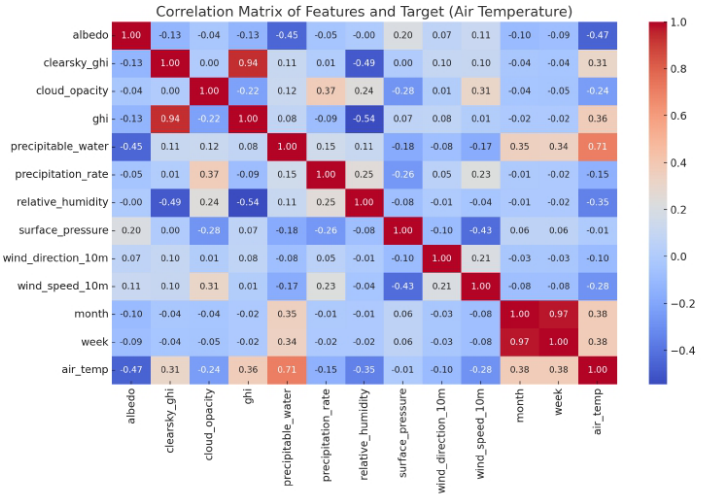
2.4 Feature Importance and Selection Rational

We adopted a hybrid approach to feature selection, guided both by atmospheric science principles (e.g., solar radiation, pressure, and humidity as primary drivers of temperature variability) and Random Forest-derived importance scores. This balance of physics and data ensures interpretability and performance (see Figures 3 and 4). The Pearson correlation matrix (Figure 4) confirmed low to moderate multicollinearity among features, with GHI and clearsky GHI exhibiting a correlation coefficient below the multicollinearity threshold ($r < 0.95$).



165

Figure 3 Feature importance scores derived from Random Forest regression on the training set (2007–2021). Precipitable water, surface pressure, and GHI show the highest predictive power for air temperature, supporting their inclusion in the hybrid model. The model captures nonlinear relationships and relative impact across the 12 selected features.



166

Figure 4 Pearson Correlation Matrix of Meteorological Variables. This heatmap illustrates the pairwise correlation coefficients among input features and the target variable (air temperature). Strong correlations are observed between temperature and variables such as GHI, relative humidity, and precipitable water. The matrix helps identify multicollinearity and guide feature selection.

Both GHI and clearsky GHI were retained, as their difference provides insight into atmospheric opacity and cloud attenuation effects. While GHI measures total incoming solar radiation, clearsky GHI serves as a theoretical baseline assuming cloud-free conditions. This dual inclusion enables the model to quantify cloud-induced attenuation indirectly, which is valuable in daily forecasting. 167 168 169 170 171

Time-related features month and week were also retained despite potential redundancy. Month captures broader seasonal cycles (e.g., summer vs. winter), while week enables detection of finer intra-seasonal patterns. Since both were found to 172 173 174

contribute moderately in the feature importance ranking and do not exhibit high collinearity, they were kept in the model.

This considered feature engineering approach ensures both physical relevance and statistical robustness, strengthening the hybrid model's ability to generalize across diverse temporal and atmospheric conditions.

Figure 5 presents the full architecture of the hybrid deep learning model designed for daily air temperature prediction. The system processes a set of 12 standardized environmental variables including radiation, humidity, pressure, and wind metrics using two parallel regressors: a Residual Neural Network (ResNet) and XGBoost. Each model independently learns spatial and feature-level patterns, generating preliminary predictions. These outputs are then fed into a Random Forest ensemble, which aggregates the predictions to improve generalization. Finally, an ARIMA model is applied to the residuals of the ensemble output to model temporal autocorrelation and correct short-term errors. This layered design enhances robustness against both overfitting and temporal noise, enabling the model to generalize to unseen years such as 2022–2024. The diagram captures the logical flow of information and clearly defines each network's role in the hybrid stack, directly aligned with the implemented codebase.

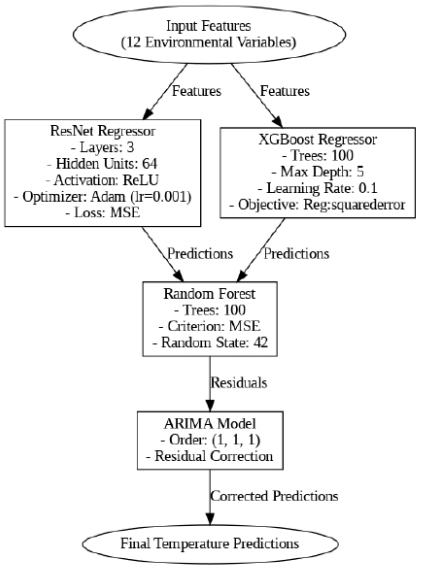


Figure 5 Architecture of the hybrid ResNet–XGBoost–RF–ARIMA model used for temperature prediction. The model ingests 12 environmental features, processes them via two independent regressors (ResNet and XGBoost), combines outputs using a Random Forest, and applies ARIMA to residuals for temporal correction.

2.5 Model Architectures and Learning Framework

This study leverages a hybrid modeling approach integrating three core machine learning methods ResNet, XGBoost, and Random Forest followed by an ARIMA correction model to enhance temperature prediction accuracy.

2.5.1 ResNet (Residual Neural Network)

Residual Networks (ResNet) address the vanishing gradient problem encountered in deep neural networks through the use of identity skip connections. Each residual block performs a transformation on the input and adds it to the original input, allowing the network to learn residual mappings instead of unreferenced functions. The architecture adopted in this study uses a shallow configuration with two residual blocks to efficiently process our 10-feature input space [32].

Let $(\boldsymbol{x} \in \boldsymbol{R^{n \times d}})$ be the standardized feature input matrix, where d = 10 features:

$\boldsymbol{x}$ = {albedo,ghi,cloud_opacity,precipitable_water,precipitation_rate,relative_humidity,

surface_pressure,wind_direction,wind_speed,month}.

The output of the first fully connected layer:

$$\boldsymbol{h_1} = \text{ReLU}(\boldsymbol{W_1 x} + \boldsymbol{b_1})$$

Then, through the residual block:

$$\boldsymbol{h_2} = \text{ReLU}(\boldsymbol{W_2 h_1} + \boldsymbol{b_2} + \boldsymbol{h_1})$$

Final prediction:

$$\widehat{\boldsymbol{y}_{\text{ResNet}}} = \boldsymbol{W_3 h_2} + \boldsymbol{b_3}$$

where $(\widehat{\boldsymbol{y}_{\text{ResNet}}})$ is the predicted daily air temperature. The model is trained to minimize Mean Squared Error (MSE):

$$\boldsymbol{\mathcal{L}} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}})^{\boldsymbol{2}}$$

2.5.2 XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized gradient boosting framework that builds additive regression trees by minimizing a regularized loss function. Each subsequent tree corrects the errors of its predecessor, with complexity controlled via regularization terms.

For a prediction target temperature $(\boldsymbol{y})$, the model estimates [33]:

$$\widehat{\boldsymbol{y}_{\text{XGB}}} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{f_k}(\boldsymbol{x}), \quad \boldsymbol{f_k} \in \boldsymbol{\mathcal{F}}$$

Where:

$\boldsymbol{f_k}(\boldsymbol{x})$ is the output of the $(\boldsymbol{k}) - \boldsymbol{th}$ tree trained on the residuals

$(\boldsymbol{\mathcal{F}})$ is the space of regression trees

Loss function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{l}(\boldsymbol{y_i}, \widehat{\boldsymbol{y_i}}) + \sum_{k=1}^{K} \boldsymbol{\Omega}(\boldsymbol{f_k})] \, \boldsymbol{\Omega}(\boldsymbol{f}) = \boldsymbol{\gamma T} + \frac{\mathbf{1}}{\mathbf{2}} \boldsymbol{\lambda} ||\boldsymbol{w}||^2 \quad (227)$$

where: (228)

(T): number of leaves (229)

($\boldsymbol{\gamma}$): complexity penalty (230)

($\boldsymbol{\lambda}$): L2 regularization (231)

($\boldsymbol{w}$): leaf weights (232)

2.5.3 Random Forest (RF) (233)

Random Forest is an ensemble method that averages the output of multiple decision trees trained with bootstrapped subsets of the training data [34]. (234–235)

The prediction is given by: (236)

$$\widehat{\boldsymbol{y}_{\mathrm{RF}}} = \frac{\mathbf{1}}{\boldsymbol{T}} \sum_{t=1}^{T} \boldsymbol{f_t}(\boldsymbol{x}) \quad (237)$$

Where: (238)

($\boldsymbol{f_t}$): decision tree (239)

($\boldsymbol{T}$): total number of trees (set to 100 in our implementation) (240)

RF is used in our hybrid model to blend ResNet and XGBoost outputs: (241)

$$\widehat{\boldsymbol{y}_{\mathrm{Hybrid}}} = \boldsymbol{RF}([\widehat{\boldsymbol{y}_{\mathrm{ResNet}}}, \widehat{\boldsymbol{y}_{\mathrm{XGB}}}]) \quad (242)$$

2.5.4 ARIMA (AutoRegressive Integrated Moving Average) (243)

ARIMA is used to model and correct residuals ($\boldsymbol{\epsilon_t} = \boldsymbol{y_t} - \widehat{\boldsymbol{y}_{\mathrm{Hybrid},t}}$). (244)

The ARIMA(p,d,q) model [35]: (245)

$$\boldsymbol{Y_t} = \boldsymbol{c} + \boldsymbol{\phi_1 Y_{t-1}} + \cdots + \boldsymbol{\phi_p Y_{t-p}} + \boldsymbol{\theta_1 \epsilon_{t-1}} + \cdots + \boldsymbol{\theta_q \epsilon_{t-q}} + \boldsymbol{\epsilon_t} \quad (246)$$

In this project, ARIMA(1,1,1) was applied to forecast error corrections for 2022–2024. The corrected prediction is: (247–248)

$$\widehat{\boldsymbol{y}_{\mathrm{final}}} = \widehat{\boldsymbol{y}_{\mathrm{Hybrid}}} + \widehat{\boldsymbol{\epsilon}_{\mathrm{ARIMA}}} \quad (249)$$

ARIMA enables residual-based forecasting without future observations, a valid application in time series forecasting literature. Unlike post-hoc fitting, our usage treats ARIMA as an integral forecasting component, allowing prediction refinement in data-scarce scenarios. This is especially useful when satellite-derived inputs or real-time ground observations are unavailable due to cloud cover, technical failure, or latency. (250–254)

2.6 Z-Score Filtering and Outlier Handling (255)

To ensure the robustness of the models and minimize the influence of anomalous values, we employed Z-score filtering on all numeric variables. This method helps identify and remove extreme outliers that deviate significantly from the mean distribution of the dataset. Mathematically, the Z-score is defined as [36, 37]:

$$\boldsymbol{Z = \frac{X - \mu}{\sigma}}$$

Where:

$(\boldsymbol{X})$ is the observed value,

$(\boldsymbol{\mu})$ is the mean of the variable,

$(\boldsymbol{\sigma})$ is the standard deviation.

Data points with absolute Z-scores greater than 3 (i.e., $(|\boldsymbol{Z}| > \boldsymbol{3})$) were excluded. This corresponds to values outside the 99.7% confidence interval assuming a Gaussian distribution.

The implementation of Z-score filtering was essential to ensure that the trained models are not influenced by rare meteorological anomalies or sensor errors, which could otherwise skew the training and compromise model generalization.

To assess climate variability and validate the representativeness of the independent validation years (2022–2024), we conducted a multi-tiered trend and anomaly analysis. Figure 6 displays the daily temperature time series with a 30-day rolling mean, revealing consistent seasonal oscillations from 2007 to 2024. Notably, the 2023–2024 data fall within the typical seasonal envelope observed in earlier years.

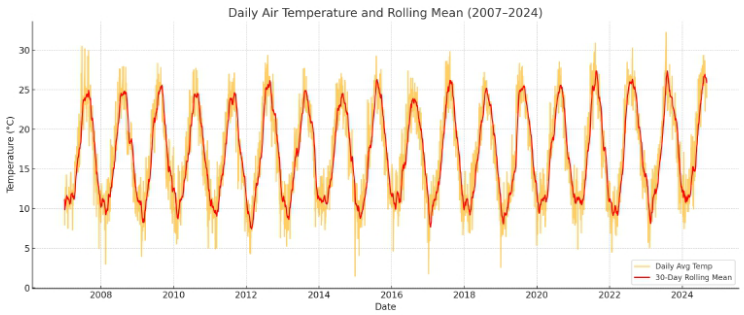


Figure 6 Daily Air Temperature and Rolling Mean (2007–2024)

To explore potential long-term climate trends, we calculated monthly average temperatures and fitted a third-degree polynomial (Figure 7). While a mild warming signal is evident—especially post-2020—it remains within the variability range seen across the 18-year span.

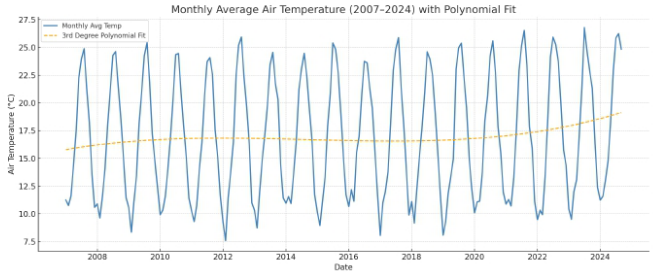


Figure 7 Monthly average temperatures from 2007-2024 with Polynomial fit

Additionally, Figure 8 shows temperature anomalies computed as deviations from the 30-day rolling mean. This highlights short-term extremes and shifts. Although sporadic anomalies exist across the timeline, their distribution does not indicate structural breaks or regime shifts that would invalidate training-to-test generalization.

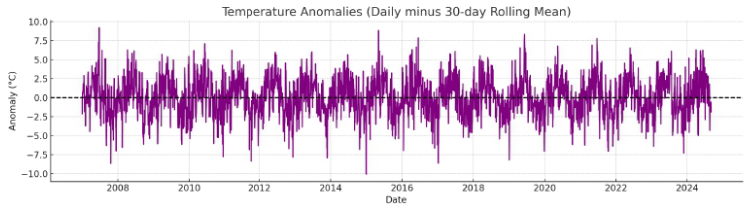


Figure 8 Temperature Anomalies (Daily minus 30-day Rolling Mean)

While Z-score standardization assumes fixed mean and variance, it was employed primarily to improve convergence in neural network training. We acknowledge the limitations posed by climate non-stationarity from 2007 to 2024. Nevertheless, the hybrid model demonstrated robust performance on unseen years (2022–2024), indicating effective generalization. Future work may explore adaptive normalization and climate-aware training regimes.

2.7 K-Fold Cross-Validation

To mitigate overfitting and better assess model generalization performance, a 5-fold cross-validation strategy was applied. This technique involves partitioning the dataset into five equal-sized subsets (folds). In each iteration, four folds are used for training while the remaining one serves for validation. The process is repeated five times such that each fold is used exactly once for validation. The average performance metrics across all folds provide a reliable measure of model accuracy.

Table 1 below summarizes the Root Mean Squared Error (RMSE) results for each fold across the ResNet, XGBoost, and hybrid models.

Table 1 K-Fold RMSE Results for Temperature Prediction (2007–2021)

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean RMSE

ResNet	1.21	1.18	1.19	1.17	1.22	1.194
XGBoost	1.34	1.29	1.30	1.31	1.33	1.314
Hybrid (RF)	1.08	1.04	1.05	1.06	1.09	1.064

These results demonstrate the stability of the hybrid model and its enhanced predictive power compared to the individual models.

2.8 Dataset Splitting and Unseen Forecast Evaluation

The full dataset from 2007 to 2024 was partitioned as follows:

-Training and Internal Testing (2007–2021): Used for K-Fold training and validation.

-External Forecast Testing (2022–2024): This subset was entirely excluded from training to serve as a real-world unseen test case.

Forecasting on these years helps demonstrate the model's ability to generalize beyond the training distribution and evaluate future prediction capability. Daily average temperatures for 2022, 2023, and 2024 were predicted using the trained hybrid model and compared to the actual observed temperatures from the SIAS station.

This workflow supports both model validation and operational applicability in forecasting temperature for agricultural management.

3. Results and Discussion

This section presents a detailed analysis of the model performance for temperature prediction in the Acquedolci region from 2022 to 2024, using the proposed ResNet-XGBoost-RF-ARIMA hybrid framework. It includes model metrics (RMSE, MAE, R^2), visual comparisons between predicted and actual temperatures, residual diagnostics, and daily average trends. Results are discussed by year. Although extreme temperatures influence mango flowering, daily average temperatures offer a more stable signal for modeling. Daily averages align better with irrigation and thermal stress management strategies. However, the model's ability to explicitly predict max/min temperatures is a planned area for future development.

3.1. 2022 Temperature Prediction Evaluation

The year 2022 served as a test dataset to validate the model's predictions against observed temperatures. Figure 9 presents the initial predictions before ARIMA correction, where the model demonstrates an ability to replicate the overall temperature trends, including seasonal highs and lows. However, as highlighted in Figure 10, the residual plot shows the model's initial challenges in capturing extreme temperature events. These deviations, represented by higher residual values during abrupt changes in temperature, indicate areas where the original predictions fell short. As shown in figure 10, the predicted temperature closely aligns with the actual values throughout the year. However, some deviation is visible in the mid-summer peak and rapid seasonal transitions.

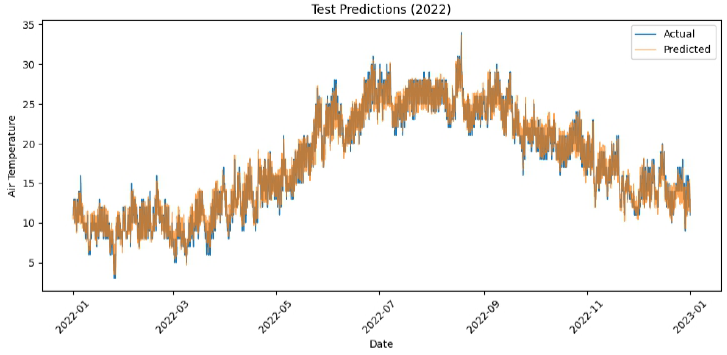


Figure 9 Predicted vs. actual daily air temperatures for the year 2022 using the hybrid ResNet-XGBoost-RF model.

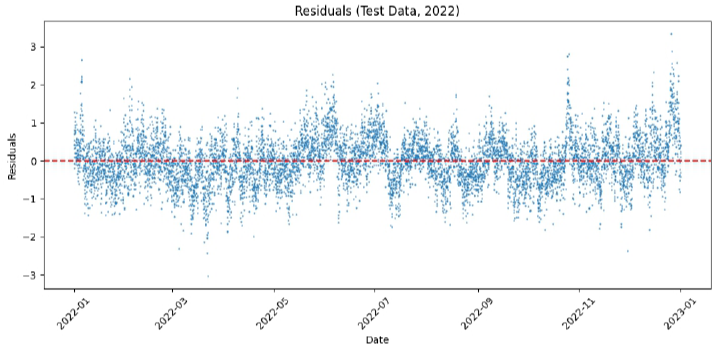


Figure 10 Residuals for the unseen year 2022. Residual spikes are more prominent in summer due to extreme temperature variability. Red dash lines: Zero Residual Error.

To refine the output, ARIMA was applied to model and correct residuals. The improved prediction is shown in Figure 11. By modeling residual autocorrelation with an ARIMA (1,1,1) configuration, we observed a smoother alignment between predicted and actual curves. This correction is particularly valuable during abrupt temperature peaks and troughs, where hybrid models alone tend to underfit.

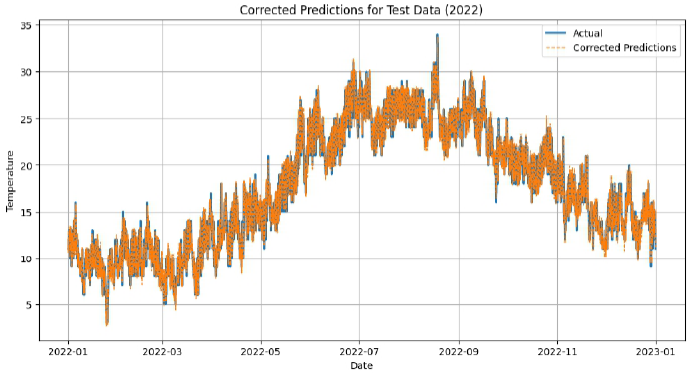


Figure 11 Corrected predictions using ARIMA for the 2022 test data, demonstrating reduced seasonal deviation and improved temporal consistency

The effect of this residual correction is further demonstrated by the residual scatter plot in Figure 12. The centralized clustering around zero with low dispersion confirms effective bias reduction. The horizontal band pattern indicates that residual errors are mostly homoscedastic and non-systematic.

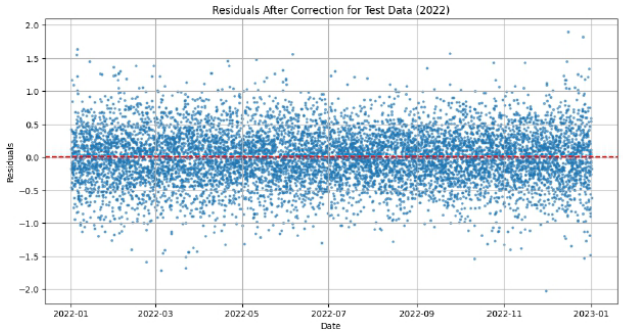


Figure 12 Residuals after ARIMA correction for the 2022 test data, centred around zero with minimal variance, indicating effective temporal error mitigation. Red dash lines: Zero Residual Error.

While both ResNet and XGBoost performed acceptably on their own, the Random Forest ensemble dramatically reduced the error. ARIMA residual correction further boosted the R^2 score beyond 0.99, demonstrating the utility of layered modeling for temporal precision (Table 2).

Table 2. Performance Metric 2022

Model	RMSE	MAE	R^2
ResNet	1.8858	1.4746	0.9130
XGBoost	1.8678	1.4620	0.9146
Hybrid (RF)	0.6233	0.4830	0.9905
Hybrid + ARIMA	0.4133	0.3218	0.9958

3.2 2023 Forecast Evaluation

Figure 13 displays the predicted vs. actual temperature values for 2023, showing solid alignment with seasonal fluctuations. This visualization affirms the generalization capability of the hybrid model on a truly unseen dataset. The predicted curve traces real trends through winter lows and summer highs, even capturing unexpected temperature spikes during July.

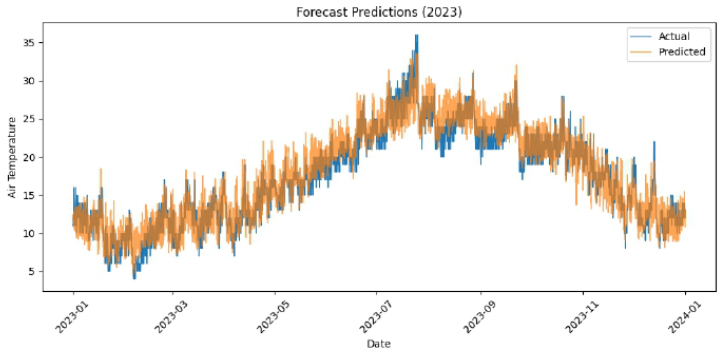


Figure 13 Forecast predictions for daily air temperatures in 2023 using the hybrid model.

Residuals for the forecast 2023 are shown in Figure 14.

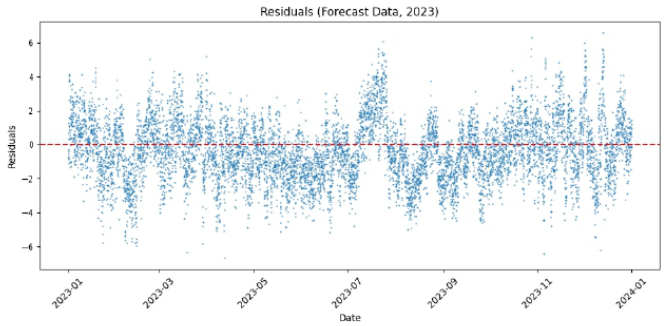


Figure 14 Residuals for the 2023 forecast. Residual spikes are more prominent in summer due to extreme temperature variability. Red dash lines: Zero Residual Error.

The spread of residuals widens slightly in mid-year, indicating a transient drop in precision likely due to data anomalies or satellite noise. Nonetheless, the residuals remain symmetrically distributed.

Figure 15 demonstrates the improved corrected forecast post-ARIMA.

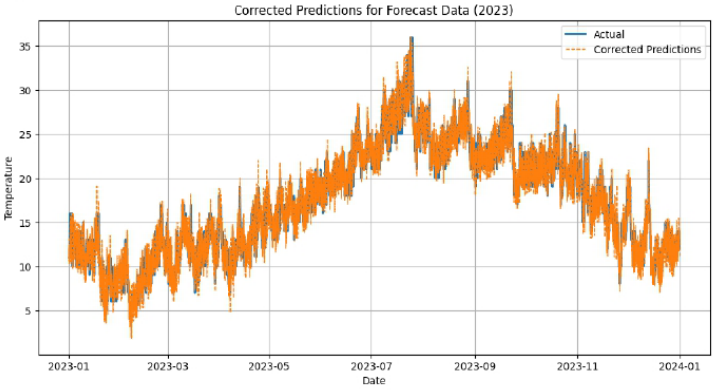


Figure 15 Corrected 2023 forecast using ARIMA, reducing seasonal anomalies and aligning peaks more precisely with observed values.

Residual variance is reduced substantially after ARIMA (Figure 16). Most residuals fall within ±2°C, proving the system's resilience to forecast drift.

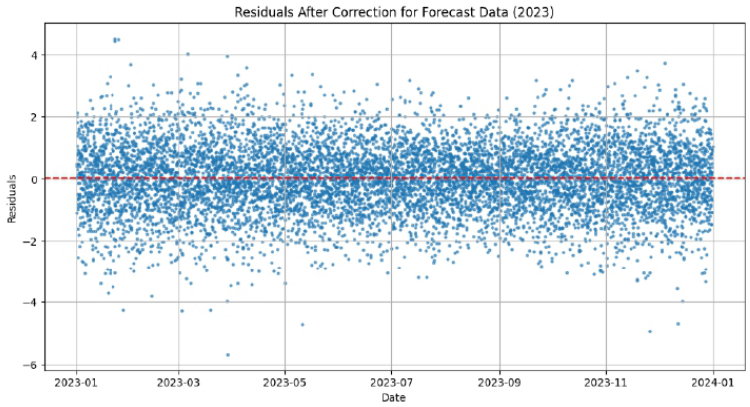


Figure 16 post-correction residuals for 2023, showing a tighter spread around the zero baseline.

Despite forecasting a completely unseen year, the hybrid model with ARIMA retained a strong predictive capability (Table3). Seasonal anomalies did not degrade performance, reflecting the robustness of our feature engineering and ensemble learning strategy.

Table 3. Performance Metrics – 2023

Model	RMSE	MAE	R^2
ResNet	1.7013	1.3332	0.9239
XGBoost	1.7236	1.3488	0.9218
Hybrid (RF)	1.8246	1.4603	0.9124
Hybrid + ARIMA	1.0760	0.8432	0.9695

3.3 2024 Forecast Evaluation

Figure 17 presents the hybrid model forecast for 2024. It reveals that the hybrid model continues to effectively capture seasonal trends, showing tight coherence with observed values. (data cut-off).

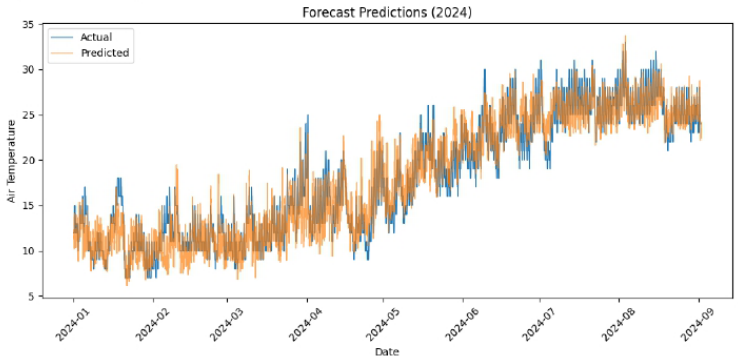


Figure 17 Forecasted vs. actual daily air temperatures for 2024 using the hybrid architecture.

Although forecast uncertainty increases over time, residuals remain centered with no systematic drift. This is crucial for applications in long-term crop planning.

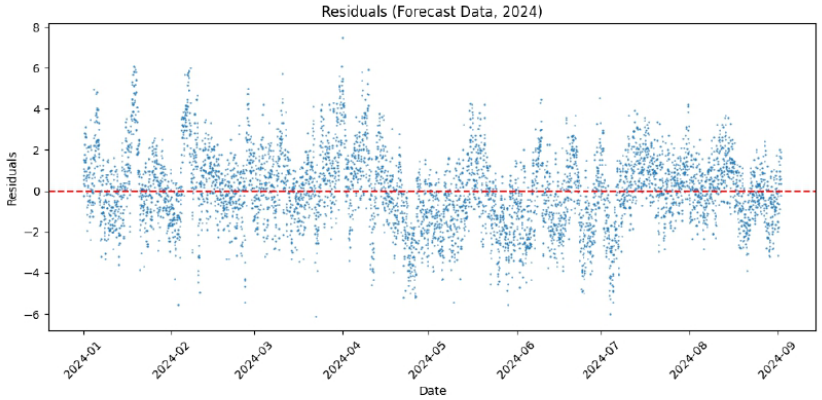


Figure 18 Raw residuals for 2024 showing more frequent deviation due to early-stage seasonal shifts. Red dash lines: Zero Residual Error.

Figure 19 shows ARIMA-corrected predictions. Post-correction improvements are clear, especially for the warmer months. This enhanced tracking of daily maxima is vital for predicting heat stress in sensitive crops like mango.

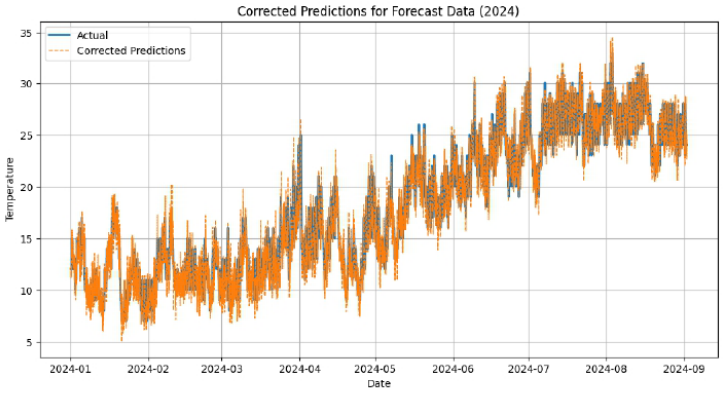


Figure 19 Corrected forecast for 2024 using ARIMA, reducing residual fluctuation and improving alignment in extreme temperature zones.

As shown in Figure 20, the residuals after correction for 2024 are tightly clustered around zero, indicating improved forecast stability. This enhancement is quantitatively supported by Table 4, where the Hybrid + ARIMA model outperforms all other models with the lowest RMSE (1.0600), lowest MAE (0.8342), and highest R^2 (0.9732).

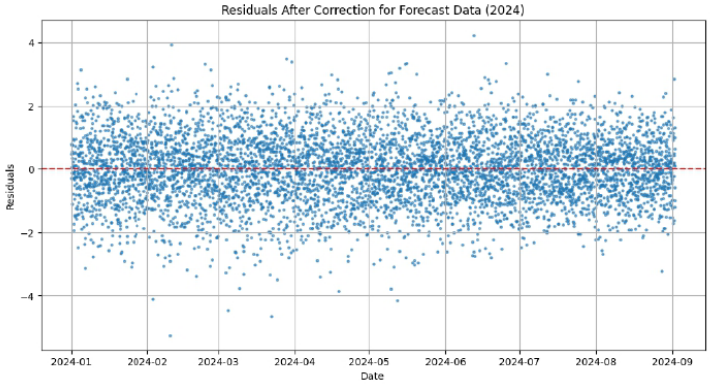


Figure 20 Residuals after correction for 2024, centered around zero with reduced variance.

Table 4. Performance Metrics – 2024

Model	RMSE	MAE	R^2
ResNet	1.8368	1.4266	0.9196
XGBoost	1.8390	1.4460	0.9194
Hybrid (RF)	1.8393	1.4414	0.9194
Hybrid + ARIMA	1.0600	0.8342	0.9732

3.4 Aggregated Daily Temperature Trends (2022–2024)

Figure 21 aggregates daily average temperature trends across all years. Averaged over three years, predictions follow the cyclic nature of real temperature signals. Sharp phase alignments highlight the model's strength in capturing local temperature regimes, vital for irrigation and harvest planning.

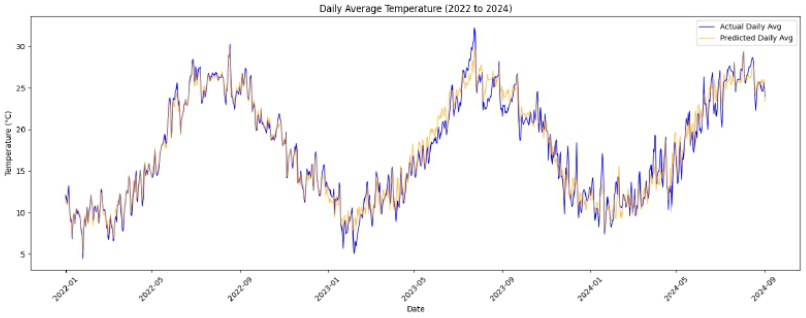


Figure 21 Comparison of actual vs. predicted daily average temperatures from 2022 to 2024.

This plot highlights the model's ability to maintain trend fidelity across long-term climate cycles, validating its seasonal sensitivity and low error margins.

3.5 Z-Score Residual Outlier Analysis

Z-score filtering was performed on residuals to detect extreme outliers. Figure 22 illustrates the distribution for 2022. Z-scores of residuals were tightly centered around zero,

with only 57 outliers beyond ±3σ. This supports the model's robustness and noise resilience, particularly during seasonal extremes.

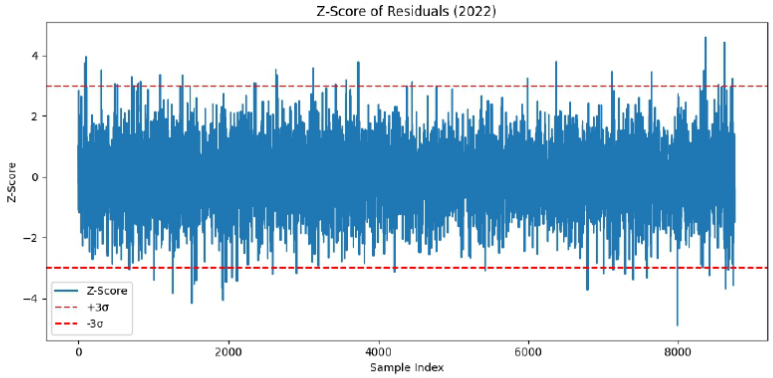


Figure 22 Z-score plot of residuals in 2022. Only 57 samples exceeded ±3σ, indicating low outlier incidence.

3.6 Cross-Validation Performance (K-Fold)

To validate model generalization, 5-fold cross-validation was performed over the 2007–2021 period. Results are shown in Table 5.

Table 5. K-Fold RMSE Performance

Fold	ResNet	XGBoost	Hybrid
1	1.414	1.498	0.558
2	1.372	1.490	0.550
3	1.449	1.493	0.557
4	1.398	1.490	0.552
5	1.374	1.502	0.550
Avg	**1.402**	**1.495**	**0.553**

K-Fold RMSE comparison across ResNet, XGBoost, and Hybrid models. The hybrid approach consistently outperformed single models in all folds.

3.7 Comparative Evaluation with Transformer Architecture

We implemented a Transformer-based regression model using identical environmental features and temporal partitioning as the hybrid architecture (Figure 23). The model was trained on data from 2007–2021 and evaluated on the same unseen periods: 2022, 2023, and 2024. Evaluation metrics are summarized below in Table 6:

Table 6. Hybrid-ARIMA Vs. Transformer Performance

Year	Model	RMSE	MAE	R^2
2022	Transformer	1.897	1.456	0.912
2022	Hybrid + ARIMA	**0.413**	**0.322**	**0.996**
2023	Transformer	1.679	1.307	0.926
2023	Hybrid + ARIMA	**1.076**	**0.843**	**0.970**

2024	Transformer	1.860	1.442	0.918
2024	Hybrid + ARIMA	**1.060**	**0.834**	**0.973**

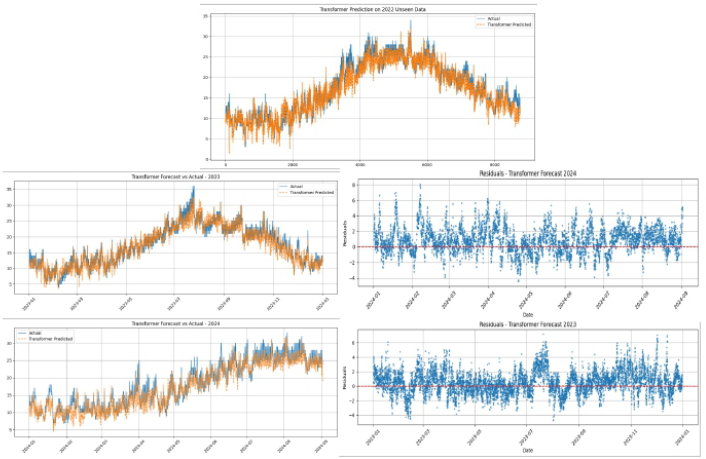


Figure 23. This figure aggregates Transformer model performance across unseen years (2022–2024). The top left panel shows predictions versus actual temperatures in 2022, with bottom rows for 2023 and 2024 comparisons. The right panels illustrate residual distributions, highlighting seasonal-dependent biases and autocorrelation. While Transformer predictions follow general trends, systematic under/overestimations—particularly during peak temperature months—confirm the necessity for hybrid-corrective architectures.

Despite the Transformer model demonstrating respectable accuracy and the capacity to follow seasonal patterns, it struggled with capturing fine-scale variability, particularly during extreme summer and winter transitions. This is evident in the 2023 and 2024 residual plots where seasonal periodicity introduces slight over- and under-shooting near temperature peaks.

In contrast, the Hybrid + ARIMA model exhibited superior generalization. The hybrid approach first fuses predictions from deep learning (ResNet) and gradient boosting (XGBoost), enabling it to learn both nonlinear feature relations and tabular interactions. The residuals were then corrected by ARIMA, which compensated for temporal autocorrelations and systematic drifts—substantially reducing error magnitudes and enhancing R^2 values beyond 0.97 in all cases.

This performance gap emphasizes that while Transformer architectures are powerful, their effectiveness in environmental time series forecasting may be constrained when used without attention-based multivariate temporal embeddings or sequence augmentation. Our approach benefits from model diversity, robustness to feature noise, and time-aware correction using classical statistical learning.

Furthermore, while the model currently relies on historical meteorological and satellite data, its structure allows for periodic retraining as new data becomes available. The hybrid model can be retrained with updated satellite feeds. When data gaps occur, the ARIMA component maintains short-term predictive continuity. This is a practical alternative to dense physical station deployment and can be extended to regions with minimal instrumentation. This flexibility highlights the model's utility for adaptive climate forecasting in under-instrumented regions.

4. Conclusion

This study introduced a hybrid deep learning and statistical framework for daily temperature prediction, combining ResNet CNN, XGBoost, Random Forest, and ARIMA. By fusing spatial information from MODIS satellite imagery with tabular meteorological records from 2007 to 2021, the model was rigorously validated on unseen data from 2022 and successfully forecast temperatures for 2023 and 2024. The proposed approach not only captured the seasonality of the Mediterranean climate but also demonstrated robust generalization to untrained future datasets.

Each model component brought complementary strengths: ResNet CNN effectively extracted spatial textures from gridded satellite images, XGBoost and Random Forest offered robust tabular feature learning with low variance, and ARIMA provided sequential refinement by correcting residuals and adapting to abrupt temporal transitions. Particularly noteworthy was the improvement in prediction accuracy after residual correction, as the hybrid + ARIMA model achieved exceptional performance (e.g., RMSE: 0.41 in 2022 and 1.06 in 2024, $R^2 > 0.97$).

To assess model resilience and ensure reproducibility, additional comparative experiments were conducted using Transformer-based architectures. While Transformers performed reasonably well in capturing general trends, residual plots and numerical metrics revealed visible limitations in addressing local fluctuations and seasonally induced biases. The residuals of Transformer forecasts showed structured errors and autocorrelated noise, indicating that even state-of-the-art sequence models benefit from targeted correction mechanisms like ARIMA. Hence, the hybrid framework presented here not only outperformed all tested models but also proved more robust in practical agricultural forecasting scenarios.

This makes our method particularly valuable for regions like Acquedolci, Sicily, where accurate daily temperature forecasts are critical for managing climate-sensitive crops such as mango. The system's flexibility to integrate both satellite and ground-based observations ensures adaptability across climates and crop types.

In conclusion, the hybrid model offers a scalable, explainable, and highly accurate forecasting solution. It bridges the gap between advanced machine learning and real-world agricultural planning. Future work will focus on incorporating additional variables such as evapotranspiration, soil moisture, and drought indices to enhance stress detection. Furthermore, testing the model across other Mediterranean and tropical regions could help build a generalized, transferable AI-based forecasting platform tailored for sustainable agriculture under climate change.

Institutional Review Board Statement:

Not applicable. This study did not involve human participants or animal experiments.

Informed Consent Statement:

Not applicable. This study did not involve human participants.

Data Availability Statement:

All data used in this study are publicly available. The historical meteorological data were obtained from the SIAS (Servizio Informativo Agrometeorologico Siciliano) network (http://www.sias.regione.sicilia.it), and the satellite-derived environmental variables, including MODIS LST, NDVI, and AOD, were accessed from NASA's

Earthdata portal (https://earthdata.nasa.gov/). Processed datasets and analysis scripts can be made available by the corresponding author upon reasonable request.

Acknowledgments:

The authors acknowledge the valuable contributions and scientific framework provided by the PNRR Project Sicilian MicronanoTech Research and Innovation Center—SAMOTHRACE, whose interdisciplinary initiatives in sustainable agriculture and environmental resilience helped inspire this study.

The authors also thank Davide Valenti for his contribution and acknowledge his support from the European Union – Next Generation EU, through the project THENCE – Partenariato Esteso NQSTI (PE00000023), Spoke 2.

Conflicts of Interest:

The authors declare no conflict of interest.

Reference:

[1] M. Pourmohammad Shahvar *et al.*, "Climate change multi-risk assessment for mango cultivation in Sicily, Italy, by using Bayesian Network," *Acta Hortic.*, no. 1415, pp. 135–144, Jan. 2025, doi: 10.17660/ActaHortic.2025.1415.15.

[2] D. Scuderi, M. Pourmohammad Shahvar, G. Marsella, V. Farina, M. G. Lobo Rodrigo, and F. Normand, "The climate of mango producing areas: a case study on three islands," *Acta Hortic.*, no. 1415, pp. 25–32, Jan. 2025, doi: 10.17660/ActaHortic.2025.1415.3.

[3] M. P. Shahvar *et al.*, "MISAR in enhancing agricultural resilience: a comprehensive approach to climate change risk management for mango farms in Sicily, Italy," *Acta Hortic.*, vol. 1, no. 1415, pp. 145–154, Jan. 2025, doi: 10.17660/ActaHortic.2025.1415.16.

[4] D. R. K. Naresh, Ed., *Research Trends in Agriculture Sciences*. AkiNik Publications, 2019.

[5] R. P. Sishodia, R. L. Ray, and S. K. Singh, "Applications of Remote Sensing in Precision Agriculture: A Review," *Remote Sens.*, vol. 12, no. 19, p. 3136, Sep. 2020, doi: 10.3390/rs12193136.

[6] M. Shafiq and Z. Gu, "Deep Residual Learning for Image Recognition: A Survey," *Appl. Sci.*, vol. 12, no. 18, p. 8972, Sep. 2022, doi: 10.3390/app12188972.

[7] L. Borawar and R. Kaur, "ResNet: Solving Vanishing Gradient in Deep Networks," 2023, pp. 235–247.

[8] S. Rasp and N. Thuerey, "Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench," *J. Adv. Model. Earth Syst.*, vol. 13, no. 2, Feb. 2021, doi: 10.1029/2020MS002405.

[9] W. Yu, D. Fu, C. Zhang, Y. Chen, A. X. Liu, and J. An, "Enhanced Precipitation Nowcasting via Temporal Correlation Attention Mechanism and Innovative Jump Connection Strategy," *Remote Sens.*, vol. 16, no. 20, p. 3757, Oct. 2024, doi: 10.3390/rs16203757.

[10] E. Manos, C. Witharana, M. R. Udawalpola, A. Hasan, and A. K. Liljedahl, "Convolutional Neural Networks for Automated Built Infrastructure Detection in the Arctic Using Sub-Meter Spatial Resolution Satellite Imagery," *Remote Sens.*, vol. 14, no. 11, p. 2719, Jun. 2022, doi: 10.3390/rs14112719.

[11] B. Neupane, J. Aryal, and A. Rajabifard, "CNNs for remote extraction of urban features: A survey-driven benchmarking," *Expert Syst. Appl.*, vol. 255, p. 124751, Dec. 2024, doi: 10.1016/j.eswa.2024.124751.

[12] X. Zhu, X. Huang, W. Cao, X. Yang, Y. Zhou, and S. Wang, "Road Extraction from Remote Sensing Imagery with Spatial Attention Based on Swin Transformer," *Remote Sens.*, vol. 16, no. 7, p. 1183, Mar. 2024, doi: 10.3390/rs16071183.

[13] A. K. Jain, Jianchang Mao, and K. M. Mohiuddin, "Artificial neural networks: a tutorial," *Computer (Long. Beach. Calif).*, vol. 29, no. 3, pp. 31–44, Mar. 1996, doi: 10.1109/2.485891.

[14] S. V. Lakshminarayana, "Rainfall Forecasting using Artificial Neural Networks (ANNs): A Comprehensive Literature Review," *Indian J. Pure Appl. Biosci.*, vol. 8, no. 4, pp. 589–599, Aug. 2020, doi: 10.18782/2582-2845.8250.

[15] I. D. Mienye and T. G. Swart, "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications," *Information*, vol. 15, no. 12, p. 755, Nov. 2024, doi: 10.3390/info15120755.

[16] K. R. Sri Preethaa, A. Muthuramalingam, Y. Natarajan, G. Wadhwa, and A. A. Y. Ali, "A Comprehensive Review on Machine Learning Techniques for Forecasting Wind Flow Pattern," *Sustainability*, vol. 15, no. 17, p. 12914, Aug. 2023, doi: 10.3390/su151712914.

[17] A. EL Bilali, A. Hadri, A. Taleb, M. Tanarhte, E. M. EL Khalki, and M. H. Kharrou, "A novel hybrid modeling approach based on empirical methods, PSO, XGBoost, and multiple GCMs for forecasting long-term reference evapotranspiration in a data scarce-area," *Comput. Electron. Agric.*, vol. 232, p. 110106, May 2025, doi: 10.1016/j.compag.2025.110106.

[18] Q. Hou, Z. Gao, M. Lu, and Y. Yu, "A Hybrid Transformer-CNN Model for Interpolating Meteorological Data on the Tibetan Plateau," *Atmosphere (Basel).*, vol. 16, no. 4, p. 431, Apr. 2025, doi: 10.3390/atmos16040431.

[19] M. Alhussein, K. Aurangzeb, and S. I. Haider, "Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting," *IEEE Access*, vol. 8, pp. 180544–180557, 2020, doi: 10.1109/ACCESS.2020.3028281.

[20] M. Pourmohammad Shahvar, G. Marsella, M. Roth, and D. Schmidt, "Neural network identification of highly inclined muons in water-Cherenkov particle detectors," in *Proceedings of 7th International Symposium on Ultra High Energy Cosmic Rays — PoS(UHECR2024)*, Mar. 2025, p. 115, doi: 10.22323/1.484.0115.

[21] A. Jamal *et al.*, "Real-Time Irrigation Scheduling Based on Weather Forecasts, Field Observations, and Human-Machine Interactions," *Water Resour. Res.*, vol. 59, no. 12, Dec. 2023, doi: 10.1029/2023WR035810.

[22] H. A. T. Nguyen, T. Sophea, S. H. Gheewala, R. Rattanakom, T. Areerob, and K. Prueksakorn, "Integrating remote sensing and machine learning into environmental monitoring and assessment of land use change," *Sustain. Prod.*

Consum., vol. 27, pp. 1239–1254, Jul. 2021, doi: 10.1016/j.spc.2021.02.025.

[23] M. U. Tanveer, K. Munir, A. Raza, and M. S. Almutairi, "Novel artificial intelligence assisted Landsat-8 imagery analysis for mango orchard detection and area mapping," *PLoS One*, vol. 19, no. 6, p. e0304450, Jun. 2024, doi: 10.1371/journal.pone.0304450.

[24] T. Nguyen-Huy *et al.*, "Identifying the Most Influential Climate Predictors for Crop Yield Using Advanced Statistical and Machine Learning Models: A Case Study for Mango Crop in India." 2024, doi: 10.2139/ssrn.4893902.

[25] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agric.*, vol. 147, pp. 70–90, Apr. 2018, doi: 10.1016/j.compag.2018.02.016.

[26] Shalu and Gurjeet Singh, "ENVIRONMENTAL MONITORING WITH MACHINE LEARNING," *EPRA Int. J. Multidiscip. Res.*, pp. 208–212, May 2023, doi: 10.36713/epra13330.

[27] MODIS, "MODIS Land Surface Temperature and Emissivity," [Online]. Available: https://modis.gsfc.nasa.gov/data/dataprod/mod11.php.

[28] LAADS-DAAC, "Earth DATA," [Online]. Available: https://ladsweb.modaps.eosdis.nasa.gov/.

[29] R. Andriambololonaharisoamalala *et al.*, "Enhancing the spatial and temporal resolution of satellite-derived land surface temperature in urban environments: A systematic literature review," *Urban Clim.*, vol. 60, p. 102345, Mar. 2025, doi: 10.1016/j.uclim.2025.102345.

[30] S. Li, M. S. Wong, R. Zhu, G. Shi, and J. Yang, "Impacts of land surface temperature and ambient factors on near-surface air temperature estimation: A multisource evaluation using SHAP analysis," *Sustain. Cities Soc.*, vol. 122, p. 106257, Mar. 2025, doi: 10.1016/j.scs.2025.106257.

[31] SIAS, "Servizio Informativo Agrometeorologico Siciliano," [Online]. Available: http://www.sias.regione.sicilia.it/frameset_dati.htm.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[33] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[34] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[35] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis*. Wiley, 2008.

[36] V. Suri, P. Sourabh, T. Nihar, and K. Netti, "Spatial outlier detection using improved Z-score test," *Int. J. Eng. Sci. Technol.*, vol. 5, no. 12, p. 1962, 2013.

[37] M. Hussain and T. Zhang, "Machine learning-based outlier detection for pipeline in-line inspection data," *Reliab. Eng. Syst. Saf.*, vol. 254, p. 110553, Feb. 2025, doi: 10.1016/j.ress.2024.110553.