Adaptive AI for Climate Resilience

This Food Agility CRC project identified a robust and repeatable framework so AI models can be used to understand the implications of climate change and related risks on agricultural crop outcomes.

By Dr Victor Chu, Dr Boyuan (Bryan) Zheng, A/Prof Zhidong Li, Nino Eclarin, Dr Evan Webster, Ashley Rootsey

foodagility UTS YAMAHA AGRICULTURE

ISBN 978-1-7646726-1-0
Project Name: Adaptive AI for Climate Resilience
Project No: FA128

Citation: Chu, V., Zheng, B., Li, Z., Eclarin, N., Webster, E., Rootsey, A. (2026), Adaptive AI for Climate Resilience Final Report, Food Agility CRC, Sydney.

IMPORTANT INFORMATION

Author contact details
Name: Victor Chu
Email: WingYan.Chu@uts.edu.au

Food Agility CRC contact details
Level 14, 5 Martin Place
Sydney, NSW 2000

+61 2 8001 6119
hello@foodagility.com
www.foodagility.com

Electronically published by Food Agility CRC at www.foodagility.com in May 2026.

The Food Agility CRC is funded under the Australian Government Cooperative Research Centres Program.

Abstract

The objective of this project is to identify a robust and repeatable framework by which agricultural Artificial Intelligence (AI) models can be used to understand the implications of climate change and related risks on agricultural crop outcomes, both now and into the future. In particular, the DSI research team investigated this explainability problem from three perspectives: i) to understand the needed AI explainability in agriculture, ii) to handle climate features and short-term extreme events in an explainable AI framework, and iii) to cater for long-term climate behaviours (e.g., arisen from climate change) in our proposed framework. Specifically, the explainability of our framework targets the impacts of extreme weather events and climate change on crop outcomes. In our experiments, we used apples, grapes, and kiwifruit yield data and their relevant climate data from the United States, Australia, and New Zealand to investigate and establish our proposed framework. During our investigation, we noticed that the problem of feature correlations is a major challenge, causing explanation issues. Several factors may contribute to this problem. For example, insufficient features can lead to causal confusion, and the significance of individual features may be diluted by other highly correlated features, rendering their explanations inaccurate. To address this problem, we adopted Convolutional Neural Network (CNN) to handle correlated features by exploiting the spatial and temporal relationships in observations through convolutional layers. Attention maps were used to visualise the climate featural-temporal effects on harvest yields, while integrated gradients were used to validate the impact of climate features. This final report summarises our research findings.

Keywords: Explainability, Agriculture, Artificial Intelligence, Machine Learning, Extreme Weather, Climate Change

Contents

List of Tables

List of Figures

Part I
Introduction

The security of food supply stands as a pivotal challenge confronting the global com-munity due to ever increasing food demand. The dynamics and often adverse impacts of changing climatic conditions further amplify our concerns. Nowadays, an important avenue to tackle food supply security is to model the system by using Artificial Intelligence (AI) and/or Machine Learning (ML) techniques. AI models and their associated processes are often per-ceived as "black boxes" creating a barrier for human to comprehend and trust. Instead, they should be something understandable to and shareable with humans allowing wider benefits to mankind. For example, the knowledge gains from understanding the impacts of weather events on agricultural productivity is a necessity for enhancing the explainability of agricul-ture predictive models. Moreover, the knowledge could assist the development of targeted interventions to mitigate the adverse effects of such events on food security. Examples of critical climate events are droughts, floods, frosts, etc., which can severely impact crop yields; and consequently, impacting the food supply chains and ultimately what we can have on our dining tables.

Identifying these critical climate events from historical data also presents its own challenges, primarily due to the variability and unpredictability of weather patterns. The task is further complicated when attempting to establish a relationship between sparse climate events and crop yields. The sporadic nature of extreme weather events, coupled with the often incomplete or noisy agricultural data, poses serious methodological and analytical challenges to data scientists. These unavoidable real-world problems underscore the need of sophisticated and explainable AI techniques, and the acquisition of comprehensive datasets to accurately determine their connections.

The objective of this project is to identify a robust and repeatable framework by which Agricultural AI models can be used to understand the implications of climate change and other risks on agricultural crop outcomes, both now and into the future. In particular, the DSI research team investigated this explainability research question from three perspectives. They are outlined below with their expected outcomes:

1. Understand the needed AI explainability in agriculture – to improve the methodologies of agriculture AI prediction models to better explain their outputs,
2. Handle climate features and short-term extreme events in an Explainable Artificial Intelligence (XAI) framework – to improve adaptiveness of XAI prediction models to short-term weather shocks, and
3. Cater for long-term climate behaviours (e.g., arisen from climate change) in our proposed framework – to understand long-term changes to climate conditions and their implications to XAI.

This research project was a collaboration among the team members outlined in Tables 1 and 2, where the project partners are:

- Data Science Institute (DSI), University of Technology Sydney (UTS),
- The Yield Technology Solutions (now Yamaha Agriculture), and
- Food Agility CRC (FACRC).

Table 1: Current Project Team Members.

Name	**Bio**	**Contact**
Dr Victor W. Chu	Victor W. Chu is a data science researcher specialising in agricultural AI, explainable AI (XAI), remote sensing, climate and extreme weather modelling, carbon and emissions modelling, and AI ethics. He currently leads the A-Theme at the Data Science Insti-tute, covering livestock, horticulture and environmental modelling. His work sits at the intersection of machine learning, spatio-temporal analytics, environmental modelling and responsible AI, contributing to national priorities in digital agrifood innovation and climate resilience.	WingYan.Chu@uts.edu.au
Dr Boyuan Zheng	Bryan is skilled in using deep learning models to solve problems and employs explainable AI techniques for post-hoc analysis of model predictions. Bryan's academic journey started with a bachelor's degree in Marine Technology from Ocean University of China, followed by a master's degree in Computer Science with a focus on AI from the Australian National University. He completed his PhD at the University of Technology Sydney, where he focused on explaining imitation learning in multimodal environments.	Boyuan.Zheng@uts.edu.au
A/Prof Zhidong Li	Zhidong Li earned his Ph.D. degree from the University of New South Wales, Sydney, Australia. Presently, he holds the position of Associate Professor at the Data Science Institute, University of Technology Sydney, Australia. Prior to this, Zhidong served as a senior engineer at Data61, CSIRO (Commonwealth Scientific and Industrial Research Organisation), the federal government agency for scientific research in Australia.	Zhidong.Li@uts.edu.au
Mr Nin˜o Eclarin	Nin˜o Eclarin is a Machine Learning Engineer with extensive experience in delivering production-grade ML and backend systems. He currently applies ML and backend engi-neering at Yamaha Agriculture, continuing after its Yield acquisition in 2024 and focusing on API design, data processing, and scalable AWS deployments. Previously at Hacarus, he built backend interfaces, led ML model development and MLOps, and mentored junior team members on projects ranging from cloud-based visual inspection to safety moni-toring.	Nino.Eclarin@yamaha-agriculture.com

Table 2: Departed Project Team Members.

Name	Bio
Dr Evan Webster	Evan Webster was the Data Science and Research Manager at Yamaha Agriculture. He completed his PhD and post-doctoral research at UNSW, focussing on large-scale environmental issues including catchment water yields, wildfire management, and river basin water management. At Yamaha Agriculture, Evan managed the development and operation of commercial weather and harvest models across the Wine, Kiwifruit, Apple, and Berry industries. Dr Webster left Yamaha Agriculture before the completion of this project.
Mr Ashley Rootsey	Ashley Rootsey was the AI & Robotics Lead at Food Agility CRC, the primary project sponsor and administrator. Ashley has over 7 years of agricultural and food technology R&D management experience and has worked on an expanse of industries and technology types over this time. Prior to this role Ashley graduated from University of Sydney with a 1st Class Hons Degree in Food Science & Agribusiness. Mr Rootsey left Food Agility CRC before the completion of this project.

The rest of this project report is organised as follows:

- ❑ Part II of this report presents a systematic survey on XAI in crop yield prediction, and a discussion of our findings on XAI for agriculture. It is followed by a summary of our case studies of climate impacts on apples, grapes, and kiwifruit.
- ❑ Part III outlines the methods that we used to collect the necessary data and explain the rationale behind our selection of datasets. Specifically, we focus on the collection of New Zealand climate and kiwifruit yield data, as well as Australian and the United States' climate and crop yield data.
- ❑ Part IV focuses on evaluating our XAI baseline models for crop yield prediction, including demonstration and testing of the latest XAI concepts. Based on our extensive evaluations of our experiment results, we identified the limitations of the existing XAI methods on climate in agriculture yield prediction and defined opportunities for the later stages of this project.
- ❑ Part V focuses on extreme weather events which have been making significant impacts on agriculture. We conducted a multi-faceted approach to further our investigation involving extreme weather events data retrieval, severity quantification, and alignment evaluation.
- ❑ Part VI presents our proposed XAI Frameworks to handle correlated climate features. Many traditional XAI methods implicitly assume feature independence. However, this assumption is often violated and hence they tend to distribute importance diffusely or inconsistently across correlated features leading to unstable or misleading explanations that obscure true causal drivers. In Part VI, we proposed to apply Convolutional Neural Networks (CNNs) enhanced with temporal attention mechanisms to predict crop yields and interpret the underlying temporal patterns.
- ❑ Part VII presents our analysis of longer term data and the extension of our frameworks to better cater for long-term climate behaviours.
- ❑ Part VIII is our overall conclusion for the project.

- Lastly, Part IX is the appendix, which records the manuscripts that we have prepared on specific topics in this project.

Part II
Literature Review & Horizon Scan

1 Part II Introduction

The objective of this project is to identify a robust and repeatable framework by which agricultural Artificial Intelligence (AI) models can be used to understand the implications of climate change and other risks on agricultural crop outcomes, both now and into the future. In particular, the DSI research team investigates this explainability research question from three perspectives: i) AI explainability, ii) extreme weather (short-term shocks), and iii) climate change (long-term trends).

Part II of this report summarises the state-of-the-current literature to prepare for the development of an AI Explanation framework for agricultural. The explainability of our framework targets the impacts of extreme weather events and climate change on crop outcomes in apples, grapes, and kiwifruit industries. During our literature review, we noticed that the problem of feature correlations is a major challenge causing the issue of counterfac-tual explanations. Several factors may contribute to this problem. For example, insufficient features can lead to causal confusion, and the significance of individual features may be diluted by other highly correlated features rendering their explanations inaccurate.

We organise Part II as follows. Section 2 presents our systematic survey on Explainable Artificial Intelligence (XAI) in crop yield prediction. It is followed by Section 3 in which we discuss our findings on XAI methodologies and frameworks in Agriculture. Next, in Section 4 we summarise our case studies of climate impacts on apples, grapes, and kiwifruit. Lastly, Section 5 concludes our literature review and horizon scan.

2 A survey of XAI in Crop Yield Prediction

We followed the guidelines proposed by Keele et al. [7] to conduct a systematic survey. This survey aims to identify the current research focus at the intersection between Explainable Artificial Intelligence (XAI) and the domain of agriculture, specifically in crop yield prediction. To assist our literature searching process, we first outlined four important research questions:

- What kind of machine learning algorithms are used in crop yield prediction?
- Which features are involved in the prediction problems?
- What are the common target crops? Is the fruit industry properly investigated?
- What kind of XAI techniques are used? and how they get evaluated?

Next, we proceed to define our search criteria by identifying the keywords commonly associated with these research questions. Our primary literature database is on the Web of Science platform[1]. To navigate the sheer amount of literature within the broad domains of artificial intelligence and machine learning, we narrowed our search using keywords such as "XAI", "explainable", "interpretable", "crop yield", etc. These keywords were combined with

[1]https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/

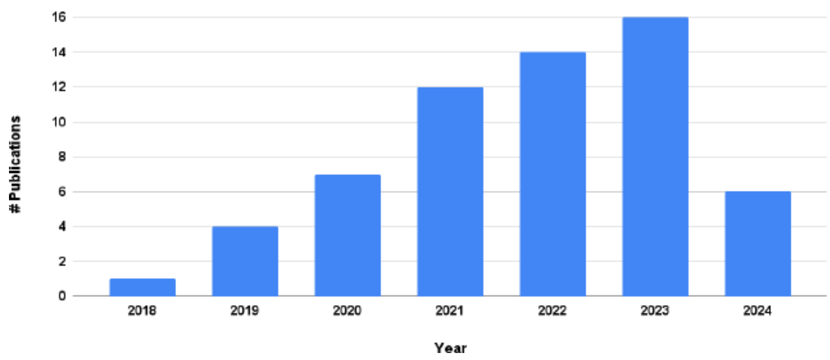


Fig. 1: Yearly counts of shortlisted publications in XAI and agriculture.

appropriate logical connectors to form our queries. We observed a considerable amount of research work related to crop disease classification, which falls outside our scope. Therefore, we excluded literature containing the term "disease" in their titles or abstracts. The final query string is as follows: (TS=("XAI") OR TS=("explainable") OR TS=("interpretable")) AND (TS=("crop yields") OR TS=("yield prediction") OR TS=("crop yield")) NOT TS=("disease"), where TS refers to topic search which includes searching through their titles, abstracts, and indexes. The query resulted in 60 pieces of shortlisted publications. We grouped the publications by year and summarised their counts in Fig. 1. We can observe that publications related to both XAI and agriculture are continuously increasing.

We differentiate the crop yield estimation approaches among these 60 publications by utilising the taxonomy introduced by He et al. [8], which is a comprehensive review of the yield estimation and prediction methods. At the top level, we can group these 60 publications into direct yield estimation and indirect yield prediction methods. Direct yield estimation involves employing cutting-edge object detection algorithms to detect crops and calculate quantity statistics to achieve yield estimation, which is widely applied to orchards. Indirect yield prediction utilises larger scale data, such as meteorological/climate data and county-level geological data, to predict crop yield. We excluded 21 papers with poor relevance to indirect crop prediction, leaving with 39 publications. We will next present our findings in four key aspects of the shortlisted publications: i) their coverage of crop types, ii) their considered features, iii) their adopted AI/Ml approaches, and iv) their model evaluation metrics. To better organise this report, we introduce our abbreviations together with their full name in Table 3. The abbreviations are used extensively in Part II of this report.

2.1 Coverage of Crop Types

First, we summarise the crop types from all the evaluated literature in Table 4. Majority of the research focuses on predicting the yield of cereals (primarily maize and wheat), which are the primary elements providing the highest calories and protein for the global food supply among wheat, rice, and maize [10]. Moreover, soybean and cotton are also widely investigated worldwide. These crops have relatively short farming cycles making them more suitable for analysing the annual influence of environmental variables and farm management practices (compared to most fruits). Furthermore, the types of crops evaluated are strongly correlated with geographic locations which in turn could be climate specific. The relationship between country and crop, depicted in Fig. 2, reveals a notable concentration of research in developed

Table 3: List of abbreviations used in this report.

Abbreviations	Full Names
AG	Agriculture/Agricultural
ANN	Artificial Neural Network
BNN	Bayesian Neural Network
CNN	Convolutional Neural Network
DL	Deep Learning
DT	Decision Tree
ENSO	El Niño-Southern Oscillation
GBDT	Gradient Boosting Decision Tree
GBM	Gradient Boosting Machine
IG	Integrated Gradients
IPO	Interdecadal Pacific Oscillation
KNN	k-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LIME	Local Interpretable Model-agnostic Explanations
LR	Linear Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
NN	Neural Network
NSE	Nash-Sutcliffe Efficiency
PDP	Partial Dependence Plots
R^2	Coefficient of Determination
RF	Random Forest
RMSE	Root Mean Square Error
SHAP	SHapley Additive exPlanation
SOTA	State-Of-The-Art
SVR	Support Vector Regression
XAI	Explainable Artificial Intelligence
XGB	Extreme Gradient Boosting

Table 4: Counts of crop types considered in evaluated literature. Note that a single paper could have multiple kinds of crop evaluated.

	Counts	Publications
Wheat	9	[4, 9–16]
Maize	9	[4, 9, 11, 13, 17–21]
Soybean	5	[9, 11, 19, 22, 23]
Cotton	3	[11, 24, 25]
Sunflower	2	[13, 15]
Others	17	[15, 16, 18, 26]

agricultural nations. The data also showcase intriguing spatial and country-specific patterns. For instance, countries located close to the equator, such as Malaysia, often study fruits and vegetables. The composition of crop types evaluated in research also reflects the role of these crops in a country's agricultural productions. For example, in Australia, wheat and cotton are investigated more frequently than other crops, which is consistent with the fact that wheat and cotton are principal crops in Australian agriculture.

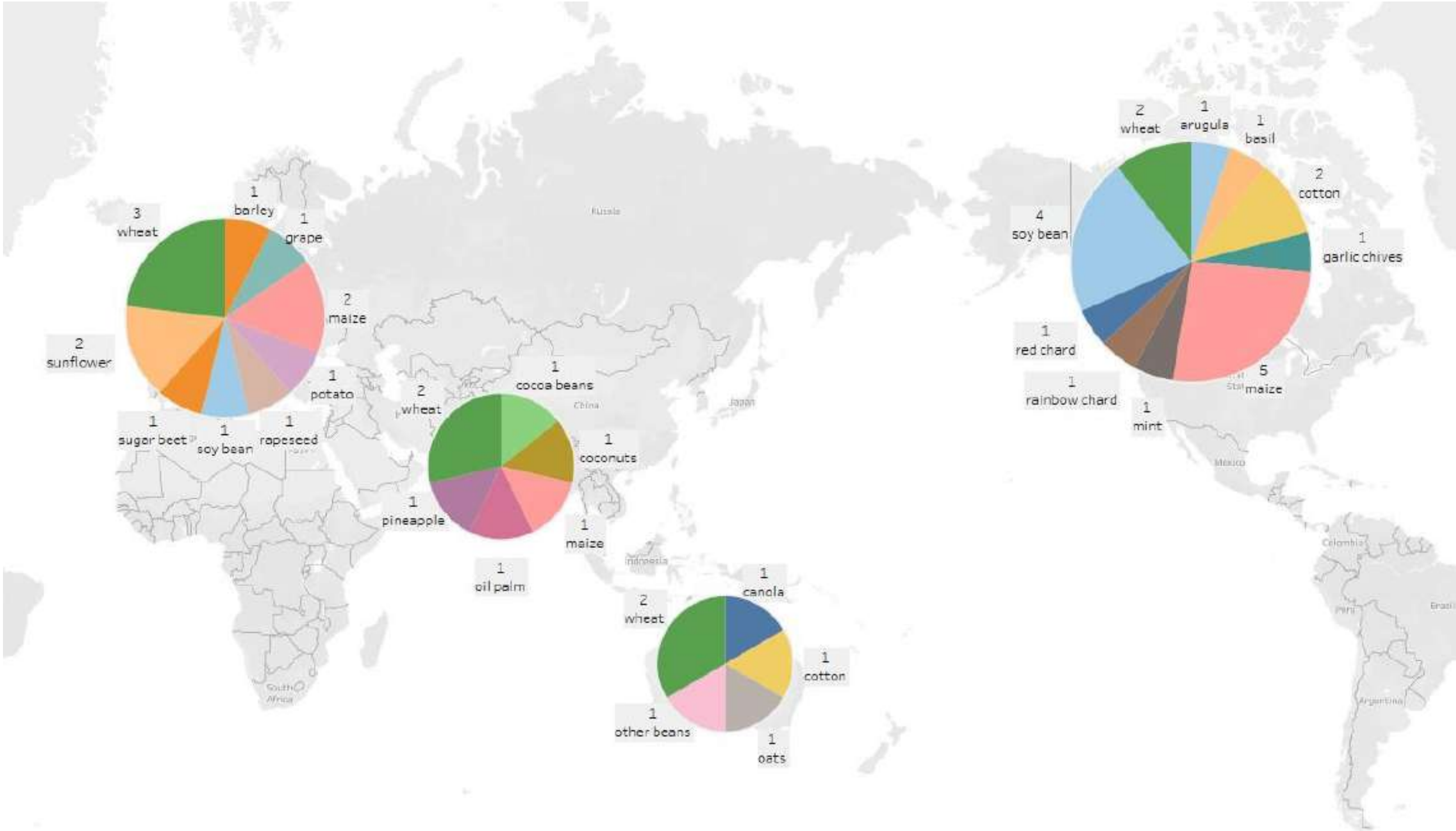


Fig. 2: Number of times a crop type is evaluated in various countries in evaluated literature. Note that a single paper could have multiple kinds of crop evaluated.

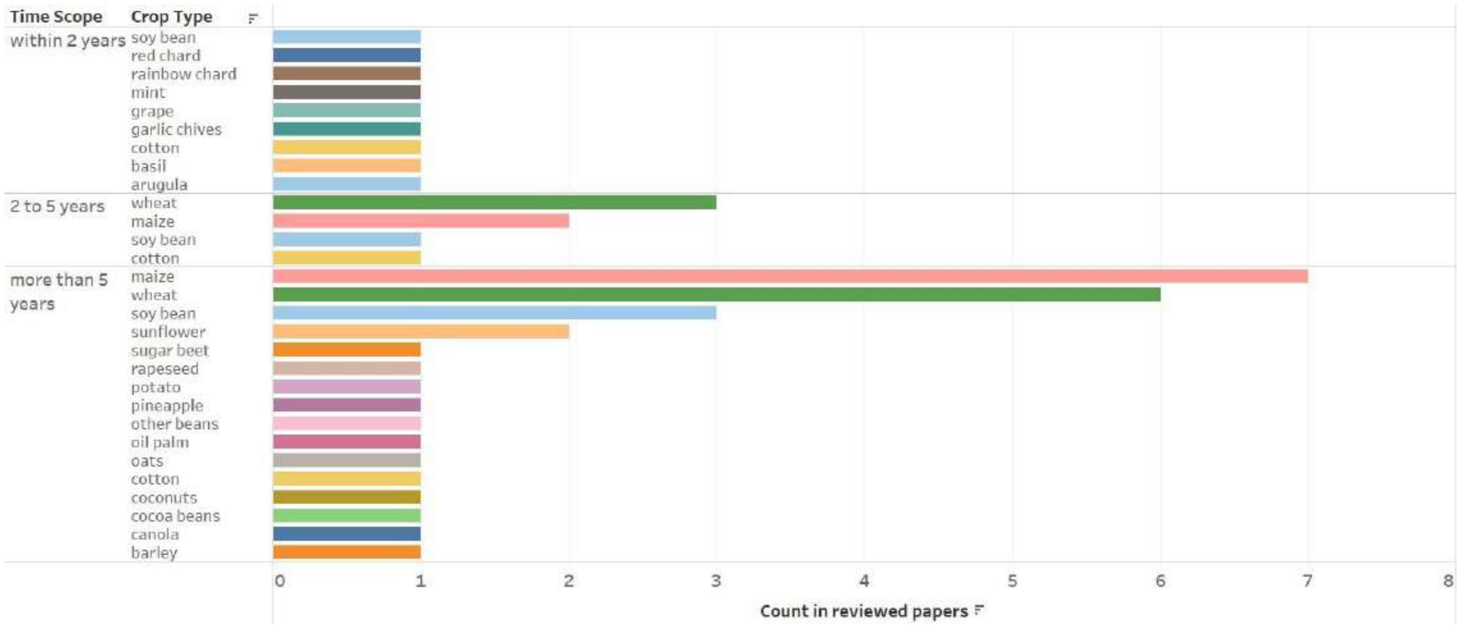


Fig. 3: Number of times a crop type is evaluated in different time scopes in evaluated literature. Note that a single paper could have multiple kinds of crop evaluated.

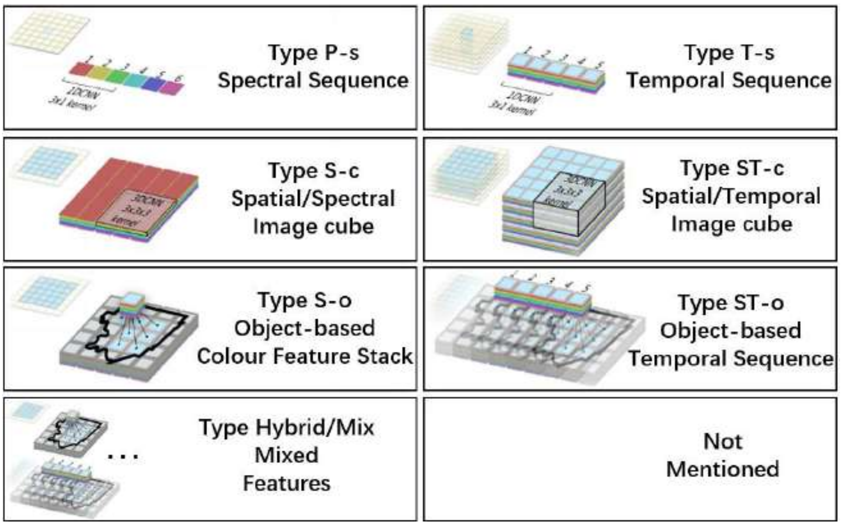


Fig. 4: Different interpretations of features (adapted from [1]).

We also categorised the duration of their input datasets into three time scopes: i) within 2 years, ii) from 2 to 5 years, and iii) more than 5 years, since we are interested in examining the relationship between crop types and time scope. Crops with longer farming cycles are expected to be investigated over a longer time span compared to those with shorter farming cycles. Fig. 3 presents the statistics of different types of crops evaluated under various time scopes. It is evident that popular crops such as wheat and maize have been extensively examined across various time scopes, indicating the significance of these crops and the accessibility of corresponding datasets. In addition to these principal crops, we observe that the temporal data for fruit yield prediction commonly spans more than 5 years, which is in line with our expectations.

2.2 Considered Features

Numerous features(variables) influence crop yield, rendering yield prediction an interdisciplinary subject characterised by diverse focuses and input features across the literature. A common taxonomy for these variables categorises them into management features and environmental features. Management features encompass variables that can be controlled by humans, such as fertiliser and pesticide usage, irrigation-related factors, crop density, and the like. Environmental features, on the other hand, stem from nature and can be further subdivided into four categories: i) meteorological/climate features, ii) soil-related features, iii) geographical features (such as slope and altitude), and iv) other features (e.g., vegetation indices obtained via remote sensing, crop duration, carbon dioxide concentration, etc.). Both management and environmental features are represented in various approaches across the literature. In this regard, we adopt the representation taxonomy proposed by Victor et al. [1], where input data can be interpreted in ten different ways (see Fig. 4). The first column of interpretations lacks a temporal component, whereas the second column incorporates temporal aspects. Each row interprets input data as a stack, sequence, image, image cube, and aggregated data based on territory, respectively.

Table 5: Distribution of feature representation – the counts of papers using a type of representation for a particular class of feature are captured in this table, where each is followed by references (in brackets). The numbers in bold are the most popular ones.

	Climate Features	Soil Features	Manipulable Features	Geo Features	Other Features
P-s	-	-	-	-	1 ([26])
S-c	-	1 ([25])	1 ([25])	-	-
S-o	2 ([12, 24])	3 ([12, 21, 27])	**10** ([12, 18, 24])	**12** ([9, 11, 21])	1 ([12])
T-s	1 ([26])	-	-	-	-
ST-c	2 ([10, 28])	-	-	-	-
ST-o	**22** ([4, 16, 19])	**6** ([9, 19, 24])	6 ([16, 19, 29])	1 ([16])	**10** ([4, 9, 20])
Mixed	-	4 ([4, 11, 29])	-	-	-
Not mentioned	2 ([25, 30])	14 ([17, 22, 23])	12 ([4, 9, 14])	16 ([10, 20, 27])	17 ([11, 17, 24])

We aim to elucidate the prevailing interpretations of various types of features in the field of crop yield prediction. To advance this inquiry, we document the interpretation categories for the five types of input features (four environmental features and one management feature). Given that certain papers emphasise specific perspectives while excluding the others, and the features within the same category may possess diverse interpretations; we expand upon the original ten classes by introducing two additional categories: hybrid and not mentioned. Table 5 summarises the interpretations in the reviewed literature of these five kinds of input features. Evidently, the majority of existing research aggregates spatial-axis inputs, focusing on county-level or even wider country-level/region data. Conversely, from a temporal perspective, various time frequencies are examined, spanning from hourly and daily to monthly and annual intervals.

The majority of interpretation proportions align with our expectations, as evidenced by several observations. For instance, meteorological/climate features are predominantly categorised as spatial-temporal features, while geographical features are perceived as relatively static. Moreover, management features are commonly overlooked, consistent with findings from prior research [13, 25]. However, two observations deviate from our initial expectations. Firstly, soil features are often considered near-static, with only minimal changes over time. Nevertheless, a significant proportion of research interprets soil features as temporal, primarily due to the inclusion of soil features related to water, such as soil water content. Excluding water-related soil features reveals that most included soil features are time-irrelevant or exhibit limited changes over time, such as apparent electrical conductivity and pH. Notably, soil features are the only category to contain a hybrid interpretation, as some studies consider both water-related soil features and near-static soil features. Secondly, another noteworthy observation is that other features are more prevalent than both management and soil features. This disparity arises from the inclusion of vegetation indices within this category. Vegetation indices, obtained from satellite sensors and measured at specific time frequencies, contribute significantly to this trend. However, in this review, we assign less emphasis to vegetation indices. For further insights into the significance of vegetation indices in yield prediction, we refer readers to [31].

We categorise climate features into four sub-classes: i) water-related features encompassing precipitation, humidity, and climate water balance; ii) temperature features including minimum air temperature, land surface temperature, and growing degree days; iii) radiation features consisting of downward short-wave radiation flux and photosynthetic active radiation; and iv) other climate features such as vapour pressure, evaporation, and day length. Fig. 5 provides a summary of the frequency of inclusion of each climate feature in the literature.

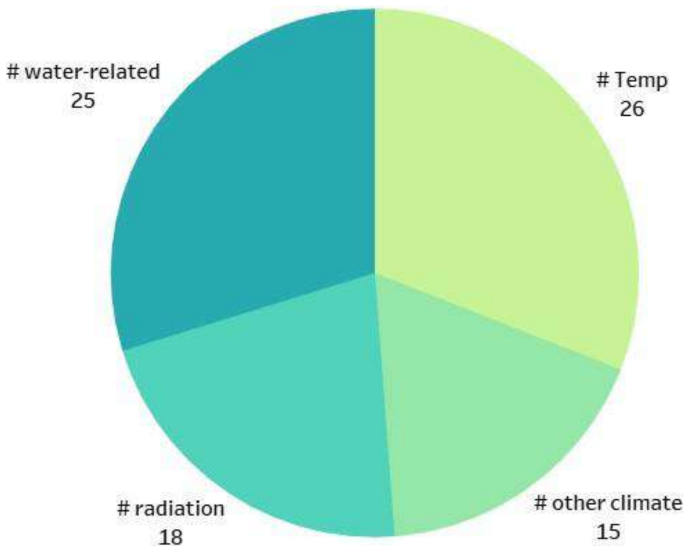


Fig. 5: Weather features evaluation frequency.

The data indicates that temperature features are the most commonly studied. Numerous studies highlight minimum temperature as a significant meteorological factor influencing crop yield [12, 13, 19, 23]. Conversely, maximum temperature is theorised to indicate any potential damage to pollen protein during the anthesis period, which is a concept widely accepted in the research community. However, empirical evidence varies, with some models assigning less weight to maximum temperature and even finding a positive effect [13]. This discrepancy may stem from spatial differences and crop variance. Additionally, Growing Degree Days (GDD) are widely examined as an indicator of accumulated heat during crop growth. Beyond its role in quantifying thermal accumulation, GDD is closely linked to crop development stages, as many physiological events from emergence to maturity occur after specific heat thresholds are met. This enables GDD to not only assess the suitability of a region for particular crops but also predict phenological milestones, maturity, and potential yield. Even when calendar-based measures fail to capture biological progress, GDD provides a unified framework for aligning phenological timing across regions and years. Sihi et al. identified GDD as the most influential climatic factor affecting yield across four evaluated crops [11]. However, excessive accumulation of GDD can also signal prolonged exposure to high temperatures, potentially accelerating phenological progression, shortening reproductive stages, and intensifying water stress, thereby reducing final yield [32–34].

Among water-related variables, precipitation emerges as the most frequently investigated feature. It is important to note that irrigation is counted as a manipulable feature (instead of a climate feature) as it involves human activities. Similar to temperature features, diverse conclusions arise from various studies. Precipitation exhibits a larger influence in the later stages of crop growth, from anthesis or flowering to harvest [15, 19, 35]. However, some research suggests its influence is relatively small compared to other climatic factors based on experimental results [21–23].

In comparison to water-related and temperature features, radiation features receive less attention in the reviewed literature. Nevertheless, radiation features may be crucial during the vegetative phase, particularly before flowering, and play a significant role for crops like maize. Other climatic features are sparsely discussed, and their perceived roles are more like indirect indicators of the climatic variables we discussed before.

The reviewed literature contains limited analysis of extreme events, while frost and drought are the primary concerns discussed. Although the frequency of frost is gradually decreasing with the global warming proceeds, frost after a warm beginning of a season could be "dangerous". Frost during springs and early summers significantly impacts bud-break and flowering phases, potentially causing irreparable damage to crops. Conversely, while drought is less harmful than frost, it still poses challenges that can be somewhat mitigated by irrigation and genetic crop benefits. Under a condition of strong and effective irrigation, mild drought may even be beneficial, as warm, dry conditions promote crop growth, whereas warm, humid environments increase disease incidence. Additionally, hail, flooding, and extreme winds are noteworthy. Hail directly damages crops, as evidenced by a 2014 event in New Zealand that resulted in a $45 million loss in the kiwifruit industry [36]. Extensive rainfall can also damage crops in areas with poor drainage, such as the 2022 flood in New Zealand, which caused an $80 million loss in the kiwifruit industry [37]. While breezes can reduce disease risk by keeping air moving, extreme winds can break crops, leading to significant production losses. El Niño and La Niña events have diverse impacts across regions, lasting from months to years. Developing strategies to anticipate and mitigate the economic impacts of these prolonged extreme events is imperative.

2.3 Adopted AI/ML Approaches

We compiled all the assessed AI/ML models from the reviewed literature in Fig. 6, encompassing both the proposed methods and baseline approaches. Based on our observations, all of our reviewed papers belong to the domain of supervised learning. Supervised learning learns a function from the labelled dataset, and its performance is highly related to the quality of the dataset. The spectrum of evaluated models, as illustrated in Fig. 6, ranges widely from traditional linear regression models to intricate deep learning architectures. Despite the effectiveness of deep learning is well-established in yield prediction tasks, our examination reveals that the majority of research focuses on conventional methods such as Random Forest (RF), Support Vector Regression (SVR), and the Least Absolute Shrinkage and Selection Operator (LASSO). However, numerous studies demonstrated that deep learning models can outperform these conventional approaches by a large margin and alleviate the need for laborious feature selection at the same time [19, 20, 26]. In the subsequent paragraphs, we will provide an overview of the mainstream methods.

Linear regression. Linear regression stands as one of the most traditional approaches for regression tasks, founded on the assumption of a linear relationship between independent and dependent variables. This method estimates coefficients of the target function based on observed data points. Its accessibility renders it relatively easy to comprehend and imple-ment compared to other machine learning techniques, offering inherent transparency and explainability to both technical and non-technical users. Consequently, it has emerged as the predominant baseline method in regression and explanation tasks. Despite its strengths, linear regression assumes linearity — a condition often not met in real-world scenarios — and its predictive capacity is comparatively limited compared to advanced techniques like neural networks and deep learning models. To address these limitations, several extensions have been proposed. Ridge regression and LASSO mitigate overfitting by introducing penalty terms that constrain the coefficients of independent variables. Ridge regression shrinks the sum of coefficients toward zero, while LASSO, more aggressively, eliminates highly correlated variables by setting their coefficients to zero, effectively serving as both a regularisation and

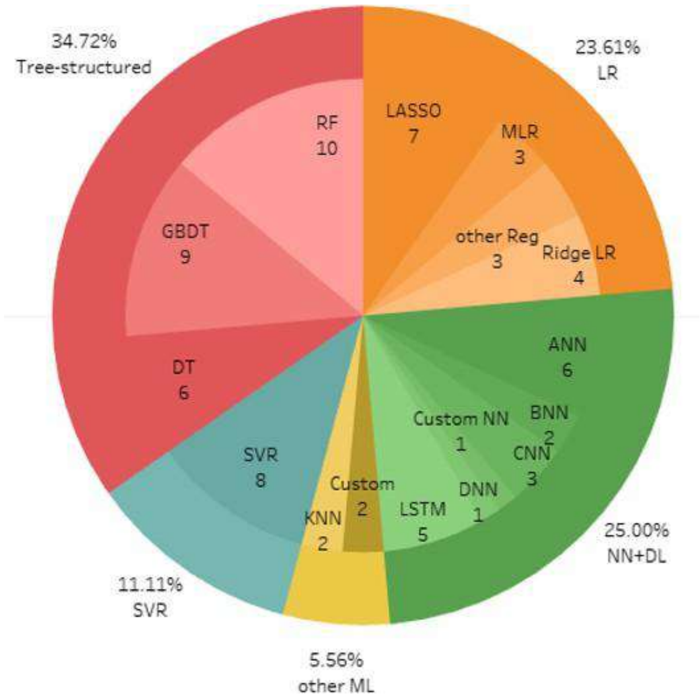


Fig. 6: Evaluated models. Note that one paper could evaluate multiple models.

a feature selection method. Partial Least Squares Regression (PLSR) combines regression analysis with dimensionality reduction techniques like Principal Component Analysis (PCA), demonstrating effectiveness in managing high-dimensional datasets.

Tree-structured ML. Tree-structured models, akin to linear regression, often eliminate the need for input rescaling. They iteratively divide features into subsets, irrespective of feature types, and excel at addressing nonlinear problems. Decision Trees (DT), the fundamental tree-structured method, remain foundational in many studies, prized for their explainability and resilience to outliers. However, their predictive power is modest and susceptible to overfitting with increasing tree depth. Random Forests (RF), an ensemble form of DT, generalise better by aggregating results from multiple trees and gauging feature importance. Nonetheless, this comes at the cost of interpretability and heightened input requirements. These methods epitomise the classic representatives of tree-structured machine learning models with their extensions explored across various research endeavours. Notable among these extensions is XGBoost, which is a special implementation of Gradient Boosting Decision Trees (GBDT). Unlike the parallel structure of RF, GBDT sequentially implements trees with each subsequent tree rectifying the errors of its predecessor. Widely examined in literature, XGBoost demonstrates promising predictive power, and efficiency through parallelisation, and generalisation. Breiman-Cutler Random Forests (BCRF) is another one that encompasses randomness into RF during feature selection to mitigate high correlation issues, thereby achieving superior generalisation.

Support Vector Regression. Support Vector Regression (SVR) is also a prevalent method which parallels classical SVM principles but diverges in objectives. Leveraging kernel functions like the Radial Basis Function (RBF) kernel, SVR captures nonlinear patterns and aims to find a hyperplane that best fits inputs within a predefined tolerance threshold. It exhibits robustness across dimensionalities and to outliers. However, drawbacks include limited explainability and sensitivity to parameter tuning. Notably, it has fewer extensions evaluated in existing literature compared to other paradigms of machine learning methods.

Neural Network and Deep Learning. Neural networks, mirroring biological neuron structures, have garnered significant attention in recent decades. They offer competitive predictive power, excel on nonlinear problems, generalise well to diverse data forms, and able to autonomously discover patterns from input features, alleviating the need for extensive feature engineering. However, their complexity often renders them akin to "black boxes" lacking interpretability. Moreover, their performance hinges on hyperparameter selection necessitating extensive experimentation for optimal tuning. With advancing computational power, neural network structures have evolved to handle increasingly complex tasks. Convolutional Neural Networks (CNNs) are prevalent in encoding images and videos, leveraging hierarchical convolutional and pooling layers to extract features. Long Short-Term Memory (LSTM) networks, a prominent Recurrent Neural Network (RNN) variant, employ gate units to discern meaning-ful patterns from sequential inputs. Bayesian Neural Networks (BNNs) advocate probabilistic interpretations, valuable for tasks requiring uncertainty quantification. Transformer-based models, the forefront of recent advancements, employ attention mechanisms for parallelised learning of sequential inputs, particularly suited for long-range dependencies.

Others. In addition to the aforementioned machine learning paradigms, K-Nearest Neighbours (KNN) represents another noteworthy approach. Operating as an instance-based method, KNN relies on the proximity of data points to make estimations, leveraging either the majority class or average value without the need for explicit model training. This sim-plicity endows KNN with the versatility to handle diverse input types while maintaining a high level of explainability. However, its efficacy diminishes when datasets fail to encompass all potential scenarios, limiting its performance in such situations.

2.4 Model Evaluation Metrics

Lastly, we summarised prevalent metrics assessed in the literature we reviewed in Fig. 7. We found that Root Mean Square Error (RMSE) and Mean Squared Error (MSE) are the primary metrics employed in regression tasks. In contrast to Mean Absolute Error (MAE), RMSE and MSE exhibit greater sensitivity to outliers. The second most common metric is R^2, which quantifies the goodness of fit by assessing how accurately the predicted points align with the ground truth. Additionally, metrics such as Nash–Sutcliffe Efficiency (NSE) and Mean Absolute Percentage Error (MAPE) find extensive application in agricultural studies and Explainable Artificial Intelligence (XAI). Table 6 provides a summary of prevalent evaluation metrics along with their mathematical expressions.

3 XAI Methodologies & Frameworks in Agriculture

Explainability stands as a central focus of this project and so as this literature review and horizon scan. We classify the XAI methods under review according to the hierarchical taxonomy proposed by Arrieta et al. [2] with some modifications. Methods leveraging widely-used, state-of-the-art XAI techniques such as SHapley Additive exPlanations (SHAP), Integrated Gradients (IG), and Local Interpretable Model-agnostic Explanations (LIME) are directly categorised under these established methods. Alternatively, if the study employs

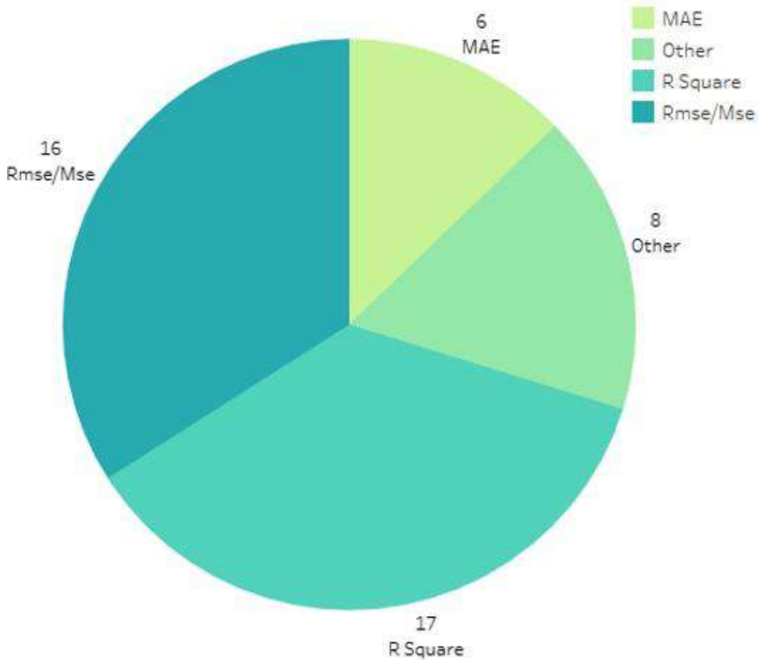


Fig. 7: Evaluation metrics. Note that one paper could use multiple metrics.

Table 6: Mathematical expression about various prevalent metrics

Metrics	Equations
RMSE	$\sum_{i=1}^{N}\sqrt{\frac{(y-\hat{y})^2}{N}}$
R^2	$\left(\frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}\right)^2$
MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert y-\hat{y}\rvert$
NSE	$1-\frac{\sum_{i=1}^{N}(y-\hat{y})^2}{\sum_{i=1}^{N}(y-\bar{y})^2}$
MAPE	$\frac{1}{N}\sum_{i=1}^{N}\frac{y-\hat{y}}{y}$

other methods falling within the post-hoc model-agnostic category, we broadly categorise them as perturbation-based, gradient-based or others, which could be instance-based, model simplification and so on (see Fig. 8).

There are versatile properties that an explanation could possess. Gilpin et al. mentioned that interpretability and fidelity are both necessary components of explainability [38]. Zhou et al. of Data Science Institute (DSI), University of Technology Sydney further decomposed these into five properties: interpretability encompasses clarity, broadness, and

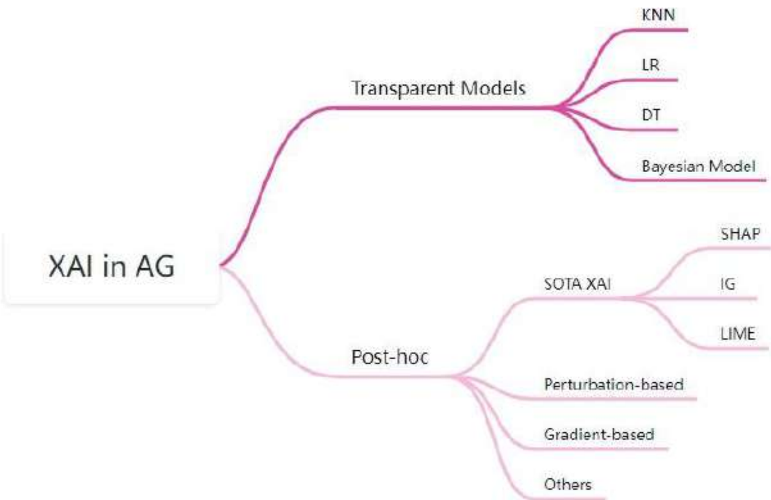


Fig. 8: XAI methods taxonomy in agriculture, modified from [2].

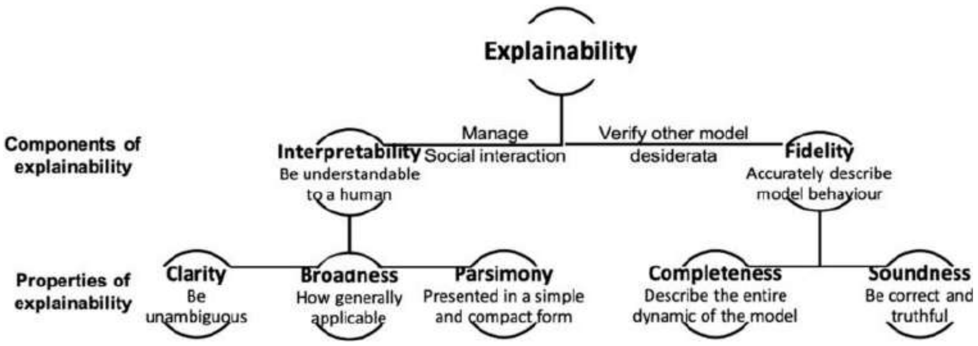


Fig. 9: Definition of AI explainability and related properties, adapted from [3].

parsimony, while fidelity includes completeness and soundness. These properties range from the comprehensiveness to the correctness of the explanation [3] (refer to Fig 9).

In the reviewed literature intersecting XAI and agriculture, almost all studies focus on attribution-based explanations which rank or measure the explanatory power of input features, and primarily concentrating on the property of soundness which reflects how accurately the explanation captures the correct relationship. This observation aligns with the results in their research. However, such ground truth is commonly not accessible. For instance, Hu et al. theoretically validated target models' fidelity by leveraging self-generated data from known functions [17]. They used machine learning models to estimate the functions, examining the accuracy compared to the ground truth. They then applied the theoretically validated model to the yield prediction problem and provided explanations. Similarly, Paudel et al. attempted to obtain the ground truth of explanations via domain experts by gathering and combin-ing expert ranks and using them as ground truth for the SHAP explanation [15]. Beyond yield prediction, other research investigates different properties. For example, Chhetri et al.

utilised a knowledge graph to present more intuitive explanations on cassava disease for better clarification [39].

In general, XAI methods utilised in agriculture can be classified into two main categories: i) transparent models, such as linear regression, decision trees, and K-nearest neighbour algorithms, and ii) post-hoc methods. The former methods are widely used as baselines in the reviewed literature. In linear regression, linear relationship is approximated by minimising its error term. The magnitude and direction of each feature's contribution could be directly observed from the coefficient in the mathematical function. For example, in [17], Hu et al. proposed to use the linear regression model to predict crop yield and underscore the transparency and performance of linear regression model, and suggested that the use of complex black-box model like neural network could be dangerous. The Bayesian ensemble model they proposed successfully reveals two significant change-points in maize cultivation history: i) the year 1944 due to hybrid corn adoption and ii) the year 1976 due to a technological slowing-down. However, transparent models often exhibit limited performance compared to other black-box ML methods, attributed in part to the non-linear relationships between climate variables and crop yield, as noted by Cai et al. [10]. The latter involves explaining the model's decision-making process after model training, providing insights into how predictions are generated. Through our examination, we observe that the majority of XAI methods adopted fall into the post-hoc category. This preference may stem from the versatility of inputs available for yield prediction.

We further categorise post-hoc XAI methods into four distinct categories: state-of-the-art (SOTA) XAI methods, perturbation-based XAI, gradient-based XAI, and other model-specific methods. While these categories may have some overlapping with each other, it is crucial to emphasise the importance of employing SOTA methods for explaining yield. Under the umbrella of post-hoc explanation, there is a noticeable scarcity of model-specific XAI methods in the reviewed literature. This observation may be attributed to the diverse range of evaluated baselines, as depicted in Fig. 6, where the utilisation of model-specific XAI methods might not ensure fair evaluation across various models.

SHAP, grounded in cooperative game theory, quantifies feature contributions by comparing predictions with and without each feature across all possible combinations. However, due to its exponential time complexity concerning the number of features, SHAP is typically employed in research work with smaller input feature spaces. For instance Paudel et al. compared a Neural Network (NN) in deep learning without feature selection and a Gradient Boosting Decision Tree (GBDT) with expert feature selection over wheat and maize. Their results show that the NN can achieve statistically similar or better performance compared to conventional machine learning methods [4]. SHAP is used to accumulate individual feature contributions, highlighting important features such as water holding capacity and water-limited dry weight storage organs. Furthermore, the NN with SHAP identifies the negative effect of climate water balance on soft wheat in the early season, which is an insight not anticipated by experts. However, the study noted certain limitations in capturing the influence of extreme events.

On the other hand, Integrated Gradients (IG) is a model-specific method that integrates the model's gradient across input features from a baseline to the actual input. Mateo-Sanchis et al. compared monthly features' contributions using both IG and SHAP across three kinds of features. They found that the explanations between SHAP and IG align in direction but differ in magnitude [9]. IG's usage is expected to be limited, as most reviewed research leverages tree-structured models, linear regression, and support vector regression, which do not heavily rely on gradients for optimisation. Additionally, Paudel et al. stated that IG is less stable than SHAP in their study [4].

LIME, a well-known local explanation method belonging to perturbation-based methods, employs simple transparent models to interpret individual predictions made by black-box models. It perturbs feature values around the target prediction's inputs to provide insights. Ahmed et al. leveraged LIME to identify the feature importance for individual sample over three types of rice and concluded that the most impactful features are a combination of meteorological, agro-chemical, and soil physio-graphic factors [40]. Similarly, Ryo also pointed out using LIME to reveal locally important variables can differ from the global ones because of the spatial condition difference [21]. Other perturbation-based methods examine trained model behaviour by altering inputs to discover patterns for explainability without further employing any other transparent models. Perturbation here could involve changes in input feature values or the inclusion/exclusion of certain features. For instance, Nayak et al. shuffled features value and according to the magnitude of performance changes to determine the feature importance [12], while Cai et al. chose to eliminate certain type of feature and see the performance difference to rank the feature importance [10]. Notably, perturbation-based methods commonly do not require additional assumptions on model structure or optimisation processes ensuring their model-agnostic characteristics.

In contrast, gradient-based XAI is prevalent in research utilising shallow neural networks and deep learning models. Gradients are utilised as weights to calculate relevance between inputs and predictions. For example, in a trained Convolutional Neural Network (CNN) model, inputting an image of a tree produces a prediction along with its probability. By com-puting the gradient of the 'tree' class with respect to the final CNN layer, the significance of each feature map is discerned. Aggregating these contributions through weighted summa-tion and activation functions generates a heatmap, highlighting regions most pertinent to the 'tree' prediction. Not only the conventional 2D CNNs, 1D CNNs are widely employed in research, the gradient approach could also offer meaningful explanatory insights. For exam-ple, Khaki et al. backpropagated the positive gradients on CNN to find input variables, and proposed that a larger neuron's gradient indicates more important role that feature plays in model prediction [19].

Table 7 provides an overview of representative works reviewed following the taxonomy in Fig. 8. Unlike Fig. 6, the "Models" column herein only includes the primary or best-performing method, as opposed to all evaluated methods. It is evident that linear regression models inherently possess self-explanatory properties, obviating the necessity for additional techniques to achieve explainability. Although histograms are commonly employed to visualise feature importance, supplementary explanation methods are requisite for comparative analysis with other baselines, particularly black-box models. For example, Hu et al. utilised Partial Dependence Plots (PDP) to elucidate explanations obtained from alternative black-box baselines [17]. Furthermore, we observed a wider spectrum of visualisation techniques employed in various machine learning methods, excluding linear regression.

While Decision Trees are inherently interpretable models, most studies treat them merely as rudimentary baselines, focusing predominantly on Random Forests or Gradient-boosted Trees, which lack transparency and necessitate further explication. Perturbation-based XAI methods enjoy widespread adoption in these studies, primarily owing to their model-agnostic nature. Perturbation-based methods are prevalent not only in tree-structured models but also in other machine learning methods. The information summarised in Table 7 illustrates the prevalent use of perturbation-based methods, including LIME and SHAP, for achiev-ing explainability. Conversely, due to the assumptions of gradient optimisation processes, we observe the utilisation of IG and gradient-based methods solely in research involving Neural Networks and Deep Learning. Adoption of SOTA methods such as LIME and SHAP commonly corresponds to specific visualisation techniques; for instance, studies employing SHAP commonly utilise Beeswarm plots to interpret the magnitude and direction of feature

Table 7: XAI methods used in agriculture.

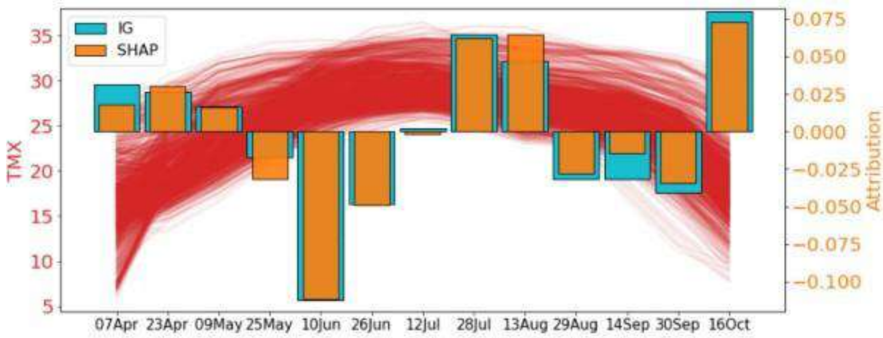
XAI Methods	Publications	Models	What Explained?		Visualizations[3]	Important Features[4]	Eval	
			FI[1]	FC[2]			H[5]	F[6]
Transparent Model	[17]	LR	✓	✓	Hist; PDP	T		✓
Transparent Model	[13]	LR		✓	Hist			
Transparent Model	[28]	LR	✓		Contours map; Table	P; T		✓
SHAP	[4]	NN&DL	✓		BsP; Hist	Biomass; WHC	✓	
SHAP	[22]	Tree	✓		BsP	Near-infrared		
IG;SHAP	[9]	NN&DL	✓		BsP; Hist	T		
LIME	[11]	Tree	✓		Hist	T; WHC		
LIME; Perturbation-based	[21]	Tree	✓	✓	BP; Hist; LC; PDP	T		
Perturbation-based	[10]	SVR	✓	✓	Hist; LC; Table	P; T		✓
Perturbation-based	[29]	Custom	✓		BP	Irrigation		
Gradient-based	[19]	NN&DL	✓		Hist	Rad; T		✓
Model-specific	[16]	Tree	✓		BP	P; T		

[1] Feature Importance.
[2] Feature Correlation.
[3] Hist denotes Histogram, PDP denotes Partial Dependency Plot, BsP denotes Beeswarm Plot, LC denotes Line Chart, BP denotes Box Plot.
[4] Important Features concluded from their research, T denotes Temperature, P denotes Precipitation, WHC denotes Water Holding Capacity, Rad denotes Radiation.
[5] H denotes Human-centered evaluation.
[6] F denotes Functionality-grounded evaluation.

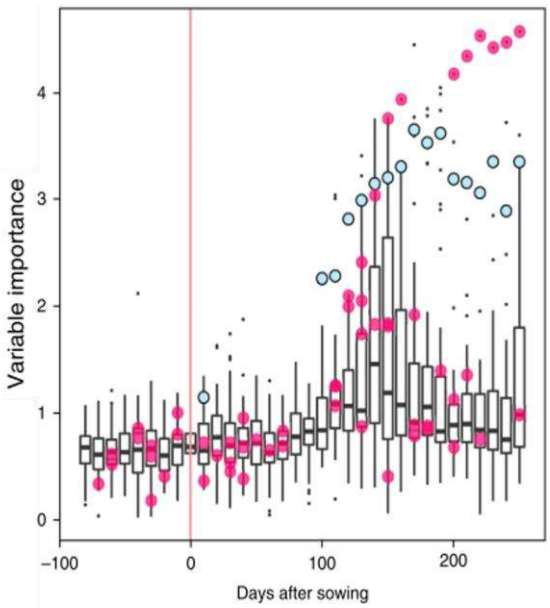
contributions concerning value changes [22, 25], while those leveraging LIME prefer positive and negative axis histograms to depict the magnitude and direction of feature contributions [11, 29].

At a high level, we categorise "What Explained?" into two types: feature importance and feature correlation. It is obvious that feature importance is widely examined across research community. Histogram and box plot are the dominant approach to visualise the feature importance. Histogram presents high-level overview about feature contributions and could averaged the contribution of features across time, space, and crop type, while box plot somewhat maintains this information. Due to the potentially interesting pattern of these features across different time, many researchers prefer to use the time span as x axis, while the importance score as the y axis, combining with their innovations, which generates many customised informative plot (see Fig. 10). We record the most important features they conclude ("Important Features" column in Table 7), if there are multiple kinds of crop evaluated, we record the most important feature for each crop. We can see that features related to temperature are the most important feature concluded by many researchers. However, the feature importance is spatial, temporal and crop-type sensitive, it is impossible to claim a feature is important without specifying the time, location and crop type. Even under the same aforementioned conditions, different representation of features could result in different outcomes. These variations make the parallel comparison across research becoming less meaningful and inaccurate.

Conversely, feature correlation receives comparatively less attention in the literature. We suspect the reason is that researchers are more interested in the temporal patterns, while advanced correlation analysis without temporal consideration, such as PDP, is already two-dimensional. Incorporating time into such analyses introduces an additional dimension to visualisation, and three-dimensional representation is hard to present in documents. Kern



(a) Customize histogram visualization, adapted from [9].



(b) Customize box plot visualization, adapted from [16].

Fig. 10: Customized visualization wrt. time.

et al. proposed to combine the feature correlation matrix with the histogram to present the feature correlation difference across a few months, while the trade-off is changes in pair-wise features and its influence on yield is ignored [13]. These ignored information could also be interesting, for instance, PDP presents in [17] demonstrated that increased precipitation and temperature correlate significantly with yield enhancement, with temperature changes showing a more pronounced effect according to their use of colour gradients.

The evaluation of XAI methods often remains overlooked in research endeavours. It is arguably out of scope to this project but briefly discussed for the sake of completeness.

Assessing explanations holds paramount importance, yielding dual benefits. First, it facilitates the comparison of available explanation methods. Despite the inherent challenge of lacking a ground truth due to the opaque nature of model internals, it enables the identification of preferred explanations. Second, evaluation serves as a litmus test for ascertaining whether explainability objectives are met within a given application context, focusing on the efficacy of provided explanations [3]. To systematically categorise and analyse the literature, we present the taxonomy delineated by both Doshi-Velez and Kim [41] and Zhou et al. of DSI [3], classifying reviewed studies into two overarching classes: Human-centered evaluation and Functionality-grounded evaluation. The latter, Functionality-grounded evaluation, relies on proxy metrics to objectively measure explanation quality without direct human involvement.

Under the umbrella of Functionality-grounded evaluation, using another transparent model to achieve explanation remains relatively scarce in the reviewed literature. Hence, there is no such related evaluation in reviewed literature. Evaluation on feature explanation (also known as attribute-based evaluation) emerges as predominant, with most studies validating their explanations by assessing model performance deviations with and without identified salient features. Notably, Hu et al. adopted a novel approach, validating explanation soundness by employing simulated inputs with ground truth variable importance [17]. Example-based explanation also features prominently in reviewed papers, albeit without explicit measures of representativeness or diversity; these examples typically complement primary explanation methods or serve as validation aids.

In contrast, human-centered evaluation, though inherently subjective, provides direct and robust evidence-based support for the effectiveness of explanations. While Rojat et al. suggested that qualitative evaluation may not be suitable for explanations involving temporal patterns [42], a notable exception exists where comprehensive human-centered evaluation has been conducted. Paudel et al. [4] engaged several domain experts in feature selection for non-neural network structured models and conducted a survey to predict the most influential features before model training. Furthermore, SHAP explanations were subsequently compared with expert empirical predictions to assess explanation quality (see Fig. 11). Domain experts also evaluated the magnitude and direction of SHAP explanations using Agree or Disagree quadrants (see Fig. 12a). Fig. 12b illustrates a quadrant evaluation on wheat, with numbers in brackets denoting the number of experts agreeing with each classification. This human-centered evaluation represents a noteworthy endeavour and could serve as a guideline for future research involving human evaluation.

4 Case Studies of Climate Impacts on Apples, Grapes, and Kiwifruit

In this section, we will conduct case studies on the effects of meteorological/climate conditions to crop yield in three types of crops, namely apples, grapes, and kiwifruit in the United States, Australia, and New Zealand respectively. Apples, grapes, and kiwifruit are woody perennial crops. Most of the prior research has indicated that the woody plant presents stronger correlation to the effects of climate change compared to the herbaceous plant. Therefore, the chosen crop types are good instances that can help us to better understand the relationships between crop yield and climate conditions. In general, similar to other herbaceous crops, apples, grapes and kiwifruit are vulnerable to extreme events like frost, especially in bud-break and vegetative phases. All the selected crops prefer warm but dry air condition while the soil remains moist. Calm, windless, and moderate temperature are preferred in their flow-ering phase. Rainfalls before or during harvest phase could help the fruits to grow larger;

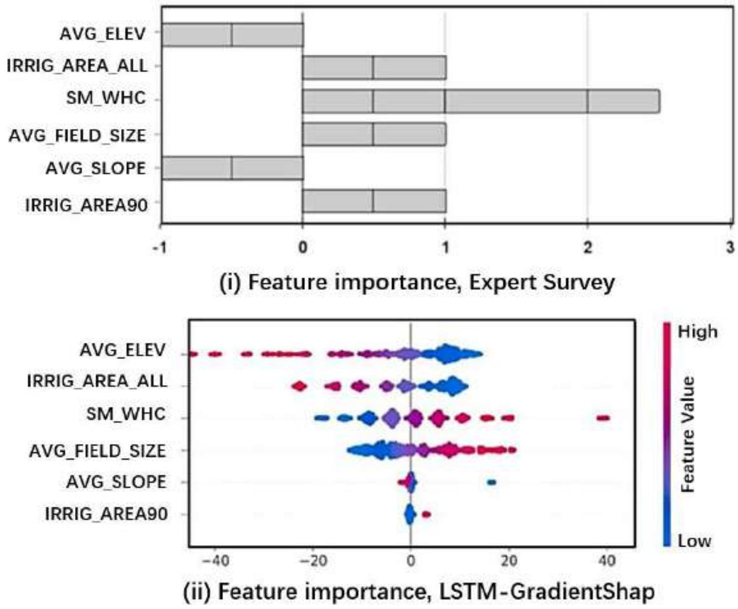
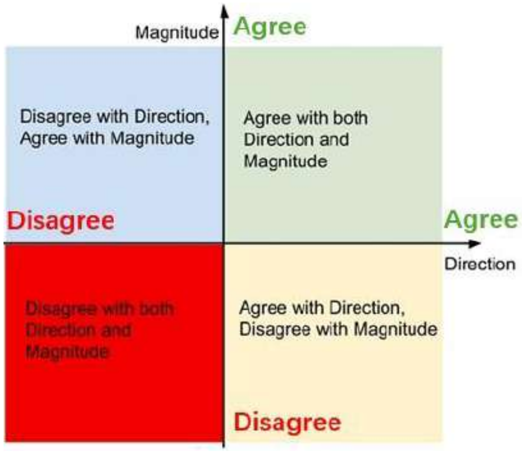


Fig. 11: Expert explanation and SHAP explanation on LSTM, adapted from [4].

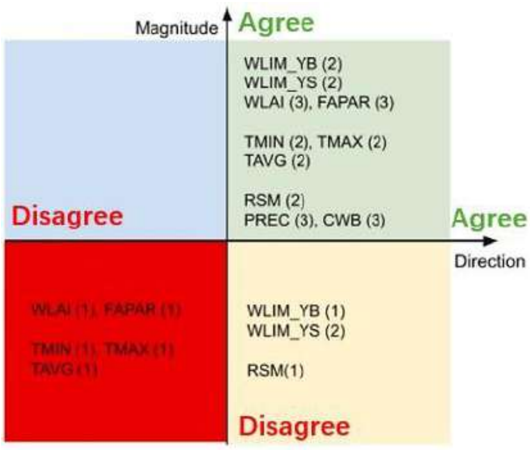
Table 8: Significant climate features for different growing phases of apples, grapes, and kiwifruit.

	Dormancy	**Budbreak**	**Vegetative**	**Flowering**	**Fruitset**	**Veraison**	**Harvest**
Apples	Sufficient chilling hours	moderate temperature + rainfall	-	Moderate temperature + rain-free + wind-free	Warm temperature + moderate rainfall	-	Cool temperature + moderate rainfall
Grapes	Sufficient rainfall	Cool temperature + moderate rainfall + less wind	Warm temperature + less rainfall	Moderate temperature + rain-free + wind-free	Cool temperature + moderate rainfall	Cool temperature + moderate rainfall	Cool temperature + moderate rainfall
Kiwifruit	Sufficient chilling hours + rainfall	Cool temperature + moderate rainfall	-	Moderate temperature + rain-free + wind-free	Warm temperature + moderate rainfall	-	Warm temperature + moderate rainfall

while dry conditions in harvest allows for concentrated flavour, and it is particularly relevant to grapes. One thing we noticed is that, different from herbaceous crops, woody crops underscore the winter chilling hours – the longer the chilling hours in winter, the better yield and flavour the crop could offer in the subsequent year. Due to the literature about apples, grapes, and kiwifruit that involves XAI are limited, we extend our investigation to the literature related to climate change, and the selected crops as well as their industrial reports. Table 8 summarises the critical climatic features affecting different growth phases across the three types of crops.



(a) Quadrants for expert evaluation.



(b) Expert quadrants result.

Fig. 12: Expert quadrants and evaluation result, adapted from [4].

4.1 Apples

Apples are one of the most common fruits in many countries. With advancements in bio-genetic and cultivation technology, the annual production of apples has gradually increased; however, climate change presents significant challenges. Over the past 30 years, insurance claims related to apple production have increased by 175%, indicating that extreme weather events are becoming more frequent [43]. Apart from the influence on quantity, research also reveals that climate change also appears to cause a reduction in fruit sugars, thinner skin and an increase in fruit acidity leading to tarter tasting fruit. The fruit in the climate change chambers were producing double the amount of ethylene, suggesting that harvesting will be advanced due to the fruit being more mature [44].

Apple growers from the United States provided insights into the effects of extreme event on yield, such as in [43]. The grower notes that frost in April and May during the spring in the United States can be detrimental to apple growth. Frost can kill young buds and shoots, leading to reduced production. In 2012, an early spring with premature heat accumulation caused an advanced flowering phase. The subsequent severe frost resulted in a 90% reduction in the apple crop, marking the most significant disaster since the 1940s [45]. Recently, from late-spring frosts in western Michigan to heatwave in the Pacific Northwest, apple growers saw a nearly 19% drop in fresh-market apple holdings in June 2021 compared with June 2020 [46].

During the bud-break and flowering phases, temperature is a critical variable. Similar to other crops, low temperatures can extend the apple growing season [27]. While research on herbaceous crops suggests that high temperatures destroy pollen proteins, studies on apples indicate that long-term low temperatures delay bud breakage and cause flowering failure, negatively impacting production [27]. Additionally, temperatures exceeding 30°C are recommended during the blooming phase. Li et al. found that not only temperature but also spring rainfall significantly influences apple yield [47]. Rainfall during the flowering phase can reduce yield, as it is often associated with low temperatures, complicating the determination of causality in this interconnected relationship.

In the summer, rainfall becomes crucial for apple growth. High summer temperature and erratic rainfall can increase environmental humidity creating a hotbed for disease and pest damage [43]. Conversely, limited rainfall poses significant challenges to irrigation systems. In 2018, some apple growers reported the need to entirely rely on irrigation due to a lack of rainfall. Despite the pressure on water usage, overall apple production was not significantly affected [48]. This scenario aligns with survey results indicating that water features are considered the least concern for apple growth [45]. However, some growers note that irrigation water can lead to smaller apples, while rainfall can promote larger apple growth [43], which aligns with the existing research finding [47]. In the United States, heatwaves in June can also impact yield and cause negative secondary effects in the supply chain. Discussions about autumn are relatively rare, but, analogous to other fruit crops, temperature is likely a critical factor. High temperatures can accelerate fruit maturity demanding higher requirements on storage and transportation. In Australia, the industry noted that mild sunburn browning injury often becomes less apparent as fruit colour intensifies toward harvest. Nevertheless, it adversely affects the quality of fruit resulting in diminished flavour and a drier texture [49]. On the other hand, and surprisingly, there is no clear connection between apple's size and quality found from the reviewed literature.

Although there are fewer activities for apple growers in winter, this season is crucial for apple growth [47]. Winter chilling hours are directly related to the next year's yield; the longer the chilling hours, the better the yield. Gallardo et al. provided similar conclusions, indicating that warm winters can favour pest and disease proliferation, reducing subsequent yields [45]. Moreover, warm winters can cause early temperature accumulation, placing early-maturing crops at risk of low temperatures and frost. Li et al. also concluded that the El Niño-Southern Oscillation (ENSO) in winter is highly correlated to the yield, which also validate that the warm winter could make negative impact yield on apples yield but exhibit spatial variations [47].

4.2 Grapes

Grapes are another common fruit consumed worldwide, and they are also the most crucial ingredient for wine making. This case study primarily focuses on the climate effects on grape production in Australia. Australia has a well-established and professional grape and wine industry, with significant spatial variance between grape-growing regions. Similar to apples, grapes are vulnerable to frost. The occurrence of frost during the bud-break and flowering phases can damage young shoots and buds leading to reduced yields [50]. In 2014-2015, non-seasonal frost events in late spring and summer resulted in substantial yield reductions. Mosedale et al. also pointed out the hail and storm could physically damage the vines and grapes, leading to immediate yield loss and long-term vineyard damage [51].

Apart from frost damage, extreme events such as ENSO and La Niña can also influence grape yield. Compared to La Niña, ENSO has a relatively smaller impact on grape yield; regions such as Riverland even benefit from ENSO teleconnections [50], indicating that grapes prefer hot and dry environments rather than cold and wet conditions. La Niña brings rainfall along with relatively cold temperatures that can adversely impact grape yield. For instance, 2023 was the wettest year since 2011 and the coldest year since 2012, resulting in $229 million less grape crush used for the wine industry compared to 2022 [52]. Persistent torrential rainfall leads to flooding as well; although not many orchards were directly damaged, the secondary effects, such as delayed spraying and chemical deficiencies, led to disease. Additionally, La Niña caused delays in harvest time and prolonged the growing duration, which led to supply chain pressures and subsequent production reductions.

Seasonal climatic features can also influence grape yield and quality, presenting diverse effects across different growth phases. Unlike apples, discussions about winter chilling hours for grapes are limited. Winter rainfall can moisten the soil in advance, which is beneficial for grape growing. However, the effect of rainfall during the bud-break phase appears to be spatially sensitive. Moderate rainfall in spring is desirable in many regions, aligning with research in Europe [28], whereas in the Barossa Valley, drier soil and warmer temperatures are preferred [34]. Similar to apples, warmer temperatures in the early stages are not always beneficial; they facilitate growth but increase the risk of frost damage [51]. Appropriate cool temperatures can delay bud-break slightly; so that as the climate warms, fewer frost days are expected and hence reducing the risk. Wind is a factor that many studies have overlooked. Windy conditions in spring may lead to uneven growth and a yield drop [34].

In summer, grapes are usually in the phases of vegetation, flowering and fruit-set. During these phases, rainfall and temperature are crucial for grape yield. Santos et al. concluded that higher temperatures and less rainfall in summer could favour higher grape yield [28]. Warmer conditions can accelerate crop growth and shorten the growth duration; but anomalously high temperatures may result in heat stress, reducing yield by causing berry shrivel or sunburn [50], and affecting wine quality due to altered chemical composition of the grapes [51]. While rainfall nourishes crops, it also prolongs growth duration and increases the risk of disease and pest invasion. The Australian wine industry also points out that the nutrients from rainfall can lead to larger canopies, which hinder fruit from receiving adequate solar radiation and require regular trimming and hedging to encourage sunlight penetration on bunches [52]. The effects of summer wind is also noteworthy and can generate two-fold effect – mild winds can prevent disease, whereas heavy winds can negatively influence yield by causing defoliation or even breaking the vines.

At the end of the fruit-set and veraison phases, preferred conditions change slightly. Moderate rainfall during these phases is appreciated as it can enlarge the fruit, but it also increases air humidity raising the risk of disease. Cooler temperatures are beneficial in these phases, as they ensure more even ripening. On the contrary, ripening under higher temperatures may result in increased water stress such that reduce grape quality and yield [51]. For example, in 2016, a long-term lack of rainfall in the Barossa Valley resulted in quicker ripening and smaller grapes [34]. On the other hand, in 2020, long time drought corresponding with bushfire and other extreme event led to small crops and reduction in yield nationally.

Regarding long-term climate change, there is a widely accepted trend that the globe is warming, with increased carbon dioxide (CO_2) emissions and reduced global rainfall. It is important to acknowledge that the influence of this trend may vary among different grape varieties. However, overall, it benefits grape growth by creating more suitable environments and enhancing photosynthesis, making more sub-optimal locations viable for grape cultivation, thereby benefiting the grape and wine industry [28, 50, 51]. However, challenges and risks also exist, requiring the grape industry to anticipate, prepare for, and adapt to these upcoming opportunities and challenges. For example, different varieties of vines requiring different climates and vines take a long time to establish (commonly 5 years+) and hence there is an inherent adaptability problem.

4.3 Kiwifruit

Kiwifruit are among the most nutritious fruits and is cultivated commercially worldwide in diverse regions such as China, Italy, and New Zealand [53]. It brings substantial profit to local communities. For instance, in New Zealand, kiwifruit is the largest horticultural commodity by value and plays a significant role in the local economy [54]. Due to geographical proximity and the woody vine crop type, similar trends are observed in specific years when compared to the grape industry in Australia, particularly during the coldest years. In 2014, for example, a widespread frost impacted the Australian grape industry, while hail caused substantial losses in the New Zealand kiwifruit industry. Additionally, in 2022 and 2023, the cold and wet climate conditions caused by La Niña led to large-scale grape yield reductions in Australia, and the kiwifruit industry in New Zealand experienced similar yield reductions, partly due to frost and hail [55]. In 2002, a severe frost in late September at the Te Puke region resulted in over 350 frost insurance claims, with crop losses ranging from 30% to 100% [56].

According to the interview in [54], extreme cold events are among the most harmful effects to kiwifruit. While hail not only affects the current yield but also impacts the subsequent year [36]. However, during years of excessive drought, there is no obvious similarity between these two crops. Survey results indicate that drought is one of the least significant concerns for kiwi growers, possibly because irrigation can compensate for limited rainfall. Conversely, the top problems growers encounter are storms, frost, and floods. The discrepancy between the survey results in [30] and interviews in [54] is notable when considering the timing of extreme events. The devastating cyclones and floods in 2022 may have led survey participants to emphasise the impact of these events, while interviews in 2014 highlighted the damage caused by frost and hail. This suggests that qualitative analysis can be subjective and event-sensitive. Furthermore, interviewed growers referred to a "lack of predictability" or "increased variability" in weather patterns, which is attributed to the effects of ENSO and IPO.

The effects of seasonal climatic features on kiwifruit yield are analogous to those observed in other crops. Temperature is the most critical determinant of yield. An increase in average temperature has a dual effect – while warmer conditions can enhance crop growth, they also introduce significant risks. Warmer temperatures can extend the growing season, support canopy development, and result in higher quality yields, such as increased dry matter [55]. Projections of climate change indicate a moderate increase in temperature across all kiwifruit growing regions, with an anticipated rise of approximately 1°C by the 2040s and 2°C by the 2090s [57]. Due to such underlying warming trends, warmer temperatures can also make previously sub-optimal regions suitable for cultivation.

However, warmer temperatures elevate the risk of invasive pests and diseases, as well as frost exposure due to early bud breaking. Warmer spring temperatures may hinder consistent bud-break and king flower production, and extreme heat during summer can cause heat stress and increase water demand for both crops and growers. Additionally, warmer winter temperatures can disrupt crop dormancy leading to reduced yields. Kiwifruit requires 950 to 1100 hours of exposure to temperatures below 7.2°C to ensure adequate bud burst in spring, similar to apples [54]. Both green and gold varieties are impacted by a reduction in winter chilling; however, green varieties typically experience a more significant decline in both fruit yield and quality [58]. To address the insufficient winter chilling hours, kiwifruit growers used chemical bud break enhancers to compensate for the winter chilling variance [58]. Although such products may be de-registered for use in the New Zealand kiwifruit industry, as has occurred in the European Union, existing research has already begun seeking alternatives to meet upcoming climate and policy challenges [59].

The influence of rainfall on kiwifruit is relatively minor compared to temperature. Rainfall during dormancy is necessary to maintain soil health. During the flowering phase, rainfall negatively impacts kiwifruit yield, as calm, windless, and rain-free conditions are preferred. Moist conditions in spring and autumn are also detrimental to kiwifruit yield [54], and excessive rain increases the risk of flooding and is closely linked to kiwifruit vine decline syndrome (KVDS) [36]. During the harvest phase, adequate rainfall or irrigation is crucial to support fruit enlargement and prevent water stress [56]. In addition to temperature and precipita-tion, solar radiation is also a significant factor. Radiation during the early stages of growing season supports photosynthesis and energy accumulation for growth, while in the late grow-ing season, it aids in cell division and the accumulation of sugars and other dry matter [56]. Cloudy weather is generally unfavourable for kiwifruit development [54]. However, it is chal-lenging to assert this effect definitively due to the high correlation between cloudy conditions and temperature.

5 Part II Conclusion

The objective of this project is to identify a robust and repeatable framework by which Agricultural AI models can be used to understand the implications of climate change and other risks on agricultural crop outcomes. During our literature review, we noticed a significant challenge that current research faces, which is counterfactual explanations due to feature correlations. For instance, Newman and Furbank observed that increased herbicide application correlates with decreased yield [16]. Following such explanations to improve yield would likely result in weed proliferation, thus reducing crop yield. The counterfactual explanation may stem from the land being in poor condition, necessitating herbicide application, thus leading to the conclusion that herbicide application reduces yield.

Several factors may contribute to these counterfactual explanations. First, insufficient features can lead to causal confusion. Expanding on the previous example, providing additional field information might enable the model to discern spatial differences, potentially rectifying causal relationships. Second, the significance of individual features may be diluted by other highly correlated features, rendering their explanations inaccurate. For example, the inclusion of soil moisture might diminish the contributions of irrigation and precipitation. A potential solution is to pre-select features by calculating mutual information or correlation to include independent variables for training.

Part III
Yield and Climate Data Preparation

6 Part III Introduction

In this project, our primary goal is to analyse the impact of extreme weather events on agricultural yields, with a specific focus on grapes in Australia, kiwifruit in New Zealand and apples in the United States. Part III of this report centres around data collection, which is critical for achieving two key project objectives:

- Identifying extreme weather events using climate data, involves establishing a relationship between climate patterns and these events through proper labelling.
- Assessing the impact of extreme weather on crop yield by integrating climate data, extreme weather event data, and crop yield data. The extreme weather event data will likely be derived from the first objective.

To achieve these objectives, we need robust datasets that encompass both climate and yield data. In Part III, we outline the methods that we used to collect the necessary data and explain the rationale behind our selection of datasets. Specifically, we focus on the collection of climate and crop yield data.

For Australia, we sourced SILO and IBM climate datasets and farm-level grape yield data from The Yield Technology Solutions (now Yamaha Agriculture), while extreme event records from the Bureau of Meteorology (BoM) were manually compiled as auxiliary information to support downstream analysis. For New Zealand, we collected daily climate data from weather stations and linked it to the yield data of kiwifruit farms from The Yield Technology Solutions (now Yamaha Agriculture). We ensured that the climate data adequately covered the same period as the yield data, making it suitable for our analysis. For the United States, we gathered both state-level and county-level climate and yield data for apple production. The state-level data provides a comprehensive overview, allowing for the analysis of climate impact across multiple states. However, variability within states required us to focus on data from the counties with the highest apple yields. At the county level, while the data offers greater granularity, it comes with limitations, such as outdated yield data and a lack of ground truth for extreme weather events.

By collecting this data, we aim to establish a strong foundation for the next phases of the project, where we will analyse extreme weather events and their effects on agricultural production.

7 Australian Data - State-Level Extreme Event

Since both the Australia climate and grape yield data were readily available, we only focus on extreme event data collocation in this section. Please note that a comprehensive data retrieval and modelling on climate extreme events will be conducted in Part V later in this report.

Australian agricultural landscape is highly sensitive to climate variability, making it essential to track extreme weather events with care. To the best of our knowledge, there is no publicly available dataset that recorded the extreme events across Australian states with comparable temporal coverage and detail. To incorporate extreme event information into our yield prediction modelling, we compiled a dataset of significant weather events at the state level using authoritative reports from the Bureau of Meteorology. While the process involved manual interpretation of descriptive and visual materials, the resulting dataset offers a structured summary of major events spanning nearly two decades. This section outlines how the data were collected, what information is included, and a brief overview of the dataset.

7.1 Data Collection Method

Extreme event records were manually compiled from the Bureau of Meteorology's Special Climate Statements[2], which provide detailed commentary on significant weather and climate anomalies across Australia. These reports, issued irregularly between 2006 and 2024, offer insights into events that are unusual in the context of local climatology. In total, 71 statements were reviewed.

All reports are published as PDF format and often rely on expert visualisation to communicate spatial impact. Since structured data tables were rarely provided, we opted for a coarse annotation approach. Specifically, we assessed each report to determine which states were affected. A state was considered impacted if the majority of its area, as described or shown in the report, was involved in the event. For each confirmed case, we manually recorded the state name, the start and end dates of the event, and the event type as described in the report.

While this method introduces a degree of subjectivity and lacks sub-state resolution, it offers a consistent framework for compiling nationwide extreme event records in the absence of machine-readable alternatives.

7.2 Datasets Information

The final dataset includes 188 records. This count exceeds the number of source reports, as many events impacted multiple states and were recorded separately for each one. Each record includes:

- State: The affected Australian state or territory.
- Start Date: The beginning date of the event within the state.
- End Date: The end date as reported.
- Event Type: Categorised as *Heat*, *Cold*, *Rain*, *Drought*, or *Storm*.
- Report ID: The associated Special Climate Statement number from BoM.

7.3 Summary

This manually assembled dataset provides a high-level view of extreme weather patterns across Australia, capturing events of national relevance over an 18-year period. Although limited by its spatial granularity and interpretive nature, the dataset reflects information drawn directly from the country's most authoritative climate reports. As such, this dataset serves as a valuable resource for linking climate extremes to agricultural or environmental outcomes, particularly when finer-resolution meteorological data are lacking.

[2] http://www.bom.gov.au/climate/current/statements/

8 New Zealand's Data - Climate & Kiwifruit Yield

In this section, we first summarise our data collection methods and then provide detailed information.

8.1 Data Collection Method

We already obtained Kiwifruit Yield Data spanning from 2016 to 2024. Additionally, using the combined information AP provided by CliFlo[3], we downloaded daily climate data from 318 stations, encompassing 11 features. CliFlo is a web-based system that provides access to New Zealand's National Climate Database, managed by National Institute of Water and Atmospheric Research (NIWA)

Upon reviewing the API, we found that it allows us to download combined information such as Sun (hours), Radiation (MJ/m^2), Cloud Cover (octas), Visibility (km), Gale occurrences, Snow, Hail, Lightning, Thunder, Fog, and more. We can also download specific data like rainfall, surface wind, and minimum and maximum temperatures. However, not all stations provide all of this information; the available features vary between stations. The climate data generally ranges from December 2008 to January 2021, covering approximately 12 years, though data availability varies between stations.

Here is a summary of the datasets:

- Kiwifruit Yield Data: This dataset includes yield data from 2016 to 2024, covering 3,437 farms.
- New Zealand's Climate Data: We have daily climate data from 318 stations, covering 11 key features: WDir(Deg), WSpd(m/s), GustDir(Deg), GustSpd(m/s), WindRun(Km), Rain(mm), Tdry(C), TWet(C), RH(%), Tmax(C), Tmin(C), and Rad(MJ/m2).

8.2 Datasets Information and Assignments

To match the climate data with the kiwifruit farms, we identified the 10 nearest stations for each of the 3,437 farms listed in the kiwifruit yield dataset and verified the availability of corresponding climate data. The results indicate that for all farms, at least one of the 10 nearest stations has associated climate data. We also calculated the distance between each farm and the nearest available station. The distribution of these distances is shown in Fig. 13. As seen in the figure, nearly all farms are located within 15 km of a station, with the majority within 10 km.

Additionally, we verified the date range for each farm's yield data and cross-referenced it with the corresponding climate stations to ensure data coverage for the same period. The results indicate that most of the required years are sufficiently covered. We visualized the coverage of climate data for both farms and stations across different years, as shown in Fig. 14, Fig. 15, and Fig. 16.

8.3 Summary

For the New Zealand climate data, we identified the 10 nearest stations for each of the 3,437 farms listed in the kiwifruit yield dataset and verified the availability of corresponding climate data CSV files. The results indicate that for every farm, at least one of the 10 nearest stations has associated climate data. We also verified the date range for each farm's yield data and

[3] https://niwa.co.nz/climate-and-weather/national-climate-database

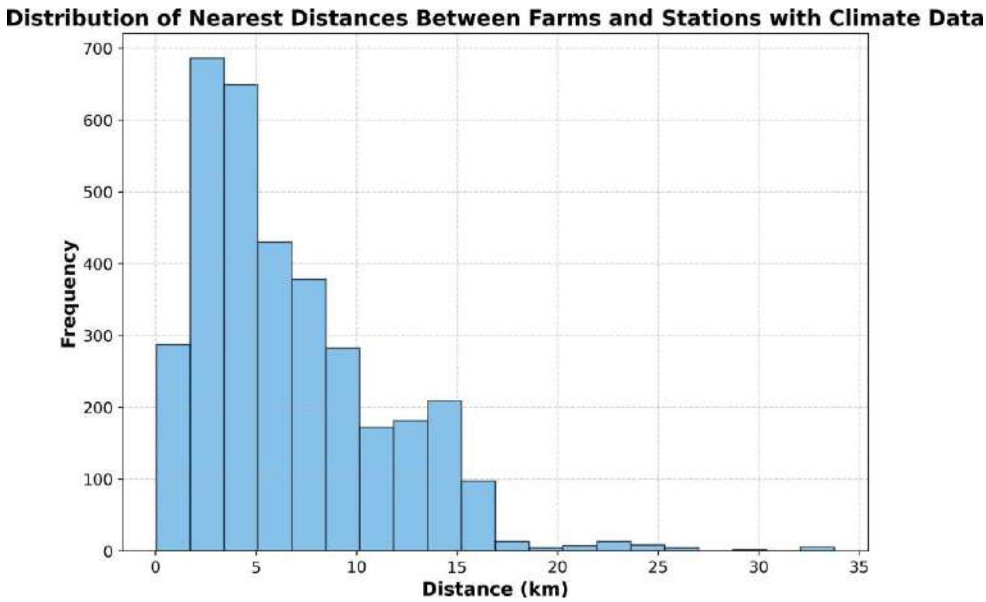


Fig. 13: Distance of the Nearest Weather Station from a farm. The histogram displays the distribution of the nearest distances between farms and their closest stations where climate data is available.

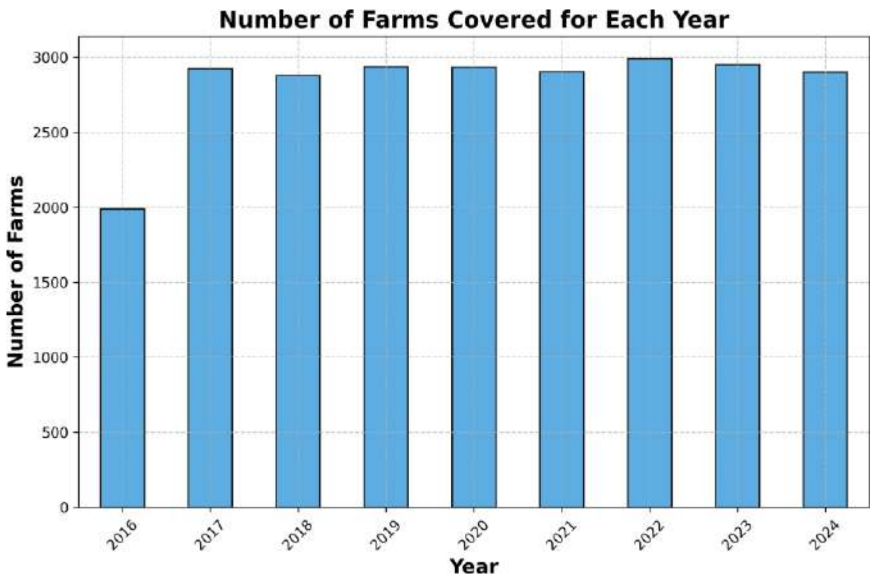


Fig. 14: Number of Farms Covered for Each Year: This chart shows the number of farms with climate data coverage per year, highlighting how many farms have full data availability over time.

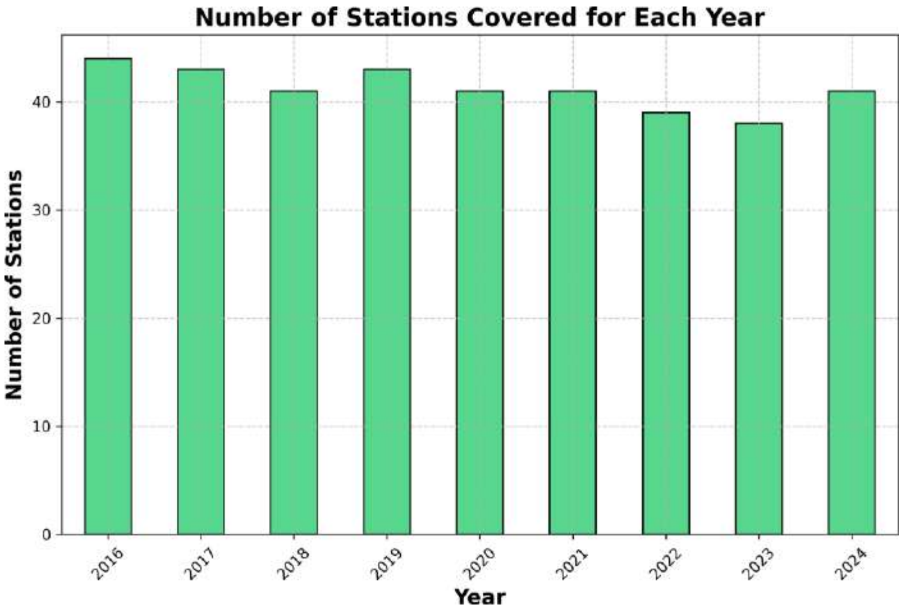


Fig. 15: Number of Stations Covered for Each Year: This chart represents the number of climate stations that have data for the respective years. It's useful for identifying if station coverage remains consistent or fluctuates over time.

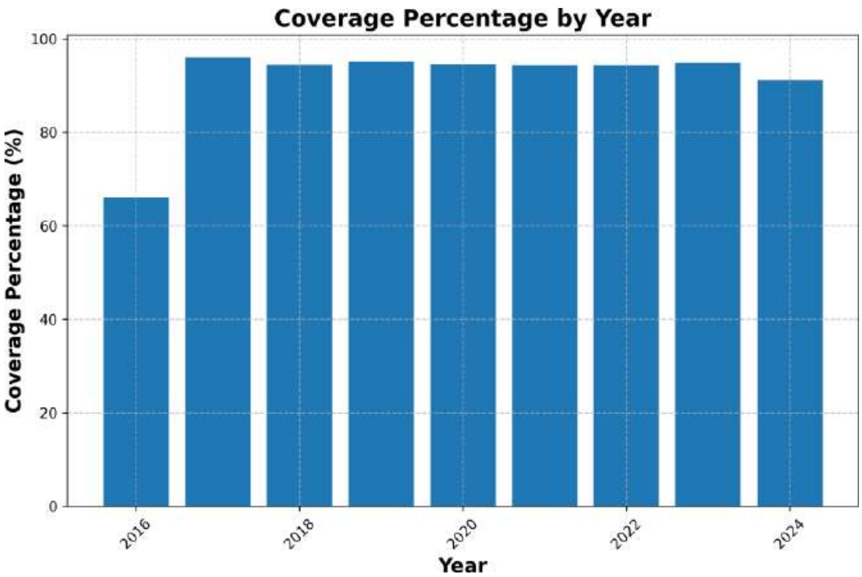


Fig. 16: Farm Coverage Across Years: The bar chart shows the percentage of coverage by year; each bar represents the percentage of data marked as "Covered" for each year.

cross-checked whether the corresponding stations provided climate data for the same years. The findings reveal that most of the required years are adequately covered. Based on these results, we conclude that the daily combined climate data are suitable for analysis.

9 The United States' Data - Climate & Apple Yield

In our project, we aim to collect yield data along with the corresponding climate data. There are two possible approaches to achieve this:

- Pairing state-level climate data with state-level yield data.
- Pairing county-level climate data with county-level yield data.

The details of these two approaches will be discussed in Sections 9.1 and 9.2, respectively.

9.1 State-Level Data Collection

In this section, we provide an overview of our state-level data collection process, followed by a detailed explanation of the method and findings, and conclude with a summary of the finalised datasets.

9.1.1 Data Collection Method

For our state-level data collection, we aimed to compile apple yield data at the state level and pair it with corresponding climate data. Our objectives were to focus on states that significantly contribute to the United States' apple production and to ensure that the years covered by the yield data overlap with those of the climate data.

Here are the steps we took for the data collection:

1. We collected state-level apple yield data for states with a high proportion of apple production in the United States.
2. We attempted to gather climate data corresponding to these states. However, we were unable to find climate data aggregated at the state level.
3. We considered whether county-level climate data could serve as a proxy for state-level climate data. Using Washington State (WA) as an example, we selected climate data from nine counties distributed across different regions of the state (the specific reasons for selecting these nine counties will be detailed later). By comparing the climate data among these counties, we assessed the degree of climatic variation. If the differences were minimal, county data could approximate state climate data; if significant, county data would not adequately represent the state's climate.
4. We analysed the trends in temperature, precipitation, wind speed, and other climatic factors among these nine counties from 2000 to 2024. We also calculated the daily and annual coefficients of variation (CV) for these counties. Our analysis revealed substantial differences among the counties, indicating that they could not collectively represent the state's overall climate conditions.
5. To minimise discrepancies between climate data and yield data, we took into account that apple production within a state is often concentrated in a few counties. Therefore, we identified the counties with the highest apple production and used their climate data to represent the state's climate. Again using WA as an example, we learned that WA has five main apple-growing regions that follow the

Table 9: State Percentage for Apple Yield Data in the United States

State	Percentage in the United States (%)
Washington (WA)	over 60%
New York (NY)	10%-15%
Michigan (MI)	6%-8%
Pennsylvania (PA)	4%-5%
California (CA)	3%-4%
Virginia (VA)	2%-3%
Oregon (OR)	2%-3%
North Carolina (NC)	1%-2%
Ohio (OH) less than	1%
Missouri (MO)	less than 1%

cool, rushing waters of the Columbia River and its tributaries. Consequently, we collected climate data from all counties within the Yakima Valley, Columbia Basin, Wenatchee Valley, Lake Chelan, and Okanogan regions, totalling 11 counties. We intend to map the climate data of these 11 counties to WA's apple yield data to establish the relationship between climate variables and apple production.

We now present the details and findings of our data collection process. For state-level data collection, our goal was to gather state-level apple yield data and pair it with corresponding the United States' state-level climate data. Initially, we identified the top 10 apple-producing states and their respective production percentages, as shown in Table 9.

Based on Table 9, we focused on collecting apple yield data from states that contribute significantly to overall the United States production. We successfully obtained state-level apple yield data from 2007 to 2023 for key states, including Washington (WA), New York (NY), Michigan (MI), Pennsylvania (PA), California (CA), Virginia (VA), Oregon (OR), and others.

Next, we aimed to obtain the corresponding climate data for these states. While we successfully gathered some state-level yield data, we encountered a lack of matching climate data at the state level. To address this gap, we explored whether county-level climate data could reasonably represent state-level climate conditions or if state climate data could be estimated using county-level data. To investigate this, we analysed county-level climate data, and the results are presented below.

We selected Washington (WA) as a case study and chose nine airports across the state: Yakima Airport, Pullman Moscow Regional Airport, Vancouver Pearson Airport, Spokane International Airport, Bellingham International Airport, Quillayute Airport, Wenatchee Pangborn Memorial Airport, Shelton Airport, and Friday Harbor Airport. These locations are marked on the map in Fig. 17. The reasons for selecting these airport locations are as follows:

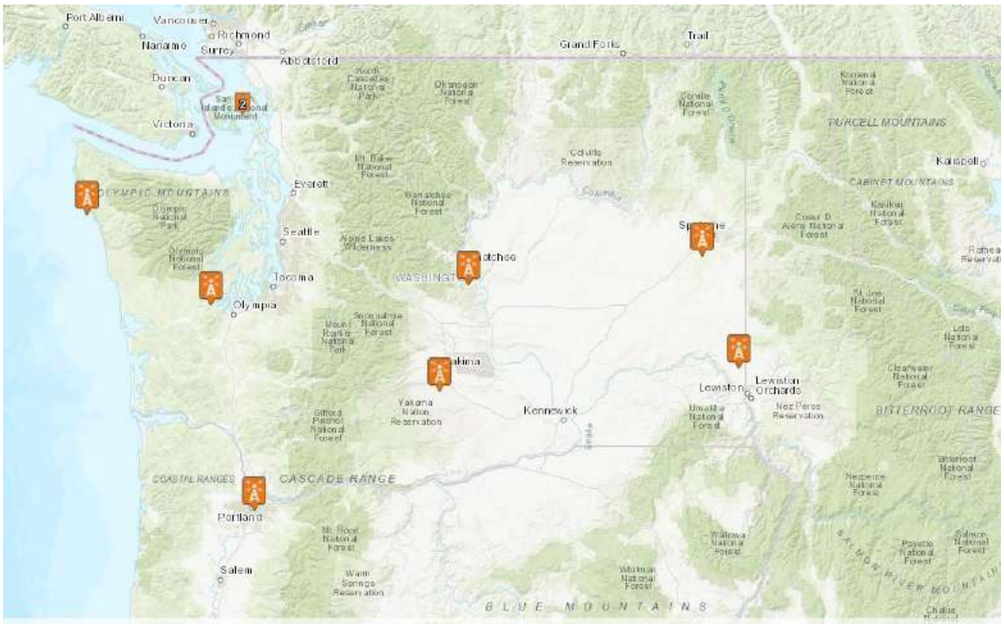


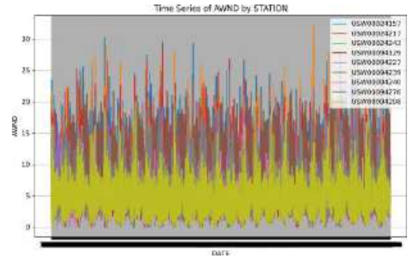
Fig. 17: The Selected Nine Counties in WA

- These stations provide data from January 1, 2000, to September 8, 2024, offering more comprehensive coverage compared to many other stations, which lack such long-term data.
- The selected locations are geographically distributed to broadly cover the entire state of Washington.
- The data from these airport stations are more complete, offering a wider range of features compared to stations in other counties.

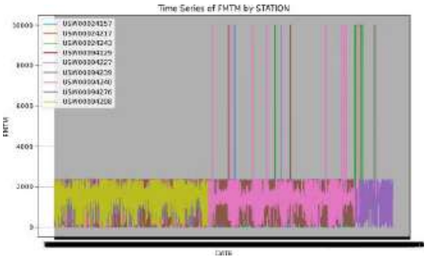
After downloading the climate data for these counties, we evaluated the variance between different locations by comparing the climate trends and calculating the coefficient of variation (CV) for each county. The dataset consists of various weather attributes recorded at different airport stations in WA, including columns such as AWND (average wind speed), FMTM (time of fastest mile or fastest 1-minute wind), PGTM (peak gust time in HHMM), PRCP (precipitation), TMAX (maximum temperature), and TMIN (minimum temperature), among others (which can be found in Fig. 46 of Part VI in this report).

Trends Comparison

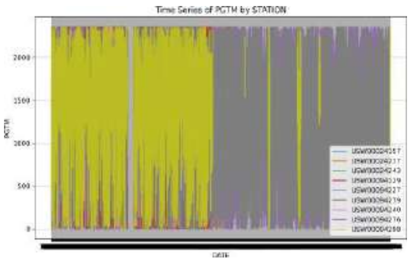
As shown in Fig. 18, we compared the climate data from nine counties within Washington (WA), focusing on six climate features: AWND, FMTM, PGTM, PRCP, TMAX, and TMIN. The trend comparison charts are presented below. Based on the analysis, it is challenging to conclude that the differences between the counties are insignificant. Therefore, to determine whether the climate data from the nine counties is similar or different, we conducted further analysis by calculating the Coefficient of Variation (CV).



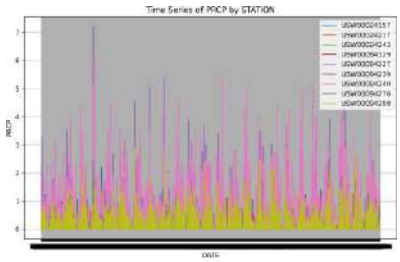
(a) AWND



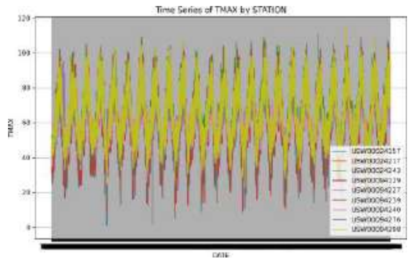
(b) FMTM



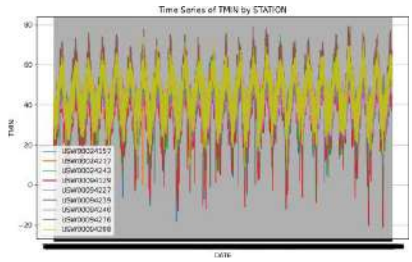
(c) PGTM



(d) PRCP

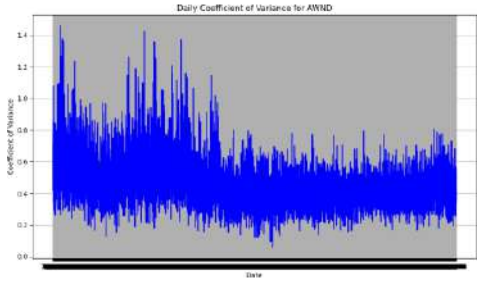


(e) TMAX

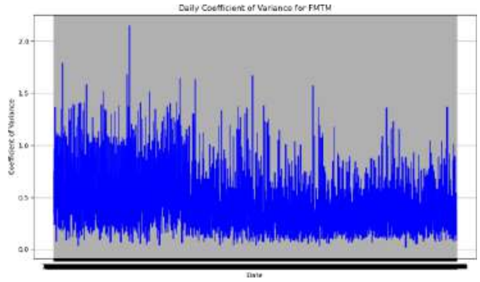


(f) TMIN

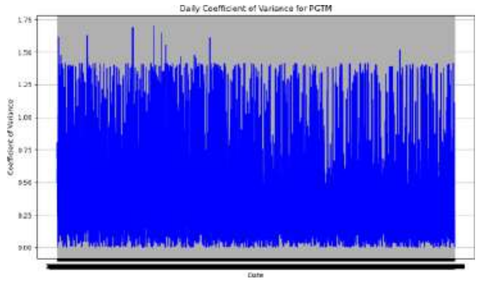
Fig. 18: Climate trends for nine counties in Washington (WA), covering the period from January 1, 2000 to September 8, 2024.



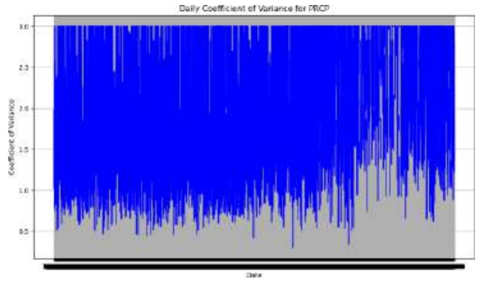
(a) AWND



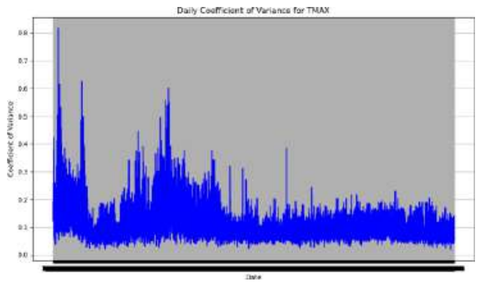
(b) FMTM



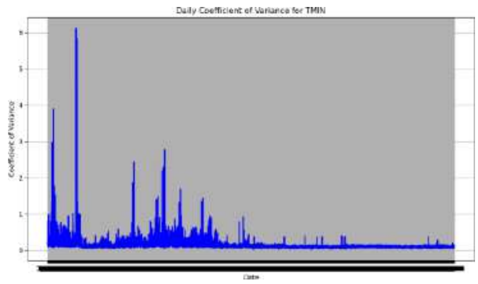
(c) PGTM



(d) PRCP

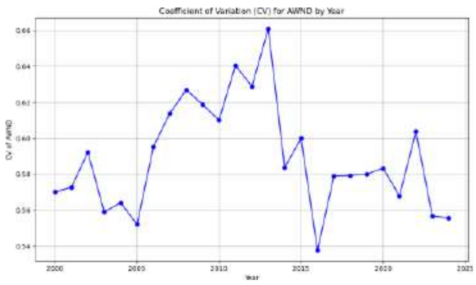


(e) TMAX

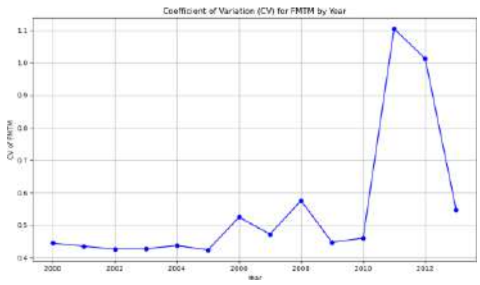


(f) TMIN

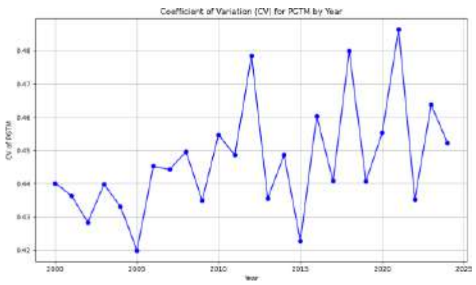
Fig. 19: CV comparison for nine counties in Washington (WA) on a daily basis, covering the period from January 1, 2000 to September 8, 2024.



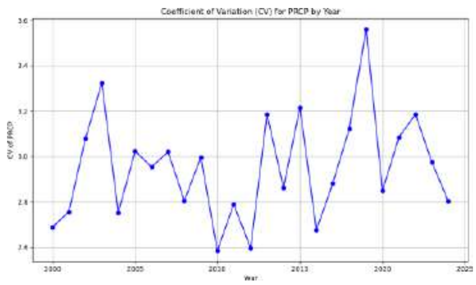
(a) AWND



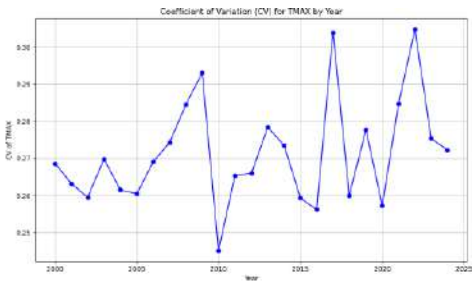
(b) FMTM



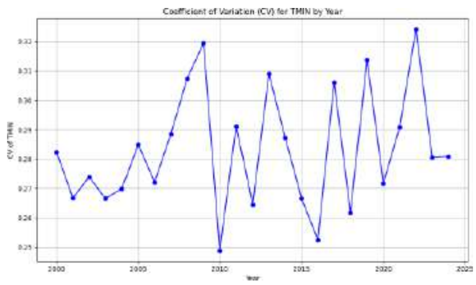
(c) PGTM



(d) PRCP



(e) TMAX



(f) TMIN

Fig. 20: CV Comparison for Nine Counties on an Annual Basis

Coefficient of Variation (CV) Comparison

We calculated the CV for nine counties on both a daily and annual basis, and the results are presented in Fig. 19 and Fig. 20, respectively. The calculated CV is notable and shows variability across the dataset. The Coefficient of Variation (CV) for each weather feature is calculated using the following formula:

$$CV_w = \frac{\sigma_w}{E\,(w_i)}$$

Where σ_w is the standard deviation of a weather feature across stations, and $E(w_i)$ is the mean of the weather feature across stations.

We computed the CV separately for six weather features: AWND, FMTM, PGTM, PRCP, TMAX, and TMIN. For each feature, we calculated the mean ($E(w_i)$) and standard deviation (σ_w) across all stations, and subsequently computed the CV for each day and each year.

Based on the results, we found the CV to be significant, indicating substantial variability. This suggests that county-level data may not adequately represent state-level climate conditions. To address this issue, we propose focusing on the key counties that contribute the most to apple production. This approach involves selecting counties with the highest apple yields, analyzing their climate data, and then mapping it to the overall state-level yield data.

9.1.2 Datasets Information

For the state-level data, we have obtained the following datasets.

- Apple Yield Data (2007–2023): We have comprehensive state-level apple yield data spanning from 2007 to 2023.
- The United States' Climate Data (2006–2023): Climate data is available for four key apple-producing states: Washington (WA), Michigan (MI), California (CA), and Pennsylvania (PA). This dataset covers a wide range of weather variables, including temperature (TMAX, TMIN), precipitation (PRCP), wind speed (AWND), and other relevant climate indicators.

This comprehensive dataset enables us to explore the relationships between climate variability, extreme weather events, and apple production across multiple states over a significant time period. By leveraging this data, we aim to develop insights that can inform future agricultural practices and policies to mitigate the impacts of climate change on apple yields.

9.2 County-level Data Collection

9.2.1 Data Collection Method

For the county-level data collection, our objective is to gather yield data at the county level paired with corresponding climate data. The data collection approach is summarised as follows, with further details and findings provided in the subsequent section.

- Collection of county-level yield data.
- Collection of climate data for the corresponding counties, ensuring the climate data spans the same years as the yield data.

For the county-level yield data, we have collected information from 171 counties across the following states: Pennsylvania, New Jersey, Virginia, South Carolina, New Mexico, and Kansas. These six states have been highlighted on the map to observe variations in yield and climate across different regions, as shown in Fig. 21.

9.2.2 Datasets Information

For the county-level data, we have obtained the following datasets:

- We have county-level apple yield data for Pennsylvania (PA), Virginia (VA), Kansas (KS), New Jersey (NJ), New Mexico (NM), and South Carolina (SC). The availability of yield data varies by county, spanning roughly from 1985 to 2012.

Fig. 21: The States with County-level Yield Data

- We have county-level climate data for PA, VA, KS, and NJ, covering approximately the same period, from 1984 to 2012. This dataset includes key weather variables relevant to agricultural productivity, such as temperature, precipitation, and wind speed.
- We do not have ground truth data for related extreme weather events at the county level. Therefore, we are considering the use of state-level climate data to account for the impact of extreme weather events.

9.3 Summary

For state-level data, we conclude the Pros and Cons as below:

- Pros: The state-level data is comprehensive, including apple yield, climate data, and corresponding ground truth for extreme weather events. This allows for a well-rounded analysis of the factors affecting apple production across multiple states.
- Cons: There is significant variation in climate conditions between different counties within the same state, making it difficult to generalise state-level climate using data from just one or a few counties. However, we have addressed this issue by focusing on the climate data from the counties with the highest apple yields and using this as a proxy for state-level climate.

For county-level data, we conclude the Pros and Cons as below:

- Pros: Both the yield and climate data are collected at the county level, allowing for a more precise estimation of how climate impacts apple yield. This granularity provides deeper insights into the relationship between local climate conditions and production outcomes.

- Cons: Except for Pennsylvania (PA), the other states (Virginia, Kansas, New Jersey, New Mexico, and South Carolina) contribute only a small percentage to the United States' apple production, making the data less representative on a national scale. What's more, the available apple yield data for these counties are relatively old, spanning from 1985 to 2012, and there is no ground truth for extreme weather events.

10 Part III Conclusion

In Part III of this project, we focused on the crucial task of data collection to meet two primary objectives: i) identifying extreme weather events through climate data and ii) assessing their impact on crop yields. By gathering both climate and yield data, we have laid the groundwork for a comprehensive analysis that will enable us to establish relationships between extreme weather events and agricultural production.

For New Zealand, we successfully matched climate data from 318 stations with yield data from 3,437 kiwifruit farms. Our analysis confirmed that for each farm, at least one of the ten nearest stations provided climate data, with most stations offering adequate coverage over the required years. These results ensure that the collected climate data is suitable for further analysis, particularly in assessing the impact of weather on kiwifruit yields. When examining the United States' apple production, we considered both state-level and county-level data. At the state level, the data is comprehensive, providing a broad overview of yield, climate conditions, and extreme weather events across multiple states. However, significant climate variation within states poses challenges for generalising state-level data. To address this, we focused on data from the counties with the highest apple yields, using this as a proxy for state-level climate. At the county level, the granularity of data allows for a more precise analysis of local climate impacts on apple production. However, the contribution of most states, except Pennsylvania, to national apple production is small, which limits the representativeness of the data. Additionally, the county-level apple yield data is somewhat outdated, spanning from 1985 to 2012, and lacks ground truth for extreme weather events.

Overall, the data collected at both the state and county levels provides valuable insights, but challenges remain in terms of data representativeness and the absence of up-to-date ground truth for extreme weather events. Despite these limitations, the current dataset forms a strong foundation for the next steps in our analysis, including labelling extreme weather events and evaluating their effects on agricultural yields.

Part IV
XAI Baseline Model Evaluation

11 Part IV Introduction

Part IV of this report focuses on evaluating our Explainable Artificial Intelligence (XAI) baseline models for crop yield prediction, including demonstration and testing of the latest XAI concepts. We completed our experiments on kiwifruit yield data with relevant New Zealand's climate data. Note that the selection of New Zealand's datasets was mainly due to the business needs at that particular stage of the project. Based on our extensive evaluations of our experiment results, we identified the limitations of the existing XAI methods on climate in agriculture yield prediction and defined opportunities for the later stages of this project.

Nowadays, model explainability has become a crucial aspect in the development and application of artificial intelligence, particularly for complex models such as deep learning and tree-based models. Explainability in XAI enhances model transparency, trustworthiness, and ability to uncover insights, which are essential for practical deployment in nearly everywhere, such as healthcare [60], finance [61], and environmental science [62], while there is still a lot of room for improvement in agriculture's applications. By providing explanations for model predictions, XAI techniques help users to understand the reasoning behind inferred results from a XAI model, thereby improving its acceptance and trustworthiness [63].

In Part IV, the goal of our experiments is to evaluate popular explanation methods, e.g., Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), and to identify their strengths and limitations on climate data. Through empirical analysis, we assessed the performance of these methods under different types of AI models, including tree-based models such as eXtreme Gradient Boosting (XGBoost), regression models such as Support Vector Regression (SVR), deep learning models such as Convolu-tional Neural Network (CNN), and probabilistic density modelling models such as Gaussian Mixture Model (GMM), and summarised their respective advantages and drawbacks.

This evaluation provides insights into the effectiveness of current XAI methods and inform the development of potential solutions for overcoming their limitations. We evaluated our baseline models on three varieties of kiwifruits: i) Gold Kiwifruit Conventional (GA Conventional), ii) Green Hayward Conventional (HW Conventional), and iii) Green Hayward Organic (HW Organic); and using three sources of climate data: i) quarterly NIWA dataset from The Yield Technology Solutions (now Yamaha Agriculture), ii) quarterly IBM dataset from The Yield Technology Solutions (now Yamaha Agriculture), and iii) daily climate data from the National Institute of Water and Atmospheric Research (NIWA) CliFlo platform[4].

12 Prediction Models & XAI Methods

12.1 Yield Prediction Methods

To review relevant AI methods comprehensively, we prepared our baseline yield prediction models using a diverse range of approaches, including both discriminative and generative models. Discriminative models focus on learning the boundaries or relationships between

[4] https://niwa.co.nz/climate-and-weather/national-climate-database

input features and target outputs to make direct predictions. In contrast, generative models aim to capture the underlying data distribution and generate new scenarios that resemble the original dataset. In our experiments, we selected discriminative models, XGBoost, SVR, and 1D-CNN, and a generative model GMM for evaluation. A brief introduction to each of the models are provided below:

- *eXtreme Gradient Boosting (XGBoost) [64]*
  XGBoost is a highly efficient and scalable implementation of gradient boosting. It builds an ensemble of decision trees sequentially, optimising a loss function at each step. Known for its speed and accuracy, XGBoost is widely used in predictive modelling tasks. In yield prediction, it can effectively handle non-linear relationships and feature interactions, making it ideal for complex agricultural datasets.
- *Support Vector Regression (SVR) [65]*
  SVR is a regression algorithm based on the principles of Support Vector Machines (SVM). The aim is to find a hyperplane that fits the data within a defined margin of tolerance. SVR is particularly effective for datasets with limited size or noise, as it minimizes overfitting by focusing only on data points that define the margin. This makes it a valuable tool for yield prediction when precise, smooth output is required.
- *1D Convolutional Neural Network (1D-CNN) [66]*
  1D-CNN is a deep learning model designed to extract features from sequential or temporal data. Convolutional operations are applied along a single dimension, making them suitable for time-series data such as climate variables or crop growth stages. For yield prediction, 1D-CNN can automatically learn complex patterns from raw data, capturing temporal dependencies that other models might miss.
- *Gaussian Mixture Models (GMM) [67]*
  GMM is a probabilistic model that assumes observations were generated from a mixture of several Gaussian distributions with unknown parameters. GMM is well-suited for clustering and density estimation, as it models the underlying probability distribution of the observations. In the context of yield prediction, GMM can help identify latent patterns or clusters in the data, such as different growing conditions or yield categories.

Performance Metric.

The R-squared (R^2) score is a key evaluation metric for the yield prediction models high-lighted above. It measures the proportion of variance in the observed data compared to the average value. An R^2 score of 1 indicates perfect prediction, 0 means the model performs no better than using the average value as prediction, and negative values signify performance worse than the average. A higher R^2 score reflects better predictive performance, indicating the model's ability to effectively capture the relationships between input climate features and yield. Table 10 provides a detailed comparison of the R^2 scores for each of the AI models obtained from our preliminary experiments, offering insights into their effectiveness in yield prediction.

The performance of the prediction models, as reflected by their R^2 scores, directly impacts the reliability of the explanation models. Explanation models interpret the features captured by prediction models, and more accurate predictions lead to more reliable explanations. When prediction models effectively represent reality, the explanations derived from them are more

Table 10: R^2 Score for Each of the AI Models on Different Types of Kiwifruit

Kiwifruit Types	AI models	Datasets		
		NIWA	IBM	Daily
GA Conventional	XGBoost	0.1853	0.1866	0.2078
	SVR	0.1603	0.1634	0.1950
	1D-CNN	0.2134	0.2314	0.1953
	GMM	0.1511	0.1189	0.1953
HW Conventional	XGBoost	0.2883	0.2917	0.2338
	SVR	0.2874	0.3011	0.2372
	1D-CNN	0.3484	0.3553	0.2343
	GMM	0.2984	0.2819	0.2338
HW Organic	XGBoost	0.3269	0.3160	0.2894
	SVR	0.2757	0.3041	0.2920
	1D-CNN	0.1857	0.3450	0.3554
	GMM	0.2570	0.195	0.2834

likely to align with domain knowledge or expert insights. This consistency fosters trust in both the explanations and the AI models.

Why We Included the R^2 Score in Our Investigation?

The R^2 score is a crucial metric for comparing the performance of various AI models, helping us identify which experimental results warrant further analysis. Given the extensive number of experiments conducted, it is not feasible to include an analysis of all results within a single report. Instead, we focus on a subset of experiments for detailed analysis across four dimensions, which we will cover in Section 14 and Section 15. For instance, Table 10 shows that CNN generally performs best in most scenarios. Therefore, when fixing one AI model while varying other parameters, we selected CNN for our analysis.

Additionally, comparing AI models with similar parameters but varying performances provides valuable insights into how performance impacts the quality and consistency of explanations. An important finding from our analysis is that the explanations generated by the same XAI method can vary significantly depending on the performance of the AI model, even under identical parameter settings (it will be discussed in Section 14.1). This high-lights the interplay between AI model performance and explanation quality, underscoring the importance of selecting high-performing models for robust and reliable explanations.

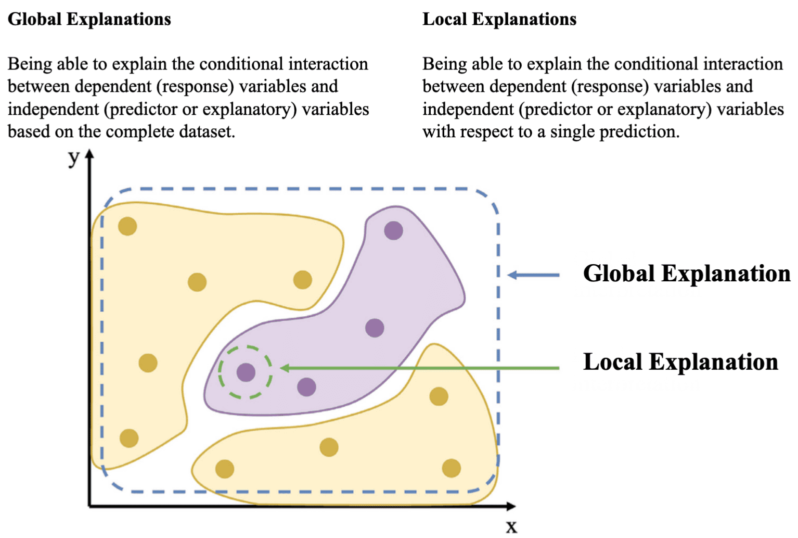


Fig. 22: Global Explanations VS Local Explanations [5]. In this figure, we provide a brief illustration of global explanations and local explanations, respectively.

12.2 XAI Methods

XAI is a critical component in artificial intelligence, which is particularly suitable to use in the agriculture domain such as yield prediction. By understanding how AI models capture the influence of climate features on yield, researchers and practitioners gain confidence in the predictions of these models. This transparency does not only enhance trust but also facilitates better decision-making, especially in agriculture, where the relationship between environmental factors and crop yield is often complex. In this context, XAI bridges the gap between sophisticated AI techniques and practical agricultural insights.

XAI methods can be broadly categorised into global explanations and local explanations, as shown in Fig. 22. Global explanations provide an overarching view of how a model makes predictions across the entire dataset, enabling an understanding of general patterns and feature importance. Examples include feature importance scores and SHapley Additive exPlanations (SHAP) summaries. Local explanations, on the other hand, focus on specific predictions, offering insights into how individual inputs lead to particular outcomes. Techniques like Local Interpretable Model-agnostic Explanation (LIME) and instance-level SHAP fall under this category.

Pros and Cons of Global and Local Explanations.

Global explanations excel at uncovering general trends and ensuring the model aligns with domain knowledge, making them ideal for high-level insights. However, they may overlook nuances and fail to address individual predictions. Local explanations, while excellent for understanding specific cases, can be computationally intensive and may not generalise well across the dataset. Combining both global and local explanation methods allows stakeholders to gain a comprehensive understanding, ensuring robust and reliable AI applications in yield prediction. Next, we will introduce SHAP and LIME, respectively.

Local Interpretable Model-agnostic Explanations (LIME) [68].

LIME is a widely used model-agnostic explanation method designed to interpret the predictions of complex AI models. Unlike global explanation techniques that provide insights into the overall behaviour of a model, LIME focuses on local explanations, offering insights into how specific features influence individual predictions.

LIME explains a prediction by perturbing the input data and observing how the model's predictions change. It does this by first creating slightly modified versions of the input, such as by removing or altering features. Then, it trains a simpler, interpretable surrogate model, such as linear regression, to approximate the behaviour of the complex model within the neighbourhood of the original input. Finally, LIME uses this surrogate model to identify and rank the features that are most influential for the prediction, providing a clear explanation of their contributions. In particular, it offers the following three properties:

- Model-Agnostic: LIME treats the AI model as a "black box", making it compatible with any predictive model.
- Instance-Level Explanations: It provides explanations specific to individual predictions, helping users understand the model's behaviour on a case-by-case basis.
- Flexibility: LIME can handle various data types, including tabular, text, and image data.

LIME offers several advantages, including its simplicity, as it uses interpretable models like linear regression for local explanations, making it easy to understand. It is also highly versatile, applicable to a wide range of models and data types without requiring modifications to the original model. Additionally, LIME allows customisation, enabling users to specify the neighbourhood size, distance metric, and surrogate model to tailor explanations to their needs. However, LIME has some limitations. The surrogate model may not fully capture the behaviour of the original model, leading to less accurate explanations. Its explanations can also vary with different perturbations or parameter settings, affecting stability. Further-more, the computational cost of perturbing the input and training surrogate models can be significant, particularly for large datasets.

SHapley Additive exPlanations (SHAP) [69].

SHAP is a widely used post-hoc model explanation method designed to provide insights into complex AI models. It is based on Shapley values from game theory and provides a fair and consistent way to attribute a model's predictions to its input features. SHAP works by calculating how much each feature contributes to a prediction, either positively or negatively.

SHAP supports both global explanations, which explain overall feature importance, and local explanations, which explain individual predictions. It offers multiple implementations tailored to specific model structures and types, such as TreeSHAP for tree-based models, DeepSHAP for neural networks, and KernelSHAP for model-agnostic scenarios.

The key advantages of SHAP include its solid theoretical foundation, flexibility, and powerful visualisation tools. However, it can be computationally expensive, especially for large datasets or high-dimensional models. Despite this shortcoming, SHAP is widely used in practice due to its interpretability and versatility across diverse AI applications. Different from LIME, it offers the following properties:

- SHapley: Represents the calculation of the Shapley Value for every feature in each sample.
- Additive: Indicates that the Shapley values for all features in a sample sum up additively to the model's output.

Table 11: Comparison of SHAP Methods

SHAP Method	AI Models	Gradient-Based	Description
LinearSHAP	Linear/Logistic regression	No	Optimised for linear and logistic regression models, calculates exact Shapley values efficiently.
PermutationSHAP	Model-agnostic	No	Uses feature permutation to measure impact on predictions; simple and efficient but less accurate for complex models.
TreeSHAP	Tree ensemble models (e.g. XGBoost)	No	Leverages tree structure properties for efficient Shapley value computation in tree-based models.
GradientSHAP	Deep learning models	Yes	Uses gradients to compute feature attributions, suitable for complex neural networks.
KernelSHAP	Model-agnostic	No	Based on sampling methods, applicable to any model but computationally intensive.

- exPlanation: Refers to the explanation of individual samples, illustrating the impact of each feature on the model's prediction.

In real-world applications, the SHAP library[5] in Python, offers several powerful features for interpreting AI models. It provides clear interpretability and rich visualisations, making it easier to understand model behaviour. SHAP includes optimised implementations, such as TreeSHAP for tree ensembles and DeepSHAP for neural networks, ensuring efficient and accurate computations. It explains the contribution of each feature to individual predictions, highlighting both positive and negative impacts. Additionally, SHAP reveals feature interactions and supports detailed comparisons between models, enabling a deeper understanding of model performance and decision-making processes.

When the type of complex model is known, such as tree ensemble models or deep learning models, more efficient methods can be used to compute the Shapley Value for each feature in each sample. The computation methods for Shapley Values vary by model type. For example:

- LinearSHAP: Optimised for linear regression, logistic regression, and similar models.
- PermutationSHAP: A simple, model-agnostic approach that computes Shapley values by permuting feature values and measuring their impact on predictions. It is computationally inexpensive but may be less accurate for highly complex models.
- TreeSHAP: Utilises the properties of tree structures for efficient computation, specifically tailored for tree-based models.
- GradientSHAP: Designed for deep learning models supported by gradient-based frameworks.
- KernelSHAP: A model-agnostic method based on sampling, applicable to a wide range of models.

[5] https://shap.readthedocs.io/en/latest/

The differences between these SHAP frameworks are summarised in Table 11. Depending on the prediction method used, we choose the appropriate framework to compute the SHAP values.

In summary, SHAP has several strengths, including its strong theoretical foundation based on Shapley values, which ensures fair and consistent feature attribution. It supports both global and local explanations, providing insights into overall feature importance and individual predictions. SHAP is versatile, offering model-specific implementations like TreeSHAP and DeepSHAP, as well as model-agnostic methods like KernelSHAP. Additionally, it pro-vides powerful visualisations for interpreting feature contributions and interactions. However, SHAP has some limitations, such as its high computational cost, especially for KernelSHAP and high-dimensional datasets. It can also suffer from scalability issues and approximation errors in sampling-based methods, and its implementation may require technical expertise like DSI, making it challenging for general users. However, it was adopted in our investigation due to the reasons outlined below.

The Adoption of KernelSHAP [69].

Since our explanation method needs to be applicable to all types of models, regardless of the specific framework or implementation (such as deep learning, tree models, or linear models), the most suitable choice is KernelSHAP. KernelSHAP combines the theoretical foundation of SHAP with the practical approach of LIME, effectively making it a hybrid method. This combination enables KernelSHAP to provide theoretically grounded explanations while retaining the flexibility of model-agnostic methods. Hence, we have chosen KernelSHAP as our baseline.

KernelSHAP is a model-agnostic explanation method that approximates Shapley values using a sampling-based approach. By leveraging the concepts from both SHAP and LIME, KernelSHAP perturbs feature values to observe their effect on model predictions and estimates the contribution of each feature. From SHAP, KernelSHAP inherits the fairness and consistency of Shapley value theory, while its sampling strategy resembles the localised, instance-level explanation approach of LIME. Let us summarise the Pros and Cons of using KernelSHAP:

- Pros:
  - Model-Agnostic: KernelSHAP can be applied to any model, regardless of the underlying structure, making it a versatile tool for explaining diverse AI models.
  - Theoretical Foundation: By incorporating Shapley value theory, KernelSHAP ensures fair and consistent feature attributions across models and predictions.
  - Localised Explanations: Like LIME, KernelSHAP focuses on individual predictions, providing detailed insights at the instance level.
  - Flexibility: Its sampling-based approach allows it to adapt to a variety of models without requiring specific modifications.
- Cons:
  - Computational Cost: KernelSHAP's reliance on sampling and model evaluations for every subset of features can make it computationally intensive, especially for high-dimensional datasets or large models.
  - Scalability: As the number of features increases, the computation time grows significantly, making it less efficient for models with a large feature set.

Table 12: Overview of Experiments for Kiwifruit Types, AI Models, XAI Methods, and Climate Datasets

AI Models	Kiwifruit Types	XAI Methods and Climate Datasets
XGBoost, SVR, GMM	All	LIME, PermutationSHAP, KernelSHAP (NIWA, IBM, Daily)
1D-CNN	All	LIME, GradientSHAP, KernelSHAP (NIWA, IBM, Daily)

– Approximation Limitations: Since KernelSHAP estimates Shapley values using sampling, its accuracy depends on the number of samples, requiring careful parameter tuning to balance accuracy and computation time.

In short, KernelSHAP's combination of SHAP's theoretical rigour and LIME's model-agnostic flexibility makes it a powerful and practical choice for explaining AI models. Its ability to work across different types of models, along with its fairness and interpretability, makes it an ideal baseline explanation method.

13 Implementation of Our Baseline Evaluation

13.1 Experiment Setup

This section describes the experimental setup for generating explanations for four AI models using explainability methods: LIME, SHAP, and KernelSHAP. Our experiment involves controlling four parameters: 4 types of AI models, 4 types of XAI methods, 3 kiwifruit varieties, and 3 climate types. The experiment is designed by keeping three parameters constant while varying the remaining one. Each parameter is independently controlled to obtain the experimental results. Table 12 summaries the predictive models, their corresponding XAI methods, kiwifruit varieties, and climate data.

LIME

- Explanation Generation:
  - Explanations are generated for each testing sample using the LIME framework.
  - To optimise space in the report, we analyse and present the results of one testing sample as an example.

SHAP

- Background Data Selection:
  - For SHAP: The first 100 samples from the training data are used as the background dataset.
  - For KernelSHAP: The background dataset is created by selecting 100 clusters using shap.kmeans.
- Explanation Generation:

Table 13: Kiwifruit Yield Data Split Details

	# training	# validation	# Test
GA Conventional	13289	1661	1662
HW Conventional	14121	1765	1766
HW Organic	1096	137	137

- Beeswarm Plots: Revised version: Beeswarm plots are generated for each dataset to illustrate global feature importance. In our experiments, we use 100 test samples to create the Beeswarm plots, which are analysed in detail in the experimental results section. For each Beeswarm plot, we selected the top 10 features based on their SHAP values, aggregated the SHAP values of the remaining features into a single category, and visualised the resulting 11 items using the shap.summary_plot function. The top features were displayed first, with the aggregated category representing the remaining features positioned at the bottom.
- Waterfall Plots: To analyse individual predictions, we analyse SHAP and KernelSHAP Waterfall plots for one testing sample. Additional Waterfall plots for five testing samples are provided in the supplementary materials.

This experimental setup ensures a systematic approach to generating and analysing explanations, while providing reproducibility and supplementary materials for extended analysis.

13.2 Datasets

The datasets used in this study are categorised based on the type of kiwifruit and their cultivation methods: GA Conventional, HW Conventional, and HW Organic. Each dataset is split into 80% for training, 10% for validation, and 10% for testing. The details of the data split are summarised in Table 13. For each kiwifruit type, we employed four AI models to predict yield based on three climate datasets. Correspondingly, we applied 3 or 4 XAI methods to interpret the predictions of these AI models.

In our experiments, we utilised three sources of climate data in two types:

- Seasonal Climate Data: CSV files summarised by The Yield Technology Solutions (now Yamaha Agriculture) which are derived from two sources – NIWA and IBM. Long-term average (denoted as _lta) climate features have been excluded.
- NZ CliFlo Daily Climate Data: Downloaded directly from the NIWA CliFlo platform, which provides daily climate information for New Zealand. More detailed analysis of the NZ CliFlo daily climate data can be found in Part III of this report.

14 Analysis Based on Different AI Approaches

This section examines the differences in explanations derived from various AI models and XAI methods. Following discussions with The Yield, their primary interest lies in GA Conventional and HW Conventional kiwifruits based on the NIWA climate data. Consequently, this report focuses on analysing HW Conventional kiwifruit based on NIWA climate data as primary examples. The analysis considers two key perspectives: i) explanations produced by different AI models using the same XAI method and ii) explanations generated by different XAI methods for the same AI model. The experimental results are visualised as follows:

- Beeswarm Plots: As shown in Fig. 23, the Beeswarm plots highlight the feature importance rankings obtained from SHAP and KernelSHAP for different AI models. These plots allow for comparing how various models identify key climate features affecting kiwifruit yield.
- Waterfall Plots: Fig. 24 and Fig. 25 present detailed Waterfall plots that illustrate the contributions of individual features to model predictions. These plots compare explanations generated using SHAP and KernelSHAP across various AI models (XGBoost, SVR, 1D-CNN, and GMM). The results are demonstrated for sample #6.
- LIME Explanation Plots: Fig. 26 provides detailed LIME explanation plots showcasing the contributions of individual features to model predictions. This plot compares LIME-derived explanations across AI models (XGBoost, SVR, 1D-CNN, and GMM). Sample #6 is used for demonstration.

14.1 Comparison Across Different AI Models on a XAI Method

To evaluate the impact of different AI models on the same XAI method, we analysed results from four AI models — XGBoost, SVR, 1D-CNN, and GMM — using KernelSHAP Beeswarm and Waterfall plots, as well as LIME explanations.

From Fig. 23, the second column (KernelSHAP Beeswarm Plots), the following findings are observed.

- XGBoost: The Beeswarm plot highlights that the features *niwa pressure Budbreak* and *niwa_pressure Harvest* are the most influential features in the XGBoost model, with pressure being the primary driver of yield predictions, while relative humidity and radiation features play a secondary role.
- SVR: The Beeswarm plot indicates that the features *niwa pressure Budbreak_delta* and *niwa_pressure Budbreak* are the most impactful features, significantly influencing the model's predictions. Pressure-related features dominate the contributions, while humidity and chill days features play secondary roles.
- 1D-CNN: The Beeswarm plot highlights the features *niwa pressure_FruitSet* and *niwa_pressure Budbreak delta* as the most influential features, showing significant contributions to the model's predictions. Pressure-related features dominate, while temperature and humidity features play supporting roles.
- GMM: The Beeswarm plot identifies the features *niwa_temperature max Dormant delta* and *niwa gdd FruitSet* as the most impactful features, significantly influencing the model's predictions. Temperature-related features are the primary drivers, while growing degree days and pressure features provide secondary contributions.

The Beeswarm plots reveal that feature importance varies significantly across model architectures. Pressure-related features dominate in XGBoost, SVR, and 1D-CNN, while temperature-related features are the primary drivers in GMM. These findings highlight

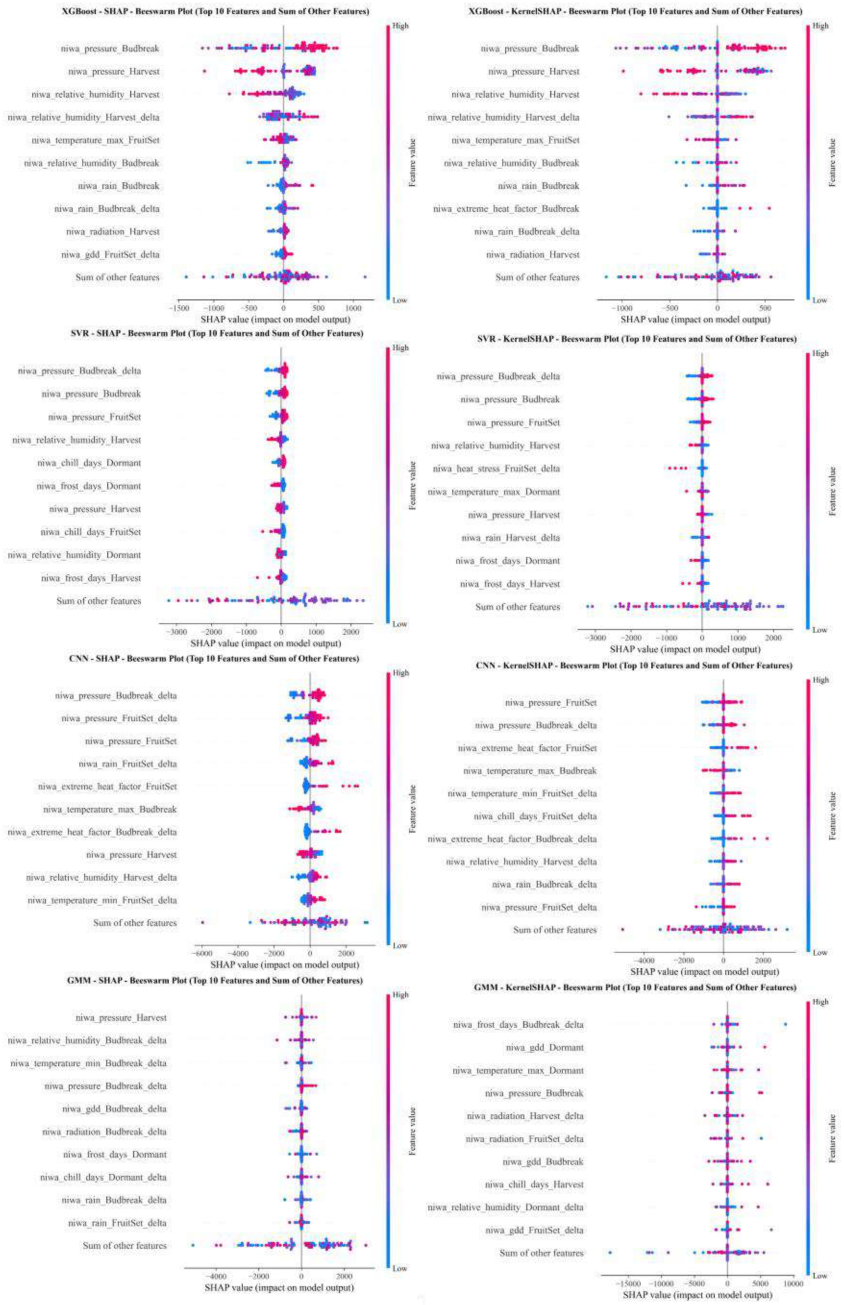


Fig. 23: Beeswarm Plots for Different Models (SHAP and KernelSHAP). These results are generated using a background of 100 training samples and 100 test samples for the HW Conventional type based on NIWA climate data.

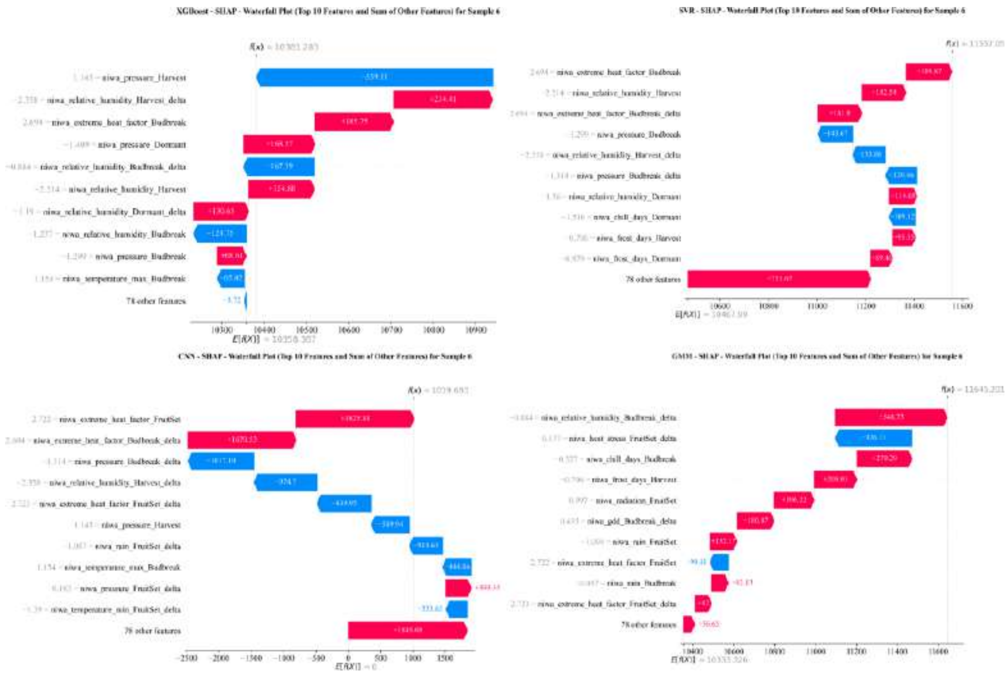


Fig. 24: SHAP Values Waterfall Plots for Different Models. In this figure, we demonstrate sample #6. These results are generated using a background of 100 training samples and 100 test samples for the HW Conventional type based on NIWA climate data.

how model architecture influences feature prioritisation, even with the same XAI method (KernelSHAP).

From Fig. 25, the following conclusions can be drawn. KernelSHAP reveals that models differ in feature prioritisation, sensitivity, and impact directions for the same sample.

- XGBoost: The feature *niwa pressure_Harvest* significantly decreases the prediction, while *niwa_chill_days Dormant* notably increases it, highlighting their dominant influence on the model's output.
- SVR: The features *niwa_extreme_heat_factor_FruitSet* and *niwa_extreme heat factor Budbreak* significantly increase the prediction, highlighting their strong positive influence on the model's output.
- 1D-CNN: The feature *niwa temperature_min_FruitSet_delta* significantly decrease the prediction, while the features *niwa extreme heat_factor FruitSet* and *niwa radiation FruitSet delta* strongly increase it, indicating their dominant impact on the model's output.
- GMM: The feature *niwa rain Budbreak_delta* generally has a small negative impact on predictions, but in this specific sample, it significantly reduces the prediction by -1259.58. Conversely, the feature *niwa pressure FruitSet_delta*, typically has a slight positive influence globally, showing the context-dependent nature of feature contributions.

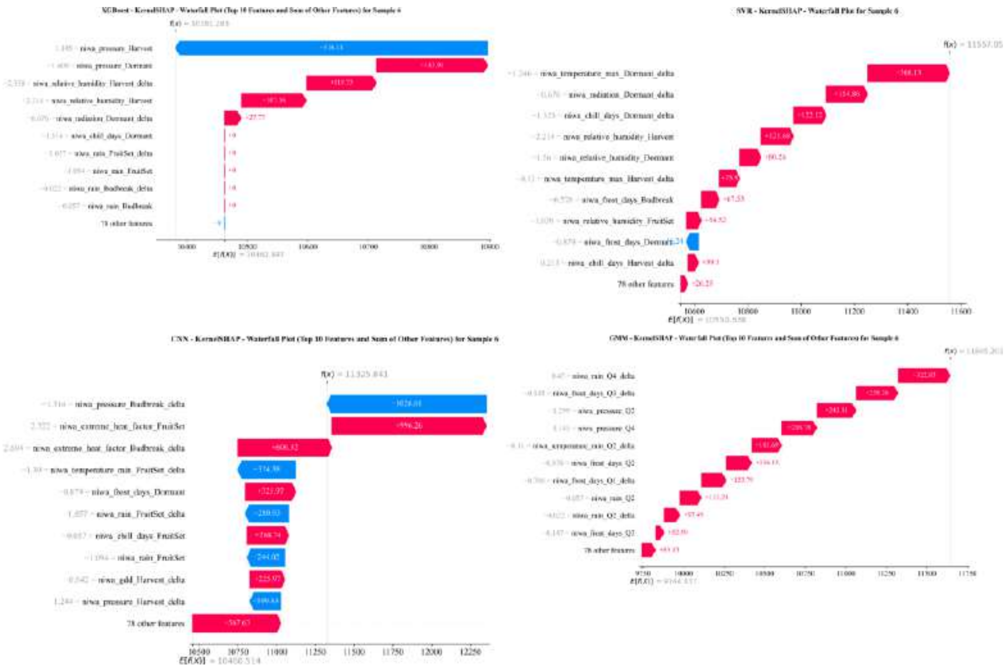


Fig. 25: KernelSHAP Waterfall Plots for Different Models. In this figure, we demonstrate sample #6. These results are for the HW Conventional type based on NIWA climate data.

From Fig. 26, the following conclusions can be drawn. LIME explanations reveal that models rely on distinct features and relationships, even for the same sample.

- XGBoost: Dominated by the feature *niwa pressure Budbreak*, with few significant contributing features.
- SVR: Focuses on the features *niwa pressure FruitSet_delta* and *niwa extreme heat factor_Budbreak*, with evenly distributed contributions.
- 1D-CNN: The feature *niwa pressure Harvest* increases the prediction significantly, while the feature *niwa_pressure Budbreak* reduces it, highlighting the key role of pressure variables.
- GMM: The feature *niwa rain Budbreak_delta* slightly increases the prediction, while most features have negative influences.

14.2 Comparison Across Different XAI Methods on a AI Model

To evaluate the impact of different XAI methods on the same AI model, we analysed results from SHAP Beeswarm plots and KernelSHAP Beeswarm plots across four AI models: XGBoost, SVR, 1D-CNN, and GMM. Based on the rows of Fig. 23, we summarise the key findings below and highlight differences between the two XAI methods for each model.

The comparison of SHAP and KernelSHAP reveals that while both methods generally agree on the most important features, there are subtle differences in feature rankings and the inclusion of additional interactions. These variations underline the need to interpret XAI

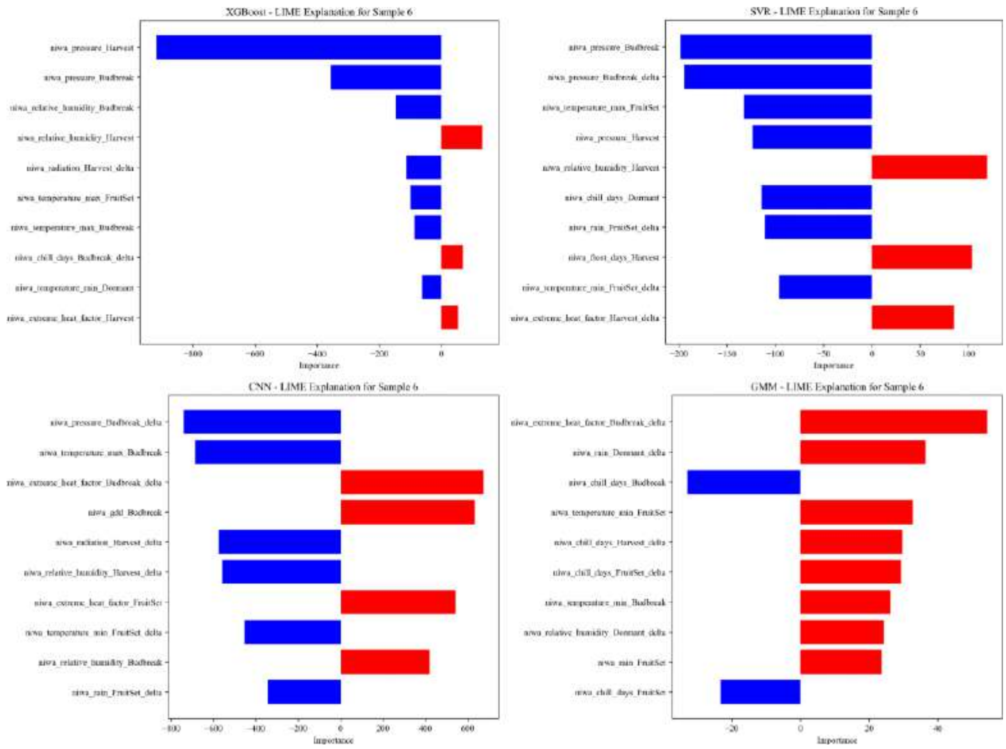


Fig. 26: LIME Explanations for Different Models. In this figure, we demonstrate sample #6. These results are generated for the HW Conventional type based on NIWA climate data.

results in the context of the chosen explanation method, as different methods may emphasise different aspects of feature contributions.

- XGBoost
  - SHAP: The most influential features are *niwa pressure Budbreak* and *niwa pressure Harvest*, with pressure-related features being the primary drivers. Secondary contributions come from the features *niwa relative humidity Harvest* and *niwa radiation Harvest*.
  - KernelSHAP: The results are consistent with SHAP, identifying the features *niwa pressure Budbreak* and *niwa pressure Harvest* as dominant features. However, the features *niwa rain Budbreak* and *niwa extreme heat factor Budbreak* also show increased importance, suggesting slight variability in feature rankings between methods.
- SVR
  - SHAP: The features *niwa pressure Budbreak delta* and *niwa pressure Budbreak* are the top contributors, with pressure-related features dominating the

predictions. Humidity (*niwa relative humidity Dormant*) and chill days (*niwa chill days Dormant*) features play minor roles.
  - KernelSHAP: The findings align with SHAP, highlighting pressure-related features as dominant. However, the features *niwa temperature_max_Dormant* and *niwa_heat stress FruitSet delta* gain slightly more prominence, indicating that KernelSHAP captures additional interactions.

- 1D-CNN
  - SHAP: The features *niwa pressure FruitSet* and *niwa pressure Budbreak_delta* are the most impactful features. Pressure remains the dominant fac-tor, while temperature (*niwa temperature max Budbreak*) and humidity (*niwa relative humidity Harvest*) features have smaller impacts.
  - KernelSHAP: Pressure-related feature (*niwa pressure FruitSet*, the feature *niwa pressure Budbreak delta*) are still dominant. However, the features *niwa extreme heat factor FruitSet* and *niwa chill days FruitSet delta* gain signif-icance.

- GMM
  - SHAP: Temperature-related features (*niwa temperature max Budbreak*) and growing degree days (*niwa_gdd_FruitSet*) feature dominate the predictions. Pressure and radiation features have minor contributions.
  - KernelSHAP: The results are consistent, with the features *niwa temperature max Budbreak* and *niwa gdd_FruitSet* remaining the top contributors. However, the features *niwa chill days FruitSet delta* and *niwa extreme heat_factor FruitSet* gain more prominence under KernelSHAP.

From Fig. 24, Fig. 25, and Fig. 26, the comparison of SHAP, KernelSHAP, and LIME for each AI model reveals the following key insights:

- XGBoost
  - SHAP and KernelSHAP: Both highlight the feature *niwa pressure_Harvest* as a significant negative contributor and the feature *niwa_chill_days_Dormant* as a major positive contributor, with minor differences in feature ranking and magnitude.
  - LIME: Focuses on the feature *niwa pressure_Budbreak* as the dominant feature, showing fewer but highly influential contributors.

- SVR
  - SHAP and KernelSHAP: Consistently emphasise the features *niwa extreme heat_factor Budbreak* and *niwa extreme heat factor_FruitSet* as dominant positive contributors, with slight variation in secondary features.
  - LIME: Highlights a broader range of features, such as the feature *niwa_pressure_FruitSet_delta*.

- 1D-CNN

- SHAP and KernelSHAP: Both identify the feature *niwa temperature min FruitSet delta* as a key negative contributor and the feature *niwa extreme heat factor FruitSet* as a dominant positive feature, differing slightly in magnitude.
- LIME: Shifts focus to the feature *niwa chill days Dormant delta* as a significant negative contributor, with a mix of additional features.

• GMM

- SHAP and KernelSHAP: Agree on the feature *niwa rain Budbreak delta* as a major negative contributor and the feature *niwa pressure FruitSet delta* as a strong positive feature, with slight differences in other rankings.
- LIME: Contrasts with SHAP methods, showing the feature *niwa rain Budbreak delta* as a minor positive contributor and highlighting different feature priorities.

SHAP and KernelSHAP provide consistent global explanations with nuanced differences, while LIME offers localised insights that often shift feature prioritisation. These differences underscore the complementary nature of XAI methods, each offering unique interpretability perspectives.

14.3 Our Findings

This section compares the impact of different AI models (XGBoost, SVR, 1D-CNN, and GMM) and XAI methods (SHAP, KernelSHAP, LIME) on explanations using Beeswarm and Waterfall plots. Notably, the selections of AI model architecture and XAI method signifi-cantly influence explanations, highlighting the need to carefully select both for interpretable results. Our analysis demonstrates that 1D-CNN captures broader feature interactions, while XGBoost focuses on a few dominant features, such as pressure-related variables. GMM and SVR show more balanced contributions across features. Among XAI methods, SHAP and KernelSHAP reveal complex, cumulative feature interactions, whereas LIME provides concise and simplified explanations by focusing on key features.

Key Insights: Different AI Models

The comparison of different AI models reveals distinct patterns in feature prioritisation. XGBoost and SVR emphasise pressure-related features, with XGBoost relying on broad seasonal metrics and SVR balancing stress-related and environmental factors. 1D-CNN focuses on pressure and heat-related metrics, capturing complex interactions, while GMM priorities temperature and growing degree days, reflecting its focus on growth-related factors.

These findings reveal that while XGBoost, SVR, and 1D-CNN consistently prioritise pressure-related features, GMM diverges by emphasising temperature and growth factors. The differences highlight the influence of model architecture on feature selection and sen-sitivity, demonstrating the need for model-specific interpretations in climate-based yield predictions.

Key Insights: Different XAI Methods

The comparison of different XAI methods reveals that SHAP and KernelSHAP provide consistent global explanations, emphasising similar key features with slight variations in rankings and interactions. For instance, both methods highlight pressure-related features for XGBoost and SVR, while KernelSHAP captures additional feature interactions for models like 1D-CNN and GMM. LIME, on the other hand, focuses on localised insights, often shifting feature prioritisation and highlighting fewer but highly influential features.

These results demonstrate the complementary nature of XAI methods: SHAP and KernelSHAP excel in global interpretability, while LIME provides instance-specific insights. The choice of method should align with the specific interpretability goals of the analysis.

15 Analysis Based on Different Datasets

In this section, we focus on analysing the experimental results of the 1D-CNN model, with attention to the two XAI methods: LIME and KernelSHAP. The objective is to examine how differences in kiwifruit types and climate data sources affect the explanations. The experimental results are visualised as follows:

- Fig. 27: LIME Explanations This figure visualises feature contributions for 1D-CNN models using the LIME method across different datasets. LIME highlights key features and their individual contributions to model predictions.
- Fig. 28: Beeswarm Plots (KernelSHAP). This figure displays KernelSHAP Beeswarm plots to show the distribution of SHAP values for 1D-CNN models across different datasets. The plots reveal the importance and distribution of features, as well as their impact directions on model predictions.
- Fig. 29: Waterfall Plots (KernelSHAP). This figure illustrates KernelSHAP Waterfall plots to detail feature contributions for 1D-CNN models. Waterfall plots break down individual predictions, showing how each feature contributes positively or negatively.

15.1 Comparison Across Different Kiwifruit Types

To understand the impact of kiwifruit type on predictions and explanations, we selected a specific AI model and a XAI method (e.g., 1D-CNN with LIME and 1D-CNN with KernelSHAP). The kiwifruit types that we analysed are GA Conventional, HW Conventional, and HW Organic.

From Fig. 27, the differences between GA Conventional (first column) and HW Conventional (second column) can be summarised as follows. The 1D-CNN model's feature prioritisation and impact directions vary significantly between GA Conventional and HW Conventional, driven by the different characteristics of the datasets.

- GA Conventional
  - NIWA: Key features include *niwa gdd FruitSet* and *niwa temperature max FruitSet delta*, which show strong negative contributions, while *niwa radiation Dormant* provides a moderate positive impact. The importance is evenly distributed across multiple factors, reflecting a balanced reliance on various factors.
  - IBM: Key features include strong negative contributions from *ibm_high_humidity_sum_Budbreak* and *ibm air temperature_min_Budbreak_delta*, with positive contributions dominated by the feature

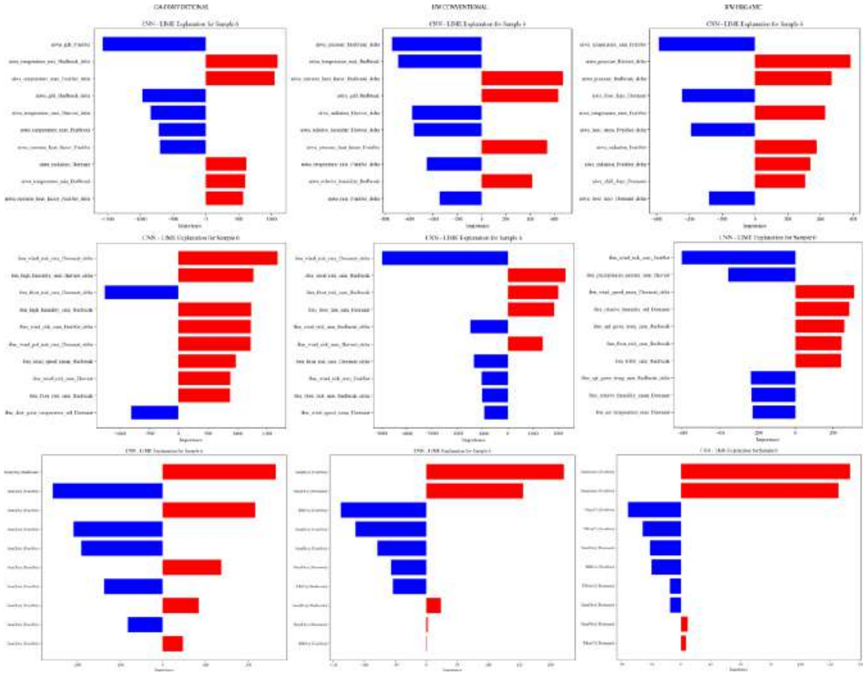


Fig. 27: LIME Explanations for Different Datasets. We present the results for sample #6 across the three kiwifruit types, using three different climate datasets.

ibm wind speed mean Budbreak. The model shows balanced reliance across multiple variables, such as humidity, temperature, and wind speed.

– Daily: Key features include strong negative contributions from *Sun(Hrs) FruitSet* and *Sun(Hrs) FruitSet*, with positive contributions from *Sun(Hrs) FruitSet* and *Sun(Hrs) FruitSet*. The model shows a balanced reliance on multiple sunlight metrics, reflecting broader variability.

- HW Conventional
  - NIWA: Key features such as *niwa relative humidity Dormant*, *niwa chill days Dormant delta*, and *niwa radiation Harvest* dominate the model, with larger positive and negative contributions. This model places greater emphasis on a few highly influential factors, particularly humidity and chill days.
  - IBM: Dominated by negative contributions from the features *ibm relative humidity mean Dormant* and positive contributions from the fea-ture *ibm wind speed mean Budbreak*. The model focuses on fewer key features, particularly relative humidity and wind speed.

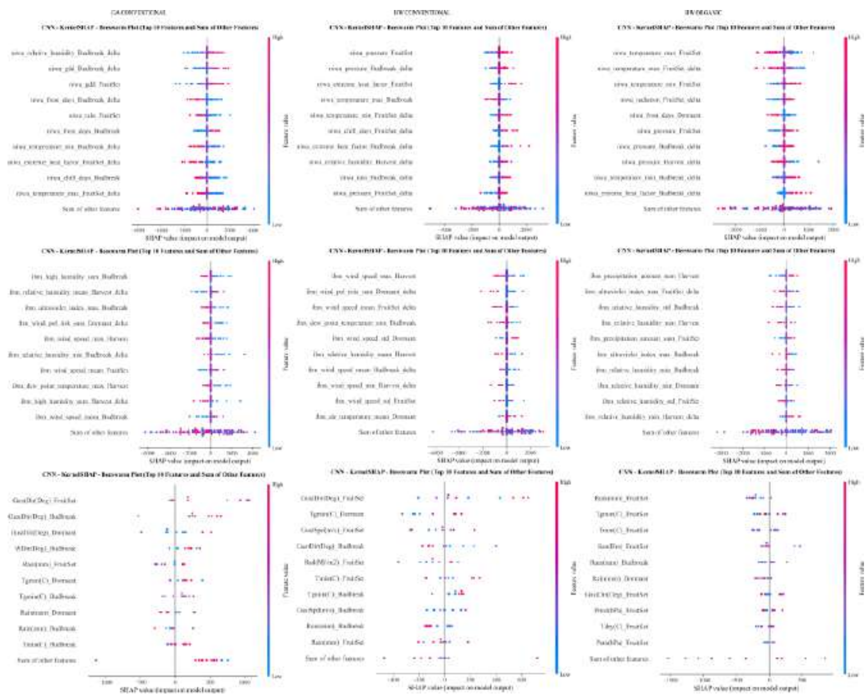


Fig. 28: Beeswarm Plots for Different Datasets: These results are generated using a background of 5 training samples and 10 test samples for the three kiwifruit types, based on three climate datasets. When using a background of 100 training samples and 100 test samples with the daily climate dataset, the system ran out of memory on an Apple M1 system-on-a-chip (SoC) chip with 16GB unified memory or a workstation equipped with 64 GB RAM and an NVIDIA A2 GPU (24 GB VRAM).

  - Daily: Key features include negative contributions from *Sun(Hrs)_FruitSet* and *RH(%)_FruitSet*, and positive contributions from *Sun(Hrs)_Dormant* and *Sun(Hrs)_Dormant*. The model focuses on fewer metrics, with an emphasis on specific sunlight periods and relative humidity.

- Key Differences
  - GA Conventional: Relies on temperature, radiation, and diverse sunlight metrics with balanced feature importance.
  - HW Conventional: Focuses strongly on humidity, chill days, specific sunlight periods, and wind, with larger feature contributions.

From Fig. 27, the differences between HW Conventional (the second column) and HW Organic (the third Column) can be summarised as follows. Note that the 1D-CNN model adapts differently for HW Conventional and HW Organic.

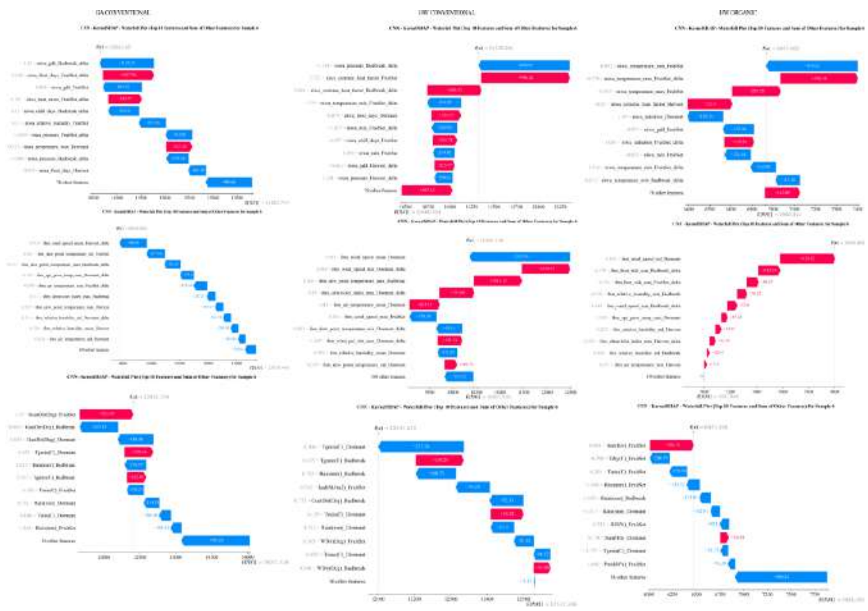


Fig. 29: Waterfall Plots for Different Datasets. In this figure, we demonstrate sample #6. These results are generated using a background of 5 training samples and 10 test samples for the three kiwifruit types, based on three climate datasets. When using a background of 100 training samples and 100 test samples with the daily climate dataset, the system ran out of memory on an Apple M1 system-on-a-chip (SoC) chip with 16GB unified memory or a workstation equipped with 64 GB RAM and an NVIDIA A2 GPU (24 GB VRAM).

- HW Conventional
  - NIWA: Strong negative contributions are observed from *humidity* and *chill days* (e.g., *niwa_chill_days_Dormant delta*), while positive contributions come from *radiation* and *heat stress* factors.
  - IBM: Strong negative contributions are observed from the feature *ibm wind risk sum Dormant delta*, with positive contributions from *ibm wind speed mean Budbreak delta*. The model primarily focuses on wind risk and humidity.
  - Daily: Key features include negative contributions from *Sun(Hrs)_FruitSet* and *RH(%)_FruitSet*, with positive contributions from *Sun(Hrs)_Dormant* and *Sun(Hrs)_Dormant*. The model emphasises specific sunlight durations and relative humidity, dominated by negative impacts.
- HW Organic

- NIWA: Negative contributions are dominated by *temperature* and *pressure* (e.g., *niwa temperature min FruitSet*), with positive contributions focused on *radiation* and *pressure* metrics.
- IBM: Negative contributions are dominated by the feature *ibm precipitation amount sum Dormant delta*, with positive contributions from the feature *ibm opt grow temp sum FruitSet delta*. The model emphasises temperature and precipitation metrics.
- Daily: Key features include negative contributions from *Rain(mm) FruitSet* and *WDir(Deg) Dormant*, with positive contributions from *Tdry(C) Budbreak* and *Sun(Hrs) Budbreak*. The model focuses on temperature, rainfall, and sunlight metrics, with more balanced contributions.

- Key Differences
  - HW Conventional: Emphasises humidity, chill days, wind risk, and sunlight metrics, with stronger negative contributions dominating the model.
  - HW Organic: Focuses on radiation, pressure, temperature, precipitation, and rainfall metrics, with more balanced contributions across features.

Using experimental results based on NIWA climate data as an example, we analyse the differences in Fig. 28. A comparison between the first column (GA Conventional) and the second column (HW Conventional) reveals the following results:

- GA Conventional Key features include *niwa gdd Budbreak delta*, *niwa gdd FruitSet*, and *niwa rain FruitSet*, with additional contributions from humidity metrics like *niwa relative humidity Budbreak delta*. The model broadly relies on temperature, rainfall, and humidity metrics, reflecting sensitivity to seasonal growth factors.
- HW Conventional Dominant features include *niwa pressure FruitSet*, *niwa pressure Budbreak delta*, and *niwa extreme heat factor FruitSet*, alongside temperature metrics such as *niwa temperature max Budbreak*. The model emphasises pressure, heat stress, and temperature, reflecting sensitivity to extreme climatic conditions.
- Key Differences
  - GA Conventional: Broad reliance on seasonal growth metrics like degree days and rainfall.
  - HW Conventional: Focused on stress-related metrics like pressure and heat.

From Fig. 28, a comparison between the second column (HW Conventional) and the third column (HW Organic) reveals the following findings.

- HW Conventional (Left Panel) Key features include *niwa pressure FruitSet*, *niwa pressure Budbreak delta*, and *niwa extreme heat factor FruitSet*, with additional contributions from temperature and humidity metrics. The model emphasises pressure and heat stress, reflecting sensitivity to atmospheric conditions.
- HW Organic (Right Panel) Key contributions come from the features *niwa temperature max FruitSet*, *niwa temperature max FruitSet delta*, and *niwa temperature min FruitSet*, with focus on pressure and radiation as well. The model relies on temperature and radiation, highlighting growth-related factors.

- Key Differences
  - HW Conventional: Prioritises pressure and heat stress, reflecting sensitivity to climatic extremes.
  - HW Organic: Focuses on temperature and radiation, emphasising growth-related metrics.

From Fig. 29, a comparison between the first column (GA Conventional) and the second column (HW Conventional) is summarised below, using experimental results based on NIWA climate data as an example.

- GA Conventional: Contributions from the features *niwa relative humidity Budbreak delta* and *niwa temperature max FruitSet* reflect a balance between humidity and temperature metrics. Highlights seasonal factors, including rainfall and chill days, with balanced contributions.
- HW Conventional: Contributions from the features *niwa extreme heat factor FruitSet* and *niwa temperature min FruitSet delta* emphasise heat stress and temperature extremes. Priorities stress-related metrics, such as heat and temperature, with stronger contributions.
- Key Differences
  - GA Conventional: Relies on diverse climatic features, including rainfall and chill days.
  - HW Conventional: Focuses on heat and temperature with stronger contributions.

From Fig. 29, the comparison between the second column (HW Conventional) and the third column (HW Organic) is summarised below.

- HW Conventional: Contributions from the feature *niwa extreme heat factor FruitSet* highlight sensitivity to heat stress, while negative contributions from the feature *niwa temperature min FruitSet delta* reflect its reliance on temperature extremes. Emphasises heat stress and temperature metrics.
- HW Organic: Positive contributions from the features *niwa pressure Budbreak delta* and *niwa radiation FruitSet delta* underscore its focus on atmospheric pressure and radiation. Prioritises pressure and radiation with less emphasis on heat stress.
- Key Differences
  - HW Conventional: Focuses on heat stress and temperature extremes.
  - HW Organic: Shifts focus to atmospheric factors like pressure and radiation.

15.2 Comparison Across Different Climate Datasets

Using the same AI model (1-D CNN) and XAI methods (KernelSHAP or LIME), we analysed predictions and explanations across three climate data sources: NIWA Seasonal Climate Data, comprising seasonal data provided by NIWA; IBM Seasonal Climate Data, which consists of summarised seasonal data from IBM; and NZ CliFlo Daily Climate Data, which includes daily climate data retrieved from the NIWA CliFlo platform. In this section, we focus on the experimental results of HW Conventional as an example.

From Fig. 27, we obtain the following results. Each climate data source offers unique insights, with NIWA focusing on seasonal trends, IBM on wind and humidity, and NZ CliFlo on daily sunlight and humidity variations.

- NIWA Seasonal Climate Data (Top Panel): Key negative contributions are observed from the feature *niwa_chill_days_Dormant_delta*, with positive contributions from the feature *niwa_radiation_Harvest*. This dataset emphasises seasonal humidity and chill days.
- IBM Seasonal Climate Data (Middle Panel): Strong negative contributions are dominated by the feature *ibm_wind_risk_sum_Dormant_delta*, while positive contributions come from the feature *ibm_wind_speed_mean_Budbreak delta*. The focus is on wind risk and seasonal humidity.
- NZ CliFlo Daily Climate Data (Bottom Panel): Negative contributions are driven by the feature *Sun(Hrs)_FruitSet*, with positive contributions from the feature *Sun(Hrs)_Dormant*. This dataset highlights daily sunlight hours and humidity.
- Key Differences
  - NIWA vs. IBM: NIWA focuses on seasonal humidity and chill days, while IBM emphasises wind risk and seasonal humidity.
  - NIWA vs. NZ CliFlo: NIWA captures broad seasonal metrics (e.g., chill days, radiation), while NZ CliFlo focuses on daily variations in sunlight and humidity.
  - IBM vs. NZ CliFlo: IBM highlights wind risk and seasonal humidity, while NZ CliFlo captures daily sunlight and humidity.

From Fig. 28, by analysing each row, we obtain the following conclusions.

- NIWA Climate Data (Top Panel): Significant contributions from the features *niwa pressure FruitSet* and *niwa extreme heat_factor FruitSet* highlight the impor-tance of pressure and heat stress. Emphasises seasonal factors such as pressure, heat stress, and temperature.
- IBM Climate Data (Middle Panel): Contributions from the features *ibm_air_temperature mean Dormant* and *ibm_wind_speed std_Harvest* demonstrate reliance on temperature and wind speed metrics. Highlights seasonal temperature and wind dynamics.
- NZ CliFlo Climate Data (Bottom Panel): Significant contributions from the features *GustDir(Deg)_FruitSet* and *Rad(MJ/m2)_FruitSet* indicate the importance of wind direction and radiation on specific days. Captures fine-grained daily variations, especially wind and radiation.
- Key Differences
  - NIWA: Focuses on seasonal pressure and heat stress.
  - IBM: Highlights seasonal temperature and wind speed.
  - NZ CliFlo: Captures daily wind direction and radiation variations.

From Fig. 29, the following conclusions are drawn by analysing each row.

- NIWA Climate Data (Top Panel): Positive contributions from the feature *niwa extreme heat_factor FruitSet* highlight the impact of heat stress, while negative contributions from the feature *niwa temperature min_FruitSet_delta* reflect sensitivity to temperature extremes. Emphasises seasonal heat stress and temperature metrics.
- IBM Climate Data (Middle Panel): Positive contributions from the feature *ibm_air_temperature max Dormant* and negative contributions from the feature

ibm_relative_humidity_max_Dormant underscore the importance of temperature and humidity. Highlights temperature metrics with additional influence from humidity.
- NZ CliFlo Climate Data (Bottom Panel): Positive contributions from the feature *Rain(mm)_FruitSet* and negative contributions from the feature *GustDir(Deg)_FruitSet* show sensitivity to daily rainfall and wind dynamics. Captures fine-grained daily weather variations.
- Key Differences
  - NIWA: Focuses on seasonal heat stress and temperature extremes.
  - IBM: Highlights temperature with additional sensitivity to humidity.
  - NZ CliFlo: Captures detailed daily variations, particularly rainfall and wind.

15.3 Our Findings

This analysis compares different kiwifruit types (GA Conventional, HW Conventional, and HW Organic) and climate datasets (NIWA, IBM, Daily) using 1D-CNN models and multiple XAI methods (KernelSHAP, LIME). Below is the summary.

Key Insights: Different Kiwifruit Types

The analysis of different kiwifruit types using a specific AI model and XAI methods reveals distinct feature prioritisation results:

- GA Conventional
  - Relies on diverse seasonal growth metrics, including degree days, rainfall, and sunlight.
  - Feature contributions are balanced across multiple climatic variables.
- HW Conventional
  - Focuses heavily on stress-related metrics such as heat stress, pressure, and temperature.
  - Reflects strong sensitivity to extreme climatic conditions.
- HW Organic
  - Shifts focus to growth-related metrics like temperature and radiation.
  - Demonstrates balanced reliance on atmospheric factors such as pressure and precipitation.

These findings show how dataset characteristics influence both model predictions and XAI explanations, with each kiwifruit type exhibiting unique patterns of feature importance and contributions.

Key Insights: Different Climate Data

The comparison of climate datasets (NIWA, IBM, and NZ CliFlo) using the same AI model and XAI methods reveals distinct patterns of feature prioritisation and focus:

- NIWA Climate Data
  - Focuses on seasonal metrics such as pressure, heat stress, chill days, and radiation.

– Demonstrates sensitivity to broad climatic conditions influencing seasonal trends.

- IBM Climate Data
  - Highlights temperature, wind speed, and wind risk as key factors.
  - Balances reliance on seasonal humidity and temperature metrics, emphasising wind dynamics.
- NZ CliFlo Climate Data
  - Captures fine-grained daily variations in sunlight, wind direction, and rainfall.
  - Reflects localised weather dynamics with a focus on specific daily conditions.

These findings demonstrate that NIWA emphasises broad seasonal trends, IBM focuses on wind and humidity with a seasonal perspective, and NZ CliFlo captures detailed daily weather variations. The choice of dataset should align with the specific analysis goals, whether to understand broad seasonal trends or detailed daily conditions.

16 Discussion

Why Did We Choose Different Yield Prediction Methods?

The choice of multiple yield prediction methods in this study is driven by the uncertainty regarding the specific AI models that end-users may adopt. To address this, we selected four distinct models spanning both generative (e.g., GMM) and discriminative (e.g., 1D-CNN, XGBoost, SVR) approaches. This diversity ensures comprehensive coverage of different model paradigms commonly used in real-world applications.

Furthermore, by comparing the explanations generated under the same XAI method (e.g., SHAP or LIME) across these models, we can assess how prediction accuracy impacts explainability. This approach allows us to evaluate the robustness of explanations and their dependency on the underlying AI model, thereby providing deeper insights into the relationship between predictive performance and explainability. Such comparisons are crucial for guiding users in selecting the most appropriate AI and XAI models for their specific use cases.

Why Did We Choose Different XAI Methods?

The choice to employ multiple XAI methods in this study is motivated by the distinct characteristics each method offers. Different XAI techniques, such as SHAP, KernelSHAP, and LIME, provide varying perspectives on feature importance and contribution, allowing for a more diverse understanding of the model's behaviour.

By comparing the explanations generated by different XAI methods for the same AI model, we can gain a more comprehensive and nuanced interpretation of the predictions. This comparative approach enables us to identify consistent patterns across methods while highlighting method-specific insights. Such an analysis not only enhances interpretability but also offers valuable guidance for selecting the most suitable XAI method for specific models and applications.

Why Did We Choose Different Datasets?

The decision to use multiple datasets in this study stems from the aim to investigate whether variations in kiwifruit types and climate data sources influence model explanations. By incorporating datasets representing different kiwifruit types (HW Organic, HW Conventional, GA

Conventional) and climate data sources (IBM, NIWA, Daily), we ensure that the analysis captures the potential impact of these factors on interpretability.

This approach allows us to evaluate how model explanations differ across diverse environmental and crop-specific conditions, providing valuable insights into the robustness and adaptability of AI and XAI models. Understanding these variations is crucial for tailoring models to specific datasets and ensuring reliable interpretations in practical applications.

Why Different XAI Methods Provide Different Explanations?

Different XAI methods provide varying explanations due to their underlying mechanisms and assumptions. For example, SHAP and KernelSHAP use game-theoretic principles to allocate feature contributions, aiming for consistency and additive explanations, while LIME approximates the model locally with simpler interpretable models. These methodological differences lead to variations in how features are ranked, their impact magnitudes, and the distribution of contributions.

The choice of XAI method also influences the granularity and focus of the explanation. For instance, SHAP tends to offer more detailed and globally consistent insights, whereas LIME provides localised explanations that can vary based on the specific sample analysed. As a result, the explanations generated by different methods may highlight complementary aspects of the model's behaviour, making it essential to consider multiple XAI techniques for a comprehensive interpretation of predictions.

16.1 Limitations and Opportunities

Through the analysis of this study, several challenges were identified, which highlight the complexities of combining multiple AI models, XAI methods, and datasets for yield prediction and explanations:

- Variability Across XAI Methods: Different XAI methods produce varying explanations for the same AI model, making it challenging to consolidate insights and identify consistent patterns.
- Confusion Due to Feature Dependencies and Correlations: Climate features often exhibit dependencies and/or temporal correlations, leading to similar SHAP values across correlated features [70, 71]. This makes it difficult to clearly identify the specific impact of each feature on the predictions.
- Limited Applicability of SHAP to Non-tree-based or Non-additive Models: While SHAP performs well with tree-based models (e.g., XGBoost, Random Forest) and linear models, it is less effective for other types of predictive models.
- Computational Limitations: The computational cost of certain XAI methods (e.g., KernelSHAP) rises sharply with larger datasets, limiting scalability.
- Alignment of Explanations with Domain Knowledge: There is a need to develop effective methods to align and compare model explanations with domain knowledge. Establishing a reliable validation framework to ensure the consistency and accuracy of explanations is critical.

These challenges underline the need for thoughtful selection of AI models, XAI methods, and datasets, as well as further investigation on how to enhance the robustness and scalability of interpretability techniques. Addressing these limitations will be critical to ensuring that the explanations provided by XAI methods are both trustworthy and actionable for stakeholders in agriculture.

17 Part IV Conclusion

Part IV presents a comprehensive evaluation of XAI baseline models for kiwifruit yield prediction, conducted across three kiwifruit types (HW Organic, HW Conventional, and GA Conventional), four AI models (XGBoost, SVR, 1D-CNN, and GMM), and three XAI methods (SHAP, KernelSHAP, and LIME), using multiple climate datasets (IBM, NIWA, and Daily). Our experiments aim to systematically investigate how model architecture, explanation technique, and climate data source jointly affect interpretability and predictive behaviour.

Our experiment results demonstrate that both model structure and XAI methodology substantially influence explanation outcomes. Tree-based and regression models tend to yield concentrated attributions dominated by pressure-related variables, whereas deep learning and probabilistic models capture broader interactions among temperature, radiation, and growth indicators. Similarly, SHAP-based methods provide globally consistent and additive explanations, while LIME produces more localised and instance-specific insights. These findings highlight that interpretability is inherently dependent on both the modelling approach and the characteristics of the input data, including its temporal resolution, feature composition, and climatic variability.

Our analysis further identified several limitations in current XAI approaches for agricultural applications. Variability among XAI methods, computational constraints with high-dimensional datasets, and challenges in aligning data-driven explanations with agronomic domain knowledge all remain open issues. These limitations emphasise the need for more robust and scalable interpretability frameworks capable of producing consistent, domain-aligned insights across models and datasets.

Overall, this study establishes a foundational benchmark for evaluating explainability in climate–agriculture analytics. The findings underscore that trustworthy agricultural AI requires a careful integration of model design, data resolution, and interpretability method.

Part V
Extreme Weather Event Retrieval & Evaluation

18 Part V Introduction

Extreme weather events, such as droughts, heavy rain, and floods etc., have been making significant impacts on agriculture, infrastructure, and many other aspects in society. Reliable detection, severity quantification, and validation of these events are critical for risk assessment and decision-making. However, the scarcity of standardised extreme weather event datasets and the difficulties in extracting consistent and aligned information from news, reports and/or other narrative sources pose considerable obstacles.

To address these challenges, in Part V of this project, we conducted a multi-faceted approach to further our investigation. It involves three components:

- **Extreme Weather Events Data Retrieval**
  We collected extreme weather event data for Australia (AU), New Zealand (NZ), and the United States (US) from both structured and unstructured data sources.
- **Extreme Weather Events Severity Quantification**
  We quantified the severity of extreme weather events by developing a continuous-scale representation based on sentiment analysis and semantic similarity scoring.
- **Extreme Weather Events Alignments Evaluation**
  We evaluated the alignment between extreme weather events extracted by Large Language Models (LLMs) from news articles and climate reports of the National Institute of Water and Atmospheric Research (NIWA) in NZ, and our detected anomalies from NZ's daily climate data.

Our methodologies and processing details in each of these components are summarised in Section 19, Section 20, and Section 21, respectively.

19 Extreme Weather Events Data Retrieval

In this section, we present our data retrieval of extreme weather events for Australia (AU), New Zealand (NZ), and the United States (US). We first consolidated our data from multiple sources, including publicly available event databases, and climate research and harvest annual reports, to ensure we have a comprehensive coverage across regions and event types. The collected dataset serves as our database for subsequent severity quantification and alignment evaluation.

Again, our primary data collection efforts focus on what happened in New Zealand, utilising multiple sources such as NIWA[6] and Industry Partner[7] reports; while the extreme weather event labels were obtained by the following four approaches:

[6] https://niwa.co.nz/publications
[7] https://www.zespri.com/en-NZ/publications/annualreports

1. **Using NIWA Daily Climate Data**
   The NIWA daily climate dataset provides records of extreme weather categories, including Gale, Snow, Hail, Lightning, Thunder, and Fog. However, this dataset is sparse, with only approximately 2.4% of entries containing extreme weather event information.
2. **Using ChatGPT to Generate Extreme Weather Data**
   To complement the official records, we employed ChatGPT[8] to generate extreme weather event data for regions corresponding to the kiwifruit yield dataset. The model was prompted to identify extreme events for specified locations, where its reply and reference links were captured for further analysis. However, our critical evaluation revealed that the generated data was largely inaccurate.
3. **Extracting Extreme Weather Events from NIWA News Reports**
   We further compiled extreme weather event data by data extraction techniques on NIWA news reports (2019–2023) and NIWA Seasonal/Annual Climate Summaries (2015–2024). Using a LLM, relevant event descriptions were extracted and organised.
4. **Analysing Industry Partner Reports for Extreme Weather Impacts on Kiwifruit Yield**
   Similarly, climate-related information was extracted from an industry partner's Annual Reports (2017–2023) using ChatGPT and other LLMs. The extracted data were analysed to assess the impacts of climate factors on kiwifruit yield. Furthermore, extreme weather events were manually extracted and analysed highlighting the effects of different event types on kiwifruit production.

Collectively, these datasets provide a comprehensive database of raw information for quantifying the severity of extreme weather events and evaluating their alignment across different extraction methods. Next, in Section 20, we will cover our proposed method for severity quantification.

20 Extreme Weather Events Severity Quantification

This section presents our methodology for quantifying the severity of extreme weather events. Beyond a na¨ıve binary event labelling (i.e., yes or no), we established a more perceptive method to assign a continuous severity scale (ranging from 0.0 to 1.0) to each of the events based on the tone of their textual descriptions through sentiment analysis and semantic evaluations among those descriptions. This continuous representation enables finer-grained analysis and enhances the interpretability of model outputs in downstream applications.

In our investigation, two complementary approaches were explored to derive severity scores:

1. **Sentiment Analysis and Semantic Similarity Scoring**
   We employed a sentiment analyser and a semantic similarity model to evaluate the textual descriptions of extreme weather events. The sentiment analyser is based on the distilbert-base-uncased-finetuned-sst-2-english[9] model, while the

[8] https://chatgpt.com
[9] https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

Fig. 30: Word cloud for the highest severity events.

(a) Lowest severity event example 1.

(b) Lowest severity event example 2.

Fig. 31: Word clouds for two examples of the lowest severity events.

semantic similarity model is all-MiniLM-L6-v2[10]. Sentiment analysis was conducted to assess the overall emotional tone of each description, while semantic similarity scoring captured the contextual relevance of event details. The final severity score was obtained by combining sentiment and semantic similarity scores with assigned weights based on their relative importance, where a higher score indicates greater event severity.

To further analyse the textual characteristics of extreme weather event descriptions, word clouds were generated. Fig. 30 presents the word cloud for the highest severity events, while Fig. 31 shows two examples of word clouds for the lowest severity events. Fig. 32 is the overall word cloud, which was constructed by averaging the word weights across all events, providing a general summary of the dataset.

We summarise our key observations from the word clouds as follows:

- **Highest Severity Events (Fig. 30):** Dominant terms such as *heavy*, *rain*, *potentially*, and *flooding* indicate that high-severity events are often characterised by intense rainfall and localised flood risks.
- **Lowest Severity Events (Fig. 31):** Terms like *small*, *new*, *hotspot*, and *formed* suggest the presence of minor or emerging conditions, such as small drought hotspots, rather than large-scale impacts.

[10]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

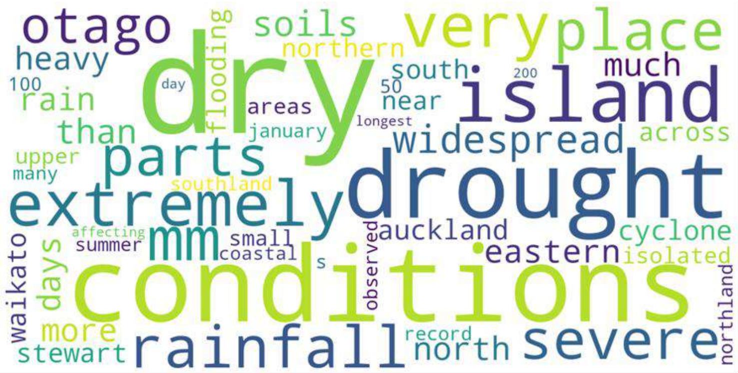


Fig. 32: Word cloud summarising all events.

- **All Events Summary (Fig. 32):** The frequent occurrence of terms such as *dry*, *conditions*, *drought*, and *rainfall* reflects the predominance of drought-related impacts and rainfall variability across the entire dataset.

These visualisations reinforce the textual patterns identified through the severity quantification process and provide qualitative support for the assigned severity scores.

2. **Direct Semantic Similarity Computation Using LLMs**

   As a complementary approach, we employed a LLM-based similarity scoring method to directly assess the severity of event descriptions. The intensity scores are also ranging from 0.0 to 1.0, where higher scores correspond to more severe events. These scores were derived based on assessments of the detailed description of the events.

The outcomes of these two methods offer complementary perspectives on the severity of extreme weather events, providing a robust foundation for the subsequent alignment evaluations presented in Section 21.

21 Extreme Weather Events Alignments Evaluation

This section evaluates the alignment between i) extreme weather events extracted by LLMs from NIWA news articles and climate reports, and ii) climate anomalies detected from NZ daily climate data. The objective is to assess the consistency between narrative-based extrac-tions and observation-based anomaly detections, providing insights into the reliability of automated extraction methods for extreme event monitoring.

21.1 Alignment Evaluation Methodology

We evaluated the alignment between extracted extreme weather events and detected climate anomalies through a three-step process:

1. **Event Extraction:** Extreme weather events were extracted from NIWA reports, including event type, start time, and end time (see Section 19).

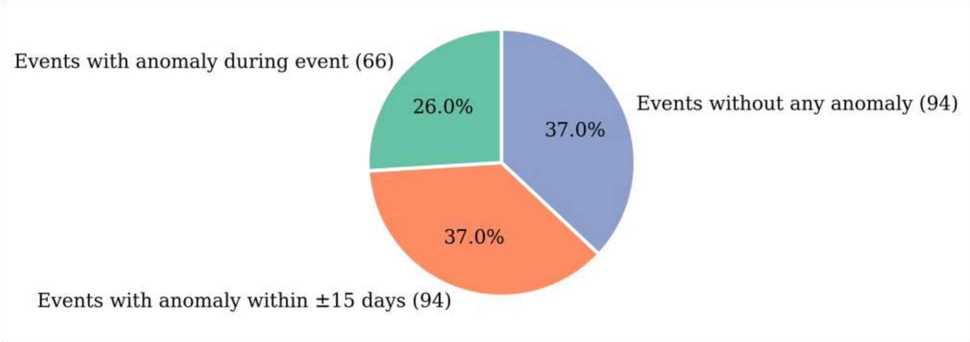


Fig. 33: Distribution of events by anomaly detection status (Total: 254 events).

2. **Anomaly Detection:** Isolation Forest — a machine learning method for anomaly detection — was applied to NZ daily climate data, labelling each value as 0 (normal), 1 (positive anomaly, above the threshold), or -1 (negative anomaly, below the threshold), retaining only 1 and -1 for further analysis.
3. **Matching Procedure:** Events were matched to detected anomalies from weather stations within a 20 km radius, during a specified time window.

21.2 Overview of Events and Temporal Distribution

We first provide an overview of the extracted events, presenting their alignment proportions and overall temporal distribution.

Alignment Proportions

Among the 254 extreme events, the anomaly detection results are summarised as follows:

- **66 aligned events (25.9%)**: Anomalies were detected precisely at the event time.
- **94 window-aligned events (37%)**: Anomalies were detected within a ±15-day window around the event.
- **94 non-aligned events (37%)**: No anomalies were detected either at the event time or within the ±15-day window.

Fig. 33 illustrates the distribution of events across these three categories, highlighting that while a portion of extreme events corresponded directly with detected anomalies, many required broader temporal consideration, and some events were not captured by anomaly detection at all.

Temporal Distribution

The timeline of the 254 extreme events, categorised by anomaly detection outcomes, is presented in Fig. 34. Events are grouped into aligned, window-aligned, and non-aligned categories based on their anomaly detection status. This visualisation reveals temporal clustering patterns and highlights periods with higher or lower anomaly alignment.

Aligned events were mainly concentrated during the summer months (December–February), window-aligned events showed a broader temporal spread, and non-aligned events were more evenly distributed across the year.

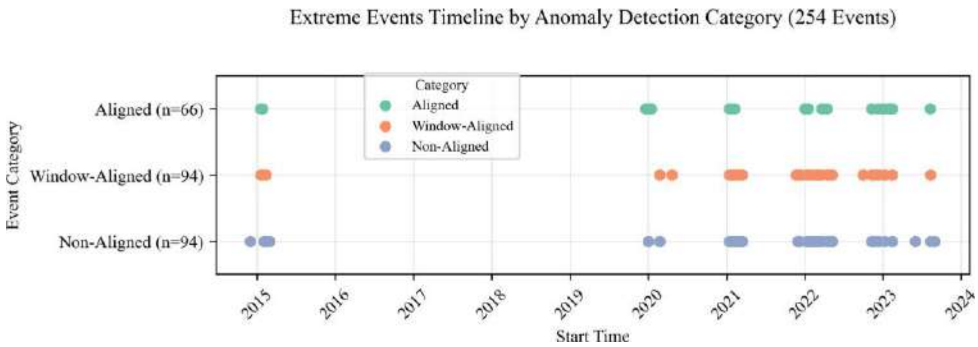


Fig. 34: Timeline of the 254 extreme events classified by anomaly detection results. Events are categorised into three groups: Aligned (anomalies detected exactly at event time), Window-Aligned (anomalies detected within a ±15-day window), and Non-Aligned (no anomalies detected).

Extreme Event Anomaly Detection

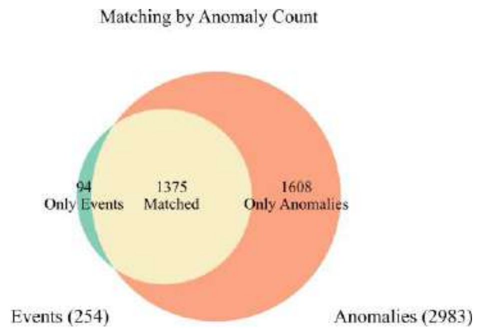
To detect extreme events in daily climate records, we applied the Isolation Forest algorithm to each climate feature individually, treating the task as univariate anomaly detection. This approach allowed us to identify feature-specific deviations without being influenced by inter-variable correlations. The contamination parameter, which determines the proportion of outliers, was set according to the highest observed ratio between known extreme events and total daily observations. Anomalies were then assigned directional labels: +1 for the most extreme high values, -1 for the most extreme lows, and 0 for all other cases. The identified outliers for each feature will later be cross-examined against events extracted from reports and news articles, with the aid of LLMs, to validate and contextualise the anomalies within real-world occurrences.

Event-Anomaly Matching Analysis

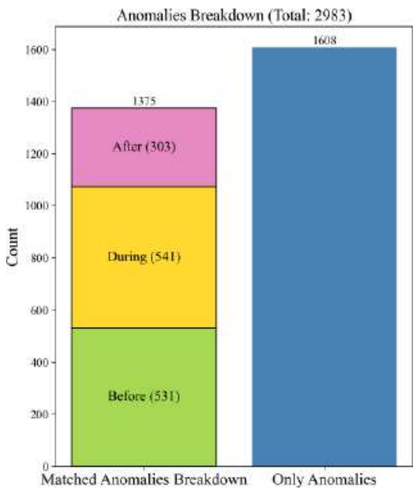
To evaluate the correspondence between reported extreme weather events and detected anomalies, we present a joint visualisation of overlap and breakdowns in Fig. 35a. Out of a total of 2,983 anomalies and 254 recorded events, 1,375 anomalies were matched to events, indicating a moderate alignment (overlap). However, 1,608 anomalies (approximately 54%) remained unmatched, while 94 events (37% of all events) did not align with any anomaly.

Fig. 35b further categorises the 1,375 matched anomalies based on their temporal alignment with the associated event. A majority of matched anomalies occurred *before* (531) or *during* (541) the event period, supporting the hypothesis that anomalies often precede or coincide with extreme events. An additional 303 anomalies were observed *after* the event, which may reflect lingering climate disturbances or post-event recovery periods. The remaining 1,608 unmatched anomalies may represent undetected or unreported events, or noise in the anomaly detection process.

On the event side, Fig. 35c provides a breakdown of the 94 unmatched events. Among the records, 23 events lacked station information, and 33 were associated with stations that did not have corresponding climate data. Only 38 unmatched events had climate data available, suggesting that most unmatched cases are attributable to data availability issues. For the



(a) Venn diagram showing event-anomaly overlap



(b) Temporal distribution of matched anomalies

(c) Data availability and matching types for events

Fig. 35: Event and Anomaly Matching Analysis.

160 matched events, 66 were strictly aligned with anomalies, while 94 were matched through a relaxed window-based approach.

To better understand the unmatched events that still have available climate data (38 in total), we further analysed their characteristics in terms of event duration and event type (Fig. 36).

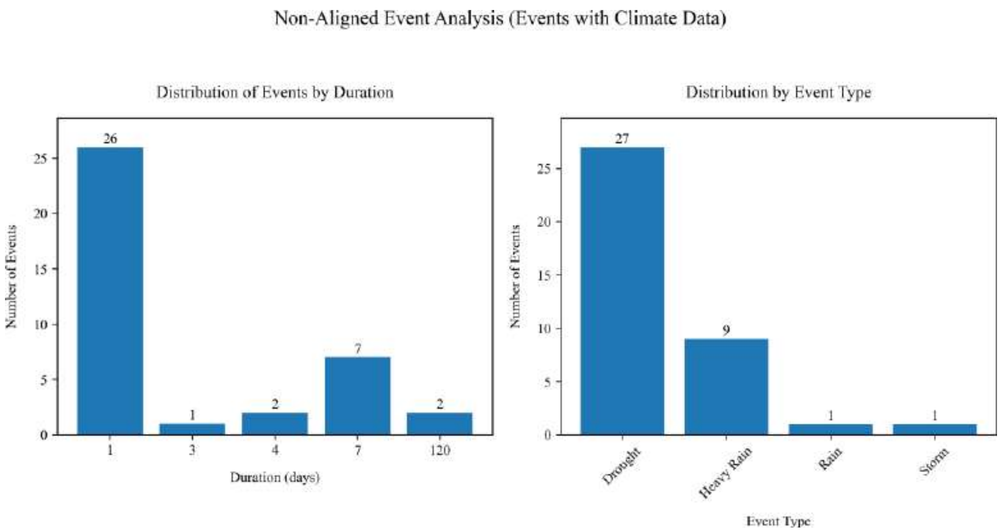


Fig. 36: Distribution of unmatched events with climate data by duration (left) and event type (right).

Duration Analysis: A large majority of the non-aligned events are short in duration, with 26 out of 38 events (68%) lasting only a single day. A few events span longer periods, including 7 events lasting 7 days and 2 events lasting up to 120 days. The short duration of most events may contribute to the difficulty in identifying corresponding anomalies, especially if the anomaly window or detection threshold is not sufficiently sensitive to brief disturbances.

Event Type Analysis: Most non-aligned events are droughts (27), followed by heavy rain events (9), with only one case each of rain and storm. This suggests that while droughts are temporally prolonged phenomena, their climate signal may be subtle or distributed across time, making them harder to match within fixed anomaly windows. Conversely, heavy rain events typically have stronger signals, but their short duration might not coincide precisely with detected anomalies.

Our Verdict: The lack of alignment for these 38 events can largely be attributed to temporal resolution mismatches, subtle or distributed anomaly signatures (as in droughts), and potential thresholding limitations in the anomaly detection algorithm. Future work may explore adaptive window sizes or anomaly scoring methods that better capture subtle yet meaningful deviations in climate variables.

21.3 Seasonal Distribution Analysis

This section analyses the seasonal occurrence patterns of extreme events, first presenting the overall distribution, followed by distributions across different alignment categories.

Overall Seasonal Distribution

Fig. 37 presents the seasonal distribution of all 254 extreme events according to customised seasonal definitions for the crop. The Fruit Set period (December–February) recorded the highest number of events, while the Dormant period (June–August) had the fewest. This pattern suggests a strong concentration of extreme events during the fruit development stage.

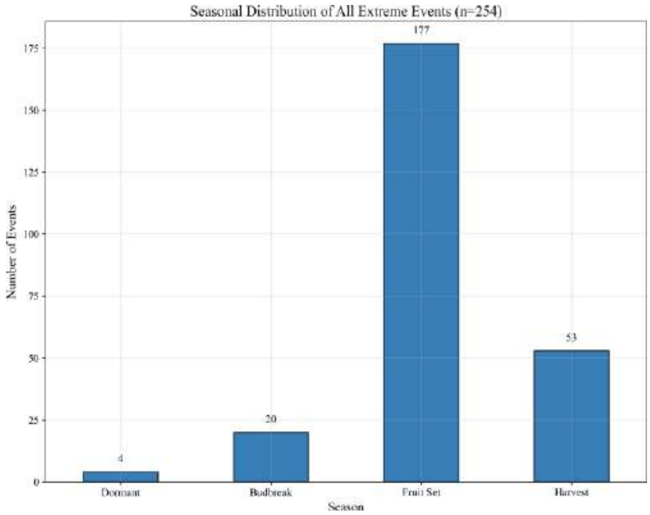


Fig. 37: Seasonal distribution of all 254 extreme events. Seasonal distribution of aligned events based on customised seasonal periods: Dormant (June–August), Budbreak (September–November), Fruit Set (December–February), and Harvest (March–May).

Seasonal Distribution by Alignment Category

Fig. 38a shows the seasonal distribution of aligned events. A strong concentration was observed during the Fruit Set stage (December–February).

Fig. 38b presents the seasonal distribution of window-aligned events based on customised seasonal definitions. Similar to aligned events, a large proportion of window-aligned events occurred during the Fruit Set period (December–February), suggesting that extreme events during this stage may exhibit slight temporal offsets from detectable anomalies.

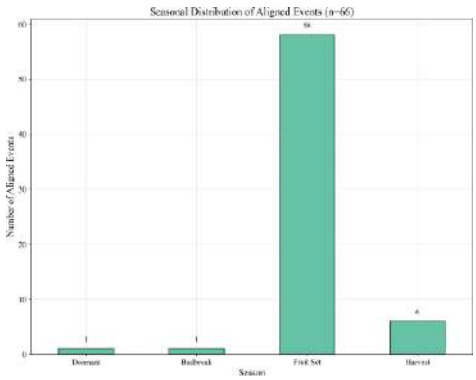
Fig. 38c presents the seasonal distribution of non-aligned events. Despite the absence of detected anomalies, a large proportion of non-aligned events still occurred during the Fruit Set (December–February) and Harvest (March–May) periods.

In addition to temporal patterns, it is important to understand the types of extreme weather events that occurred. The following section analyses the distribution of event types across different alignment categories.

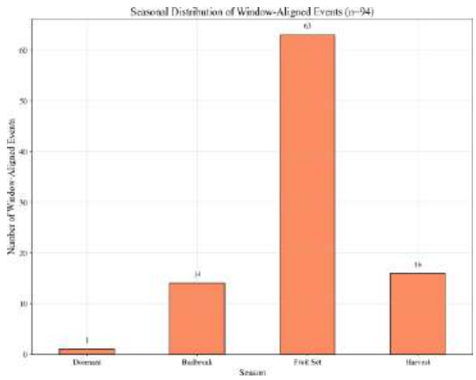
21.4 Event Type Distribution Analysis

In this section, we analyse the types of extreme weather events across different alignment categories. Understanding the distribution of event types provides insights into which kinds of events are more likely to align with detected climate anomalies.

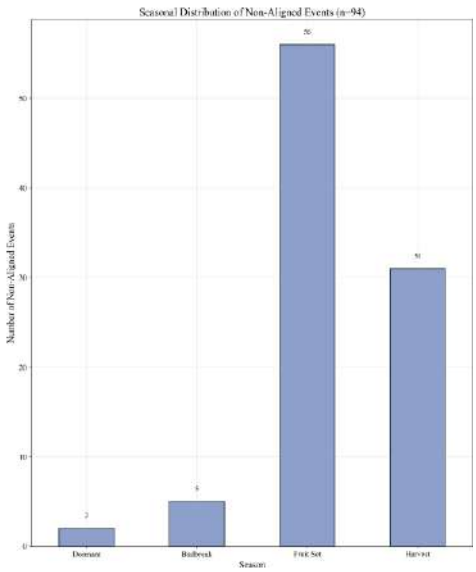
The event type distribution is presented in Fig. 39a. Droughts, heavy rain, and floods were the most frequent event types among aligned events. The event type distribution for window-aligned events is shown in Fig. 39b. Drought events dominated this category, followed by heavy rain events. The distribution of event types among non-aligned events is shown in Fig. 39c. Drought events were predominant, accounting for the majority of the non-aligned category.



(a) Aligned Events

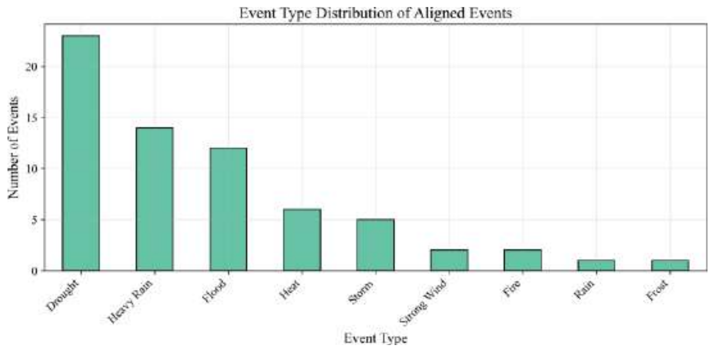


(b) Window-Aligned Events

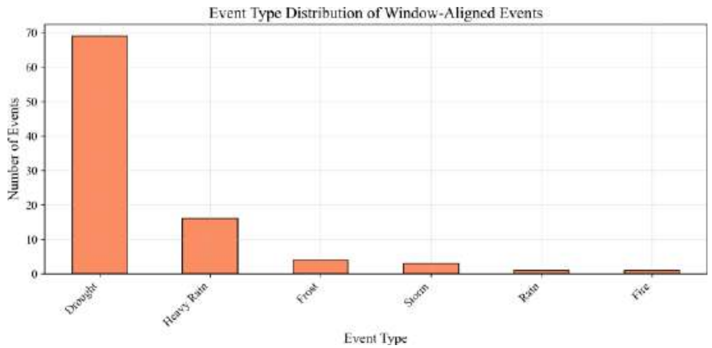


(c) Non-Aligned Events

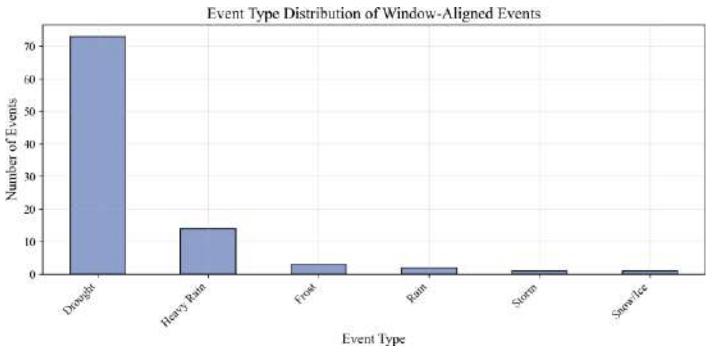
Fig. 38: Seasonal distribution across different alignment categories. Seasonal distribution of aligned events based on customised seasonal periods: Dormant (June–August), Budbreak (September–November), Fruit Set (December–February), and Harvest (March–May).



(a) Aligned Events



(b) Window-Aligned Events



(c) Non-Aligned Events

Fig. 39: Event type distribution across different alignment categories.

Across all categories, droughts were the most common type of extreme event, especially dominating the non-aligned group. Floods and heavy rain events were more frequently observed among aligned and window-aligned events, indicating stronger correspondence between these event types and observable climate anomalies.

In addition to event type patterns, spatial distribution provides critical insights into the regional vulnerabilities to extreme weather events. The next section examines the geographic distribution of events across different alignment categories.

21.5 Location Distribution Analysis

We next examine the geographic distribution of extreme events across alignment categories.

The spatial distributions of extreme weather events are summarised by the bar charts in Fig. 40. Fig. 40a and Fig. 40b present the top 15 locations with the highest frequencies of aligned and window-aligned events, respectively. Northland and Auckland consistently ranked among the most affected regions across all alignment categories, suggesting persistent susceptibility to extreme weather impacts. In contrast, Fig. 40c highlights the top 15 locations for non-aligned events, identifying areas that may warrant further investigation to assess potential data coverage limitations or unique localised event characteristics.

21.6 Anomaly Feature Patterns Analysis

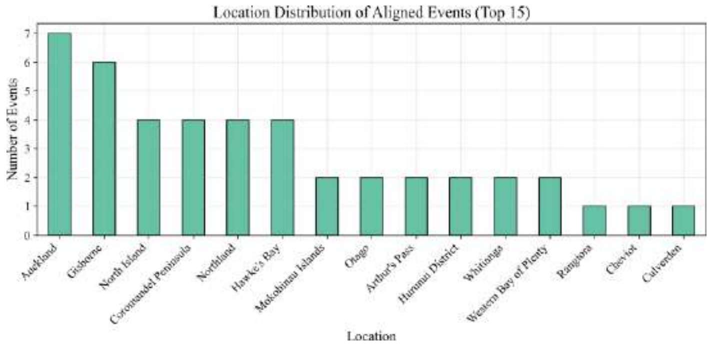
Finally, we examine the anomaly feature patterns associated with aligned and window-aligned events. Aligned events were associated with direct temperature and rainfall anomalies, while window-aligned events exhibited meaningful pre-event signals, suggesting potential for early warning detection.

Analysis of Aligned Events

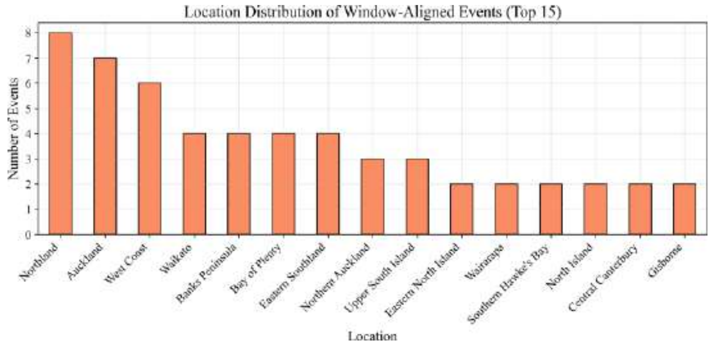
The heatmap in Fig. 41 summarises the anomaly patterns associated with different types of aligned extreme weather events. Key observations from the heatmap include:

- **Drought events** exhibited strong associations with high-temperature anomalies:
  - Positive anomalies were observed in `Tmax(C)` (67 instances), `ET20(C)` (30 instances), and `ET10(C)` (31 instances), suggesting intensified drought conditions.
- **Flood events** and **Heavy rain events** were closely associated with positive anomalies in `Rain(mm)`:
  - 28 instances of rainfall anomalies for floods and 32 instances for heavy rain events were recorded.
- **Negative anomalies in relative humidity (**`RH(%)`**)** were frequent in drought events:
  - 38 instances of negative `RH(%)` anomalies were detected, reinforcing the presence of arid conditions.

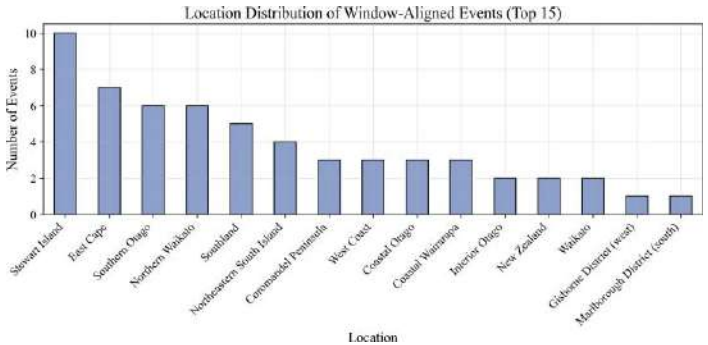
Overall, the results demonstrate meaningful correspondences between detected meteorological anomalies and reported extreme events.



(a) Aligned Events



(b) Window-Aligned Events



(c) Non-Aligned Events

Fig. 40: Top 15 locations for extreme events across different alignment categories.

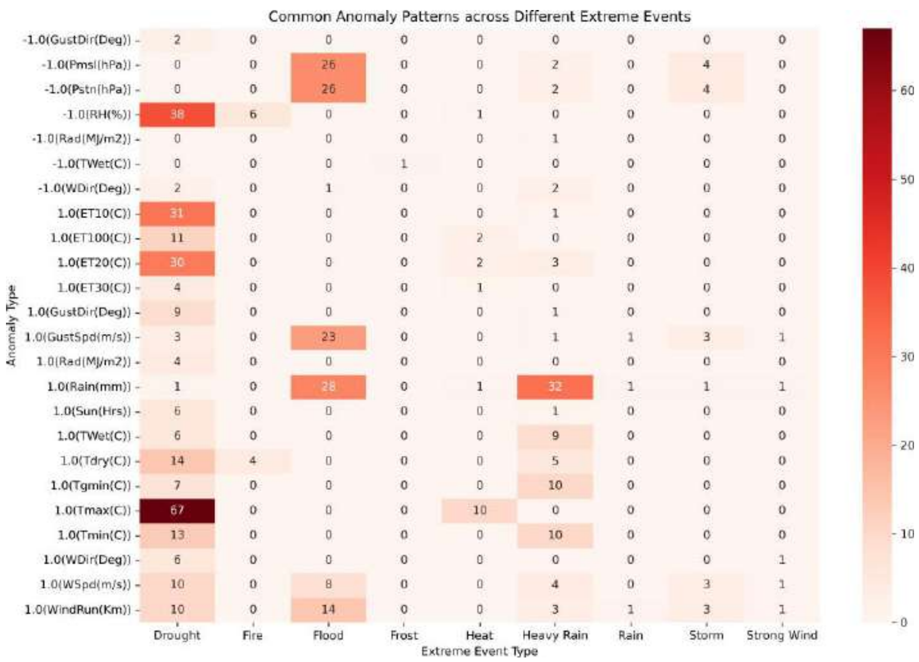


Fig. 41: Heatmap illustrating anomaly patterns aligned with extreme weather events. Droughts were strongly associated with high-temperature and low-humidity anomalies; floods and heavy rain events corresponded to positive rainfall anomalies; and fire events were linked to low-humidity anomalies. (Note: In this heatmap, feature names prefixed with 1 indicate positive anomalies (above standard values), and -1 indicate negative anomalies (below standard values).)

Analysis of Window-Aligned Events

We further analysed extreme events that did not directly align with detected anomalies. To expand the temporal matching window, we considered anomalies occurring within 15 days before or after the reported event period. In the corresponding heatmap, feature names are prefixed with 1 or -1 to indicate positive and negative anomalies, respectively, and anomalies are labelled based on their temporal relation to the event (#1 for before, #-1 for after). Fig. 42 visualises the anomaly patterns observed around window-aligned events. Key findings from the non-aligned events heatmap analysis include:

- **Drought events** continued to show strong associations with temperature-related anomalies:
  - Positive anomalies were observed in Tmin(C) (51 instances) and Tdry(C) (30 instances).
  - Anomalies in Tgmin(C) and Tgmax(C) were also notable.
- **Heavy rain events** were often preceded by anomalies in radiation and rainfall features:

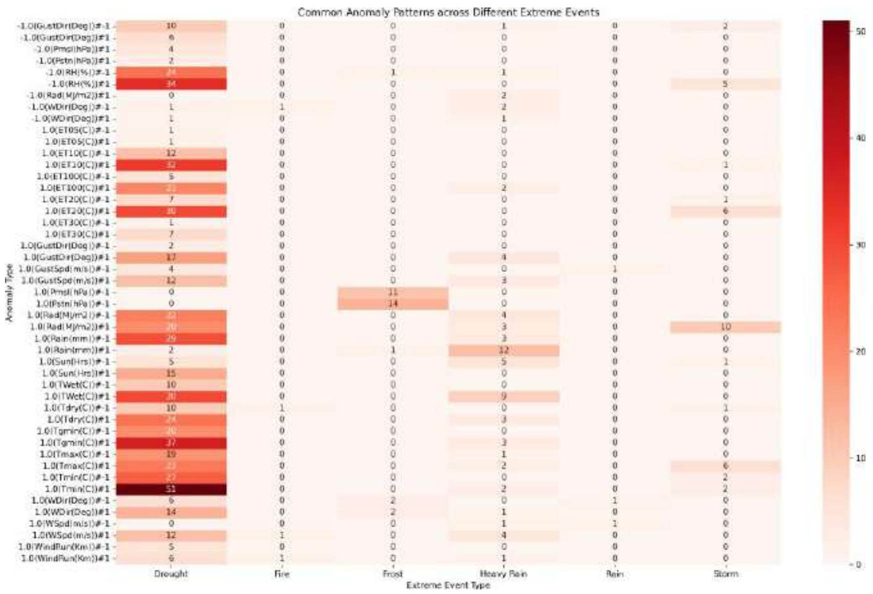


Fig. 42: Heatmap illustrating anomaly patterns associated with window-aligned extreme events. (Note: Feature names prefixed with 1 or -1 indicate positive and negative anomalies relative to standard thresholds. Anomalies are labelled as #1 if occurring before the event and #-1 if occurring after the event.)

  - Positive anomalies in Rad(MJ/m^2), TWet(C) and Rain(mm) were identified.

- **Frost events** showed fewer but relevant associations, mainly in pressure anomalies.
- **Fire events** exhibited prior anomalies in wind-related features such as WDir(Deg) and WindRun(Km).

These results suggest that even when direct temporal alignment is not observed, extreme events are often preceded by detectable meteorological anomalies, highlighting the potential for early warning signals based on anomaly detection.

22 Extensibility Verification - Case Study on US Data

We applied our methods to a new dataset from the United States to demonstrate the extensibility of our extreme event data retrieval framework. This case study includes a multi-faceted severity quantification analysis and alignment evaluation between detected anomalies and recorded extreme events. We extended our framework with two distinct severity scoring methods:

1. **Anomaly-Based Severity Quantification**
   Anomalies (both positive and negative) were clustered based on their values to

establish feature-specific severity thresholds. Events were then evaluated based on the severity levels of their associated anomalies.

2. **NOAA-Based Severity Quantification**
   Using direct indicators of impact from the Storm Events database[11] of National Oceanic and Atmospheric Administration (NOAA)—such as deaths, injuries, and economic losses—we constructed a rule-based and weighted scoring system to assess event severity.

Following severity quantification, we conducted a detailed comparative analysis of the two scoring methods, assessing differences in severity classification and exploring their correlation. We then analysed the alignment between anomalies and events, including event type distribution and anomaly frequency heatmaps categorised by event type.

22.1 Anomaly-Based Severity Quantification

This method evaluates the severity of extreme events based on the intensity and frequency of meteorological anomalies. A two-stage process was implemented: i) anomaly clustering for each variable to determine severity thresholds, and ii) rule-based aggregation to derive event-level severity scores. This approach captures deviations from normal climate conditions and emphasises the meteorological extremity of events.

Anomaly Clustering and Threshold Derivation

We separately process positive anomalies (anomaly = 1) and negative anomalies (anomaly = -1) for each meteorological variable. Only features with sufficient sample size (at least three values) are eligible for clustering to ensure reliability. For features with limited unique values, a customised classification logic is applied:

- **Single-value feature:** All instances are classified as *low severity*.
- **Two-value feature:** The smaller value is labelled *low*, the larger *medium*.
- **Three or more values:** K-means clustering ($k = 3$) is applied. Cluster labels are assigned based on centroid order:
  - Smallest centroid → *low severity*
  - Middle centroid → *medium severity*
  - Largest centroid → *high severity*

The derived thresholds for each feature are stored for use in downstream analysis. For instance:

- **Wind-related features:**
  - AWND (Average Wind Speed): anomaly > 9.84 → *low*, > 12.53 → *medium*
  - WSF2 (2-min Gust): anomaly > 25.1 → *low*, > 29.1 → *medium*
- **Temperature features:**
  - TAVG: anomaly > 79.0°F → *low*, > 84.0°F → *medium*
  - TMIN: anomaly > 58.0°F → *low*, > 63.0°F → *medium*

[11] https://www.ncei.noaa.gov/stormevents/

- **Weather-type indicators (WT*):** Treated as *high severity* if any anomaly is present.

Event-Level Severity Evaluation

After deriving severity labels for anomalies, we perform event-level assessment by aggregating anomaly signals over the event duration. The evaluation process is as follows:

- **For weather-type features (WT*):** Any anomaly triggers a *high severity* label for the event, due to the direct association with hazardous conditions (e.g., freezing rain, snow, tornado).
- **For other meteorological features:** Severity is determined based on anomaly frequency:
  - ≥ 3 anomalies → *high severity*
  - 2 anomalies → *moderate severity*
  - 1 anomaly → *mild severity*

In cases where multiple features show anomalies, a cumulative severity judgement is made by prioritising WT* presence and summing anomaly occurrences across features.

Special Case Handling

- **Missing thresholds:** Features without valid thresholds are treated as *low severity* by default.
- **No anomaly records:** Events with no detected anomalies are automatically labelled *low severity*.
- **Compound anomalies:** Events with concurrent anomalies across multiple features may be upgraded in severity, depending on the anomaly count and type.

Outputs and Validation

The final outputs include:

- A severity label (*low*, *medium*, *high*) for each event
- A distribution summary of severity levels across the dataset
- Visual examples highlighting the clustering results and representative severe events

This method provides a quantitative and meteorologically grounded way to assess event severity. It complements impact-based scoring systems by capturing the degree of deviation from normal climate conditions.

22.2 NOAA-Based Severity Quantification

To complement the anomaly-based method, we implemented a scoring system based on the direct impacts of each event as recorded in the NOAA Storm Events database. Using fields such as deaths, injuries, and property or crop damage, we designed a multi-level rule-based and weighted system to classify event severity. This approach focuses on the real-world consequences of extreme events.

Severity Determination Criteria

The classification system consists of both rule-based thresholds and a weighted scoring scheme, structured as follows:

1. **High Severity (direct assignment)**:
   - At least one direct death (DEATHS_DIRECT > 0)
   - At least one direct injury (INJURIES_DIRECT > 0)
   - Property damage ≥ $200,000 (DAMAGE_PROPERTY_NUM)
   - Crop damage ≥ $150,000 (DAMAGE_CROPS_NUM)
2. **Moderate Severity (direct assignment)**:
   - Property damage ≥ $30,000
   - Crop damage ≥ $10,000
3. **Low-to-Moderate Severity (weighted scoring)**:
   For events not covered by the above conditions, we calculate a severity score using weighted contributions from property and crop losses:
   - **Property Damage Score** (weight 0.7):
     - 3 points: ≥ $50,000
     - 2 points: ≥ $20,000
     - 1 point: ≥ $3,000
     - 0 points: < $3,000
   - **Crop Damage Score** (weight 0.3):
     - 3 points: ≥ $50,000
     - 2 points: ≥ $10,000
     - 1 point: ≥ $3,000
     - 0 points: < $3,000
4. **Final Severity Label (based on weighted sum)**:
   - *High Severity*: total score ≥ 1.0
   - *Moderate Severity*: total score ≥ 0.4
   - *Low Severity*: total score < 0.4

Design Rationale

This scoring system reflects the following priorities:

- Human casualties are given the highest priority; the presence of any death or injury results in immediate classification as high severity.
- Significant economic damages are directly classified into moderate or high severity levels.
- For smaller-scale events, a fine-grained scoring system allows nuanced differentiation, with a higher weight placed on property damage (0.7) compared to crop damage (0.3).

22.3 Comparison of Severity Scores

To evaluate the consistency and divergence between the two severity quantification meth-ods, we conducted a comparative analysis across matched extreme weather events from the anomaly-based and NOAA-based datasets.

Event Matching Strategy

Events were matched using the following criteria:

- Identical event dates
- Same state of the US
- Approximate matching of event type descriptions using fuzzy string similarity

This matching procedure yielded:

- 578 records in the anomaly-based dataset
- 445 records in the NOAA dataset
- 578 successfully matched events (within time range)
- 0 unmatched events (outside time range)

Severity Score Comparison

Both scoring systems used the same three-level severity scale:

$$\text{low} = 1, \quad \text{medium} = 2, \quad \text{high} = 3$$

However, their distributions showed substantial differences:

- **Anomaly-Based Severity Distribution (578 events):**
  - High: 377 events (65.2%)
  - Medium: 67 events (11.6%)
  - Low: 134 events (23.2%)
- **NOAA-Based Severity Distribution (445 events):**
  - High: 25 events (5.6%)
  - Medium: 47 events (10.5%)
  - Low: 373 events (83.8%)

We computed the Pearson correlation coefficient between the severity scores from the two systems across the 578 matched events. The result was:

$$r = -0.020$$

This near-zero value indicates virtually no linear correlation between the two scoring methods. The slight negative trend further suggests disagreement between the two approaches in how they rank the same events.

Interpretation

The discrepancy highlights the fundamental difference in focus between the two systems:

- **NOAA-based method** prioritises human and economic impact, capturing the societal severity of events.

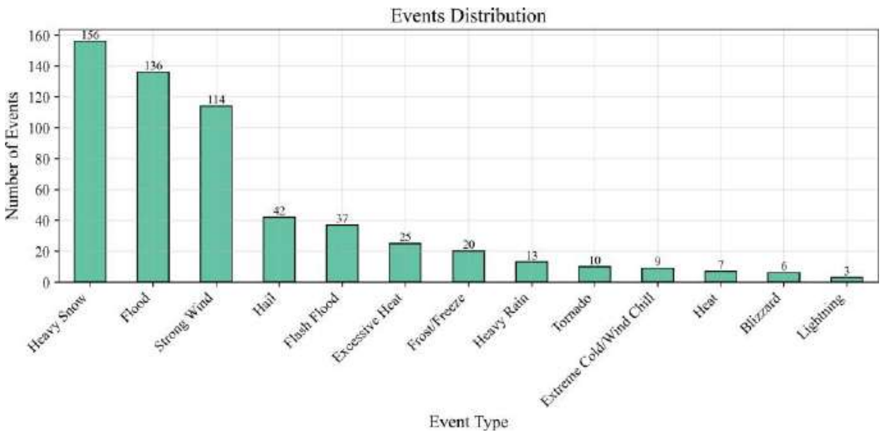


Fig. 43: Distribution of matched events by event type.

- **Anomaly-based method** reflects the meteorological intensity, focusing on the extent of deviation from climatological norms.

As a result, an event may be considered meteorologically extreme but socially inconsequential (e.g., high wind without damage), or vice versa (e.g., minor meteorological anomaly causing major local impact due to infrastructure vulnerability).

This comparison suggests that the two severity frameworks serve different but complementary roles. Depending on the application context (e.g., disaster planning vs. climate modelling), one may be more appropriate than the other, or both can be integrated to achieve a more comprehensive severity assessment.

22.4 Event–Anomaly Alignment Analysis

In this section, we evaluate the alignment between recorded extreme events and detected anomalies. We analyse the distribution of event types within the aligned data and explore how often different types of events co-occur with anomalies. A heatmap is constructed to visualise the frequency of anomaly detection across event types, providing insights into which categories are more consistently associated with abnormal meteorological conditions.

Event Type Distribution

Fig. 43 shows the distribution of matched events across different event types. The three most common categories were *Heavy Snow* (156 events), *Flood* (136 events), and *Strong Wind* (114 events). These events constitute the majority of the dataset and therefore dominate the overall alignment statistics.

Anomaly Frequency by Event Type

To further explore how different types of anomalies are associated with specific event types, we constructed a heatmap of anomaly frequency by event type (Fig. 44). The x-axis rep-resents event types, while the y-axis shows anomaly types (denoted by feature name and direction). The colour intensity reflects the number of times a specific anomaly was observed in conjunction with a particular event type.

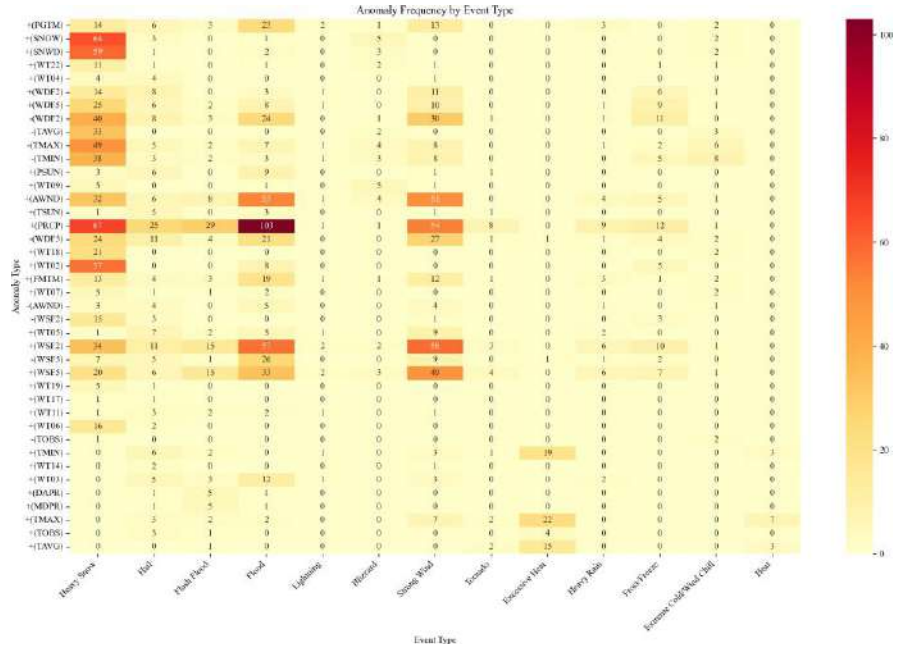


Fig. 44: Heatmap of anomaly frequency across different event types.

Several key patterns emerge from the heatmap:

- **Precipitation-related anomalies** (e.g., +PRCP, +SNOW, +SNWD) show strong associations with events like *Flood*, *Heavy Snow*, and *Flash Flood*.
- **Wind anomalies** (e.g., +WSF2, +WSF5, +AWND) are prevalent in *Strong Wind* and *Tornado* events.
- **Temperature-related anomalies** (e.g., +TMAX, -TMIN) are frequently aligned with *Excessive Heat*, *Frost/Freeze*, and *Extreme Cold/Wind Chill*.
- **Weather-type indicators** (e.g., +WT02, +WT03, etc.) are sparsely but clearly activated in several categories, often indicating presence of severe meteorological conditions.

Interpretation

The results indicate that many extreme events are accompanied by detectable meteorolog-ical anomalies. However, the nature and intensity of anomalies vary considerably by event type. While some event types, such as *Flood* and *Strong Wind*, are strongly linked with mul-tiple high-frequency anomalies, others like *Lightning* or *Heat* appear less frequently aligned, possibly due to the granularity of the anomaly detection method or data sparsity.

These findings provide evidence that anomaly-based signals can be informative for identifying or anticipating certain classes of extreme weather events, and could be integrated into event forecasting or early warning systems. Due to this motivation, we have established a multi-hazard early warning systems for agriculture and a copy of the manuscript is recorded in Part IX of this report.

23 Challenges and Limitations

Despite our achievements in extracting and analysing extreme weather event data, several challenges and limitations remain. We summarise them as follows:

1. **Lack of Ground Truth for Extreme Weather Event Data**
   There is currently no authoritative and comprehensive dataset available for extreme weather events, making it difficult to verify the accuracy and completeness of the collected data. In particular, the LLM-extracted datasets exhibit several sources of inaccuracies:
   - **Missing Identifications:** Some extreme events were not recognised by the language model during the extraction process.
   - **Extraction Errors:** Incorrect event start and end times, as well as misclassification of event types (e.g., storms misidentified as hail), were observed.
   - **False Identifications:** In certain cases, events were extracted that did not actually occur in the original source material.
2. **Lack of Ground Truth for Severity Scores**
   The absence of verified severity scores poses a major challenge for evaluating the effectiveness of the proposed severity quantification methods. Without access to ground truth labels, it is difficult to determine whether the model-generated severity scores accurately reflect the real-world impact of events. The availability of verified severity data would facilitate model calibration, validation, and more rigorous reliability assessments.

24 Part V Conclusion

In Part V, we developed a comprehensive framework for extreme event retrieval, severalty quantification, and validation of extreme weather event data. By integrating multiple data sources — including NIWA climate reports, industry annual reports, and daily climate observations — we constructed a comprehensive dataset for subsequent analysis. Severity scores were assigned to events through a combination of sentiment analysis and semantic similarity evaluation, providing a continuous measure of event intensity.

The alignment analysis between LLM-extracted events and observed anomalies revealed meaningful correspondences, though notable inconsistencies remain. These findings underscore both the potential and the limitations of current automated extraction and severity assessment methods. Addressing the lack of authoritative ground truth data will be critical for enhancing model calibration and improving reliability in future work.

Part VI
XAI Framework for Short-term Climate Effects

25 Part VI Introduction

Climate variability and extreme events pose significant challenges to agricultural productiv-ity [72, 73]. Predictive models capable of estimating crop yield under such uncertainty are vital for informed decision-making and risk mitigation. However, many traditional post-hoc Explainable Artificial Intelligence (XAI) methods, such as SHAP [74] and LIME [75], implicitly assume feature independence when attributing importance. In climate datasets, this assumption is often violated because numerous variables (e.g., temperature, humidity, radiation, and evapotranspiration) exhibit strong spatial and temporal correlations. As a result, these methods tend to distribute importance diffusely or inconsistently across correlated features, leading to unstable or misleading explanations that obscure true climate drivers.

In Part VI, we address this challenge by applying Convolutional Neural Network (CNN) with Temporal Attention mechanisms to predict crop yields and interpret the underlying temporal patterns. CNN is capable of exploiting local spatial or temporal dependencies through convolutional filters, which capture joint patterns among features and across time [76, 77]. In our journey of framework development, both 1-dimensoinal CNN (1D-CNN) and 2-dimensoinal CNN (2D-CNN) architectures were investigated. We began with 1D-CNN enhanced by a temporal attention mechanism to extract short-term dynamics and identify key growth phases that influence crop yield. Despite the 1D-CNN effectively models temporal dependencies within each feature, it cannot capture cross-feature relationships — which are crucial for understanding interactions among temperature, humidity, and other co-varying climate variables.

To overcome this limitation, we extended our modelling framework to 2D-CNN with Convolutional Block Attention Module (CBAM) [78]. Unlike traditional XAI methods, our framework can exploit both spatial and temporal relationships among correlated variables through convolutional filters, which extract local feature patterns and build hierarchical representations. By structuring climate features into a 2-dimensoinal feature-time space, our proposed 2D-CNN framework jointly learns intra-feature and inter-feature dependencies, producing more coherent, physically meaningful, and robust explanations.

Our investigation specifically focused on evaluating how extreme events influence yield by analysing their impact on temporal attention patterns and prediction performance. We integrated daily climate data and annotated extreme weather events to help identify critical time periods that influence yield outcomes while additional work can be done on feature-level attribution. Different from our earlier analyses which focus on kiwifruit's yield in New Zealand, our experiments in Part VI are centred on apple's yield prediction in California (CA) and Washington (WA) states in the United States, as well as grape's yield prediction in Coonawarra and Lake Cullulleraine regions of Australia. More information on the datasets can be found in Section 26. Our experiment results demonstrate competitive prediction accuracy while producing interpretable attention-based insights.

Table 14: Summary of data sources by region and crop type

Crop Type	Region	Climate Features	Extreme Event Types	Yield Years
Apple	California (CA)	22	9	2010–2023
	Washington (WA)	31	8	2010–2023
Grape	Coonawarra	72	5	2018–2024
	Lake Cullulleraine	82	5	2018–2024

26 Data Sources for Our Experiments

This study integrates three primary data types across multiple regions and crop types, as summarised in Table 14.

- **Climate Data:** Daily meteorological records collected across full growing seasons. The number of climate features varies by region: California (22 features), Washington (31 features), Coonawarra (72 features), and Lake Cullulleraine (82 features).
- **Extreme Events:** Manually or externally annotated extreme weather events, including strong winds, flooding, and extreme temperatures. Each region includes a distinct number of event types: California (9 types), Washington (8 types), Coonawarra (5 types), and Lake Cullulleraine (5 types).
- **Yield Records:** Ground-truth annual yield data are used as prediction targets. Apple yield records are available for California and Washington from 2010 to 2023, while grape yield records for Coonawarra and Lake Cullulleraine span from 2018 to 2024.

27 Framework #1: 1D-CNN with Temporal Attention (1D-CNN-TA)

Let us first introduce our Framework #1: Convolutional Neural Network (CNN) with Temporal Attention, named 1D-CNN-TA, aiming to extract short-term dynamics and identify key growth phases that influence yield.

27.1 Model Architecture

To capture both local temporal patterns and growing season-wide dynamics in climate sequences, we adopt a hybrid architecture combining a CNN with a temporal attention mechanism. CNN layers are effective in extracting short-term dependencies through localised filters, while the attention module enables the model to focus on the time steps that are most influential to the yield. This design allows the model to learn both fine-grained local trends and high-level temporal structure. In addition, the learned attention weights provide interpretability by highlighting key periods throughout the growing season.

The model processes multivariate daily time series as input, structured as a $C \times T$ matrix, where C is the number of input features (e.g., 22 for California climate data), and T is the number of time steps (typically 365 days). Three input configurations are evaluated: climate-only, extreme-events-only, and a combined setup.

As illustrated in Fig. 45, the input is first passed through a two-layer 1D-CNN backbone:

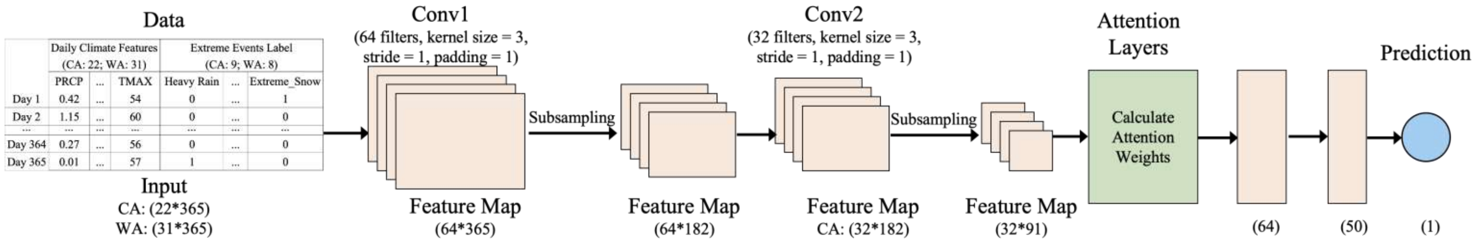


Fig. 45: Overview of our proposed Framework #1 (1D-CNN with Temporal Attention) for yield prediction.

- **Conv1:** Applies 64 filters with a kernel size of 3 over all input channels. Zero-padding preserves the temporal dimension, resulting in an output of shape $64 \times T$. The choice of 64 filters balances expressiveness and computational efficiency, providing sufficient capacity to capture diverse local temporal patterns without incurring excessive cost.
- **Conv2:** Applies 32 filters with the same kernel size over the Conv1 output, yielding a compact representation of shape $32 \times T$. Reducing the number of channels from 64 to 32 helps compress feature representations, improving generalisation while reducing model complexity and overfitting risk.

The attention module then operates on the $32 \times T$ feature map. It computes a scalar attention weight for each time step, normalised via a softmax function to form a probability distribution over time. These weights are used to compute a weighted sum across all time steps, producing a fixed-size context vector.

This context vector is finally passed through a fully connected layer to generate the yield prediction. The overall model is lightweight, end-to-end trainable, and provides temporal interpretability through visualisable attention distributions.

27.2 Input Settings and Experimental Design

To assess how different input sources affect yield prediction accuracy and model interpretability, we evaluate three input configurations:

- **Exp. 1 – Climate Only:** Daily multivariate climate features, excluding any extreme event information. Fig. 46 provides an example of the input feature set used for apple yield prediction in CA and WA.
- **Exp. 2 – Climate + Extreme Events:** Climate features concatenated with binary indicators for annotated extreme events, forming an enriched time series. This setting tests whether event labels enhance predictive power and interpretability.
- **Exp. 3 – Extreme Events Only:** Daily binary indicators representing the occurrence of extreme events, without climate data. Fig. 47 shows the types of extreme events used for CA and WA.

All configurations use the same 1D-CNN-TA model architecture described in Section 27.1. The model takes a full-season daily sequence (typically 365 days) and predicts the final crop yield. Each experiment is conducted separately for different crop types and regions:

- **Apple Yield Prediction:** California (CA) and Washington (WA) states in the US, using data from 2010 to 2023.
- **Grape Yield Prediction:** Coonawarra and Lake Cullulleraine regions in Australia, using data from 2018 to 2024.

For each crop-region pair, the three input configurations are evaluated independently. Training and testing follow a leave-one-year-out or cross-year strategy, with repeated runs to ensure result stability.

Model performance is evaluated using:

- **Prediction Accuracy:** Compared to ground-truth yields.
- **Error Reduction:** Relative to the long-term average yield baseline.

Abbreviation	Full Description	Location	Abbreviation	Full Description	Location	Abbreviation	Full Description	Location
AWND	Average daily wind speed	CA, WA	FMTM	Time of fastest mile or fastest 1-minute wind	CA, WA	PGTM	Peak gust time	CA, WA
PRCP	Precipitation	CA, WA	PSUN	Daily percent of possible sunshine (percent)	WA	SNOW	Snowfall	WA
SNWD	Snow depth	WA	TAVG	Average temperature	WA	TMAX	Maximum temperature	CA, WA
TMIN	Minimum temperature	CA, WA	WDF2	Direction of fastest 2-minute wind (degrees)	CA, WA	WDF5	Direction of fastest 5-second wind (degrees)	CA, WA
WSF2	Fastest 2-minute wind speed	CA, WA	WSF5	Fastest 5-second wind speed	CA, WA	WT01	Fog, ice fog, or freezing fog (may include heavy fog)	CA, WA
WT02	Heavy fog or heaving freezing fog (not always distinguished from fog)	CA, WA	WT03	Thunder	CA, WA	WT04	Ice pellets, sleet, snow pellets, or small hail	WA
WT05	Hail (may include small hail)	CA, WA	WT06	Glaze or rime	WA	WT07	Dust, volcanic ash, blowing dust, blowing sand, or blowing obstruction	CA, WA
WT08	Smoke or haze	CA, WA	WT09	Blowing or drifting snow	CA, WA	WT10	Tornado, waterspout, or funnel cloud	CA
WT11	High or damaging winds	WA	WT13	Mist	CA, WA	WT16	Rain (may include freezing rain, drizzle, and freezing drizzle)	CA, WA
WT17	Freezing rain	WA	WT18	Snow, snow pellets, snow grains, or ice crystals	WA	WT19	Unknown source of precipitation	WA
WT21	Ground fog	CA, WA	WT22	Ice fog or freezing fog	CA, WA			

Fig. 46: Climate features used for apple yield prediction in CA and WA.

Extreme Events	Location	Extreme Events	Location	Extreme Events	Location
Flood	CA, WA	Hail	CA, WA	Heat	CA, WA
Heavy Rain	CA, WA	Strong Wind	CA, WA	Extreme_Snow	CA, WA
Extreme_Cold	CA, WA	Frost	CA, WA	Lightning	CA

Fig. 47: Extreme event indicators used for apple yield prediction in CA and WA.

- **Attention Interpretability:** Analysed via the temporal distribution of attention weights and their alignment with known seasonal phases or extreme events.

27.3 Experiments on the Apple Data in the US

27.3.1 California (CA) State

Model Performance Summary

We evaluated all three input configurations on the California apple yield dataset from 2010 to 2023. Fig. 48 illustrates the temporal attention patterns across all three input configura-tions. Exp. 1 (climate only) focuses attention on mid- to late-season periods, such as Days 59–72, 140–147, and 348–363, aligning with key stages in the apple growth cycle. Exp. 2 (climate + extreme events) assigns the same number of high-attention days (73), but redis-tributes them more broadly, including early-season windows like Days 13–19 and 24–35, as well as mid-season periods (e.g., Days 80–103 and 159–168). This indicates that extreme event information helps the model capture more temporally diverse signals.

In contrast, Exp. 3 (extreme events only) yields no high-attention time steps, despite a comparable overall accuracy. This suggests that extreme event indicators alone lack the temporal density or granularity to guide attention in the California context.

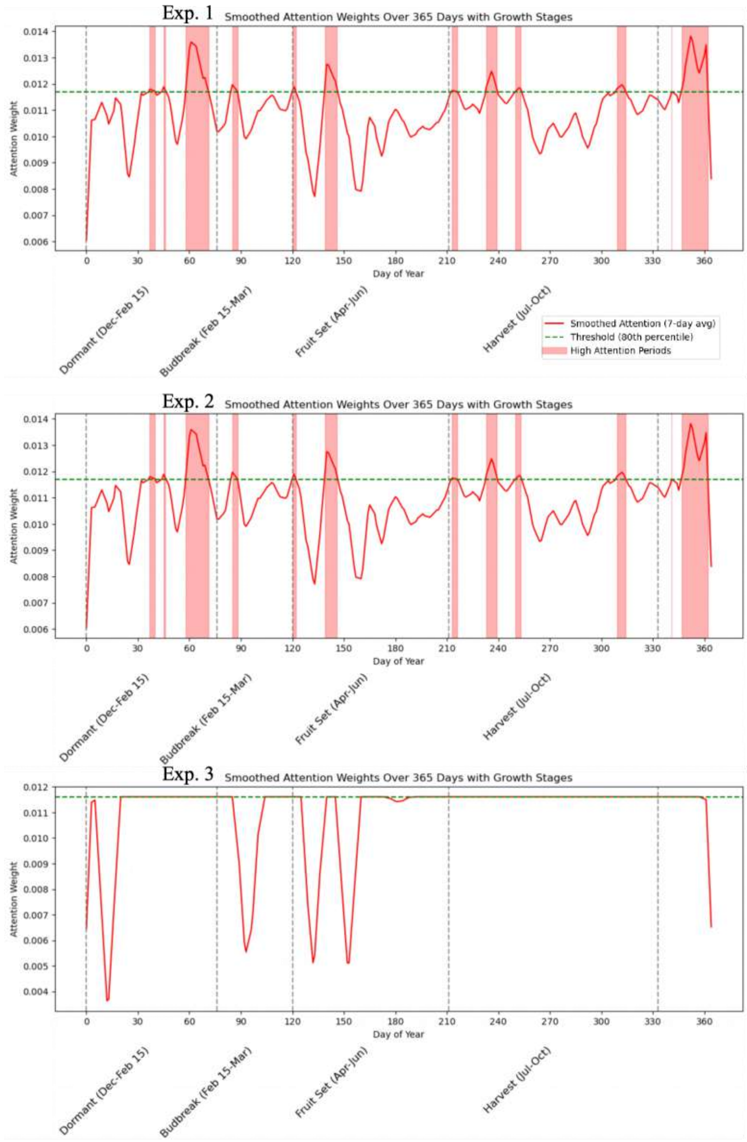


Fig. 48: Temporal attention patterns for California apple yield prediction across three input configurations. Red shading indicates high-attention periods.

Table 15 summaries the accuracy and attention characteristics of all configurations. While Exp. 2 achieves the highest accuracy and temporal coverage, the improvements over Exp. 1 are modest in prediction terms but more pronounced in interpretability. The long-term average baseline (Exp. 4) yields the lowest accuracy and provides no temporal insight.

Further analysis in Fig. 49 shows the overall distribution of attention weights and overlapping high-attention time periods. Overlap between Exp. 1 and Exp. 2 occurs at Days 131–136,

Table 15: Accuracy and high-attention periods for California apple yield under different input configurations.

Experiment	Accuracy	High Attention Points	High Attention Periods (Day-of-Year)
Exp. 1	87.65%	73 days	38–41, 46–47, 59–72, 86–89, 121–123, 140–147, 214–217, 234–240, 251–254, 310–315, 342, 348–363
Exp. 2	87.92%	73 days	13–19, 24–35, 80–103, 131–136, 159–168, 209–214, 272–274, 294–297, 354
Exp. 3	87.39%	0 days	NA
Long-Term Avg.	86.61%	NA	NA

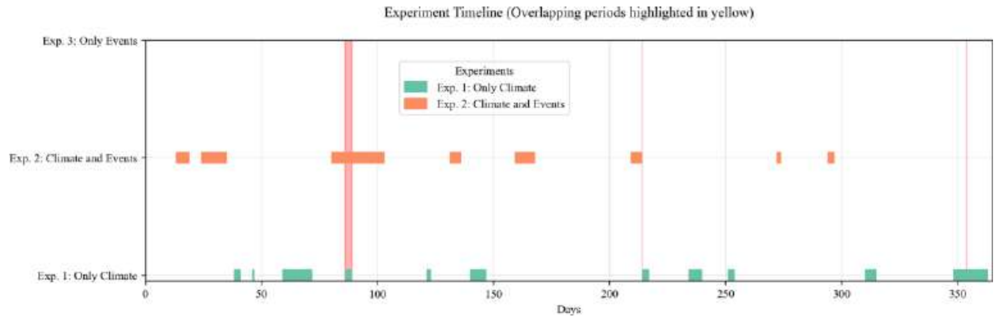


Fig. 49: Attention weight statistics and temporal overlap analysis between Exp. 1 and Exp. 2 in California.

209–214, and 294–297, indicating some consistency across settings. However, Exp. 2 reveals unique early- and mid-season windows not captured by climate-only input, demonstrating its added interpretive value.

Year-Specific Error Reduction

We further examine prediction accuracy in 2022 and 2023, two contrasting years in terms of yield. Table 16 reports the predicted yields, ground-truth yields, average yield baseline, and resulting error reduction.

The results show that both Exp. 1 and Exp. 2 significantly outperform the average base-line in both years. The margin is especially pronounced in 2022, a year with lower yield and more favourable model fit. In 2023, although the ground-truth yield was higher, the model still reduced error by more than 8%, showing robustness across yield scenarios.

Year-specific Performance: 2022 Attention and Accuracy Analysis

In 2022, all three experimental configurations (Exp. 1–3) achieved high accuracy for California apple yield prediction, each exceeding 99%. Exp. 1 (climate only) produced the highest

Table 16: Prediction performance on CA apple yield (2022–2023).

Year	Input Setting	Ground Truth	Prediction	Error Reduction
	Average Yield	18,000	17,675	NA
2022	Exp. 1	18,000	18119.27	63.30%
	Exp. 2	18,000	18119.12	44.89%
	Exp. 3	18,000	17907.7	71.60%
	Average Yield	23,500	17,675	NA
2023	Exp. 1	23,500	18076.09	6.89%
	Exp. 2	23,500	18166.01	8.43%
	Exp. 3	23,500	17907.7	3.99%

accuracy at 99.34%, followed by Exp. 3 (extreme events only) at 99.49%, and Exp. 2 (climate + extreme events) at 99.00%. The long-term average baseline (Exp. 4) yielded a lower accuracy of 98.00%.

Despite comparable accuracy, the attention distributions differ notably across models. Exp. 1 and Exp. 2 each identified 73 high-attention days, but with distinct temporal pat-terns. Exp. 1 concentrated attention on mid- to late-season periods (e.g., Days 128–144 and 151–156), aligning with known strong wind events. In contrast, Exp. 2 emphasised a broader temporal span, including early- and mid-season intervals (e.g., Days 48–54, 131–135, and 153–154), again aligning with extreme wind occurrences.

Interestingly, Exp. 3 (extreme events only) also yielded meaningful temporal interpretability, with 68 high-attention days. Many of these corresponded directly to annotated extreme events, such as Days 4–20 (flood and strong wind), and Days 127–160 (multiple strong wind periods). This suggests that event-only inputs, although lacking continuous climate trends, still carry sufficient signal for attention-based temporal modelling.

Table 17 summarises the accuracy and high-attention periods across all experimental settings. The majority of high-attention periods coincide with known climate anomalies, particularly strong wind events and temperature extremes, reinforcing the model's capacity to associate adverse conditions with yield variability – especially in climatically volatile years like 2022.

Comparison to Long-Term Average

To assess the value of using full-resolution daily inputs versus coarser temporal granularity, we compare attention-based performance across different input formats: daily data, 5-day average, 10-day average, and 20-day average. All settings use the same model architecture with combined climate and extreme event inputs.

As shown in Table 18, the full 365-day input achieves the highest attention resolution, with 73 distinct attention points spread across key seasonal phases. However, performance remains competitive even under coarser resolutions. Notably, the 20-day average yields the highest prediction accuracy (88.14%), despite resulting in far fewer high-attention points (only 4). This suggests a trade-off between interpretability and predictive accuracy: finer granularity offers richer insight into critical periods, while coarser aggregation may enhance robustness.

Fig. 50 visualises the overlapping attention periods identified under different input granularities. Despite differences in format, overlapping high-attention segments (highlighted in red) are consistently concentrated in biologically meaningful intervals such as Days 40–110

Table 17: Accuracy and high-attention periods with extreme event alignment for California apple yield prediction in 2022.

Experiment	Accuracy	High Attention Days	High Attention Periods & Extreme Events
Exp. 1	99.34%	73 days	Days 16–18, 52–56, 65, 100–102, **128–144**: Strong Wind; **151–156**: Strong Wind; 159–176, 193–199, 230–231, 258–268
Exp. 2	99.00%	73 days	Days 48–54, 113–118, **131–135**: Strong Wind; **153–154**: Strong Wind; 163–165, 177–183, 209–211, 232–236, 262–296
Exp. 3	99.49%	68 days	**Days 4–20**: Flood, Strong Wind; **87–104**: Strong Wind; **127–140**: Strong Wind; **147–160**: Strong Wind; 188–192
Long-Term Avg.	98.00%	NA	NA

Table 18: Comparison of attention patterns under different temporal granularities (CA, Exp. 2: Climate + Extreme Events).

Input Days	Accuracy	High Attention Days
365 Days (Daily)	87.92%	73 points: Days 13–19, 24–35, 80–103, 131–136, 159–168, 209–214, 272–274, 294–297, 354
5-day Average	87.63%	15 points (8–22): Days 40–110
10-day Average	87.96%	8 points (4–6, 14–15, 32–34): Days 40–60, 140–150, 320–340
20-day Average	88.14%	4 points (6–8, 16): Days 120–160, 320

and 120–160, indicating that even low-resolution inputs can preserve signal relevant to yield formation.

27.3.2 Washington (WA) State

Model Performance Summary

In Washington (WA), all three experimental configurations (Exp. 1–3) achieved competitive accuracy for apple yield prediction, with all results above 88%. Exp. 3 (extreme events only) yields the highest accuracy at 90.20%, followed by Exp. 2 (climate + events) at 89.24%, and Exp. 1 (climate only) at 88.80%. The long-term average baseline (Exp. 4) also reaches 88.80%, indicating limited improvement from climate-only features in this region.

Although similar in prediction accuracy, the three experiments exhibit distinct attention distributions. Exp. 1 focuses on mid- to late-season periods such as Days 151–161, 213–224, and 247–250, aligning with key physiological stages of fruit development and maturation. Exp. 2 produces a wider temporal focus, spanning early (Days 9–12), mid (Days 56–76), and late-season periods (Days 266–274), suggesting that the addition of extreme events enhances temporal diversity in attention.

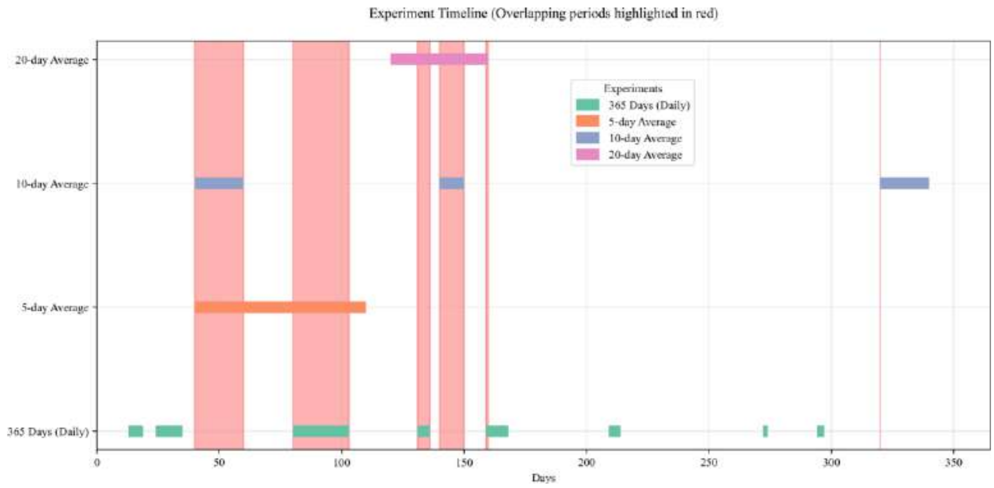


Fig. 50: Comparison of attention periods under varying input granularities. Overlapping high-attention intervals across formats are highlighted in red.

Table 19: Accuracy and high-attention periods for Washington apple yield under different input configurations.

Experiment	**Accuracy**	**High Attention Points**	**High Attention Periods (Day-of-Year)**
Exp. 1	88.80%	73 days	12–19, 52–56, 74–77, 125–133, 151–161, 188–189, 199–207, 213–224, 238–242, 247–250, 282, 291–293
Exp. 2	89.24%	73 days	9–12, 20–22, 33–50, 56–76, 83–92, 102, 115–118, 166, 178, 266–274, 293
Exp. 3	90.20%	49 days	4–8, 43–46, 259–280, 283–300
Long-Term Avg.	88.80%	NA	NA

Exp. 3 highlights 49 high-attention days using only extreme event indicators. Despite the absence of granular climate variables, this model still captures relevant late-season windows (Days 259–300), along with early-season intervals (Days 4–8), indicating the strong standalone utility of extreme event labels in WA.

Fig. 51 illustrates the attention weight distributions across the three configurations, while Fig. 52 highlights overlapping periods, particularly during the harvest window. Table 19 summarises accuracy and high-attention periods across all experiments.

Year-Specific Error Reduction

Table 20 summarises the prediction performance of all input configurations for Washington (WA) apple yield in 2022 and 2023. While all three experimental settings achieved comparable

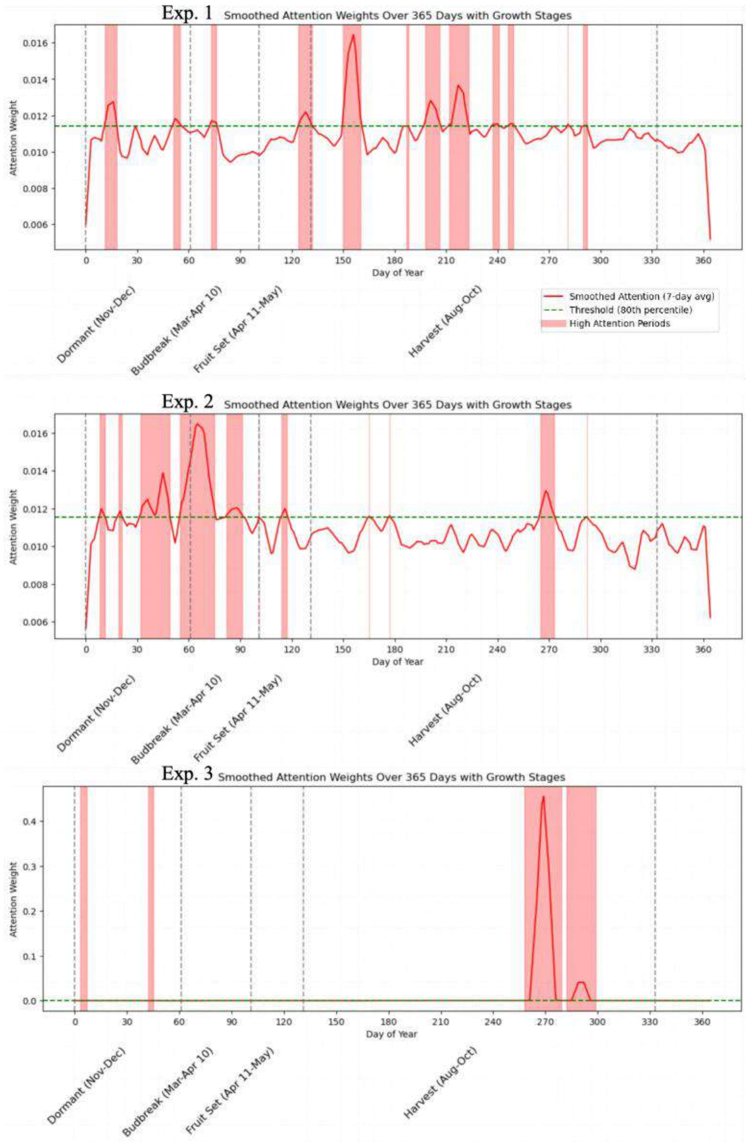


Fig. 51: Temporal attention distributions for WA across Exp. 1–3. Each configuration highlights different seasonal stages, with overlapping focus around the harvest phase.

accuracy in 2022, their error reduction over the long-term average baseline differs noticeably across years.

In 2022, Exp. 1 (climate only) achieved the largest error reduction (5.96%) compared to the average yield baseline, followed closely by Exp. 3 (events only) at 5.78%. Exp. 2 (climate + events) yielded minimal error reduction (1.01%), suggesting that in this year, climate

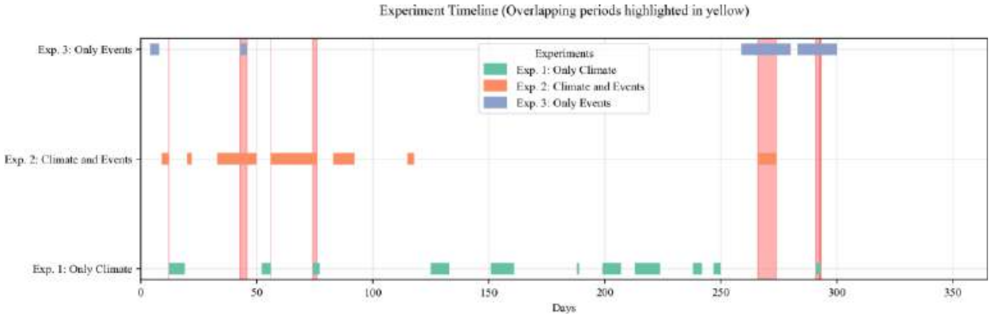


Fig. 52: Overlap of high-attention days across WA experiments. Late-season windows (Days 266–300) show consistent alignment across configurations.

Table 20: Prediction performance on WA apple yield (2022–2023).

Year	Input Setting	Ground Truth	Prediction	Error Reduction
2022	Average Yield	35,500	41,150	NA
	Exp. 1	35,500	40,813.28	5.96%
	Exp. 2	35,500	41,093.17	1.01%
	Exp. 3	35,500	40,823.26	5.78%
2023	Average Yield	44,000	41,150	NA
	Exp. 1	44,000	40,778.26	13.04%
	Exp. 2	44,000	41,356.81	7.26%
	Exp. 3	44,000	42,730.58	55.46%

features alone were sufficient for accurate predictions, and adding event data provided limited incremental benefit.

In contrast, 2023 presents a different pattern. Exp. 3 demonstrated a significant improvement with a 55.46% error reduction, substantially outperforming both Exp. 1 (13.04%) and Exp. 2 (7.26%). This suggests that extreme events played a more prominent role in shaping yield outcomes in 2023, reinforcing the importance of capturing these disruptions explicitly.

Year-specific Performance: 2022 Attention and Accuracy Analysis

In 2022, all three experimental configurations (Exp. 1–3) achieved comparable prediction accuracy for Washington apple yield. Exp. 1 (climate only) reached 85.03%, Exp. 2 (climate + extreme events) was slightly lower at 84.09%, and Exp. 3 (extreme events only) matched closely at 85.00%. The long-term average baseline (Exp. 4) recorded 84.08%.

Despite the similarity in accuracy, the attention patterns reveal meaningful differences. Both Exp. 1 and Exp. 2 produced 73 high-attention days, but their temporal distributions vary:

- **Exp. 1** concentrated attention on Days 49–59 and 263–267, aligning with documented extreme snow and heat events. - **Exp. 2** demonstrated broader temporal coverage, highlighting Days 53–57 and 159–171, again associated with snow anomalies.

Table 21: Accuracy and high-attention periods with extreme event alignment for Washington apple yield prediction in 2022.

Experiment	Accuracy	High Attention Days	High Attention Periods & Extreme Events
Exp. 1	85.03%	73 days	Days 49–59: **Extreme Snow**; 73–82, 133–134, 201–208, 217–218, 234, 244–256, **263–267**: Heat; 276–281, 296–307, 360–362
Exp. 2	84.09%	73 days	Days 29–31, 40–45, **53–57**: Extreme Snow; 85–86, 136–138, **159–171**: Extreme Snow; 179–187, 190–193, 201–207, 245–255, **273–278**: Heat; 296–297, 361–362
Exp. 3	85.00%	18 days	Days 4–8, **147–155**: Extreme Snow; 359–362
Long-Term Avg.	84.08%	NA	NA

Interestingly, even though Exp. 3 uses only binary event indicators, it identified 18 high-attention days, notably Days 147–155 (extreme snow) and 359–362 (year-end anomalies). This suggests that event-based signals alone can provide focused and interpretable information in years with dominant weather disturbances.

Table 21 summarizes the alignment between attention and extreme events across all experiments.

Comparison to Long-Term Average

To assess how attention patterns evolve across temporal granularities, we compare Experiment 2 (Climate + Extreme Events) using daily, 5-day, 10-day, and 20-day aggregated inputs. While the full 365-day setting achieves the highest accuracy (89.24%), the more aggregated versions—despite lower accuracy—reveal complementary insights into seasonal trends.

As shown in Table 22, all configurations consistently highlight periods within early and late growing seasons. The 365-day model attends to a wide span of high-risk windows, espe-cially Days 33–50 and 266–274, which are linked to early spring variability and pre-harvest stress.

The 5-day average retains strong alignment with early-season volatility (Days 30–50) and adds mid-season peaks (Days 90–135). In contrast, the 10-day and 20-day models shift attention toward the late-season period (Days 270–340 and 260–320, respectively), reinforcing the importance of post-flowering and harvest phases.

Fig. 53 visualises these attention timelines, with overlapping regions highlighted in red. Notably, the aggregation-based models detect complementary periods not captured by the daily setting alone, demonstrating how different temporal granularities can enhance interpretability by surfacing multi-phase stress events.

27.3.3 Summary of Findings & Insights

Our cross-state comparison reveals several consistent patterns and region-specific insights:

- **Consistent Attention to Late-Season Periods**
  Across both California (CA) and Washington (WA), the models consistently assign high attention to late-season time windows, particularly around Days 260–300.

Table 22: Comparison of attention patterns under different temporal granularities (WA, Exp. 2: Climate + Extreme Events).

Input Days	Accuracy	High Attention Days
365 Days (Daily)	89.24%	73 points: Days 9–12, 20–22, 33–50, 56–76, 83–92, 102, 115–118, 166, 178, 266–274, 293
5-day Average	88.77%	15 points: 6–10, 18–27 Days: 30–50, 90–135
10-day Average	88.64%	8 points: 27–34 Days: 270–340
20-day Average	88.51%	4 points: 13–16 Days: 260–320

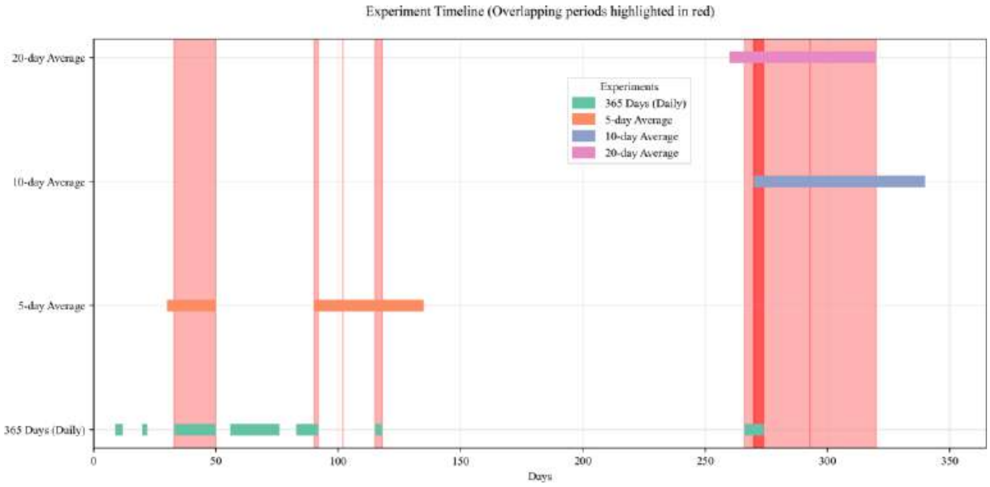


Fig. 53: Attention timeline comparison across different temporal granularities in WA (Exp. 2). Overlapping high-attention windows are shaded in red.

These periods align with known harvest stages, suggesting that the learned attention weights are reflective of biologically meaningful growth phases. This consistency underscores the model's ability to identify and generalise critical temporal dynamics across different growing environments.

- **Extreme Events Enhance Temporal Coverage**
  The inclusion of extreme event indicators (Exp. 2) leads to a broader temporal spread of attention in both states. Compared to climate-only inputs (Exp. 1), the hybrid configuration (climate + extreme events) captures earlier and more diverse phases of the growing season. This suggests that event annotations contribute complementary temporal signals, improving both interpretability and potential robustness to anomalies.
- **Region-Specific Utility of Extreme Events**
  While climate features alone provide strong baseline predictions across both regions,

Table 23: Attention weight statistics and accuracy for Coonawarra grape yield prediction (All Sites).

Experiment	Accuracy	Mean Attention Weight
Exp. 1	93.25%	0.0137
Exp. 2	93.01%	0.0137
Exp. 3	92.41%	0.0137

the standalone utility of extreme events (Exp. 3) varies by state. In WA, models trained solely on extreme event indicators achieve competitive accuracy and identify meaningful attention periods, highlighting the high informativeness of annotated events in this region. In contrast, in CA, the extreme-events-only model yields reasonable accuracy but fails to identify distinct attention points, suggesting that event data in CA may be less granular or less correlated with yield outcomes when used in isolation.

These findings collectively support the effectiveness of integrating attention-based models across diverse regions while emphasising the need for region-specific evaluation of input signal relevance.

27.4 Experiments on the Grape Data in Australia

This section evaluates whether the proposed 1D-CNN-TA framework can generalise effectively to grape yield prediction. Unlike apple datasets, which include fine-grained daily climate data, the grape datasets are constructed with coarser seasonal climate summaries. Our goal is to assess whether the same model architecture — without any structural modification — can still deliver accurate yield predictions across different regions and vineyard sites.

The experimental results demonstrate that the framework remains robust across the grape datasets, achieving consistently high accuracy across all sites and configurations. However, due to the limited temporal granularity of the input (seasonal-level rather than daily), the attention mechanism fails to identify meaningful or discriminative time points. This limitation highlights the importance of fine-grained climate inputs in enabling temporal interpretability. If full-resolution daily climate data were available for the grape regions, we expect the attention module would yield more informative attributions and further enhance model interpretability.

27.4.1 Coonawarra Region

All Sites

Table 23 summarises the model performance on grape yield prediction using climate and extreme event data from all Coonawarra stations combined. All three configurations achieve accuracy above 92%, with the climate-only model slightly outperforming the others. Because the grape datasets are provided at a seasonal (rather than daily) temporal resolution, the input sequences contain repeated values across all time steps. Consequently, the attention mechanism produces nearly uniform weights over time, and we report only the mean attention weight rather than specific high-attention days, as the latter would not correspond to meaningful temporal patterns.

Table 24: Attention weight statistics and accuracy for Coonawarra Penfolds site.

Experiment	Accuracy	Mean Attention Weight
Exp. 1	93.20%	0.0137
Exp. 2	92.63%	0.0137
Exp. 3	92.21%	0.0137

Table 25: Attention weight statistics and accuracy for Coonawarra O'Deans site.

Experiment	Accuracy	Mean Attention Weight
Exp. 1	93.39%	0.0137
Exp. 2	93.29%	0.0137
Exp. 3	91.97%	0.0137

Table 26: Attention weight statistics and accuracy for Coonawarra Ward site.

Experiment	Accuracy	Mean Attention Weight
Exp. 1	93.37%	0.0137
Exp. 2	92.71%	0.0137
Exp. 3	91.94%	0.0137

Penfolds

As shown in Table 24, performance trends at the Penfolds site are similar to the aggregated results. The climate-only configuration achieves the highest accuracy, while extreme events alone still yield over 92% accuracy, confirming their complementary predictive value.

O'Deans

At the O'Deans site (Table 25), model performance remains consistent with other locations. Climate-only and climate + events settings both exceed 93% accuracy. All models show stable attention behaviour.

Ward

Results from the Ward site (Table 26) also follow similar patterns, with accuracy across all configurations above 91.9%. The consistency in attention behaviour across experiments supports the model's robustness at the site level.

Table 27: Attention weight statistics and accuracy for Lake Cullulleraine grape yield prediction.

Experiment	Accuracy	Mean Attention Weight
Exp. 1	93.02%	0.0137
Exp. 2	92.52%	0.0137
Exp. 3	92.75%	0.0137

27.4.2 Lake Cullulleraine Region

Table 27 presents the prediction results for Lake Cullulleraine. Model accuracy across all settings exceeds 92.5%, with nearly identical attention distributions. These results confirm that both climate and event features contribute effectively to yield estimation in this region.

27.4.3 Summary of Findings & Insights

Across all vineyard sites in Coonawarra and Lake Cullulleraine, the proposed 1D-CNN-TA model achieves consistently high accuracy, exceeding 92% under all experimental settings. These results demonstrate that the framework generalises well to grape yield prediction despite the shift from daily to seasonal climate inputs.

However, because the grape datasets are constructed at a seasonal rather than daily temporal resolution, the input sequences contain repeated values across all time steps. As a result, the attention mechanism produces uniform temporal weights and cannot identify meaning-ful high-attention days. This confirms that while the model remains robust for prediction, fine-grained daily climate data are essential for deriving interpretable temporal attention patterns.

Overall, the grape experiments highlight a clear distinction: the architecture is transferable and effective for yield estimation across different regions, but the interpretability benefits of the attention module depend critically on the availability of daily climate information.

28 Framework # 2: 2D-CNN with CBAM (2D-CNN-CBAM)

In this section, we will introduced our improved framework. In contrast to traditional XAI methods and 1D-CNN, which treat features independently, the proposed 2D-CNN architecture explicitly models feature inter-dependencies through convolution across the feature dimension. This design alleviates the attribution diffusion problem in correlated climate variables and provides more stable, physically interpretable explanations of yield variability.

28.1 Model Architecture

The proposed model is a 2-dimensional Convolutional Neural Network (2D-CNN) enhanced with Convolutional Block Attention Module (CBAM) [78]. This design is aimed at jointly modelling temporal patterns and cross-feature interactions in climate data, while providing interpretable attention maps that highlight critical feature-time regions. The overall framework is illustrated in Fig. 54.

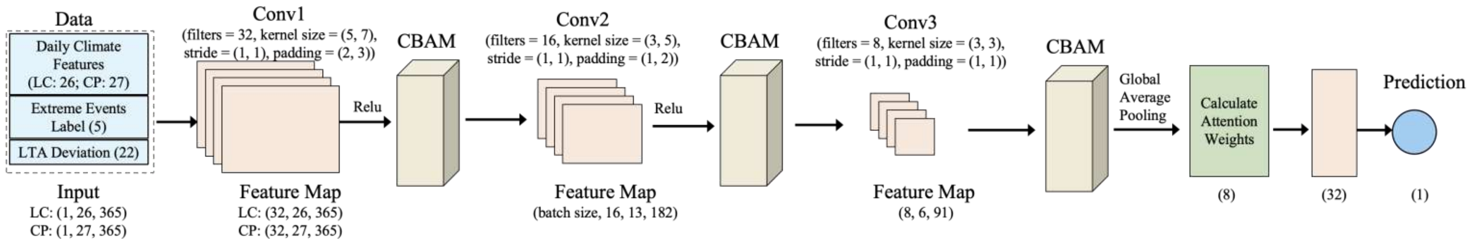


Fig. 54: Overview of our Proposed Framework #2 (2D-CNN with CBAM) – three convolutional blocks, each integrated with a CBAM. Spatial and channel attention maps are fused across layers to provide interpretable feature-time importance visualisation.

Input Representation

The input consists of daily climate records with 26 features over 365 days. Each sample is structured into a matrix of size 26 × 365, where the vertical axis corresponds to climate features (e.g., maximum temperature, humidity, etc.) and the horizontal axis corresponds to time (days of the year). This layout enables convolutional filters to jointly capture temporal trends and cross-feature interactions.

Convolutional Blocks with CBAM

The network comprises three sequential convolutional blocks. Each block performs 2-dimensional convolution, batch normalisation, ReLU activation, max pooling, and dropout for regularisation. After every convolutional layer, a CBAM module is inserted to refine the feature maps. CBAM consists of two complementary attention mechanisms:

- **Channel Attention (CA):** Learns the relative importance of intermediate convolutional filters, highlighting which latent representations (e.g., temperature trends, rainfall anomalies) contribute most to yield prediction. Outputs include $ca1$, $ca2$, and $ca3$ at different depths.
- **Spatial Attention (SA):** Generates a 2-dimensoinal attention mask over the feature-time plane (26 × 365), highlighting which climate variables are most critical at specific periods. Outputs include $sa1$, $sa2$, and $sa3$.

Fused Feature–Time Attention Map

Spatial attention maps from all three CBAM blocks are upsampled to the input resolution and averaged, producing a fused feature-time attention map. This map directly indicates which climate features and time periods drive model predictions, thereby enhancing interpretability.

Rationale for Combining Channel and Spatial Attention

Channel attention alone identifies the most informative latent representations, while spatial attention alone pinpoints the most relevant feature-time locations. By integrating both within CBAM, the model benefits from complementary strengths: CA selects the most useful learned filters, and SA emphasises critical feature-time regions. Together, they improve prediction accuracy and provide stable, interpretable climate-yield explanations.

28.2 Hierarchical Clustering with Optimal Leaf Ordering (OLO)

In CNN, the order of input features directly influences the convolutional operations, which are sensitive to the spatial adjacency of features. An arbitrary or unstable feature order can, therefore, lead to inconsistent outputs and performance variations. To mitigate this feature order dependency, we adopt a hierarchical clustering strategy combined with Optimal Leaf Ordering (OLO) [79, 80], where optimal leaf ordering[12] in Python library Scipy was used. This approach ensures that semantically similar features are arranged adjacently and that the ordering remains stable throughout training. Keeping the feature order fixed until a deliberate refresh offers a practical compromise between stability and adaptability. Moreover, reordering events may themselves serve as informative indicators of underlying shifts in climate data, potentially revealing useful signals for further analysis.

[12] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.optimal_leaf_ordering.html

Clustering Method

We enable feature ordering by setting ENABLE_FEATURE_ORDERING = True. The specific clustering method is determined via FEATURE_ORDERING_METHOD, where we evaluated four linkage options: *ward*, *complete*, *average*, and *single*. Among these, the *average* method was selected for its balanced trade-off between correlation strength and cluster flexibility. The similarity metric is controlled by FEATURE_SIMILARITY_METRIC. We compared *Pearson*, *cosine*, *Euclidean*, and *correlation* measures, and identified *Pearson correlation* as the most effective metric for our climate feature data.

With Pearson correlation fixed, the differences across linkage methods can be summarised as follows: - **Ward:** prioritises grouping features with higher absolute correlations, producing balanced clusters. - **Complete:** ensures all features within a cluster maintain consistently high correlations. - **Average:** maintains high overall within-cluster correlation but is less restrictive than complete linkage. - **Single:** may merge clusters based on a single highly correlated pair, often resulting in chaining effects.

Implementation

The procedure consists of the following steps:

1. **Feature similarity computation:** compute pairwise correlations between features across all samples and time steps, then convert similarities into a distance matrix.
2. **Hierarchical clustering:** apply agglomerative clustering with the selected linkage method.
3. **Optimal Leaf Ordering:** use scipy.cluster.hierarchy.optimal_leaf_ordering to reorder the dendrogram leaves by minimising the sum of pairwise distance between adjacent features, where pairwise distances are derived from the correlation matrix. This optimisation ensures that highly correlated features are positioned next to each other in the final order.
4. **Feature reordering:** update both training and test sets to match the derived optimal feature order, ensuring consistent alignment across experiments.

Illustration

Fig. 55 and Fig. 56 present a comparison of feature dendrograms before and after applying OLO. Fig. 55 shows the dendrogram with an arbitrary feature order, where adjacent leaves do not necessarily correspond to highly correlated features. In contrast, Fig. 56 displays the reordered dendrogram after OLO, where strongly correlated features are placed adjacently, resulting in more coherent clusters. This comparison clearly demonstrates how OLO improves spatial coherence and stabilises feature ordering.

Benefits

This hierarchical clustering with OLO yields several advantages:

- **Improved stability:** feature order remains consistent across runs, reducing variance in model performance.
- **Enhanced spatial coherence:** semantically related features are positioned adjacently, strengthening the CNN's capacity to extract meaningful patterns.
- **Interpretability:** clustering groups climate features with similar dynamics, facilitating more stable and interpretable attention maps.

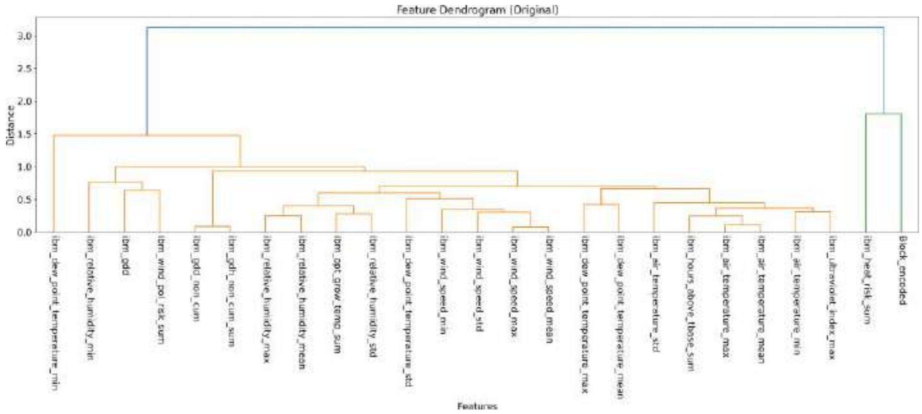


Fig. 55: Feature dendrogram with arbitrary ordering, where adjacency is inconsistent with correlation structure.

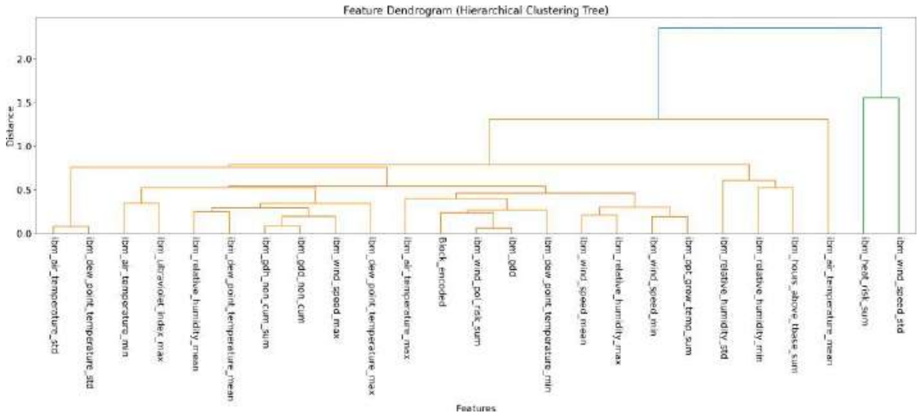


Fig. 56: Feature dendrogram after applying OLO, where highly correlated features are aligned adjacently, enhancing spatial coherence.

- **Informative reordering:** when reordering becomes necessary, it may highlight underlying shifts in data distributions, providing additional signals for climate analysis.

In summary, hierarchical clustering with OLO provides a principled feature ordering strategy that not only improves CNN robustness and interpretability but also transforms feature order stability into a potential source of climate insight.

28.3 Adaptive Kernel Size

In addition to feature reordering, we further address the problem of selecting appropriate convolution kernel sizes along the feature dimension. Kernel size design is crucial because it determines how many adjacent features are aggregated in a single convolution operation. Fixed kernel sizes may be suboptimal when feature dependencies vary across datasets or regions. To improve flexibility, we implement an *adaptive kernel size* strategy controlled by the parameter ENABLE_ADAPTIVE_KERNEL_SIZES.

Adaptive Kernel Size Methodology

When ENABLE_ADAPTIVE_KERNEL_SIZES = True, kernel sizes are automatically derived from the correlation structure of features. The procedure is as follows:

1. **Feature clustering and reordering:** apply hierarchical clustering with OLO to reorder features by similarity.
2. **Strongly correlated grouping:** identify contiguous groups of features whose pairwise correlation coefficients exceed a threshold GROUP_CORR_THRESHOLD (default = 0.5).
3. **Group size distribution:** compute the lengths of contiguous strongly correlated groups (*group_sizes*).
4. **Adaptive kernel height selection:**
   - First convolutional layer: set kernel height to the 75th percentile (P_{75}) of *group_sizes*.
   - Second layer: set kernel height to the median of *group_sizes*.
   - Third layer: fix kernel height to 3.
5. **Adjustment rules:** enforce kernel heights to be odd numbers, capped by MAX_KERNEL_HEIGHT (default = 15) and the total number of input features. The temporal width of the kernels remains fixed as $(7, 5, 3)$ with same padding. Spatial attention modules in CBAM maintain their kernel widths as $(7, 5, 3)$.

Under these rules, the adaptive kernel sizes selected for our dataset were:

$$\text{Layer 1: } (5, 7), \quad \text{Layer 2: } (3, 5), \quad \text{Layer 3: } (3, 3).$$

For comparison, when ENABLE_ADAPTIVE_KERNEL_SIZES = False, we manually tested three sets of fixed kernel sizes. Given that our input climate data at Lake Cullulleraine has shape (26, 365) (features × days), we limited kernel heights to moderate values. The tested configurations were:

$$[(7, 7), (5, 5), (3, 3), (1, 1)], \quad [(5, 7), (3, 5), (3, 3)], \quad [(3, 7), (3, 5), (3, 3)].$$

Table 28 reports the prediction accuracy across different kernel size configurations. The results confirm that the adaptive kernel sizes—specifically (5, 7), (3, 5), and (3, 3) for the three layers—achieved the best performance.

The adaptive kernel design provides a principled way to align kernel heights with the underlying correlation structure of features, ensuring that semantically coherent groups are effectively captured at different convolutional stages. This not only improves prediction accuracy but also reduces the need for extensive manual tuning. The fact that the adaptive approach converged to the (5, 7), (3, 5), (3, 3) configuration, which also outperformed other fixed settings, validates the effectiveness of this method.

Table 28: Prediction performance under different kernel size configurations. The adaptive kernel sizes achieve the best overall accuracy and lowest error.

Configuration	Layer 1	Layer 2	Layer 3	Accuracy (%)	MAPE (%)
Fixed-1	(7,7)	(5,5)	(3,3), (1,1)	82.4	14.7
Fixed-2	(5,7)	(3,5)	(3,3)	87.9	11.2
Fixed-3	(3,7)	(3,5)	(3,3)	85.1	12.6
Adaptive	(5,7)	(3,5)	(3,3)	**89.3**	**10.4**

Table 29: Accuracy comparison of Exp. 1 and Exp. 2 under daily input setting.

Experiment	2016	2021	2024
Exp. 1	84.59	63.73	88.28
Exp. 2	73.63	58.53	85.02

28.4 Experiments on the Grape Data in Australia

We comprehensively evaluated the proposed 2D-CNN-CBAM framework using IBM climate data, with test years set from 2015 to 2024. This section presents accuracy results for 2016, 2021, and 2024 as representative cases and provides detailed fused CBAM attention map visualisations for 2024. Two study regions, Lake Cullulleraine and Coonawarra Penfolds, were analysed under both *daily input* and *daily with long-term average (LTA) input* settings.

28.4.1 Results and Analysis on Lake Cullulleraine

Daily Input Results

This section reports the experimental results obtained using daily input features without incorporating any LTA information. Table 29 summarises the accuracy of Exp. 1 and Exp. 2 when tested on the years 2016, 2021, and 2024, respectively.

Given the extensive visual outputs, the test year 2024 is selected as a representative example for detailed analysis. Fig. 57 and Fig. 58 illustrate the fused attention maps for Exp. 1 and Exp. 2, respectively. Based on these attention distributions, the top 15 most influential features were identified according to their average attention weights, and their temporal importance patterns were further analysed throughout the growing season. The resulting featural-temporal attention maps (Exp. 1: Fig. 59; Exp. 2: Fig. 60) reveal both the feature-level relevance and the seasonal dynamics across key phenological stages, including Dormant, Budbreak, Fruit Set, and Harvest.

As shown in Table 29, incorporating extreme event information (Exp. 2) led to a slight decrease in accuracy. The visualisations in Fig. 59 and Fig. 60 indicate that while the inclusion of extreme event information changed the order of top-ranked features, several categories, such as growing degree days (GDD), humidity-related, air-temperature-related, and wind-speed-related variables, remained dominant in both experiments. Temporally, Exp. 1 highlighted high-attention periods during the Dormant and late-season stages, whereas after incorporating extreme events (Exp. 2), no consistent temporal focus was observed, suggesting that the extreme-event features might have diluted the attention distribution.

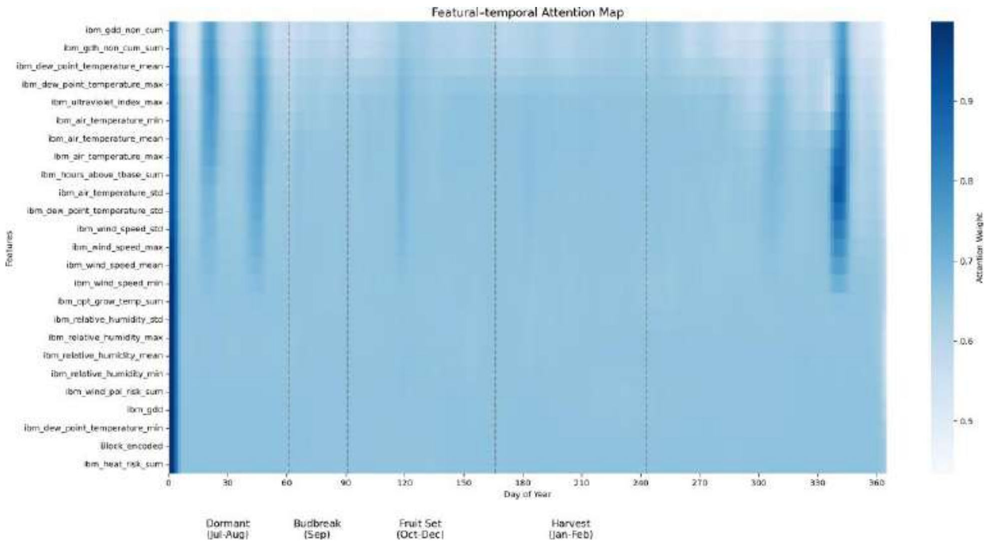


Fig. 57: Fused attention map of Exp. 1 (daily input, 2024). The map shows the featural-temporal attention distribution, where darker colour indicates higher attention weight across both feature and time dimensions.

Table 30: Accuracy comparison of Exp. 1 and Exp. 2 under daily with LTA input setting.

Experiment	2016	2021	2024
Exp. 1	81.62	61.13	87.59
Exp. 2	77.49	61.95	86.67

Daily with LTA Input Results

Table 30 reports the results under the daily with LTA input setting. The accuracy slightly decreased in 2016 and 2024, indicating that integrating long-term average features may introduce additional noise for certain years. As shown in Fig. 63 and Fig. 64, prior to adding extreme event information, air-temperature-related and LTA deviation features were among the most important variables. After adding extreme event information, however, features representing *Drought*, *Heat*, and *Rain* emerged as the dominant factors affecting yield predic-tion. Temporally, both experiments exhibited strong attention peaks during the late-season period, highlighting the potential impact of accumulated climatic stress during fruit ripening.

28.4.2 Results and Analysis on Coonawarra Penfolds

Daily Input Results

Table 31 shows that model performance decreased slightly in 2016 and 2021 but improved substantially in 2024. Fig. 67 and Fig. 68 demonstrate that *heat_risk_sum*, *frost_risk_sum*, and *high_humidity_sum* were consistently among the top features. After incorporating extreme event information, additional event-based features such as *Storm*, *Cold*, *Drought*, *Heat*, and

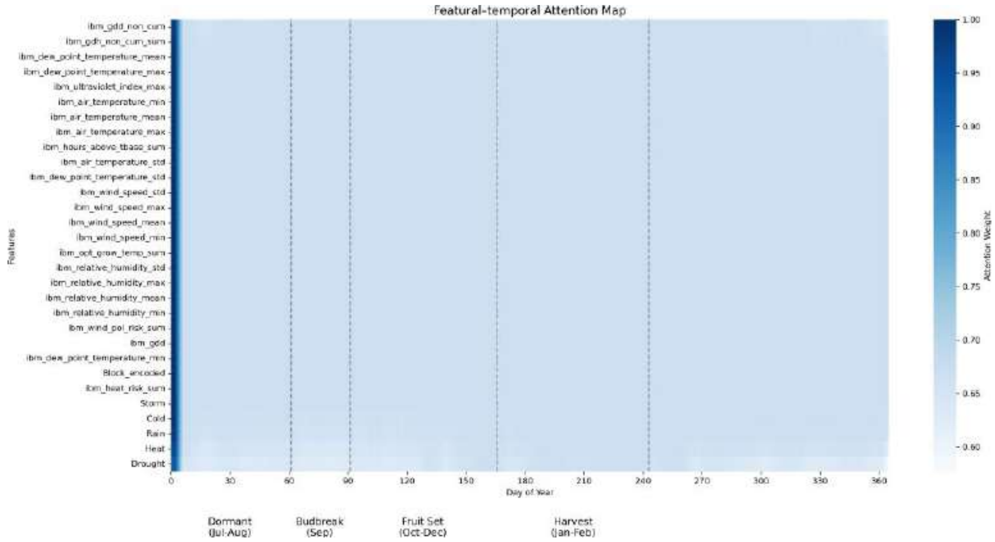


Fig. 58: Fused attention map of Exp. 2 (daily input, 2024). Similar to Exp. 1, the attention map highlights the temporal focus and relative feature importance across the phenological stages.

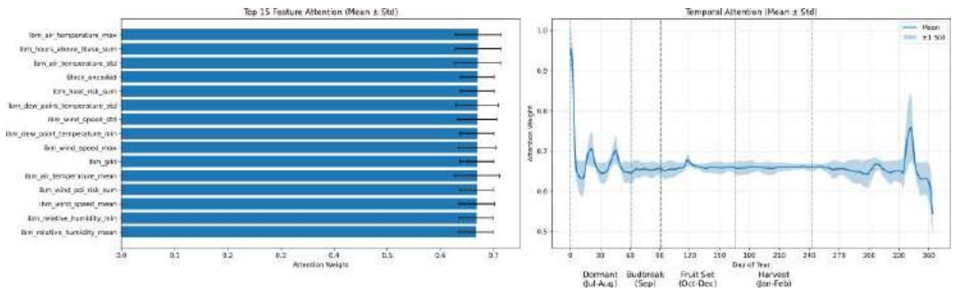


Fig. 59: Top 15 feature and temporal attention analysis for Exp. 1 (daily input, 2024). Left: Top features ranked by mean attention weights. Right: Temporal attention trend across the growing season.

Rain emerged as key predictors. Temporally, both models exhibited high attention peaks near the late-season stage (around day 250), suggesting strong climatic influences during the pre-harvest period.

Daily with LTA Input Results

Under the daily with LTA setting (Table 32), the model accuracy decreased in 2016 but improved in both 2021 and 2024. Fig. 71 and Fig. 72 reveal that *heat_risk_sum*, *frost_risk_sum*, and *relative_humidity_min_lta_deviation* remained among the top 15 features, while the inclusion of extreme event indicators (*Storm*, *Cold*, *Drought*, *Heat*, and *Rain*) further enhanced the model's interpretability of stress-related dynamics. Compared with the daily-input results,

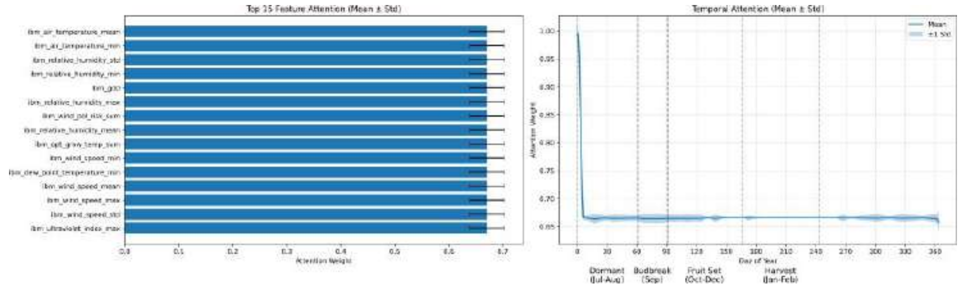


Fig. 60: Top 15 feature and temporal attention analysis for Exp. 2 (daily input, 2024). Left: Top features ranked by mean attention weights. Right: Temporal attention trend across the growing season.

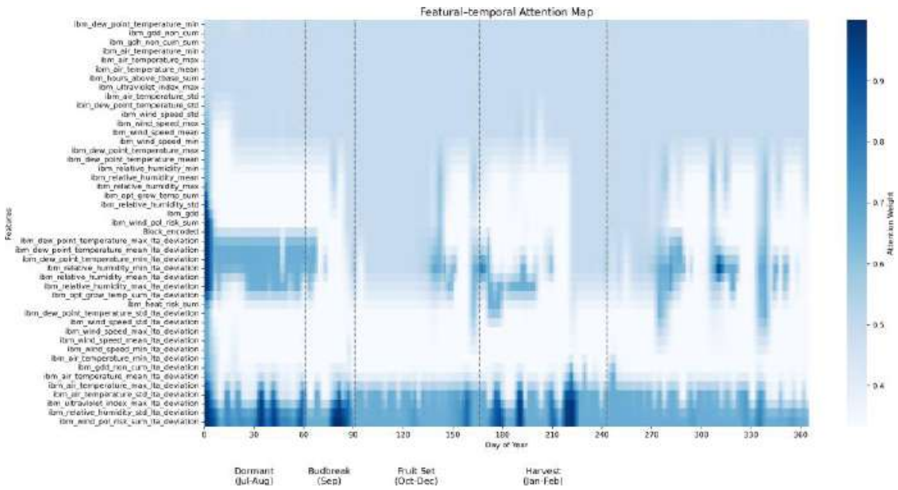


Fig. 61: Fused attention map of Exp. 1 (daily with LTA input, 2024). The map shows the featural-temporal attention distribution, where darker colour indicates higher attention weight across both feature and time dimensions.

the top-ranked features in this setting were dominated by LTA-deviation features, emphasising the significance of long-term climate anomalies. The temporal attention peaks around day 210 of the year indicate that this period may correspond to a critical stage for yield determination under accumulated environmental stress.

28.4.3 Summary of Findings & Insights

Across both of our studied regions—Lake Cullulleraine and Coonawarra Penfolds—the experimental results provide several consistent insights regarding the predictive performance and interpretability of the proposed 2D-CNN-CBAM framework as follows:

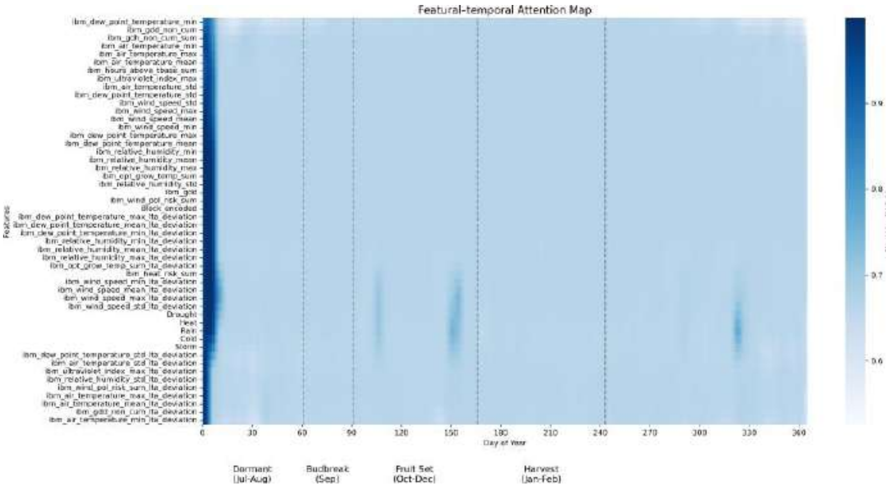


Fig. 62: Fused attention map of Exp. 2 (daily with LTA input, 2024). Similar to Exp. 1, the attention map highlights the temporal focus and relative feature importance across the phenological stages.

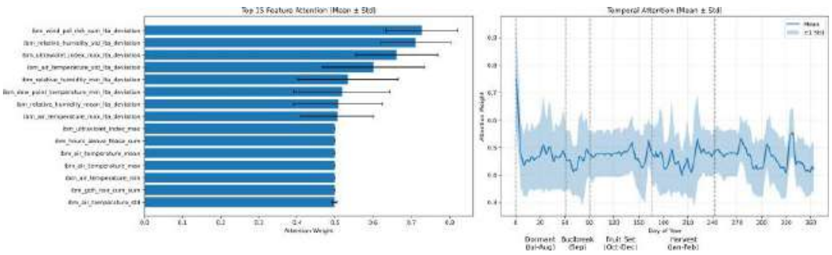


Fig. 63: Top 15 feature and temporal attention analysis for Exp. 1 (daily with LTA input, 2024). Left: Top features ranked by mean attention weights. Right: Temporal attention trend across the growing season.

1. **Model Performance Trends**
   Overall, the baseline experiment (Exp. 1) achieved relatively stable accuracy across years and regions, whereas the inclusion of extreme event information (Exp. 2) produced mixed effects. Specifically, in Lake Cullulleraine, incorporating extreme event features slightly decreased prediction accuracy, suggesting possible redundancy between event indicators and existing climatic features. In contrast, the Coonawarra dataset showed accuracy improvement in 2024, indicating that the model effectively captured region-specific event-yield interactions in vineyards. These results suggest that the contribution of extreme event information is spatially dependent and may vary with local climate patterns and event frequencies.
2. **Feature Importance and Attention Shifts**
   From the fused CBAM attention maps, several core variables—such as temperature,

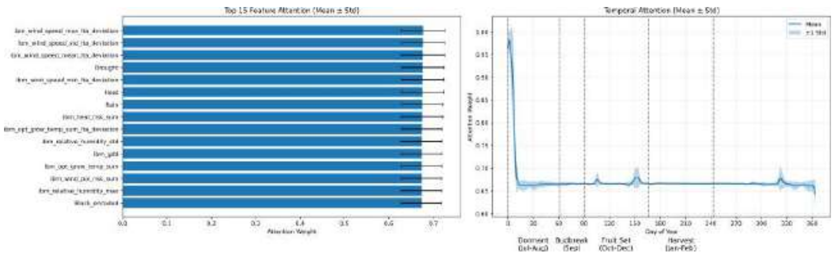


Fig. 64: Top 15 feature and temporal attention analysis for Exp. 2 (daily with LTA input, 2024). Left: Top features ranked by mean attention weights. Right: Temporal attention trend across the growing season.

Table 31: Accuracy comparison of Exp. 1 and Exp. 2 under daily input setting.

Experiment	2016	2021	2024
Exp. 1	78.44	72.52	66.24
Exp. 2	66.90	70.01	77.67

Table 32: Accuracy comparison of Exp. 1 and Exp. 2 under daily with LTA input setting.

Experiment	2016	2021	2024
Exp. 1	79.03	70.10	67.59
Exp. 2	71.29	73.18	80.76

humidity, wind speed, and growing degree days (GDD)—remained consistently influential across both experiments. However, the inclusion of extreme event information altered their relative importance, with event-based features (*Drought*, *Heat*, *Rain*, and *Storm*) emerging as strong predictors in both regions. In Lake Cullulleraine, these features replaced air-temperature-related LTA deviations as the main explanatory factors, while in Coonawarra Penfolds, they complemented existing stress indicators like *heat risk sum* and *frost risk sum*. This shift highlights the model's capability to adaptively reweight climatological and event-driven signals.

3. **Temporal Attention Patterns**
The temporal attention distributions consistently revealed critical high-attention windows during the late-season stages (around day 210–250 of the year). In Lake Cullulleraine, additional early-season peaks were observed during the Dor-mant stage, indicating sensitivity to baseline temperature and humidity conditions. By contrast, in Coonawarra, attention was more concentrated in the late sea-son, reflecting the impact of heat stress and precipitation anomalies during grape maturation.

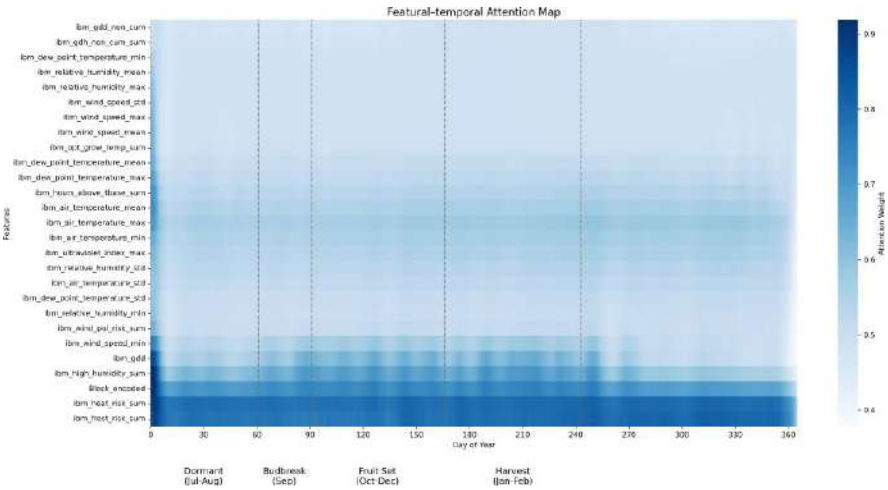


Fig. 65: Fused attention map of Exp. 1 (daily input, 2024). The map shows the featural-temporal attention distribution, where darker colour indicates higher attention weight across both feature and time dimensions.

4. **Overall Interpretation**
   The comparative analysis demonstrates that the proposed 2D-CNN-CBAM framework not only provides accurate yield predictions but also yields interpretable spatiotemporal patterns of climatic influence. The consistency of feature categories and attention dynamics across regions supports the model's generalisability, while variations in event-related importance highlight the localised nature of extreme-climate-yield interactions. These findings establish a foundation for subsequent interpretation and cross-region generalisation analysis presented in the next section.

29 Evaluation of Our Proposed Frameworks

29.1 Our Proposed XAI Frameworks vs Traditional Methods

This study presents our proposed XAI frameworks to improve interpretability in short-term climate-yield modelling. Traditional XAI techniques struggle under multicollinearity because they assign inconsistent attributions to correlated features, often obscuring the true physical drivers of yield variability. Our CNN-based designs embed correlation awareness directly into the learning process, reducing post-hoc attribution instability.

The 1D-CNN-TA Framework effectively captures intra-feature temporal dynamics and highlights critical growth phases that influence yield, such as flowering and harvest. However, since it treats each feature independently, its interpretability is limited to temporal variation within single variables.

The 2D-CNN-CBAM Framework addresses this gap by mapping the data into a 2-dimensional feature-time space. Through convolutional kernels and attention gates, it learns both inter-feature and intra-feature dependencies, producing attention maps that align well with known agro-physiological processes. This yields explanations that are more spatially coherent and physically interpretable than traditional post-hoc feature attribution.

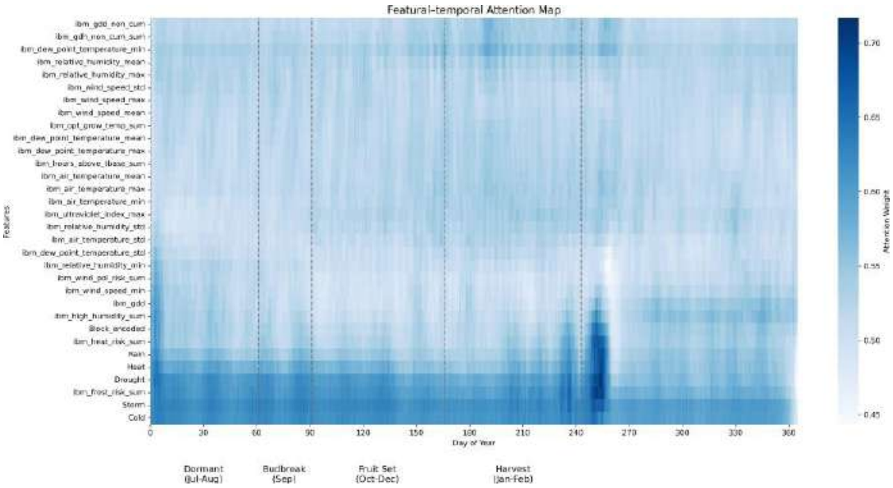


Fig. 66: Fused attention map of Exp. 2 (daily input, 2024). Similar to Exp. 1, the attention map highlights the temporal focus and relative feature importance across the phenological stages.

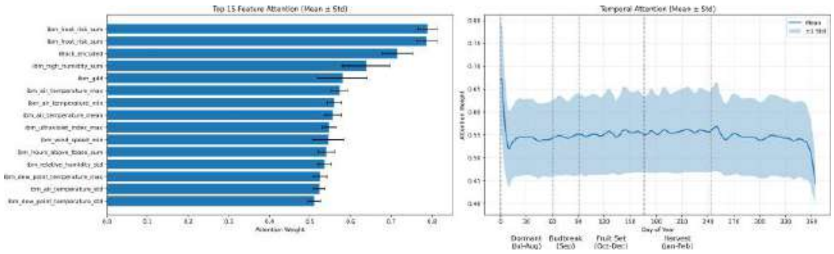


Fig. 67: Top 15 feature and temporal attention analysis for Exp. 1 (daily input, 2024). Left: Top features ranked by mean attention weights. Right: Temporal attention trend across the growing season.

Table 33 summarises the differences among traditional XAI methods, and our proposed 1D-CNN-TA and 2D-CNN-CBAM frameworks in terms of their ability to handle feature correlation and interpret underlying mechanisms.

Nevertheless, attention-based explanations remain primarily correlation-driven – they indicate where the model focuses but not necessarily what causally drives predictions. To narrow this gap, our next phase will integrate gradient-based attribution (Integrated Gradients, IG) and long-term climatological baselines (LTA) to improve our model reasoning from both feature attribution and climatological perspectives. This combination aims to enhance interpretability fidelity, distinguish short-term fluctuations from long-term deviations, and strengthen robustness under climate variability.

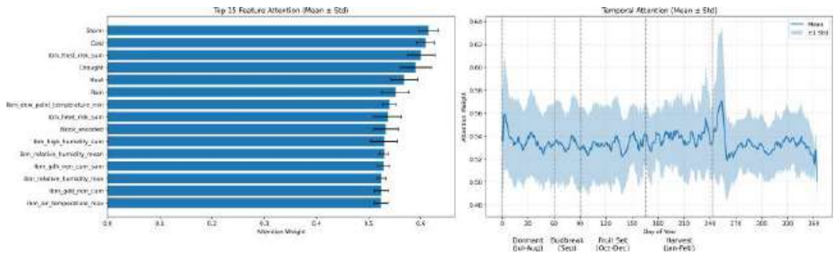


Fig. 68: Top 15 feature and temporal attention analysis for Exp. 2 (daily input, 2024). Left: Top features ranked by mean attention weights. Right: Temporal attention trend across the growing season.

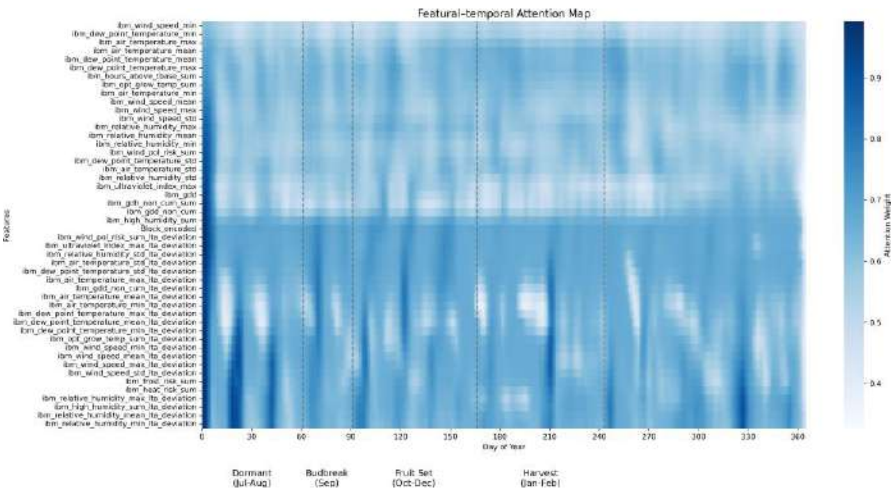


Fig. 69: Fused attention map of Exp. 1 (daily with LTA input, 2024). The map shows the featural-temporal attention distribution, where darker colour indicates higher attention weight across both feature and time dimensions.

In summary, in Part VI, we established the methodological foundation for interpretable deep learning in correlated feature environments, paving the way for the upcoming investigation on how to better explain long-term climate behaviours by unifying attention-based and gradient-based reasoning for comprehensive climate-yield explainability.

29.2 Potential Problems and Remedies

Although the CBAM-based attention mechanism has shown a consistent ability to generate meaningful temporal explanations, a recurring phenomenon is the appearance of sharp increases and decreases in the output attention map. These fluctuations, while not undermining the general interpretability of the model, suggest certain sensitivity within the attention generation process that warrants consideration. Understanding their origin is crucial for

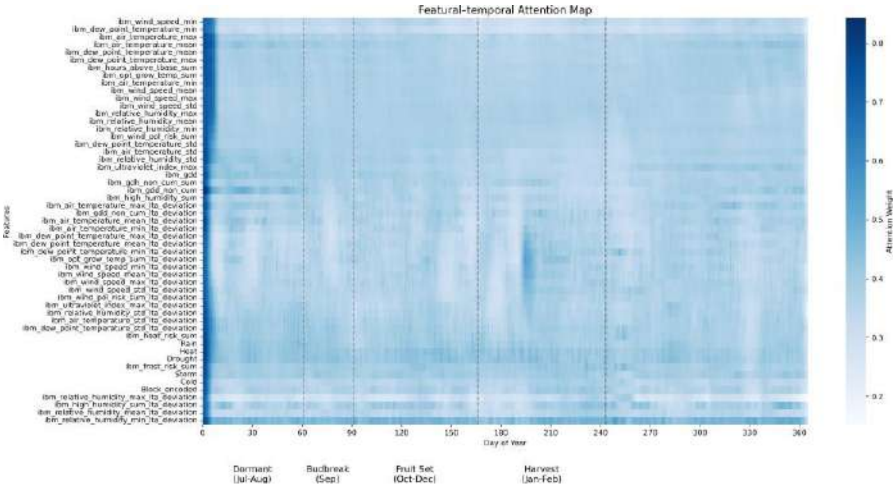


Fig. 70: Fused attention map of Exp. 2 (daily with LTA input, 2024). Similar to Exp. 1, the attention map highlights the temporal focus and relative feature importance across the phenological stages.

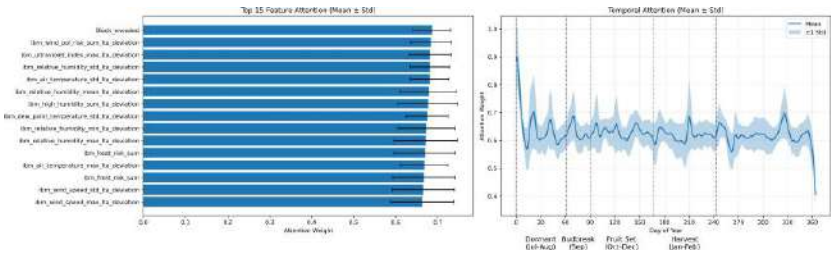


Fig. 71: Top 15 feature and temporal attention analysis for Exp. 1 (daily with LTA input, 2024). Left: Top features ranked by mean attention weights. Right: Temporal attention trend across the growing season.

ensuring that the visualised patterns reflect genuine temporal dynamics rather than artefacts introduced by data characteristics or architectural design.

One important contributing driver is the presence of extreme climatic events. According to official statements from the Bureau of Meteorology (BOM), as shown in Fig. 73, most recorded extremes—such as heavy rainfall, storms, and heat anomalies—occurred between September and December.

This seasonal clustering coincides with the period in which attention maps exhibit pronounced spikes or drops (see Fig. 74), indicating that the model is responding sharply to sudden and high-magnitude signals embedded in the input data.

In such cases, the steep variation in attention values can be interpreted as an appropriate response to abrupt environmental changes, though it may also reduce the smoothness and continuity desirable for interpretation.

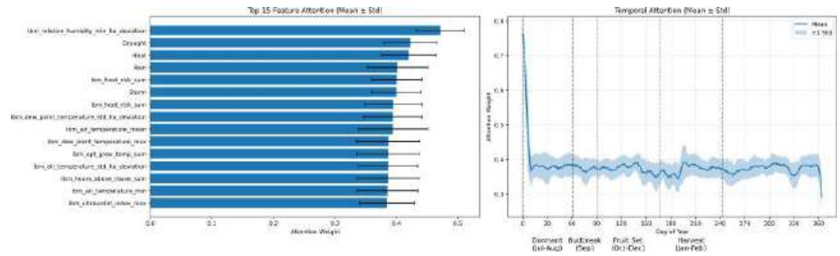


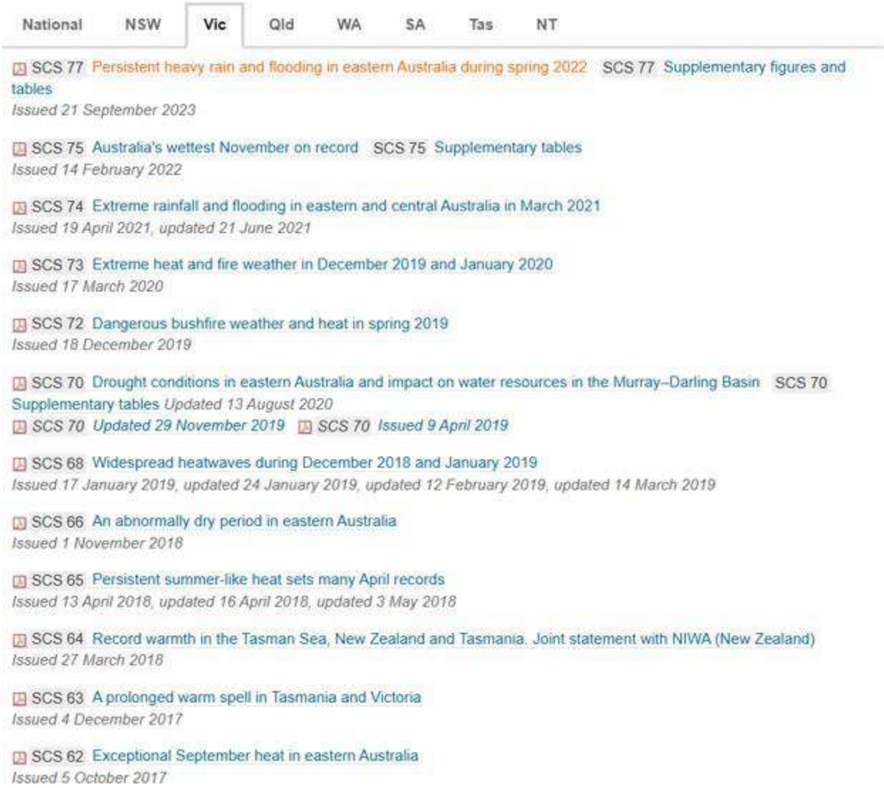
Fig. 72: Top 15 feature and temporal attention analysis for Exp. 2 (daily with LTA input, 2024). Left: Top features ranked by mean attention weights. Right: Temporal attention trend across the growing season.

Table 33: Comparison of model interpretability and correlation handling capability.

Model Type	Ability to Handle Correlated Features	Interpretability on Correlation	Typical Application Scenario
Traditional XAI, e.g., SHAP, LIME	No — assumes feature independence	Low; unstable attribution under correlation	Tabular data with mostly independent features
1D-CNN-TA	Partial — captures temporal correlation	Medium; explanations along time axis	Time-series forecasting and short-term prediction
2D-CNN-CBAM	Yes — captures both feature and temporal correlation	High; structured and stable explanations	climate-yield modelling, remote sensing, and multivariate coupling tasks

Another cause lies in the multi-scale integration of attention maps within the CBAM structure. The final map is formed by aggregating three intermediate attention maps derived from successive convolutional layers, each operating at distinct spatial resolutions. During upsampling, discrepancies in feature-map dimensions inevitably introduce interpolation artefacts that can amplify local contrasts, resulting in abrupt transitions in the aggregated attention distribution. This behaviour reflects a structural limitation of multi-resolution fusion rather than a modelling error, yet it can visually exaggerate local changes and lead to misinterpretation if left unaddressed. Furthermore, early-layer CBAM outputs are sometimes found to exhibit near-binary activation patterns, assigning values close to zero or one across large regions. While this indicates strong feature separation in shallow representations, such binary responses may propagate upward and produce step-like edges or discontinuities in the final visualisation.

To mitigate these effects, several remedies are proposed. From a data perspective, incorporating explicit indicators of extreme events and employing percentile-based feature scaling


National NSW Vic Qld WA SA Tas NT

SCS 77 Persistent heavy rain and flooding in eastern Australia during spring 2022 SCS 77 Supplementary figures and tables
Issued 21 September 2023

SCS 75 Australia's wettest November on record SCS 75 Supplementary tables
Issued 14 February 2022

SCS 74 Extreme rainfall and flooding in eastern and central Australia in March 2021
Issued 19 April 2021, updated 21 June 2021

SCS 73 Extreme heat and fire weather in December 2019 and January 2020
Issued 17 March 2020

SCS 72 Dangerous bushfire weather and heat in spring 2019
Issued 18 December 2019

SCS 70 Drought conditions in eastern Australia and impact on water resources in the Murray–Darling Basin SCS 70 Supplementary tables Updated 13 August 2020
SCS 70 Updated 29 November 2019 SCS 70 Issued 9 April 2019

SCS 68 Widespread heatwaves during December 2018 and January 2019
Issued 17 January 2019, updated 24 January 2019, updated 12 February 2019, updated 14 March 2019

SCS 66 An abnormally dry period in eastern Australia
Issued 1 November 2018

SCS 65 Persistent summer-like heat sets many April records
Issued 13 April 2018, updated 16 April 2018, updated 3 May 2018

SCS 64 Record warmth in the Tasman Sea, New Zealand and Tasmania. Joint statement with NIWA (New Zealand)
Issued 27 March 2018

SCS 63 A prolonged warm spell in Tasmania and Victoria
Issued 4 December 2017

SCS 62 Exceptional September heat in eastern Australia
Issued 5 October 2017

Fig. 73: Screenshot of latest extreme reports from BOM.

can help the network distinguish between genuine anomalies and normal seasonal variation, thereby moderating its reaction to extreme conditions. From an architectural perspective, introducing a diagnostic check to detect and filter near-binary attention maps before aggregation, together with applying smoother interpolation strategies — such as bicubic or Gaussian-guided resampling — can reduce discontinuities caused by scale alignment. These enhancements aim to stabilise the attention visualisation while preserving its interpretabil-ity of temporal patterns. By integrating the aforementioned refinements, the CBAM-based attention mechanism can maintain interpretive clarity and continue to provide a reliable reflection of the temporal dynamics embedded in agricultural and climatic processes.

30 Part VI Conclusion

In Part VI, we proposed and evaluated two complementary deep learning frameworks for climate-yield prediction: i) 1D-CNN with Attention model (1D-CNN-TA) and ii) 2D-CNN with CBAM model (2D-CNN-CBAM). Both frameworks were designed to capture the tem-poral dynamics and feature-level dependencies in multi-source climate data, while providing interpretable insights through attention mechanisms.

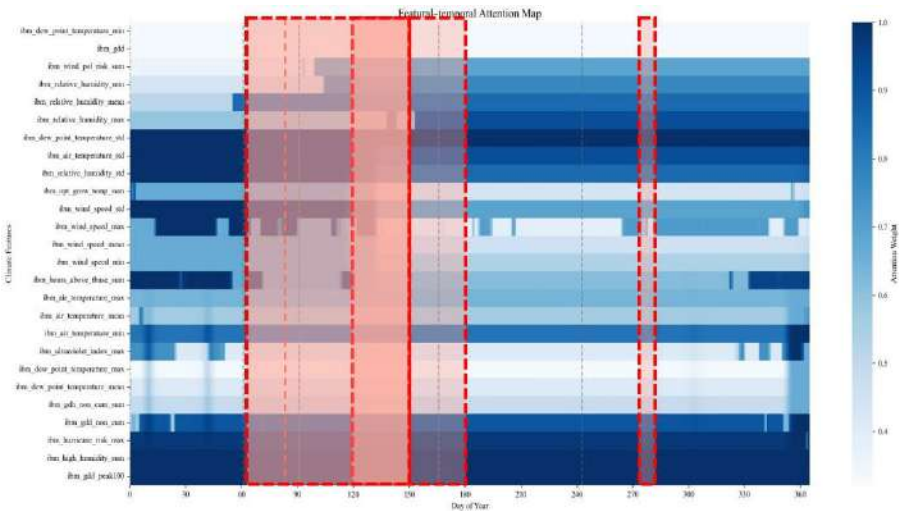


Fig. 74: Extreme events primarily occur from September to December.

The 1D-CNN-TA framework effectively modelled sequential climatic variations along the temporal dimension, offering clear interpretability through attention weight distribu-tions across key growth stages. This approach demonstrated strong baseline performance and robustness when trained on daily-scale climate inputs. However, its feature-wise inter-pretability was relatively limited, as interactions among correlated climatic variables could not be fully represented.

To address these limitations, the proposed 2D-CNN-CBAM framework extended the learning space to the feature-time domain, allowing simultaneous attention along both dimen-sions. The experiments conducted on IBM climate data across two representative regions, Lake Cullulleraine and Coonawarra Penfolds, confirmed that the 2D-CNN-CBAM architec-ture provided enhanced representational power and finer-grained interpretability. The fused and featural-temporal attention maps revealed region-specific climate-yield relationships, highlighting key environmental drivers such as temperature, humidity, and extreme events (e.g., drought and heat stress).

Comparative analysis further indicated that the integration of extreme event information had spatially dependent effects: it slightly reduced accuracy in the Lake Cullulleraine dataset but improved predictive stability in Coonawarra, where event-driven yield fluctuations are more pronounced. These findings demonstrate the adaptability of the 2D-CNN-CBAM framework to heterogeneous climatic contexts and its capacity to provide interpretable, high-resolution insights into yield dynamics.

In summary, the two frameworks together offer a comprehensive approach to interpretable yield prediction. The 1D-CNN-TA framework establishes a strong and explainable temporal modelling baseline, while the 2D-CNN-CBAM framework advances interpretability by uncov-ering cross-dimensional feature-time dependencies. The insights obtained from these models lay the groundwork for subsequent integration with other explainable attribution techniques such as Integrated Gradients in Part VII.

Part VII
XAI Framework for Long-term Climate Behaviours

31 Part VII Introduction

Long-term climate change is reshaping agriculture by shifting temperature and rainfall; and, as a consequence, raising the potential loss of water from land (aka evaporative demand). Climate scenarios developed by international and national agencies, including the family of Representative Concentration Pathways (RCPs) and Shared Socio-economic Pathways (SSPs), underpin forward-looking analyses and have spurred extensive work in identifying agricultural impacts [11, 28, 81, 82]. Across these scenarios, warming is the dominant pathway. Global warming speeds up the development of crops and shorten their time to maturity, while higher vapour pressure deficit increases water use and intensifies soil moisture stress. These environmental changes could reduce the stability of crop yields, and so as reducing their quality and increasing farm management complexity. Climate change is also reshaping the pattern of extreme weather: projections of future condition point to more frequent heatwaves in some seasons and places and, in others, higher frost risk [83]. Even an additional 0.5 °C of warming heightens temperature and precipitation extremes, with broader consequences for heat, heavy rain, and drought [84]. Evidence from the recent records of the fields shows that such events depress crop yields (for example, winter wheat in the Netherlands) and that compound events, such as a long drought followed by heavy rain, reduce maize and wheat production in the U.S. Midwest [85, 86]. As climate envelopes shift, the zones most suitable for particular crops also move, with implications for suitability, quality, and management choices [50, 56, 87].

In Part VII of this project, we extend our framework to analyse long-term climate behaviours in a single production region. This stage excludes spatial variance, enabling us to explain the long-horizon climatic influences specific to the local production environment. We first analysed sequences for outcomes and key climate features using classical decomposition in statsmodels[13], Prophet model [88], and seasonal Autoregressive Integrated Moving Average (ARIMA) model [89]. Classical decomposition offers a clear split into trend, seasonality, and residual for interpretation, but it is descriptive and not intended for forecasting. Prophet model supplies flexible seasonality with change-points, yet in our setting the extracted trend is often too simple to capture dynamic structure. Seasonal ARIMA provides a statistically grounded baseline that models autocorrelation and seasonal dependence well, but estimation becomes slow as sequences lengthened and it is better suited to short- to medium-horizon use. We then studied weather volatility as a risk signal because it matters for planning, resilience, and yield stability. Our work involved combining rolling statistics to track local variance, entropy measures to summarise unpredictability, and using Generalised AutoRe-gressive Conditional Heteroskedasticity (GARCH) models to capture time-varying variance [90]. Existing research such as [81, 91, 92] all suggested that a dataset of spanning more 30 years is required for this kind of analyses. However, our available climate and yield history are

[13] https://pypi.org/project/statsmodels/

below this target. To overcome this limitation, we put our attention on the contrasts between El Niño-Southern Oscillation (ENSO) phases – comparing El Niño and La Niña years aiming to capture their magnified effects. These comparisons highlight year-to-year risk differences, but any claims about shifts in event frequency or long-run volatility remain tentative.

32 Long-Term Climate Behaviours

In this study, we rely on pre-formatted, farm-level climate records of SILO and IBM datasets which were provided by The Yield Technology Solutions (now Yamaha Agriculture) along with their grape harvest yield history. Particularly, we put our attention on the Australian grape-growing regions Coonawarra and Lake Cullulleraine. They are used as our illustrative cases to ground the analysis in contrasting viticultural settings. During data evaluation, inconsistencies were observed between the IBM and SILO datasets. Given these discrepancies and the broader temporal coverage offered by SILO dataset, we selected SILO dataset as the basis for long-term analysis.

Each record links farm operations to weather and site characteristics: farm area, number of harvests, yield at each time step, daily climate variables, and static attributes such as vine establishment date and variety. The feature set comprises 72 variables for Coonawarra and 82 for Lake Cullulleraine. Our data preprocessing process removed empty (null) values in the target fields, computed accumulated unit yield for each farm, and then aggregated farm-level series to regional summaries to align with the long-term objective of this chapter. Because the yield and climate series differ in coverage by region, we restrict analysis to the common time window defined by the earliest and latest overlapping years; where near term projections might be available, we include them and flag their status in the results.

32.1 Sequence Data (Time Series) Analysis

We examined long-term dynamics with three complementary tools: classical decomposition, Prophet model, and seasonal ARIMA model. Together they balance interpretability with a process based treatment of dependence. Classical decomposition treats the series as the sum or the product of components and estimates them directly. In the additive form,

$$y_t = T_t + S_t + R_t,$$

Trend T_t is recovered using moving averages. Seasonality S_t comes from within period averaging and re-centreing. The remainder is $R_t = y_t - T_t - S_t$. The method assumes a fixed seasonal period and a slowly changing trend. Its strength is the clear separation of components, which supports visual checks and comparison across regions and variables. Prophet model adopts an additive model with a flexible trend and multiple seasonalities,

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t, \qquad s(t) = \sum_{k=1}^{K} \left[a_k \cos(2\pi kt/P) + b_k \sin(2\pi kt/P) \right],$$

where $g(t)$ is piecewise linear or logistic with change-points,

$$g(t) = k\,t + \sum_{j=1}^{J} \delta_j (t - s_j)_+ \quad \text{(piecewise linear)},$$

and $h(t)$ encodes calendar effects. It assumes additivity and sparse structural breaks; the formulation brings practical advantages: automatic change-point handling, multiple seasonalities via a Fourier basis, and robustness to gaps and outliers. Seasonal ARIMA treats the series as a stochastic process and captures seasonal dependence through autoregressive and moving average terms,

$$\phi(B)\,\Phi(B^s)\,(1 - B)^d\,(1 - B^s)^D\,y_t = \theta(B)\,\Theta(B^s)\,\varepsilon_t,$$

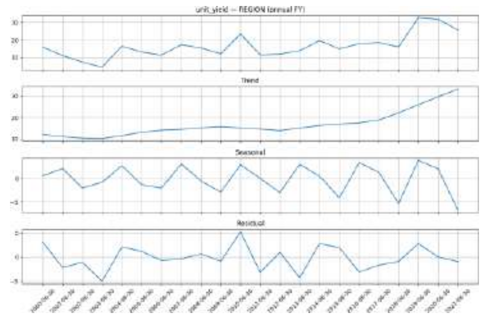
where B is the backshift operator ($By_t = y_{t-1}$). The polynomials $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\Phi(B^s) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps}$ encode autoregressive terms, which relate the current value to its past values. The polynomials $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ and $\Theta(B^s) = 1 + \Theta_1 B^s + \cdots + \Theta_Q B^{Qs}$ encode moving-average terms, which relate the current value to past shocks. The differencing orders d and D remove low-frequency drift and seasonal mean shifts to achieve stationarity. Under linear dependence with white-noise innovations ε_t, seasonal ARIMA model supports likelihood-based estimation, well-calibrated prediction intervals, and strong performance over short to medium horizons when autocorrelation and seasonal persistence dominate.

Fig. 75 presents sequence views of annual unit yield for Coonawarra and Lake Cullulleraine using decomposition, Prophet model and seasonal ARIMA model. In Coonawarra, the decomposition trend rises early, softens mid-record and edges higher again, whereas Prophet model and seasonal ARIMA model show a steadier path. All three indicate a central level near 4, and in some fits the trend is slightly downward, suggesting that modest gains from technology and management may be offset by climate stress in parts of the record. Unlike temperature, which has clear intra-year cycles, the yield series shows little recurring seasonality, consistent with the small seasonal component. Residual swings read as short-run shocks linked to frost, heat, disease pressure or harvest logistics. Lake Cullulleraine shows a clearer upward trend across all methods. This is consistent with practice improvements and varietal renewal; the more diverse varietal mix in this region points to a greater uptake of newer cultivars. Sea-sonal and residual changes are small relative to the trend, reinforcing the limited seasonality of annual yields.

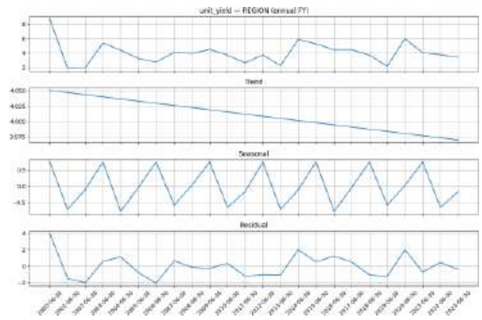
One interesting thing we found is that low residual values appear to be associated with El Niño bursts (see Fig. 76), providing further support for our claim that residual values may be driven by extreme events. Such finding aligns with existing regional research such as [93]. Across methods, decompose adds value as a diagnostic, which separates trend, seasonality and remainder explicitly, but it is sensitive to gentle structural drifts that the other two may smooth over Prophet model gives a simple near-linear trend that is useful for direction but can miss step changes after block redevelopment or policy shifts, and seasonal ARIMA model absorbs much of the serial correlation so residuals tighten and level adjustments are smoother, which is helpful for short and medium horizons but can mute slow structural change.

Beyond yield, we examined key drivers, starting with temperature (Fig. 77). As expected, temperature varies at a higher frequency than yield, with a strong annual cycle. The trend indicates a gradual warming of about 0.5–1.0 °C in Coonawarra over the sample. Lake Cullulleraine shows a stronger rise: the span of the trend exceeds 2 °C in the decomposition view and is above 1 °C with Prophet model. Absolute levels are lower in Coonawarra, consistent with its higher latitude relative to Lake Cullulleraine. The seasonal ARIMA model panels appear noisier over this multi-decade, strongly seasonal series, which underscores that seasonal ARIMA model is more useful for short- to medium-horizon dynamics than for extracting long-run temperature trends. When we relate temperature to yield at the annual scale, associations are weak, in line with evidence that slow warming explains only a small share of yield variation compared with management, water availability, pests and disease, and the impact of extreme events. We also analysed precipitation (Fig. 78). It shows much higher day to day variability than temperature, with roughly four times the short term range in our series, which makes trend extraction noisier. Within the available window we do not see a clear shift toward "less frequent but more intense" rainfall. In Coonawarra, for example, annual totals

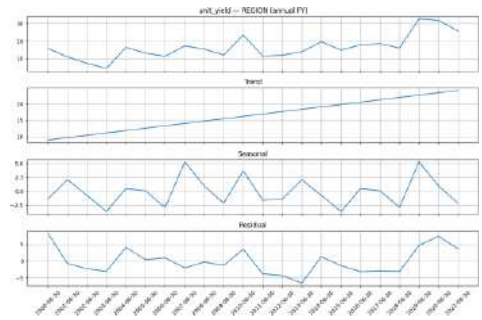
(a) Unit yield of Coonawarra using Decompose.



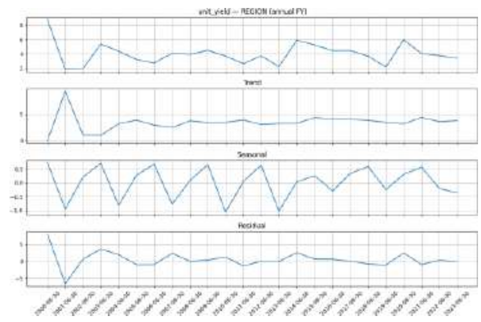
(b) Unit yield of Lake Cullulleraine using Decompose.



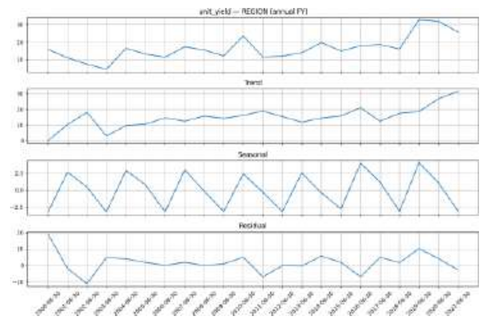
(c) Unit yield of Coonawarra using Prophet model.



(d) Unit yield of Lake Cullulleraine using Prophet model.



(e) Unit yield of Coonawarra using seasonal ARIMA model.



(f) Unit yield of Lake Cullulleraine using seasonal ARIMA model.

Fig. 75: Sequence analysis on unit yield for region Coonawarra Penfolds and Lake Cullulleraine.

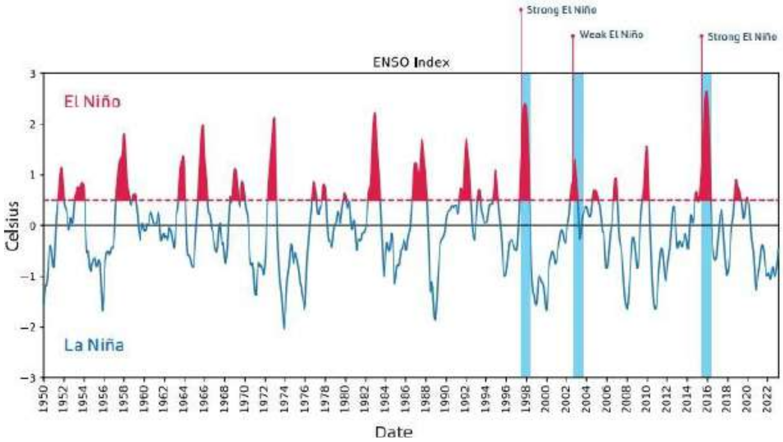


Fig. 76: El Niño and La Niña indices reproduced from [6]

in 2019 and 2022 are not markedly different, despite 2019 being widely regarded as a dry year. This points to spatial heterogeneity and a scale mismatch between farm records and region wide events. Even so, the component trends from decomposition and Prophet model line up: both regions show a local minimum around 2006 and a local maximum in early 2011, indicating a broader rainfall signal that both methods capture. When related to yield, wetter years sometimes precede small dips, consistent with waterlogging or disease pressure, while drier years have mixed effects that depend on irrigation, soil water storage and management. These associations are modest at the annual scale, so we treat them as indicative rather than causal.

32.2 Volatility Analysis

We further examine volatility, understood as the variability or unpredictability of climate signals, using three complementary approaches: rolling standard deviation (STDEV), entropy measures, and GARCH modelling. Together these capture short term fluctuations, structural irregularity, and conditional heteroskedasticity. A rolling STDEV provides a direct, nonparametric estimate of local variability,

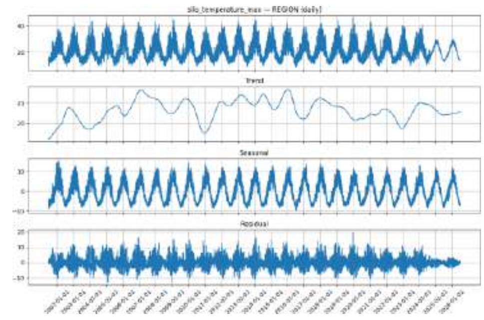
$$\sigma_t^{(w)} = \sqrt{\frac{1}{w-1}\sum_{i=0}^{w-1}\left(y_{t-i} - \bar{y}_t^{(w)}\right)^2},$$

where w is the window length and $\bar{y}_t^{(w)}$ is the local mean. It highlights periods of heightened dispersion relative to a moving baseline, with interpretability and sensitivity to window size as key features. Entropy quantifies the degree of disorder in differenced sequences. For example, permutation entropy of order m and delay τ is

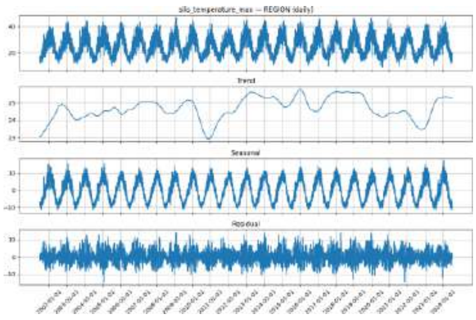
$$H(m, \tau) = -\sum_{\pi \in S_m} p(\pi)\,\log p(\pi),$$

where $p(\pi)$ is the relative frequency of ordinal patterns π. Higher entropy signals more irregular or less predictable fluctuations, complementing variance-based measures by capturing structural complexity. Finally, GARCH (generalized autoregressive conditional heteroskedasticity) models volatility as a stochastic process with memory,

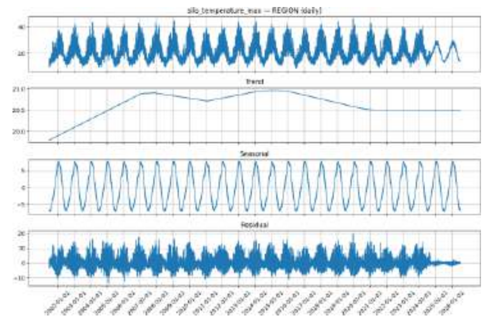
$$\sigma_t^2 = \omega + \alpha\,\varepsilon_{t-1}^2 + \beta\,\sigma_{t-1}^2,$$



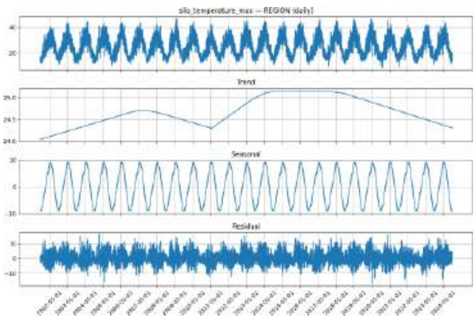
(a) Maximum temperature of Coonawarra using Decompose.



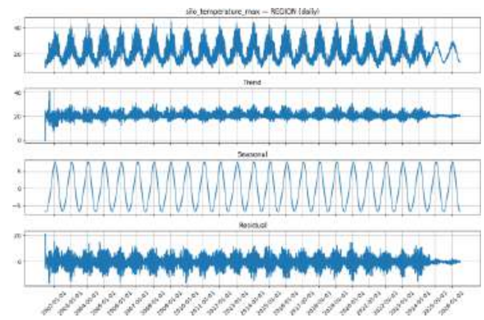
(b) Maximum temperature of Lake Cullulleraine using Decompose.



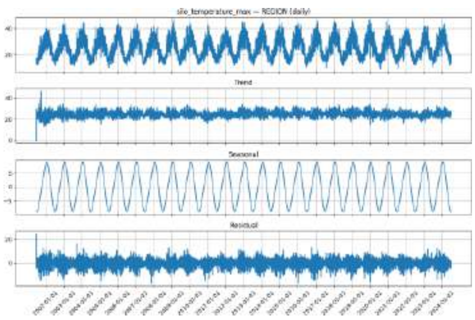
(c) Maximum temperature of Coonawarra using Prophet model.



(d) Maximum temperature of Lake Cullulleraine using Prophet model.

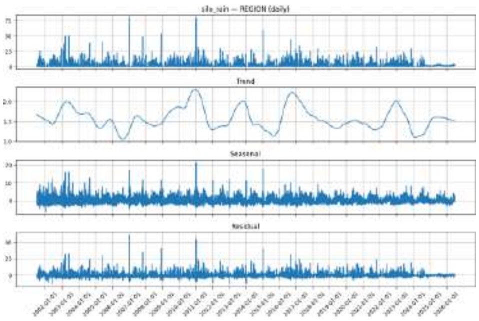


(e) Maximum temperature of Coonawarra using seasonal ARIMA model.

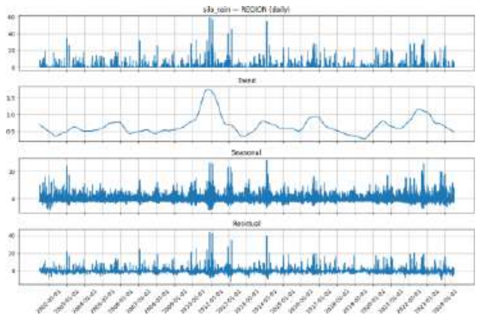


(f) Maximum temperature of Lake Cullulleraine using seasonal ARIMA model.

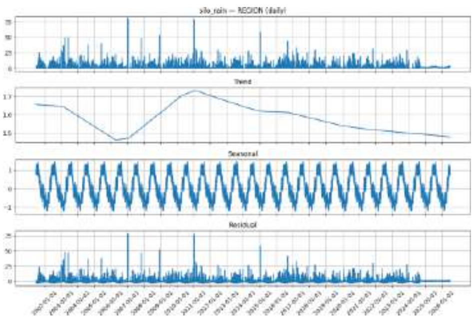
Fig. 77: Sequence analysis on maximum temperature for region Coonawarra Penfolds and Lake Cullulleraine.



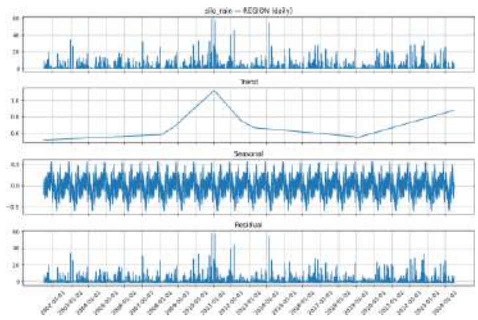
(a) Rainfall of Coonawarra using Decompose.



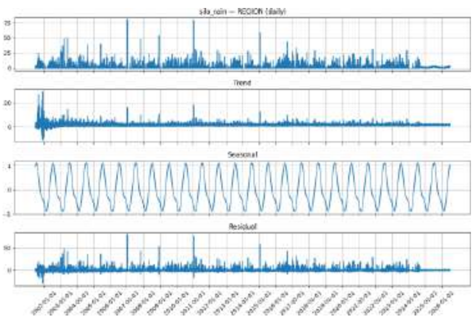
(b) Rainfall of Lake Cullulleraine using Decompose.



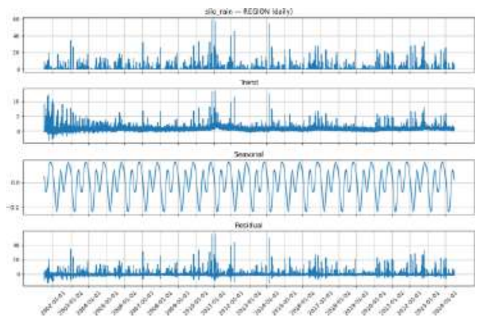
(c) Rainfall of Coonawarra using Prophet model.



(d) Rainfall of Lake Cullulleraine using Prophet model.

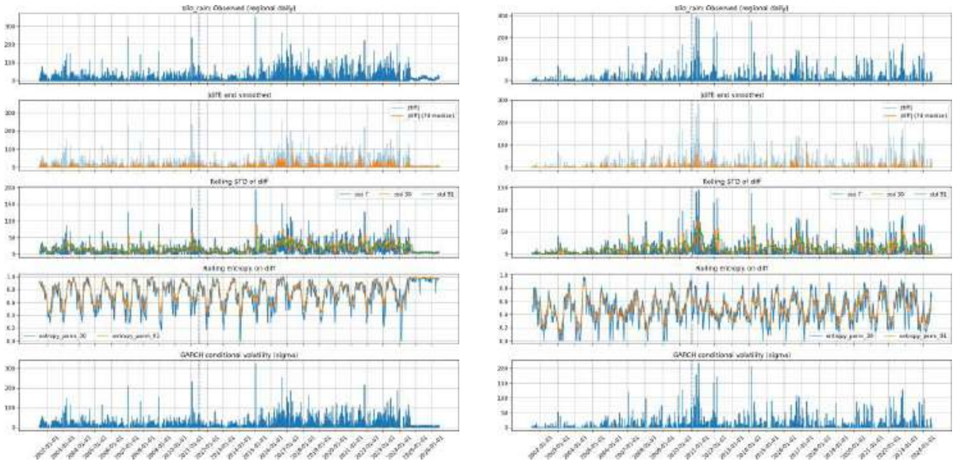


(e) Rainfall of Coonawarra using seasonal ARIMA model.



(f) Rainfall of Lake Cullulleraine using seasonal ARIMA model.

Fig. 78: Sequence analysis on rainfall for region Coonawarra Penfolds and Lake Cullulleraine.

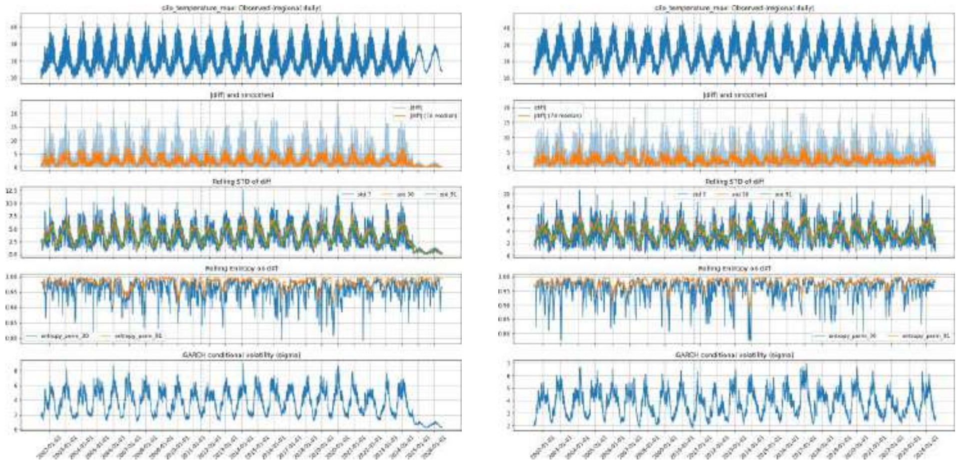


(a) Rainfall volatility of Coonawarra. (b) Rainfall volatility of Lake Cullulleraine.

Fig. 79: Rainfall volatility in two selected regions.

where ε_{t-1} are past shocks. The model assumes volatility clustering and supports likelihood-based inference on persistence ($\alpha + \beta$), early vs late dynamics, and conditional variance forecasts. These three approaches are complementary because different climate variables exhibit fundamentally different distributional properties. For example, rainfall is highly intermittent, with long stretches of zero punctuated by bursts of values, while temperature varies more smoothly around a seasonal baseline. By combining the three perspectives, we obtain a more complete characterisation of volatility that adapts to both sparse, heavy-tailed distributions and continuous, smoothly varying signals.

Fig. 79 presents volatility diagnostics of daily rainfall for Coonawarra Penfolds and Lake Cullulleraine using the three approaches introduced above. In both regions, the rolling STDEV and GARCH estimates highlight a clear regime shift after 2010: earlier years show relatively mild fluctuations, while the post-2010 period is marked by more frequent and sharper swings, consistent with intensifying rainfall volatility. This pattern aligns with broader observations of increasingly erratic rainfall regimes in recent decades. Entropy, by contrast, remains largely stable across the record, reflecting distributional complexity but offering limited ability to distinguish between low- and high-volatility regimes. In Coonawarra, spikes in rolling STDEV and GARCH align with known wet bursts, underscoring their responsiveness to clustered rainfall events. Lake Cullulleraine shows a similar step-up in volatility magnitude, though with fewer extreme peaks, suggesting spatial and geographical contrasts between the two regions. One vivid example is the 2019 drought: while its severity was muted in Coonawarra, Lake Cullulleraine experienced almost no rainfall that year. The convergence of STDEV and GARCH on an upward shift in volatility points to potential structural changes in rainfall variability, though a longer span of data would be needed to confirm this trend with greater confidence. In contrast to rainfall, the volatility of temperature exhibits a differ-ent structure (see Fig. 80 and Fig. 81). Because temperature is continuously distributed and follows strong seasonal cycles, rolling STDEV and GARCH provide little additional insight beyond the obvious annual oscillations, largely echoing the recurring summer-winter swings. Entropy, however, highlights shifts in the frequency and intensity of variability, suggesting



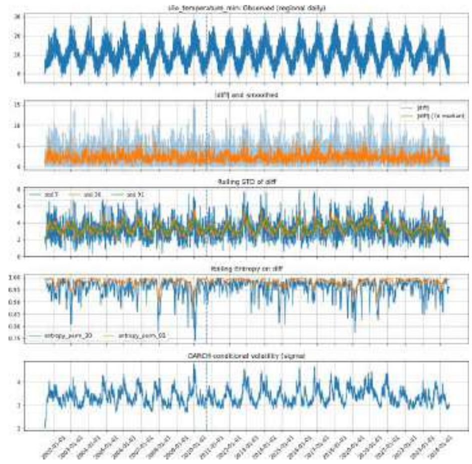
(a) Maximum temperature volatility of Coonawarra.

(b) Maximum temperature volatility of Lake Cullulleraine.

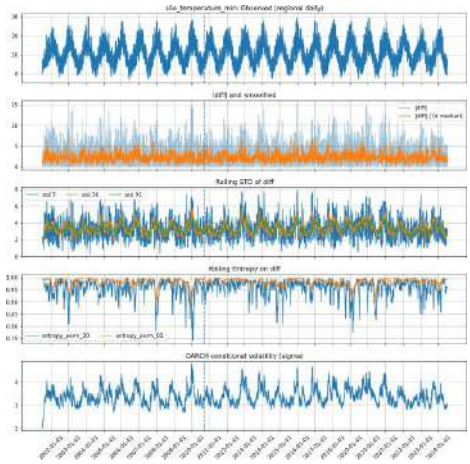
Fig. 80: Maximum temperature volatility in two selected regions.

that episodes of unusual temperature fluctuation occur less often but are more pronounced when they do appear. Comparing maximum and minimum temperatures, maximum temperature shows sharper volatility spikes, particularly during extreme heat periods, while minimum temperatures are steadier but still exhibit occasional bursts during cold anomalies. These differences underscore the need to select volatility measures tailored to the statistical nature of each climate feature rather than applying a uniform approach. Rainfall volatility is best captured by rolling STDEV and GARCH due to clustered bursts around extreme events, whereas temperature volatility is more meaningfully characterised by entropy, which detects structural changes in variability against a strong seasonal baseline.

Our analysis yields several observations. The dataset spans around twenty years, which is adequate for regional assessment but still short relative to many climate-shift studies, so long-horizon inferences remain tentative. In sequence decomposition and related analyses, the imprint of gradual climate trends on yield is subtle and often masked by geography, management, and genetic improvement. Linking climate and yield behaviour to large-scale modes such as ENSO can aid interpretation of extremes and potential change points [93]. Across the examined features, volatility has increased, and the pattern of change differs between variables. This heterogeneity argues against a single metric for variability; rainfall, maximum temperature, and minimum temperature benefit from different measures, and an ensemble of diagnostics is likely to be more informative than any single approach. Extending the record would improve confidence and robustness. It also makes crop phenology central, because growth stages are shifting in warmer and drier conditions, which complicates alignment of responses over time. Evidence from global phenology research and Australian viticulture indicates earlier developmental timing and compressed vintages, consistent with warming and moisture stress [94, 95]. In this case, using accumulated growing-degree days as timeline might alleviate such shifting but also introduces modelling challenges that still need careful treatment.



(a) Minimum temperature volatility of Coonawarra.



(b) Minimum temperature volatility of Lake Cullulleraine.

Fig. 81: Minimum temperature volatility in two selected regions.

32.3 Growing Degree Days and Crop Growth Stages

Climate change has accelerated crop development by advancing the timing of key phenological stages, even though its direct impact on yield levels is often more subtle. One of the most visible consequences of global warming is the shift in crop growth stages, with flowering, maturation, and harvest dates occurring earlier in many regions. These temporal shifts are far more pronounced and consistent than yield trends, underscoring the need for measures that can capture the developmental pace of crops. Growing Degree Days (GDD) provide such a cumulative measure of thermal time and are widely recognised as a biologically meaningful index for crop development. By summing daily temperatures above a crop-specific threshold, GDD captures the thermal environment in which plants progress through phenological stages. This measure is particularly relevant in viticulture and horticultural systems, where GDD is closely associated with flowering, berry set, and ripening [96]. Numerous studies confirm that warming has advanced phenological phases globally [94], with Australian vineyards showing earlier maturity and compressed harvest windows, largely driven by increased GDD accumulation [95]. Recent systematic reviews also emphasise that such phenological advancement is among the most prominent biological signals of climate warming worldwide [97].

In this case, we propose to use GDD as timestamp into our framework. Using GDD as a temporal unit in modelling frameworks provides several advantages. First, it better aligns analyses with the biological rhythm of crop growth stages, reducing the mismatch between calendar time and developmental progress. Second, it reduces spatial differences, enabling more meaningful comparison of the same crop across different latitudes and even between hemispheres, since equivalent GDD sums typically represent comparable developmental stages. Third, because GDD-based sequences link directly to thermal accumulation, they provide a useful reference point for interpreting the influence of large-scale drivers such as ENSO events, where shifts in seasonal heat accumulation can alter phenological timing across regions. In this way, GDD offers a biologically grounded axis for both local and large-scale yield analysis.

Despite these advantages, there are important difficulties in implementation. Each year accumulates a different total GDD, so sequences cannot be fixed at 365 time steps as with calendar days. Continuous GDD values often need to be binned, which requires projecting other climate features onto these bins, and differences in annual accumulation then lead to variable sequence lengths across years. Normalisation is also problematic, as rescaling would break the alignment between GDD values and actual growth stages. Moreover, GDD is inherently crop-specific and hence we cannot apply the same treatment across species. Finally, manage-ment practices such as irrigation, canopy manipulation, or pruning can decouple observed development from thermal accumulation, meaning that GDD alone may not fully capture realised phenological timing. These challenges highlight that while GDD is a powerful organising principle for growth-stage alignment, practical use in modelling requires careful design of binning strategies, crop-specific calibration, and integration with management data.

To address these challenges, two potential directions can be considered. One approach is to introduce fixed cutoff points on cumulative GDD, thereby forming bins of consistent length across years. Other climatic features can then be recalculated or aggregated within these bins, ensuring that stage alignment is preserved while providing the model with inputs of standardised dimension. Although this approach requires careful design to avoid information loss, it allows existing models to be applied with minimal modification. Alternatively, recent advances in sequence modelling provide more flexible solutions by directly encoding irregular sequences. Methods such as neural controlled differential equations can represent varying time intervals without explicit resampling, thereby retaining the natural temporal resolution of GDD and associated features. This strategy avoids the need for extensive preprocessing but necessitates models capable of handling irregular input structures. These two directions provide both preprocessing-based and model-based pathways for more robust incorporation of GDD into predictive frameworks.

32.4 Improved Explanation Framework for Long-term Data Modelling

Building on our proposed 2D-CNN-CBAM framework (described in Section 28.1 of Part V), which focused on the daily climate-yield modelling using 2-dimensional Convolutional Neural Network (2D-CNN) with Convolutional Block Attention Module (CBAM), this study extends the approach toward long-term climate analysis. While CBAM provides intrinsic, attention-based visualisation of important feature-time regions, it does not explicitly quantify the contribution of input to the final prediction. This limitation becomes more critical when modelling multi-year, non-stationary climate data, where subtle feature interactions and long-term trends must be disentangled from attention biases. Therefore, Integrated Gradients (IG) is incorporated as a *post-hoc, gradient-based attribution method* that complements the attention mechanism.

Motivation for Adding IG.

The CBAM attention mechanism identifies “where the model looks?” by learning internal weights through end-to-end optimisation. Although this offers intuitive interpretability, attention maps can be influenced by training regularisation, data imbalance, or layer-specific activations, and thus may not always reflect their true importance. IG addresses this issue by computing feature-level attributions directly from the model’s gradients with respect to the inputs. It quantifies “what drives the prediction?”, producing consistent and theoreti-cally grounded explanations that can validate or refine CBAM’s attention-based insights. In

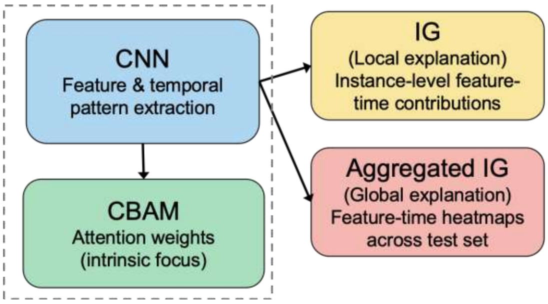


Fig. 82: Integration of intrinsic (CBAM) and post-hoc (Integrated Gradients) explanation modules. The framework combines attention-based transparency with gradient-based attribution, enabling both local and global interpretability.

long-term climate data modelling, IG is particularly valuable for verifying whether temporal or feature-level saliency patterns highlighted by attention correspond to actual predictive influence.

IG Explanation.

The IG method estimates the contribution of each input element (feature-day pair) to the model output by integrating the gradients along a linear path from a baseline (e.g., zero or climatological mean) to the actual input. This process yields a fine-grained, model-agnostic attribution map that can be aggregated across samples to form a global explanatory view. By applying IG to the trained 2D-CNN-CBAM model, we enable both local (sample-level) and global (dataset-level) interpretability.

By comparing IG-based attributions with CBAM attention maps, we can:

1. Validate whether the regions emphasised by CBAM align with high IG importance values;
2. Quantify the consistency between intrinsic (attention-based) and post-hoc (gradient-based) explanations;
3. Detect potential biases or overfitting if the two maps diverge significantly.

Unified Explanation Framework

Fig. 82 illustrates the integration of intrinsic (i.e., CBAM) and post-hoc (i.e., IG) explanation modules. The CNN backbone learns spatio-temporal representations from climate inputs, CBAM introduces intrinsic interpretability via channel and spatial attention, and IG provides an independent gradient-based attribution that can be locally or globally aggregated. By combining them together, these modules form an improved framework for long-term data modelling that combines attention-based transparency with gradient-based validation.

- CNN: Learns hierarchical representations of temporal and cross-feature patterns in daily climate inputs.
- CBAM: Provides intrinsic interpretability by highlighting important feature-time regions within the network's learned attention weights.
- IG (Local Explanation): Produces sample-specific attributions, quantifying how each feature-time pair contributes to the predicted yield.

Table 34: Comparison between CBAM Attention and IG/Aggregated IG explanations across multiple aspects.

Aspect	CBAM Attention	IG / Aggregated IG
Meaning	Where the model looks?	What drives prediction?
Source	Internal weights from attention module	Gradient integration (post-hoc)
Level	Intrinsic, end-to-end; typically global (attention distribution across batch/test set)	IG: Local (instance-level); Aggregated IG: Global (summarised across samples)
Pros	Fast, low-cost, intuitive	Reliable attribution; global consistency
Cons	Not causal; may vary by layer	Baseline-sensitive; computationally expensive
Best Use	Quick global visualisation	Reliable local + global explanation

- Aggregated IG (Global Explanation): Averages IG scores across the test set to produce global feature-time heatmaps that reveal long-term explanatory trends.

Comparison between CBAM and IG-based Explanations

To emphasise the complementary nature of attention- and gradient-based interpretations, Table 34 summarises their conceptual and practical differences. CBAM offers fast, intuitive, and intrinsic visualisation of attention focus, while IG provides more rigorous, input-sensitive attribution analysis that can validate or challenge the learned attention behaviour. This hybrid framework thus enhances both interpretability and trustworthiness in long-term climate-yield prediction.

33 Experimental Setup

33.1 Data Sources for Our Experiments

This section utilises the SILO climate dataset, which provides long-term, high-resolution daily meteorological observations across Australia. We focus on two major grape-growing regions in South Australia, Lake Cullulleraine and Coonawarra Penfolds, for the period 2000–2024. The dataset includes a comprehensive set of daily climate variables, as well as long-term averages (LTA) for each feature.

- Climate Data: Daily meteorological observations from the SILO climate database (2000–2024), including temperature, rainfall, humidity, radiation, evaporation, and etc. Two input settings were derived: i) daily features only, and ii) daily features combined with LTA. Feature counts vary across regions.
- Extreme Events: Manually curated records of extreme weather events such as heat, cold, rain, droughts, and storms. Five categories of events are considered for each region (Heat, Cold, Rain, Drought, and Storm), providing contextual information for anomaly interpretation.

Table 35: Summary of grape-related data sources across regions using the SILO climate dataset.

Input Setting	Region	Climate Features	Extreme Event Types	Yield Years
Daily	Lake Cullulleraine	19	5	2000–2024
	Coonawarra Penfolds	20	5	2000–2024
Daily + LTA	Lake Cullulleraine	36	5	2000–2024
	Coonawarra Penfolds	38	5	2000–2024

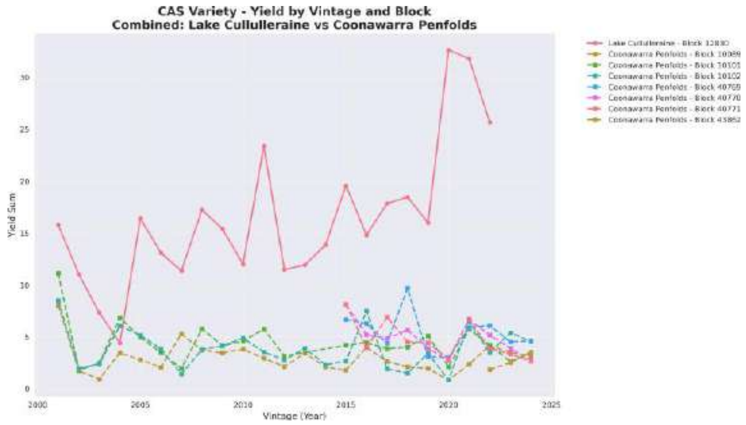


Fig. 83: CAS variety yield across Lake Cullulleraine and Coonawarra Penfolds blocks (2000–2024).

- Yield Records: Ground-truth annual grape yield data are used as prediction targets. Records are available for both Lake Cullulleraine and Coonawarra Penfolds, covering the period 2000–2024.

Three primary data modalities were integrated to support the analysis, as summarised in Table 35.

33.2 Yield Variation across Varieties, Blocks, and Regions

Fig. 83 to Fig. 87 illustrate the temporal yield dynamics (2000–2024) for five grape varieties (CAS, CHR, MER, SEMI, SHI) across different vineyard blocks and regions (Lake Cullulleraine vs. Coonawarra Penfolds). Several key observations can be made:

- CAS Variety: Lake Cullulleraine (Block 12830) consistently outperforms all Coonawarra Penfolds blocks, reaching peak yields above 30 in recent years. In contrast, Coonawarra Penfolds blocks remain below 10, suggesting strong regional constraints (Fig. 83).

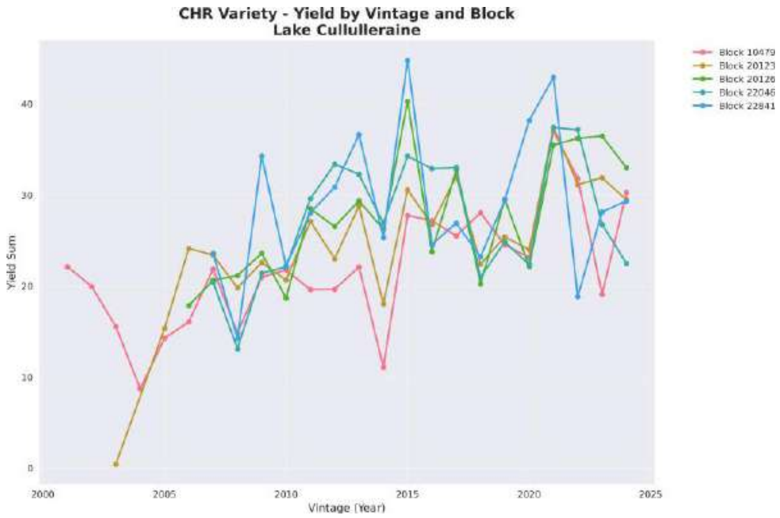


Fig. 84: CHR variety yield across Lake Cullulleraine blocks (2000–2024).

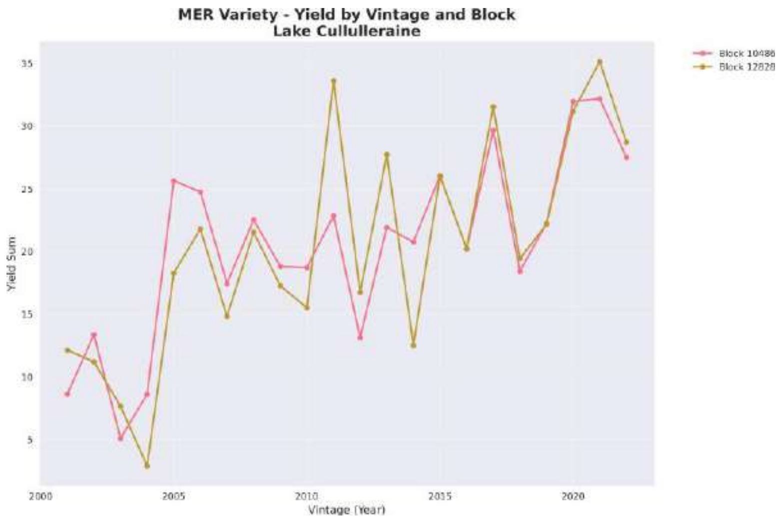


Fig. 85: MER variety yield across Lake Cullulleraine blocks (2000–2024).

- CHR Variety: Multiple Lake Cullulleraine blocks (10479, 20123, 20126, 22046, 22841) show an upward trajectory since 2005, with yields exceeding 40 in peak years, indicating high productivity potential (Fig. 84).
- MER Variety: Two Lake Cullulleraine blocks (10486, 12828) display highly synchronised trends, stabilizing around 20–30, with peaks surpassing 35 in the 2020s (Fig. 85).

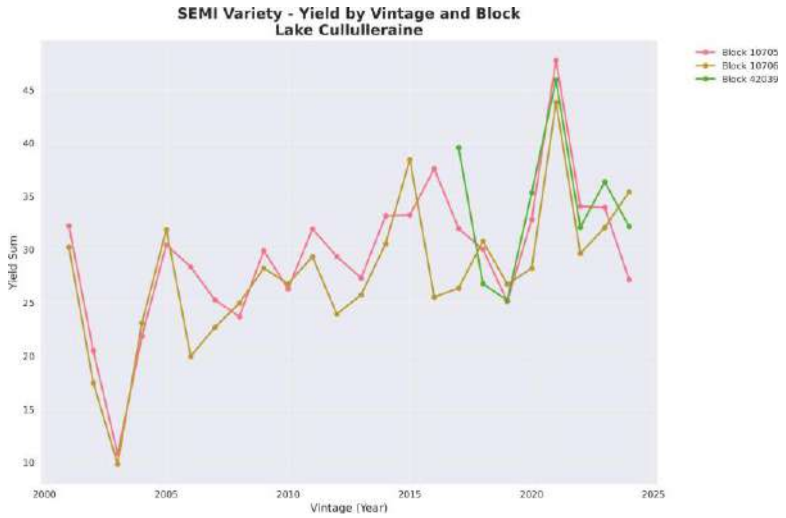


Fig. 86: SEMI variety yield across Lake Cullulleraine blocks (2000–2024).

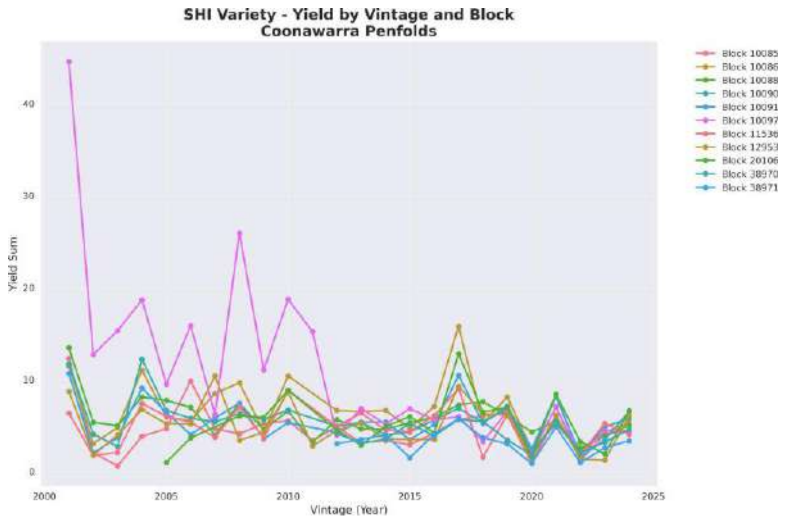


Fig. 87: SHI variety yield across Coonawarra Penfolds blocks (2000–2024).

- SEMI Variety: Yields are highly variable but show sharp increases after 2015, with some blocks exceeding 45. This suggests high responsiveness to favourable conditions but also greater inter-annual volatility (Fig. 86).
- SHI Variety: Coonawarra Penfolds blocks exhibit generally low yields (<15), with the exception of Block 10097, which had an unusually high spike in early vintages. No long-term upward trend is observed (Fig. 87).

Overall, Lake Cullulleraine vineyards demonstrate significantly higher yield potential across CHR, MER, and SEMI, while Coonawarra Penfolds blocks (CAS, SHI) remain yield-limited. This indicates strong region-variety interactions, highlighting the importance of location-specific yield prediction and management strategies.

Beyond these general trends, several additional patterns were identified:

1. Large yield differences across blocks: Significant variability exists even within the same variety and region. For instance, the SEMI blocks show substantial yield differences in 2015, and CHR yields vary widely in 2001. For SHI, yield disparities between 2001 and 2011 are particularly evident.
2. Data coverage imbalance: Some vineyard blocks have limited historical records, which may influence model stability and the robustness of temporal trend analysis. Examples include SEMI Block 42039, CHR Block 22841, and CAS Blocks 40769, 40770, and 40771.
3. Regional contrasts: Substantial yield differences are observed between regions. Lake Cullulleraine generally exhibits higher and more stable yields, while Coonawarra Penfolds yields are lower and more variable, suggesting that climate and soil factors play a major role in determining production potential.

These findings emphasise the heterogeneity of vineyard yield behaviour across both spatial (block and region) and temporal dimensions, reinforcing the necessity of modelling yield at multiple levels of granularity for accurate prediction and interpretation.

Implications for Modelling

The observed yield heterogeneity across varieties, blocks, and regions provides valuable guidance for model design. First, the large yield differences between blocks (e.g., SEMI 2015, CHR 2001, SHI 2001–2011) highlight the need for a model capable of learning localised spatio-temporal dependencies rather than relying on global averages. This justifies the adoption of a 2-dimensional convolutional framework, which can capture both temporal patterns and cross-feature interactions within each growing season.

Second, the unbalanced data coverage across blocks (e.g., SEMI Block 42039, CHR Block 22841, CAS Blocks 40769–40771) suggests that model regularisation and interpretability are essential to prevent overfitting to sparsely sampled years. The incorporation of attention mechanisms (CBAM) enables the model to adaptively focus on the most informative feature-time regions, compensating for uneven data density.

Finally, the strong regional contrasts between Lake Cullulleraine and Coonawarra Penfolds indicate that climate-yield relationships are location-specific. This reinforces the necessity of integrating both intrinsic (attention-based) and post-hoc (gradient-based) explanation modules, allowing the model not only to achieve accurate predictions but also to provide interpretable insights into region-dependent yield drivers.

In summary, the variability patterns observed in the raw yield data directly motivate the use of an interpretable deep learning framework—combining CNN, CBAM, and Integrated Gradients—to effectively model heterogeneous spatio-temporal climate-yield relationships.

33.3 Input Configuration and Experimental Design

Two input configurations were designed to evaluate the model's ability to capture both short-term and long-term climate effects on grape yield:

- Daily input: Uses only daily meteorological observations (e.g., temperature, rainfall, humidity, and radiation). This configuration reflects short-term weather dynamics within a single growing season.
- Daily with LTA input: Extends the daily inputs with corresponding LTA for each feature, allowing the model to learn deviations from climatological norms and better generalise under long-term variability.

Experimental Design

Experiments were conducted at two levels of granularity:

1. *Variety-specific experiments:* Models were trained and evaluated separately for five grape varieties — CAS, CHR, MER, SEMI, and SHI — to examine variety-dependent responses to climate variability. These experiments capture intra-varietal sensitivity and facilitate interpretability comparisons across different grape types. Among these varieties, the SEMI variety was selected as the representative example for detailed visualisation and interpretability analysis, owing to its stable yield patterns and rich temporal coverage.
2. *Combined-variety experiments:* Data from all grape varieties within each region (Lake Cullulleraine and Coonawarra Penfolds) were merged to form region-level models. This design evaluates cross-varietal generalisation and regional-scale predictive robustness.

34 Results and Analysis on SILO Climate Data

This section presents results progressively from quantitative performance to qualitative interpretability, and finally to region-specific generalisation. For each analysis aspect, both daily and daily with LTA input configurations are compared.

34.1 Overall Performance across Regions and Varieties

This subsection reports the quantitative results of yield prediction performance, measured by mean absolute error (MAE), across different grape varieties and regions.

34.1.1 Daily Input Results

Fig. 88 and Fig. 89 summarise the model's mean absolute error (MAE) under the daily input configuration, across grape varieties and years for the Lake Cullulleraine and Coonawarra Penfolds regions, respectively. Overall, Coonawarra Penfolds exhibits lower MAE values and more stable inter-annual performance compared with Lake Cullulleraine. Notably, CAS 2020 (Lake Cullulleraine) and CAS 2015 (Coonawarra Penfolds) appear as clear outliers, largely due to limited data coverage. These cases highlight the model's sensitivity to data sparsity and extreme yield anomalies.

34.1.2 Daily with LTA Input Results

To further evaluate the model's generalisation capability, we extended the input configuration from daily-only features to combined Daily with LTA features. Fig. 90 and Fig. 91 summarize the year-wise MAE across grape varieties in Lake Cullulleraine and Coonawarra Penfolds, respectively. However, the inclusion of LTA-based deviations does not consistently stabilise

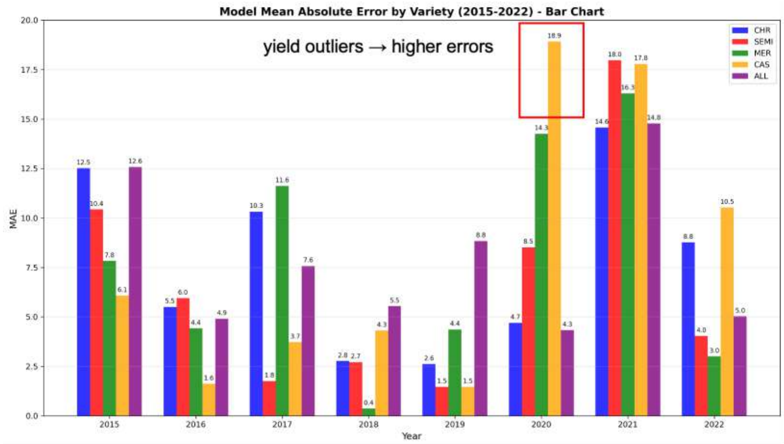


Fig. 88: Mean absolute error (MAE) by grape variety for Lake Cullulleraine (2015–2022) under daily input.

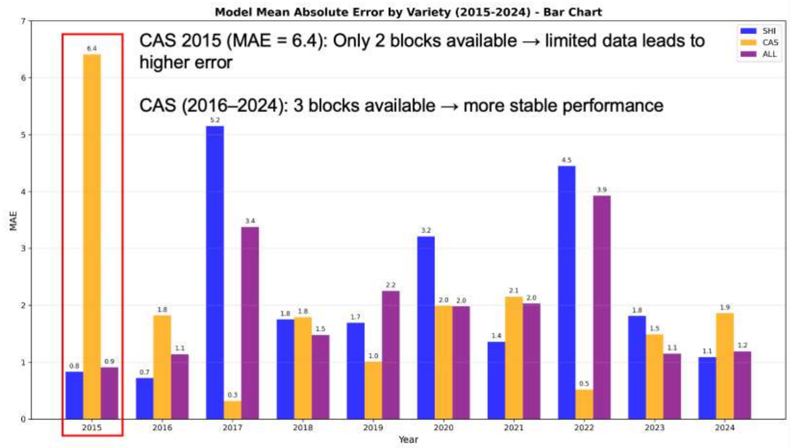


Fig. 89: Mean absolute error (MAE) by grape variety for Coonawarra Penfolds (2015–2024) under daily input.

model behaviour, indicating that long-term climatological averages provide only marginal benefits in improving robustness under varying climatic conditions.

34.2 Interpretability Analysis on Representative Variety (SEMI, 2022)

To further examine model behaviour, IG and CBAM attention maps were analysed for representative cases. We selected SEMI in 2022 as a reference example.

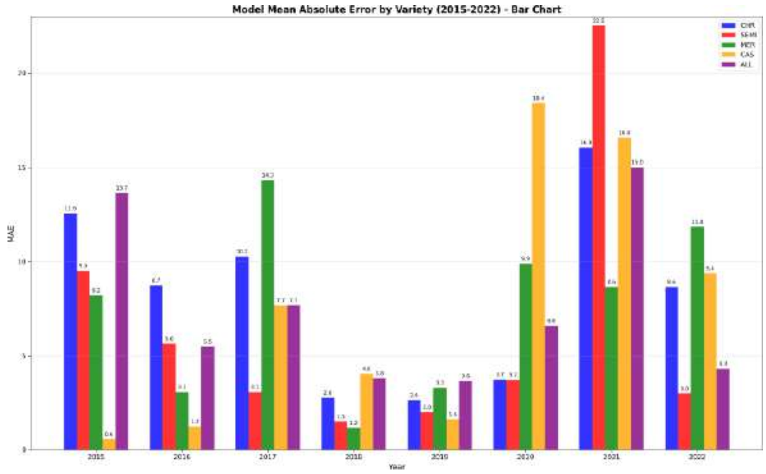


Fig. 90: Mean absolute error (MAE) by grape variety for Lake Cullulleraine (2015–2022) under daily with LTA input.

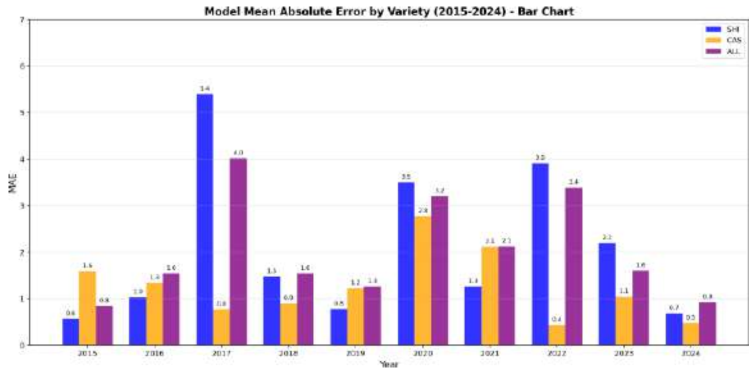


Fig. 91: Mean absolute error (MAE) by grape variety for Coonawarra Penfolds (2015–2024) under daily with LTA input.

34.2.1 Daily Input Results

Fig. 92 to Fig. 97 illustrate a progressive view of model explanations, from intrinsic attention to gradient-based attributions.

Layer-wise and Fused Attention Analysis

Fig. 92 shows the featural-temporal attention maps extracted from three CBAM layers. The first layer captures early-stage temporal fluctuations and broad feature dependencies, while deeper layers emphasise specific growth phases such as *Budbreak* and *Fruit Set*. Attention becomes increasingly localised and sharper in deeper layers, indicating stronger focus on critical features and time periods.

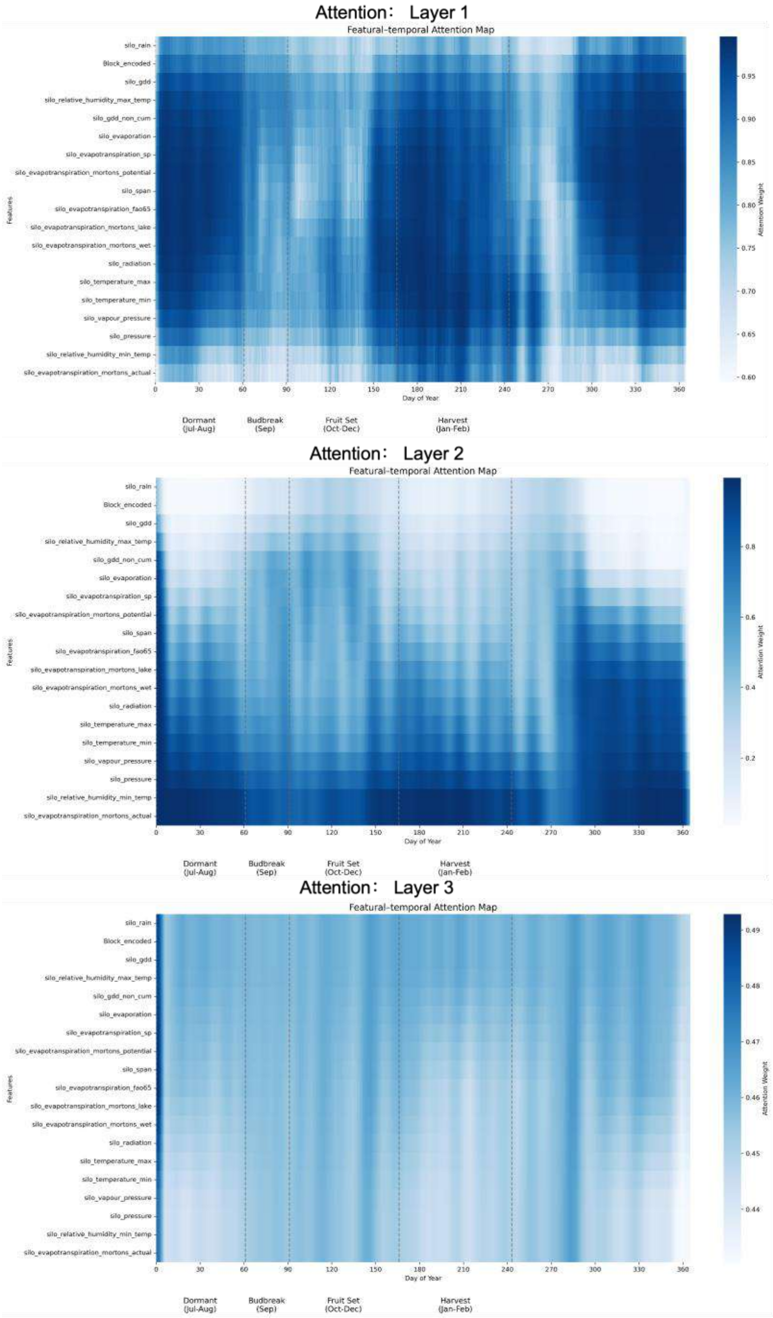


Fig. 92: Layer-wise CBAM attention maps (SEMI, 2022, daily input).

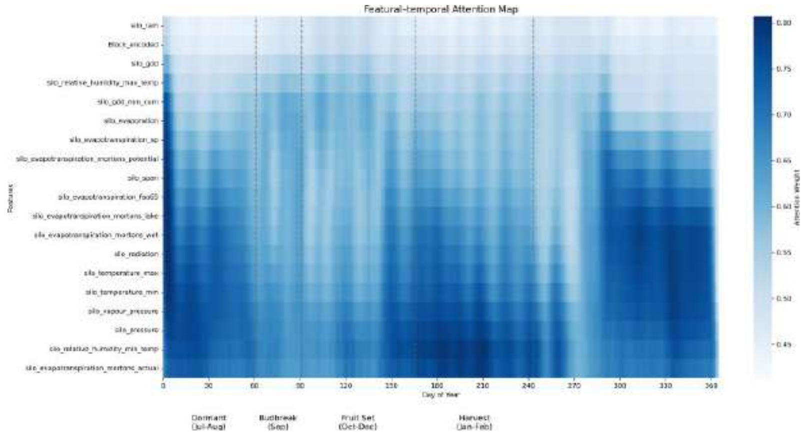


Fig. 93: Fused attention map (SEMI, 2022, daily input).

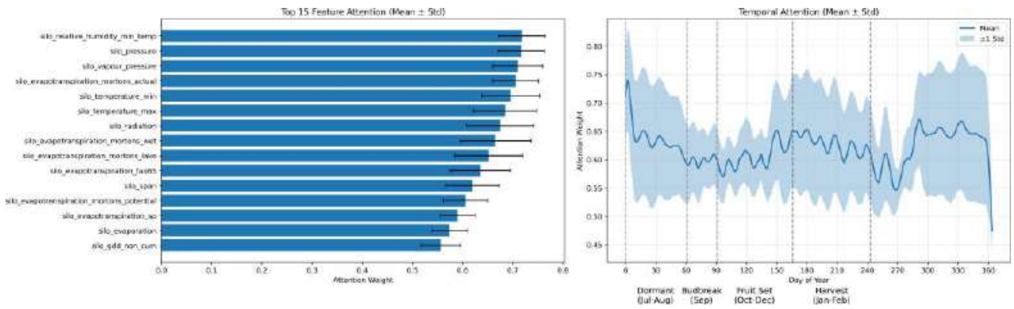


Fig. 94: Feature-wise (left) and temporal (right) summaries of fused attention map (SEMI, 2022, daily input).

After upsampling and averaging all CBAM attention layers, a fused featural-temporal attention map is obtained (Fig. 93). The fused map integrates multi-level saliency, revealing consistent high attention around the *Harvest* stage. To further quantify these patterns, Fig. 94 decomposes the fused attention into separate feature-wise and time-wise statistics. Feature-wise attention ranks *humidity*, *vapour pressure*, and *temperature* as the most influential factors, while temporal attention peaks during the late-season period.

IG and Aggregated IG Analysis

Fig. 95 to Fig. 97 show IG-based explanations for the same model. Local IG maps (Fig. 95) capture sample-specific feature-time attributions, whereas the aggregated IG map (Fig. 96) represents the global explanatory structure by averaging across all test samples. The decomposed view (Fig. 97) confirms that GDD, evapotranspiration, and humidity remain the most influential variables. Temporal attribution peaks during the Harvest and late-season period, closely matching the attention-derived saliency pattern.

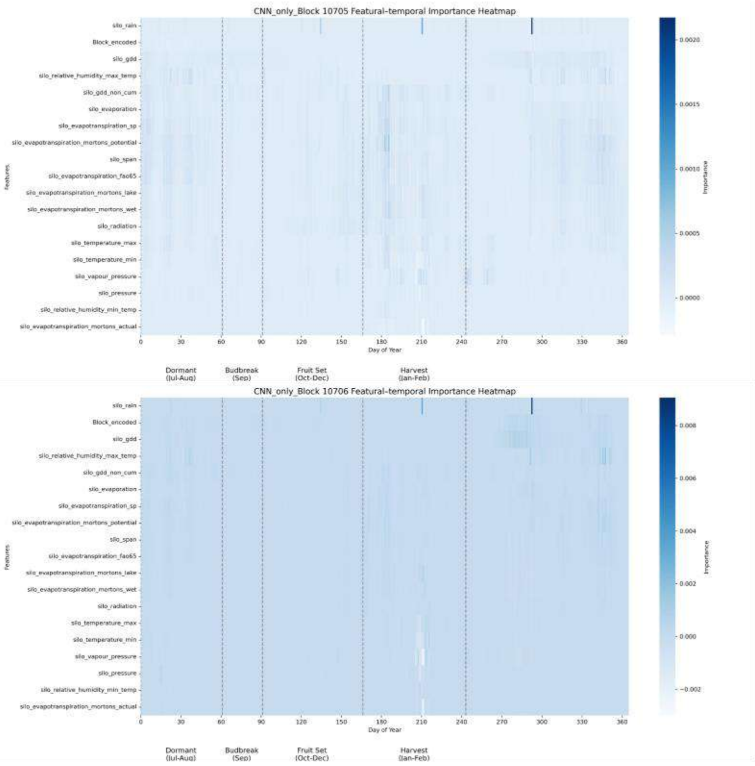


Fig. 95: Instance-level Integrated Gradients (IG) attributions. Each map corresponds to one test sample (SEMI, 2022, daily input).

Fused Attention vs. Aggregated IG Comparison

The fused CBAM attention and Aggregated IG maps provide complementary interpretability perspectives:

- CBAM Attention: an intrinsic, end-to-end focus showing where the model *looks* within the feature-time space.
- Aggregated IG: a post-hoc, gradient-based attribution showing what truly *drives* model predictions.

To better understand the consistency between the two explanation methods, we com-pared the top 15 features ranked by fused attention (Fig. 94) and aggregated IG (Fig. 97). The summary focuses on climate-related variables that exhibit a strong influence on yield prediction, while the non-climatic categorical feature *Block_encoded* was excluded from the comparison. The final summary of feature importance overlap is presented in Table 36.

Their strong spatial-temporal consistency, particularly during the Harvest stage and late-season period, demonstrates that the model's learned attention closely aligns with the attributions of predictions to input features. Minor discrepancies, such as the broader IG sensitivity to silo_gdd, silo_rain, and silo_relative_humidity_max_temp, indicate that IG

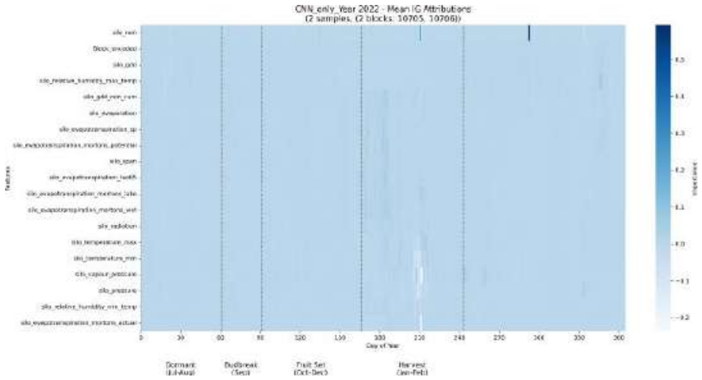


Fig. 96: Aggregated IG attributions averaged across all test samples, representing global explanatory patterns (SEMI, 2022, daily input).

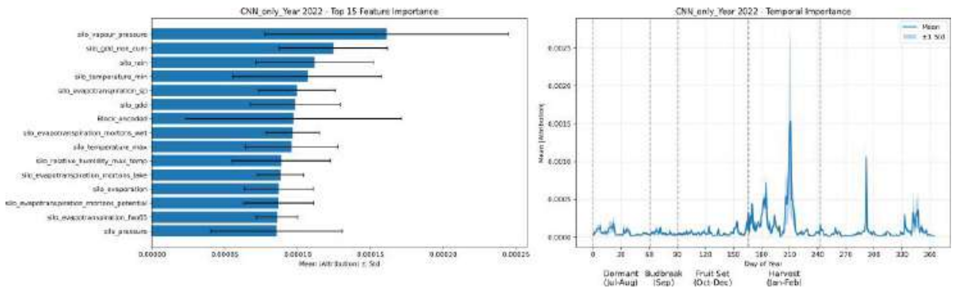


Fig. 97: Feature-wise (left) and temporal (right) decompositions of Aggregated IG attributions (SEMI, 2022, daily input).

captures additional climatic cues not strongly emphasised by CBAM. Therefore, integrating both perspectives provides complementary insights, enhancing the overall interpretability and trustworthiness of the model.

Discussion

The strong overlap between Aggregated IG and CBAM Attention confirms that both intrinsic and post-hoc explanations capture consistent yield-driving mechanisms. Yet, their distinct focuses highlight complementary interpretive strengths. CBAM reflects short-term physiological sensitivity, while IG reveals longer-term, accumulative stress patterns. Together, they offer a more comprehensive understanding of how climatic dynamics shape the variability of grape yield.

34.2.2 Daily with LTA Input Results

Layer-wise and Fused Attention Analysis

Fig. 98 to Fig. 100 present the featural-temporal attention results of the CBAM-based model trained with combined Daily with LTA inputs. Compared with the daily-only model, attention

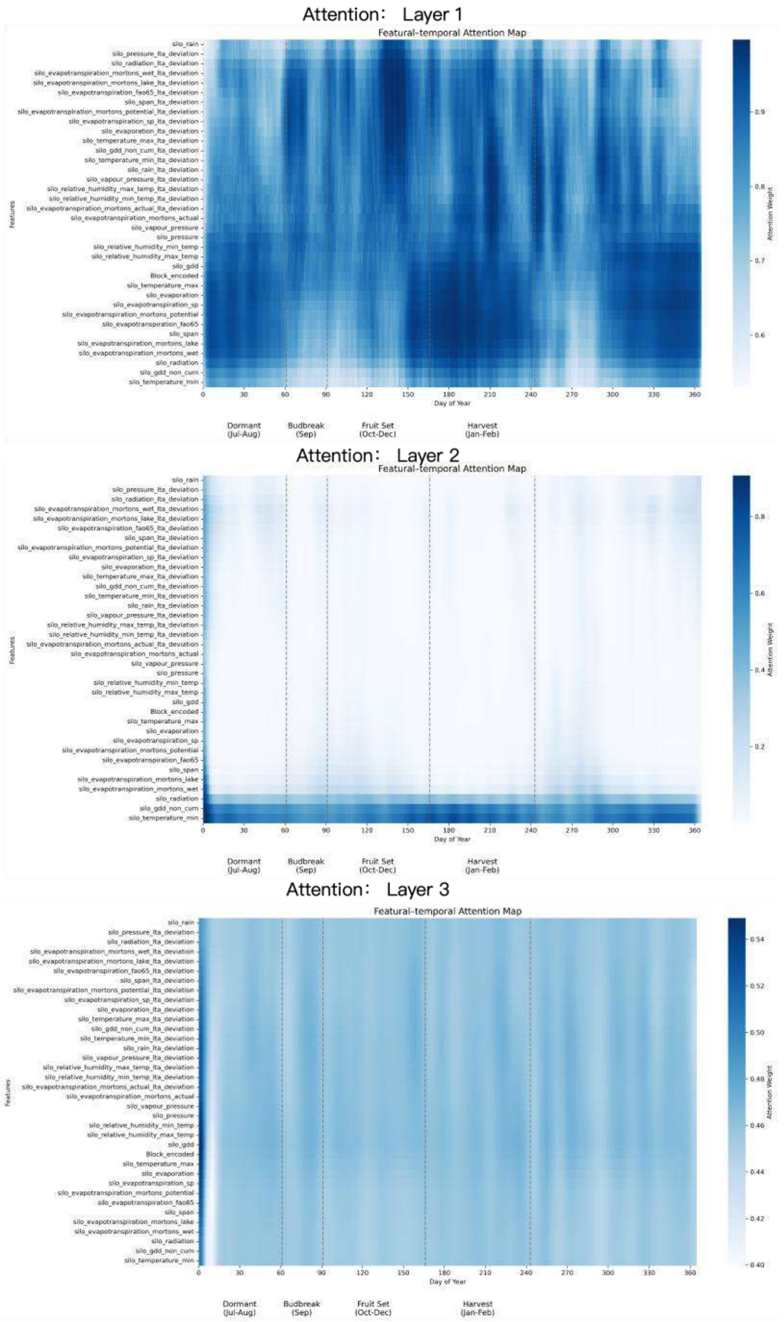


Fig. 98: Layer-wise CBAM attention maps (SEMI, 2022, daily with LTA input).

Table 36: Comparison of top-ranked climate features in aggregated IG attributions and fused attention (SEMI, 2022, daily input). "Aggregated IG Only" denotes features captured exclusively by the aggregated IG method; "Common Features" represent those shared by both explanation methods; and "Fused Attention only" refers to features highlighted solely by fused attention.

Aggregated IG Only	Common Features	Fused Attention Only
silo_gdd	silo_vapour_pressure	silo_evapotranspiration_mortons_actual
silo_rain	silo_gdd_non_cum	silo_span
silo_relative_humidity_max_temp	silo_temperature_min	silo_relative_humidity_min_temp
	silo_evapotranspiration_sp	
	silo_evapotranspiration_mortons_wet	
	silo_temperature_max	
	silo_evapotranspiration_mortons_lake	
	silo_evaporation	
	silo_evapotranspiration_mortons_potential	
	silo_evapotranspiration_fao65	
	silo_pressure	

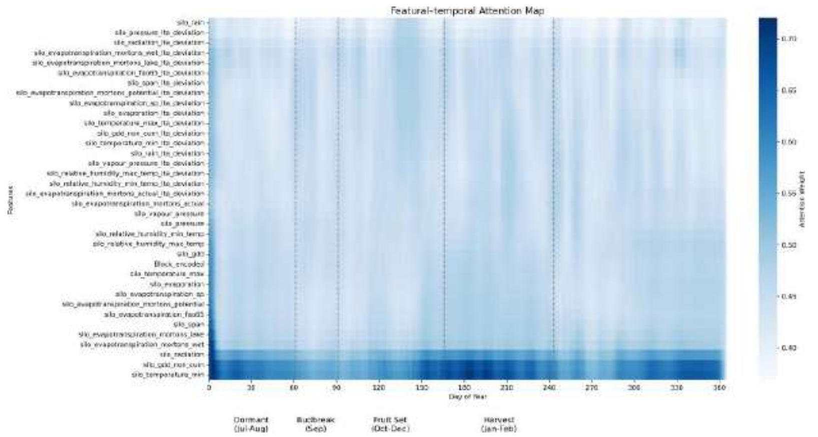


Fig. 99: Fused attention map (SEMI, 2022, daily with LTA input).

patterns become more structured and emphasise not only short-term climatic fluctuations but also long-term deviation effects.

Fig. 98 shows the layers-wise CBAM attention maps. The fused attention map (Fig. 99) integrates these multi-level representations and exhibits stronger concentration on rain-, pressure-, and evapotranspiration-related features.

Feature and time-wise attention summaries are shown in Fig. 100. Feature-wise importance highlights silo_temperature_min, silo_evapotranspiration_fa065, and silo_gdd as the most influential variables, whereas temporal attention peaks during Harvest growth stage.

IG and Aggregated IG Analysis

Fig. 101 to Fig. 103 display the IG-based explanations for the same model. Local IG attributions (Fig. 101) reveal that both daily and LTA deviation features contribute interactively during key phenological stages. When averaged across all test blocks, the aggregated IG map (Fig. 102) shows well-defined high-importance features corresponding to growth stages.

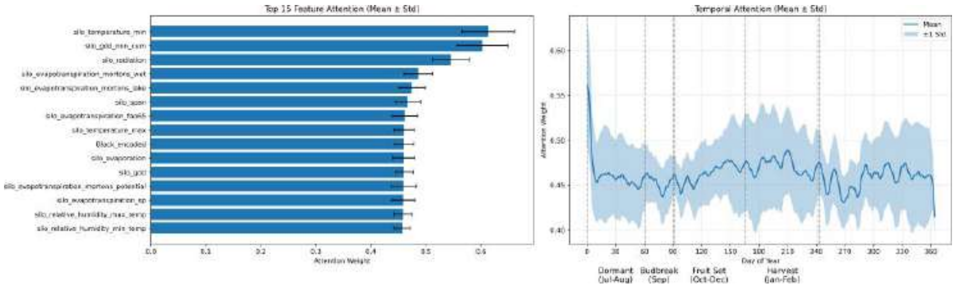


Fig. 100: Feature-wise (left) and temporal (right) summaries of fused attention map (SEMI, 2022, daily with LTA input).

The feature-wise and temporal decomposition of aggregated IG (Fig. 103) highlights the dominant explanatory role of vapour pressure deviation, rainfall deviation, and evaporation deviation, demonstrating that the model captures yield variation as a function of both short-term weather and persistent climate anomalies.

Fused Attention vs. Aggregated IG Comparison

Both interpretability methods reveal strong mid-to-late season alignment, with overlapping emphasis on evapotranspiration- and temperature-related deviation variables. Both meth-ods exhibit a consistent temporal focus around the *Fruit Set* and *Harvest* stages, suggesting that the internal attention mechanisms and the external gradient-based attributions con-verge on key phenological periods. Minor discrepancies—such as IG's broader sensitivity to deviations—indicate that IG tends to capture cumulative climatic effects, whereas attention emphasises transient but highly salient events.

To further assess the consistency between the two explanation methods, we compared the top 15 features ranked by fused attention (Fig. 100) and aggregated IG (Fig. 103). The comparison focuses on climate-related variables that strongly influence yield prediction, while the non-climatic categorical feature was excluded from the analysis. A consolidated summary of overlapping and distinct feature importance patterns is presented in Table 37. Specifically, although the top 15 features are not completely identical, most of the variables highlighted by aggregated IG correspond to the long-term deviation counterparts of the features emphasised by fused attention. This observation suggests that IG explanations tend to focus on anomalous climatic deviations, whereas attention emphasises the direct, instantaneous effects of the same underlying variables.

34.3 Regional Interpretability Analysis: Lake Cullulleraine

34.3.1 Daily Input Results

For the Lake Cullulleraine region, variety-specific predictions were aggregated to perform a region-level interpretability analysis. Since the test set contains multiple vineyard blocks and grape varieties, we focus on the fused CBAM attention and aggregated IG results to capture the dominant explanatory patterns across all samples. The goal is to identify common feature-time dependencies that drive regional yield variability under the daily-input configuration.

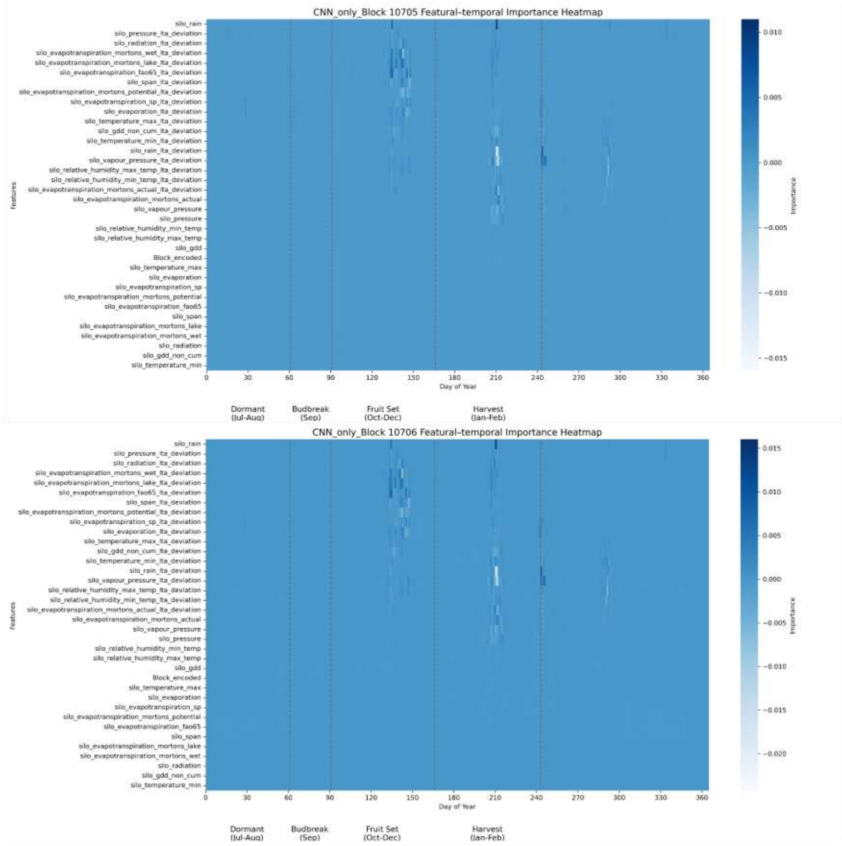


Fig. 101: Instance-level Integrated Gradients (IG) attributions. Each map corresponds to one test sample (SEMI, 2022, daily with LTA input).

Fused Attention (Intrinsic Explanation)

Fig. 104 shows the fused featural-temporal attention map derived from all CBAM layers. High attention weights are concentrated during the *Dormant* and *Harvest* periods, indicating that the model places greater focus on early and late season climate conditions that are critical for grape maturation. The feature-wise and temporal decompositions in Fig. 105 reveal that *humidity*, *vapour pressure*, and *temperature extremes* receive consistently high attention scores, reflecting the physiological importance of thermal and moisture balance in grape yield formation. Temporal attention also exhibits two distinct peaks around *Dormant* and *late-season*, suggesting sensitivity to both early vegetative growth and final stages.

Aggregated IG (Post-hoc Explanation)

The aggregated IG heatmap (Fig. 106) visualises the global attributions of each feature-day pair averaged across all test samples. Similar to attention, the IG-based importance distribution highlights *temperature extreme*, *vapour pressure*, and *humidity* as dominant explanatory variables, particularly during late-season stages. The feature-wise and temporal decomposition of aggregated IG in Fig. 107 confirms that the highest attributions occur during *Harvest*,

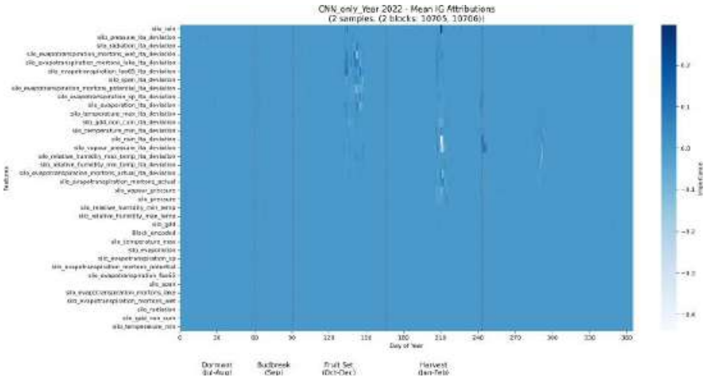


Fig. 102: Aggregated IG attributions averaged across all test samples, representing global explanatory patterns (SEMI, 2022, daily with LTA input).

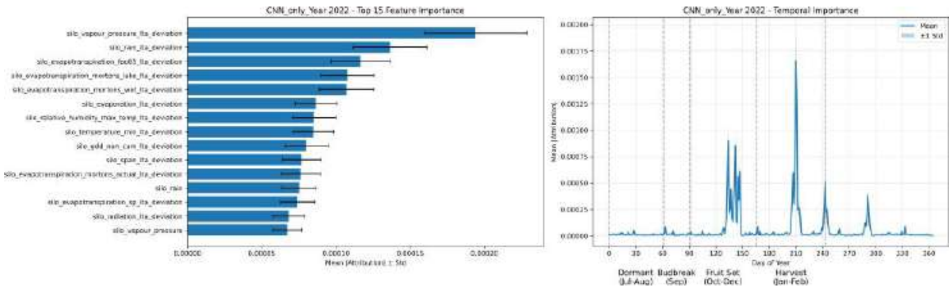


Fig. 103: Feature-wise (left) and temporal (right) decompositions of Aggregated IG attributions (SEMI, 2022, daily with LTA input).

corresponding to key developmental phases where climate variability has the largest impact on final yield.

Cross-comparison and Insights

We summarise the top 15 features ranked by fused attention (Fig. 105) and aggregated IG (Fig. 107). The comparison focuses on climate-related variables that strongly influence yield prediction, while the non-climatic categorical feature was excluded from the analysis. A consolidated summary of overlapping and distinct feature importance patterns is presented in Table 38, illustrating the degree of consistency between intrinsic (attention-based) and post-hoc (gradient-based) explanations.

Overall, both fused CBAM attention and aggregated IG exhibit consistent saliency patterns, particularly along the feature dimension, indicating strong agreement between intrinsic and post-hoc explanations and reinforcing the reliability of the model's interpretability. This alignment between intrinsic and post-hoc explanations strengthens confidence in the model's learned mechanisms and demonstrates that Lake Cullulleraine's yield dynamics are primarily

Table 37: Comparison of top-ranked climate features in aggregated IG attributions and fused attention (SEMI, 2022, daily with LTA input). "Aggregated IG Only" denotes features captured exclusively by the aggregated IG method; "Common Features" represent those shared by both explanation methods; and "Fused Attention only" refers to features highlighted solely by fused attention.

Aggregated IG Only	Common Features	Fused Attention Only
silo_temperature_min_lta_deviation		silo_temperature_min
silo_gdd_non_cum_lta_deviation		silo_gdd_non_cum
silo_radiation_lta_deviation		silo_radiation
silo_evapotranspiration_mortons_wet_lta_deviation		silo_evapotranspiration_mortons_wet
silo_evapotranspiration_mortons_lake_lta_deviation		silo_evapotranspiration_mortons_lake
silo_span_lta_deviation		silo_span
silo_evapotranspiration_fao65_lta_deviation		silo_evapotranspiration_fao65
silo_evaporation_lta_deviation		silo_evaporation
silo_evapotranspiration_sp_lta_deviation		silo_evapotranspiration_sp
silo_relative_humidity_max_temp_lta_deviation		silo_relative_humidity_max_temp
silo_rain_lta_deviation		silo_relative_humidity_min_temp
silo_vapour_pressure_lta_deviation		silo_temperature_max
silo_vapour_pressure		silo_gdd
silo_evapotranspiration_mortons_actual_lta_deviation		silo_evapotranspiration_mortons_potential
silo_rain		

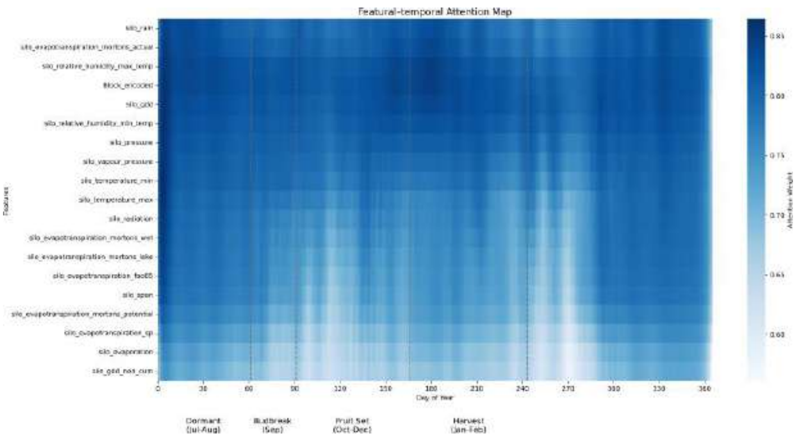


Fig. 104: Fused attention map (Lake Cullulleraine, 2022, daily input).

driven by humidity-temperature interactions and evapotranspiration processes during critical phenological stages.

34.3.2 Daily with LTA Input Results

Fused Attention (Intrinsic Explanation)

Fig. 108 presents the fused featural-temporal attention map, while Fig. 109 provides its feature-wise and temporal summaries. According to the ranked top 15 features in Fig. 109, the fused attention mainly emphasises evapotranspiration-, humidity-, and radiation-related variables, indicating that these processes dominate the model's intrinsic focus under the Daily with LTA input setting.

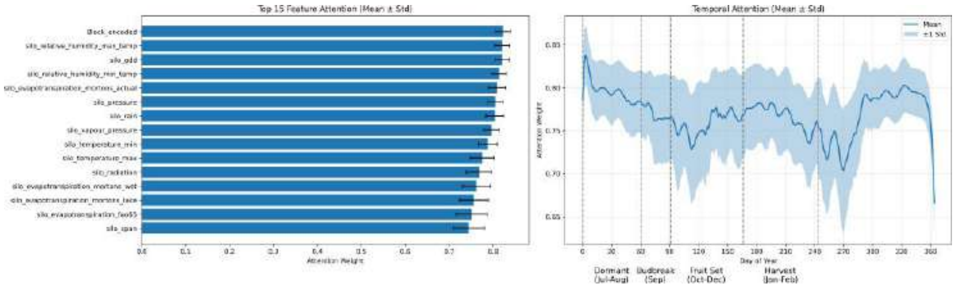


Fig. 105: Feature-wise (left) and temporal (right) summaries of fused attention map (Lake Cullulleraine, 2022, daily input).

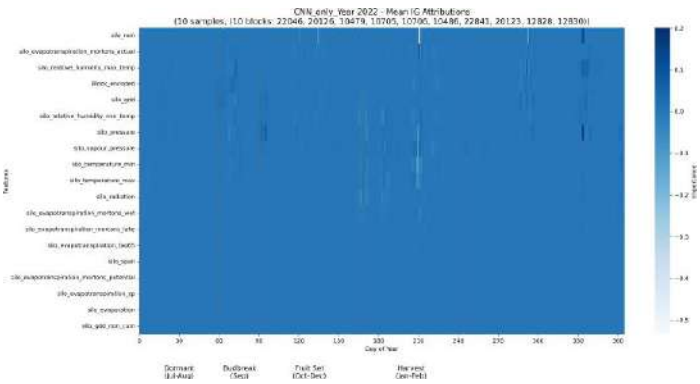


Fig. 106: Aggregated IG attributions averaged across all test samples, representing global explanatory patterns (Lake Cullulleraine, 2022, daily input).

Aggregated IG (Post-hoc Explanation)

Fig. 110 to Fig. 111 show the aggregated IG heatmap and its feature/time decompositions. The IG results reveal concentrated importance near the *Harvest* stage, with strong contribu-tions from humidity, evapotranspiration, and radiation-related LTA deviations. This pattern suggests that the model captures long-term climatic anomalies that significantly affect final yield outcomes.

Cross-comparison and Insights

We summarise the top 15 features ranked by fused attention (Fig. 109) and aggregated IG (Fig. 111). The comparison focuses on climate-related variables that strongly influence yield prediction, while the non-climatic categorical feature was excluded from the analysis. A con-solidated summary of overlapping and distinct feature importance patterns is presented in Table 39. Specifically, fused attention and aggregated IG consistently highlight evapotran-spiration, humidity, and radiation as key drivers, while showing complementary emphases: attention reflects stable season-long dependencies, whereas IG captures long-term deviation features. This complementarity suggests that incorporating LTA inputs helps disentangle

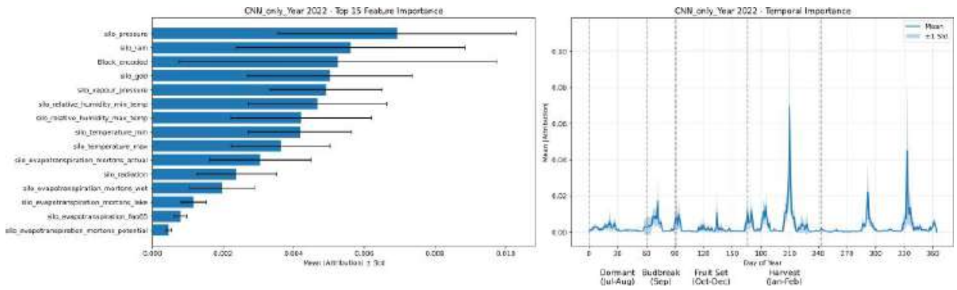


Fig. 107: Feature-wise (left) and temporal (right) decompositions of Aggregated IG attributions (Lake Cullulleraine, 2022, daily input).

Table 38: Comparison of top-ranked climate features in aggregated IG attributions and fused attention (Lake Cullulleraine, 2022, daily input). "Aggregated IG Only" denotes features captured exclusively by the aggregated IG method; "Common Features" represent those shared by both explanation methods; and "Fused Attention only" refers to features highlighted solely by fused attention.

Aggregated IG Only	**Common Features**	**Fused Attention Only**
silo_evapotranspiration_mortons_potential	silo_pressure	silo_span
	silo_rain	
	silo_gdd	
	silo_vapour_pressure	
	silo_relative_humidity_min_temp	
	silo_relative_humidity_max_temp	
	silo_temperature_min	
	silo_temperature_max	
	silo_evapotranspiration_mortons_actual	
	silo_radiation	
	silo_evapotranspiration_mortons_wet	
	silo_evapotranspiration_mortons_lake	
	silo_evapotranspiration_fao65	

baseline climatology from intra-season variability, thereby improving interpretability and robustness.

34.4 Regional Interpretability Analysis: Coonawarra Penfolds

34.4.1 Daily Input Results

Fig. 112 to Fig. 115 present the interpretability analysis results for Coonawarra Pen-folds (2022) under the daily input configuration, including the fused CBAM attention and aggregated IG visualizations.

Fused Attention (Intrinsic Explanation)

The fused CBAM attention map (Fig. 112) reveals distinct featural-temporal focus patterns compared with Lake Cullulleraine. Attention is distributed more diffusely across time, indicating that the model experiences greater uncertainty in identifying dominant growth phases. Feature-wise attention ranking (Fig. 113) highlights evapotranspiration, chill days,

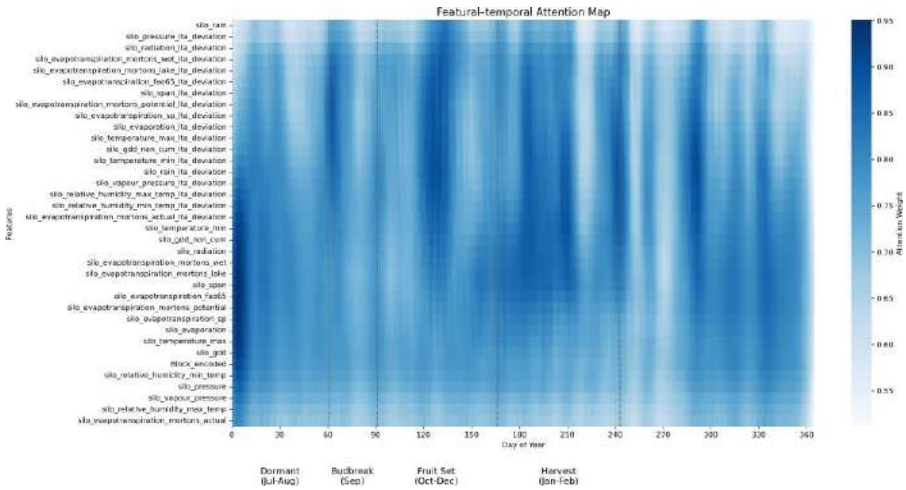


Fig. 108: Fused attention map (Lake Cullulleraine, 2022, daily with LTA input).

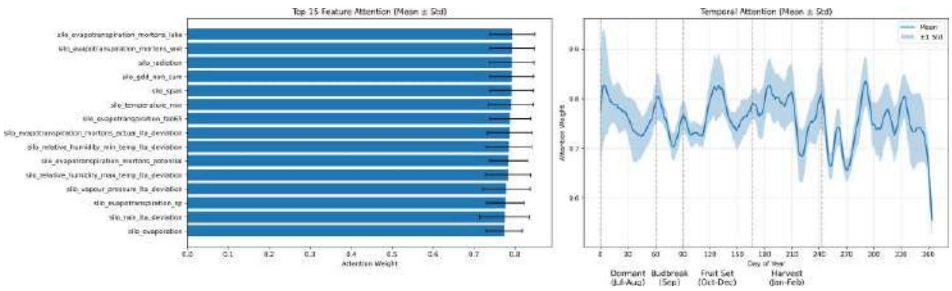


Fig. 109: Feature-wise (left) and temporal (right) summaries of fused attention map (Lake Cullulleraine, 2022, daily with LTA input).

and humidity-related variables as dominant contributors, reflecting the high climate variability characteristic of Coonawarra Penfolds. The temporal profile shows moderate attention across most of the growing season, without clear single-peak concentration.

Aggregated IG (Post-hoc Explanation)

The aggregated IG heatmap (Fig. 114 to Fig. 115) provides a complementary gradient-based perspective. The global IG distribution (Fig. 114) shows that most positive attributions concentrate around the *Harvest* period, aligning with key physiological development stages. The corresponding feature- and time-wise decompositions (Fig. 115) further reveal that humidity, vapour pressure, and temperature extremes contribute most strongly to yield variation.

Cross-Comparison and Insights

We summarise the top 15 features ranked by fused attention (Fig. 113) and aggregated IG (Fig. 115). The comparison focuses on climate-related variables that strongly influence

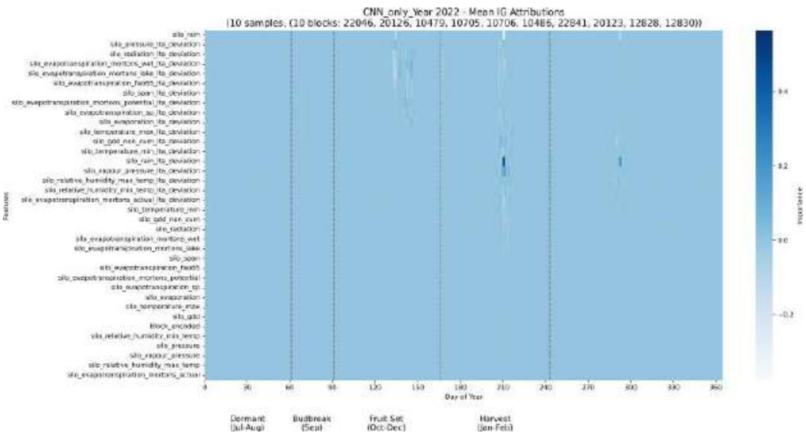


Fig. 110: Aggregated IG attributions averaged across all test samples, representing global explanatory patterns (Lake Cullulleraine, 2022, daily with LTA input).

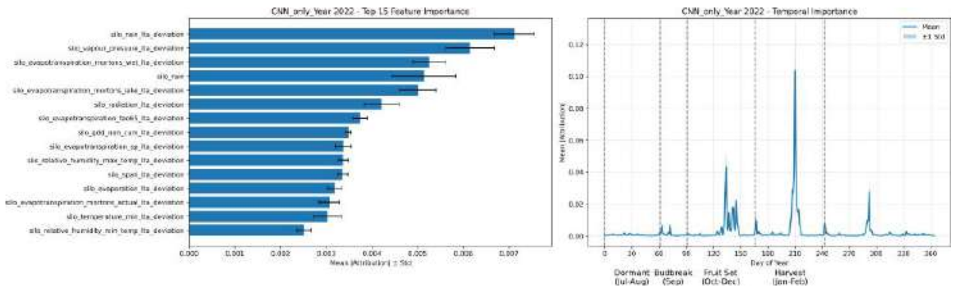


Fig. 111: Feature-wise (left) and temporal (right) decompositions of Aggregated IG attributions (Lake Cullulleraine, 2022, daily with LTA input).

yield prediction, while the non-climatic categorical feature was excluded from the analysis. A consolidated summary of overlapping and distinct feature importance patterns is presented in Table 40.

Both fused CBAM attention and aggregated IG reveal consistent climate-yield relationships, yet subtle regional differences emerge when compared with Lake Cullulleraine. Coonawarra's attention maps exhibit broader, less sharply localised temporal peaks, reflecting higher inter-annual climate variability and more complex soil-moisture interactions. Aggregated IG attributions align well with major phenological stages but extend further into the Harvest period, highlighting prolonged environmental influences.

Overall, both interpretability perspectives consistently emphasise the key roles of humidity, vapour pressure, and temperature-related variables. This convergence between intrinsic attention and post-hoc IG attribution strengthens confidence in the model's robustness and interpretability across diverse viticultural environments.

Table 39: Comparison of top-ranked climate features in aggregated IG attributions and fused attention (Lake Cullulleraine, 2022, daily with LTA input). “Aggregated IG Only” denotes features captured exclusively by the aggregated IG method; “Common Features” represent those shared by both explanation methods; and “Fused Attention only” refers to features highlighted solely by fused attention.

Aggregated IG Only	Common Features	Fused Attention Only
silo_evapotranspiration_mortons_lake_lta_deviation	silo_evapotranspiration_mortons_actual_lta_deviation	silo_evapotranspiration_mortons_lake
silo_evapotranspiration_mortons_wet_lta_deviation	silo_relative_humidity_min_temp_lta_deviation	silo_evapotranspiration_mortons_wet
silo_radiation_lta_deviation	silo_relative_humidity_max_temp_lta_deviation	silo_radiation
silo_gdd_non_cum_lta_deviation	silo_vapour_pressure_lta_deviation	silo_gdd_non_cum
silo_span_lta_deviation	silo_rain_lta_deviation	silo_span
silo_temperature_min_lta_deviation		silo_temperature_min
silo_evapotranspiration_fao65_lta_deviation		silo_evapotranspiration_fao65
silo_evapotranspiration_sp_lta_deviation		silo_evapotranspiration_sp
silo_evaporation_lta_deviation		silo_evaporation
silo_rain		silo_evapotranspiration_mortons_potential

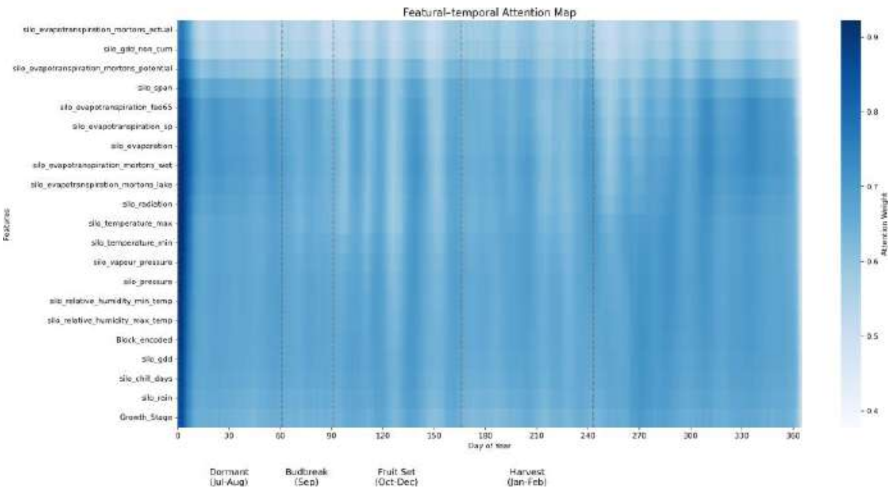


Fig. 112: Fused attention map (Coonawarra Penfolds, 2022, daily input).

34.4.2 Daily with LTA Input Results

Fused Attention (Intrinsic Explanation)

Fig. 116 shows the fused CBAM featural-temporal attention map for Coonawarra Penfolds (2022) under the daily with LTA input configuration. Compared with the daily-only model, this configuration yields stronger focus on long-term climate deviations, particularly in evapotranspiration- and humidity-related variables. The decomposed feature- and time-wise attention statistics (Fig. 117) confirm that *silo vapour pressure*, evapotranspiration indices, and humidity metrics dominate the importance ranking. Temporally, multiple peaks emerge from the Dormant stage through late Harvest, reflecting temperature and humidity fluctuations that critically influence grape development.

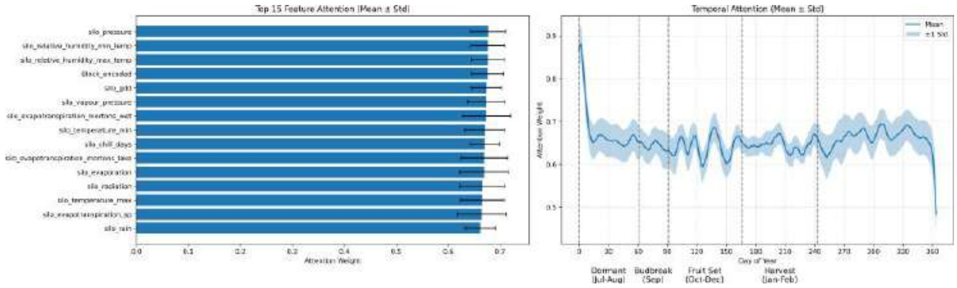


Fig. 113: Feature-wise (left) and temporal (right) summaries of fused attention map (Coonawarra Penfolds, 2022, daily input).

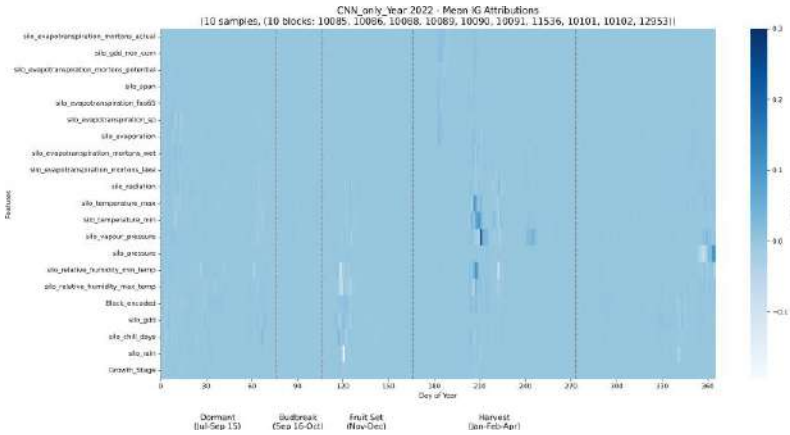


Fig. 114: Aggregated IG attributions averaged across all test samples, representing global explanatory patterns (Coonawarra Penfolds, 2022, daily input).

Aggregated IG (Post-hoc Explanation)

Fig. 118 visualises the aggregated IG attributions averaged across all vineyard blocks. Prominent attribution zones appear around *Harvest*, with strong activations for pressure-, vapour pressure-, and humidity-related variables. The decomposed feature/time profiles in Fig. 119 show that pressure-based features (*silo pressure*, *silo vapour_pressure*) and temperature-related deviations (*silo temperature_min*, *silo_temperature min lta_deviation*) exert persistent influence. Temporal IG peaks are observed during late *Fruit Set* and early *Harvest*, corresponding to stress-induced reductions in fruit set efficiency.

Cross-comparison and Insights

To further evaluate the consistency between the two explanation methods, we compared the top 15 features ranked by fused attention (Fig. 117) and aggregated IG (Fig. 119). The comparison focuses on climate-related variables that strongly influence yield prediction, while the non-climatic categorical feature was excluded. A consolidated summary of overlapping and distinct feature importance patterns is presented in Table 41.

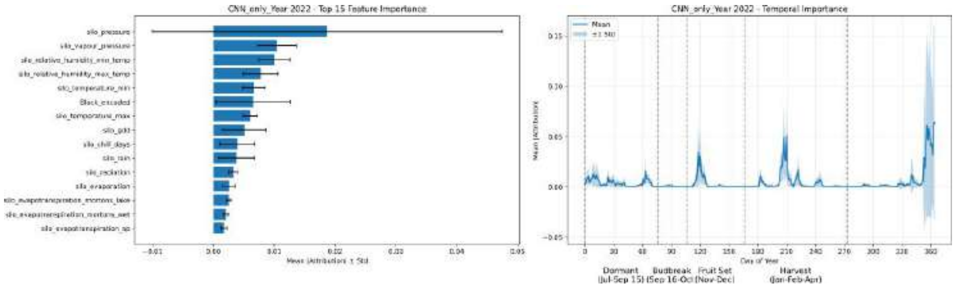


Fig. 115: Feature-wise (left) and temporal (right) decompositions of Aggregated IG attributions (Coonawarra Penfolds, 2022, daily input).

Table 40: Comparison of top-ranked climate features in aggregated IG attributions and fused attention (Coonawarra Penfolds, 2022, daily input). "Aggregated IG Only" denotes features captured exclusively by the aggregated IG method; "Common Features" represent those shared by both explanation methods; and "Fused Attention only" refers to features highlighted solely by fused attention.

Aggregated IG Only	Common Features	Fused Attention Only
	silo_pressure	
	silo_relative_humidity_min_temp	
	silo_relative_humidity_max_temp	
	silo_gdd	
	silo_vapour_pressure	
	silo_evapotranspiration_mortons_wet	
	silo_temperature_min	
	silo_chill_days	
	silo_evapotranspiration_mortons_lake	
	silo_evaporation	
	silo_radiation	
	silo_temperature_max	
	silo_evapotranspiration_sp	
	silo_rain	

The strong overlap between the top features extracted by both methods (Table 41) demonstrates stable explanatory consistency. Notably, most of the features uniquely identified by either method are associated with **evapotranspiration-related processes**, reinforcing their fundamental role in regulating water balance and yield formation.

34.5 Summary of Findings

Based on the comparative analysis of CBAM fused attention and aggregated IG across different regions (Lake Cullulleraine, and Coonawarra Penfolds) and input configurations (Daily and Daily with LTA), the following key findings are summarised.

Regional consistency under Daily input

In *Coonawarra Penfolds (Daily input)*, all top 15 features identified by CBAM and IG overlap, although their ranking orders differ. Similarly, in *Lake Cullulleraine (Daily input)*, all but one feature overlap between the two explanation methods, again with different importance

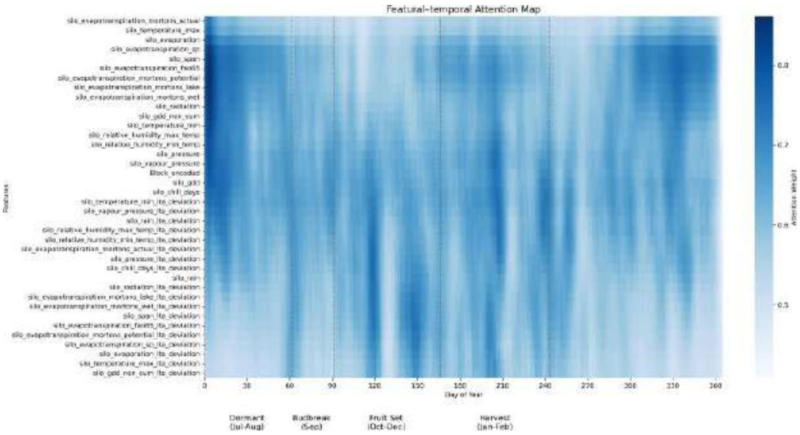


Fig. 116: Fused attention map (Coonawarra Penfolds, 2022, daily with LTA input).

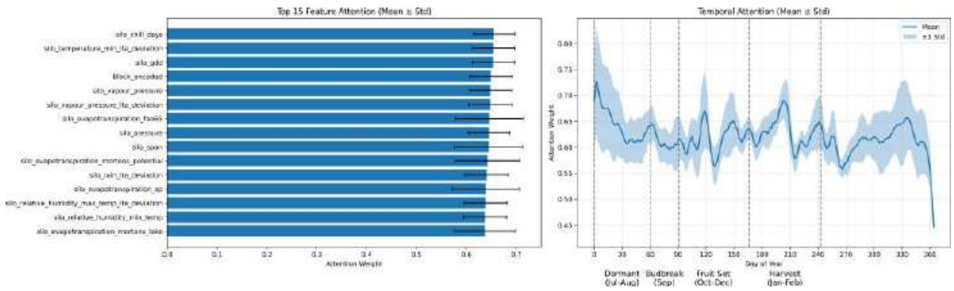


Fig. 117: Feature-wise (left) and temporal (right) summaries of fused attention map (Coonawarra Penfolds, 2022, daily with LTA input).

rankings. This indicates strong inter-method consistency in identifying key climatic drivers under the Daily configuration.

Feature alignment under LTA input

In *SEMI (2022, Daily with LTA input)*, there are no identical features between CBAM and IG in literal form, yet the IG-selected features correspond almost exactly to the LTA-deviation versions of CBAM's top 15 features. This demonstrates that when long-term deviations are introduced, both methods remain semantically aligned, capturing equivalent climatic effects expressed through anomaly-based variables.

Higher overlap in Daily-only settings

Across all regions, the *Daily-only* configuration consistently yields greater overlap between CBAM and IG compared with the *Daily with LTA* configuration. This suggests that while incorporating LTA deviations enhances anomaly sensitivity, it also diversifies the representation space, leading to fewer exact feature overlaps.

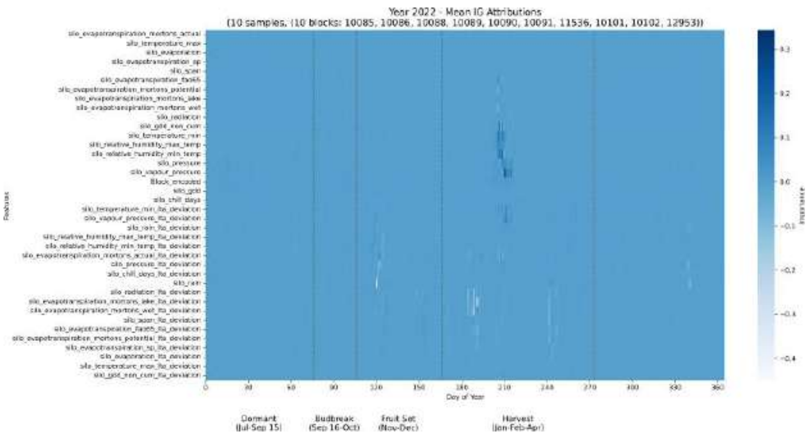


Fig. 118: Aggregated IG attributions averaged across all test samples, representing global explanatory patterns (Coonawarra Penfolds, 2022, daily with LTA input).

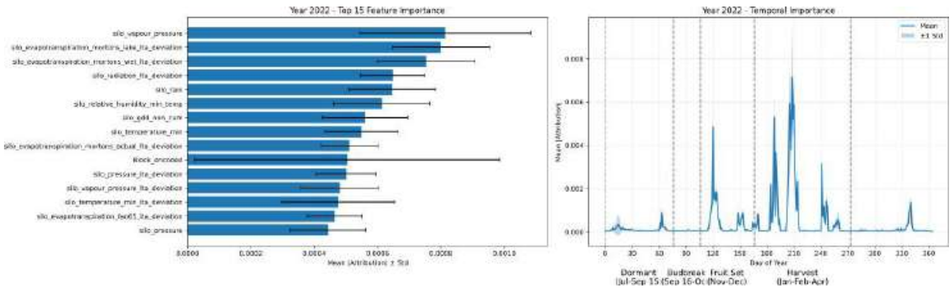


Fig. 119: Feature-wise (left) and temporal (right) decompositions of Aggregated IG attributions (Coonawarra Penfolds, 2022, daily with LTA input).

Cross-regional common features

Considering all regions and configurations together, CBAM fused attention and aggregated IG share a stable set of common top-ranked climatic features, as summarised in the Table 42. These results show that CBAM attention (intrinsic) and IG (post-hoc) converge on a consistent and climatologically meaningful set of explanatory variables. Humidity, vapour pressure, temperature extremes, and evapotranspiration together represent the most reliable and physically interpretable drivers of grape yield variability across all experimental conditions.

35 Part VII Conclusion

This chapter examined long-term climate behaviours and their implications for grape yield using SILO records from 2000–2024. Through sequence analysis, volatility diagnostics, and interpretable deep learning models, the results reveal three consistent patterns: i) a gradual warming trend with increasing rainfall variability after 2010, ii) limited but region-specific yield responses to long-term climate shifts, and iii) stable explanatory drivers, humidity,

Table 41: Comparison of top-ranked climate features in aggregated IG attributions and fused attention (Coonawarra Penfolds, 2022, daily with LTA input). "Aggre-gated IG Only" denotes features captured exclusively by the aggregated IG method; "Common Features" represent those shared by both explanation methods; and "Fused Attention only" refers to features highlighted solely by fused attention.

Aggregated IG Only	Common Features	Fused Attention Only
silo_gdd_non_cum	silo_vapour_pressure	silo_gdd
silo_evapotranspiration_fao65_lta_deviation	silo_vapour_pressure_lta_deviation	silo_evapotranspiration_fao65
silo_temperature_min	silo_temperature_min_lta_deviation	silo_relative_humidity_max_temp_lta_deviation
silo_pressure_lta_deviation	silo_pressure	silo_span
silo_evapotranspiration_mortons_lake_lta_deviation	silo_relative_humidity_min_temp	silo_evapotranspiration_mortons_lake
silo_evapotranspiration_mortons_wet_lta_deviation		silo_chill_days
silo_evapotranspiration_mortons_actual_lta_deviation		silo_evapotranspiration_mortons_potential
silo_rain		silo_rain_lta_deviation
silo_radiation_lta_deviation		silo_evapotranspiration_sp

Table 42: Common top-ranked climate features shared by CBAM fused attention and aggregated IG across regions and input types.

Feature Category	Common Features	Interpretation / Physical Meaning
Humidity and Pressure	silo_relative_humidity_min-temp, silo_relative_humidity_max_temp, silo_vapour-pressure, silo_pressure	Highlight vapour-pressure balance and moisture regulation mechanisms that control grape water status.
Temperature Extremes	silo_temperature_min, silo_temperature_max	Represent key thermal stress indicators during growth and ripening phases.
Evapotranspiration Processes	silo_evapotranspiration_mortons_actual, silo_evapotranspiration_mortons_wet, silo_evapotranspiration_mortons_lake, silo_evapotranspiration_fao65, silo_evapotranspiration_sp	Demonstrate that water and energy exchange processes dominate yield variability across all regions.
Heat Accumulation and Rainfall	silo_gdd, silo_gdd_non_cum, silo_rain	Capture the joint influence of accumulated temperature and precipitation on grapevine productivity.

vapour pressure, temperature extremes, and evapotranspiration, across all regions and input settings. By integrating CBAM attention with IG, the study demonstrates that intrinsic and post-hoc explanations converge on physically meaningful yield determinants, enhancing interpretability and trustworthiness. Overall, the findings indicate that moisture and energy balance are the primary processes governing yield variability, and that explainable deep learning provides a reliable tool for interpreting long-term agricultural climate interactions.

Part VIII
Conclusion

The objective of this project is to identify a robust and repeatable framework by which agricultural AI models can be used to understand the implications of climate change and related risks on agricultural crop outcomes. With the support of The Yield Technology Solutions (now Yamaha Agriculture) and Food Agility CRC (FACRC), the DSI research team investigated this explainability problem from three perspectives: i) to understand the needed AI explainability in agriculture, ii) to handle climate features and short-term extreme events in an explainable AI framework, and iii) to cater for long-term climate behaviours (e.g., arisen from climate change) in our proposed framework. Specifically, the explainability of our framework targets the impacts of extreme weather events and climate change on crop outcomes.

Climate variability and extreme events pose significant challenges to agricultural productivity. Predictive models capable of estimating crop yield under such uncertainty are vital for informed decision-making and risk mitigation. However, many traditional post-hoc Explainable AI methods, such as SHAP and LIME, implicitly assume feature independence when attributing importance. In climate datasets, this assumption is often violated because numerous variables (e.g., temperature, humidity, radiation, and evapotranspiration) exhibit strong spatial and temporal correlations. As a result, these methods tend to distribute importance diffusely or inconsistently across correlated features, leading to unstable or misleading explanations that obscure true climate drivers.

We addressed this challenge by applying Convolutional Neural Network (CNN) with Temporal Attention to predict crop yields and interpret the underlying temporal patterns. CNN is capable of exploiting local spatial or temporal dependencies through convolutional filters, which capture joint patterns among features and across time. In our journey of frame-work development, both 1-dimensional CNN (1D-CNN) and 2-dimensional CNN (2D-CNN) architectures were investigated. We began with 1D-CNN enhanced by a temporal attention mechanism to extract short-term dynamics and identify key growth phases that influence crop yield. Though, 1D-CNN cannot capture cross-feature relationships which are crucial for understanding interactions among climate features. We extended our modelling frame-work to 2D-CNN with Convolutional Block Attention Module (CBAM). Unlike traditional XAI methods, the extended framework can exploit both spatial and temporal relation-ships among correlated features through convolutional filters, which extract local feature patterns and build hierarchical representations. By structuring climate features into a 2-dimensional feature-time space, our proposed 2D-CNN framework jointly learns intra-feature and inter-feature dependencies, producing more coherent, physically meaningful, and robust explanations.

Next, we turned the focus to extend our frameworks to better cater for long-term climate behaviours. We first analysed sequences (time series) for outcomes and key climate features using classical decomposition in statsmodels, Prophet model, and seasonal Autoregressive Integrated Moving Average (ARIMA) model. Classical decomposition offers a clear split into trend, seasonality, and residual for interpretation, but it is only descriptive and not intended for forecasting. Prophet model supplies flexible seasonality with change-points, yet in our setting the extracted trend is often too simple to capture dynamic structure. Seasonal ARIMA provides a statistically grounded baseline that models autocorrelation and seasonal dependence well, but estimation becomes slow as sequences lengthened and it is better suited to short- to medium-horizon use. We then studied weather volatility as a risk signal because it matters for planning, resilience, and yield stability. Our work involved combining rolling statistics to track local variance, entropy measures to summarise unpredictability, and using Generalised AutoRegressive Conditional Heteroskedasticity (GARCH) models to capture time-varying variance. We also put our attention to the contrasts between El Niño – Southern Oscillation (ENSO) phases – comparing El Niño and La Niña years aiming to capture their magnified effects.

Moreover, building on our proposed framework which focused on the daily climate-yield modelling using 2D-CNN with CBAM, we investigated how to extend our framework for long-term climate analysis. While CBAM provides intrinsic, attention-based visualisation of important feature-time regions, it does not explicitly quantify the contribution of input to the final prediction. This limitation becomes more critical when modelling multi-year, non-stationary climate data, where subtle feature interactions and long-term trends must be disentangled from attention biases. Therefore, Integrated Gradients (IG) is incorporated into the framework as a post-hoc, gradient-based attribution method that complements the attention mechanism. By integrating CBAM attention with IG, we demonstrated that intrinsic and post-hoc explanations converge on physically meaningful yield determinants, enhancing interpretability and trustworthiness.

In our experiments, we used apples, grapes, and kiwifruit yield data and their relevant climate data from the United States, Australia, and New Zealand to investigate and establish our proposed framework. As this project is not only research focused but to cater for the business needs of our industry collaborator, the use of particular datasets varies during the project. For example, we put more emphasis on the New Zealand's kiwifruit data to assist our partner to establish their business propositions to their client for a commercial opportunity. Later on, we spent more time on the United States' apples and Australian grape datasets to ensure that our proposed frameworks are general enough for real-world applications. This final report summarises our investigations and research findings of this project.

Lastly, we would like to express our sincere thanks to Dr Evan Webster (project leader) and Mr Ashley Rootsey (project manager). Although both were departed early from this project due to their resignations from their employers, their support to the research team are much appreciated. We wish them all the best for their future endeavours.

Part IX
Appendix

36 Manuscripts on Specific Topics

We have prepared two manuscripts on specific topics in this project, they are:

1. "The Complexity of Extreme Climate Events on NZ Kiwifruit Industry", and
2. "Multi-Hazard Early Warning Systems for Agriculture with Featural-Temporal Explanations".

Their pre-prints are archived to arxiv.org and the current version of the papers are also attached below for completeness.

36.1 The Complexity of Extreme Climate Events on NZ Kiwifruit Industry

The Complexity of Extreme Climate Events on the New Zealand's Kiwifruit Industry

Boyuan Zheng*, Victor W. Chu*, Zhidong Li*, Fang Chen*
Evan Webster†, Ashley Rootsey† ‡
*Data Science Institute, University of Technology Sydney
†Yamaha Agriculture & ‡Food Agility CRC

***Abstract*—Climate change has intensified the frequency and severity of extreme weather events, presenting unprecedented challenges to the agricultural industry worldwide. In this investigation, we focus on kiwifruit farming in New Zealand. We propose to examine the impacts of climate-induced extreme events, specifically frost, drought, extreme rainfall, and heatwave, on kiwifruit harvest yields. These four events were selected due to their significant impacts on crop productivity and their prevalence as recorded by climate monitoring institutions in the country. We employed Isolation Forest, an unsupervised anomaly detection method, to analyse climate history and recorded extreme events, alongside with kiwifruit yields. Our analysis reveals considerable variability in how different types of extreme event affect kiwifruit yields underscoring notable discrepancies between climatic extremes and individual farm's yield outcomes. Additionally, our study highlights critical limitations of current anomaly detection approaches, particularly in accurately identifying events such as frost. These findings emphasise the need for integrating supplementary features like farm management strategies with climate adaptation practices. Our further investigation will employ ensemble methods that consolidate nearby farms' yield data and regional climate station features to reduce variance, thereby enhancing the accuracy and reliability of extreme event detection and the formulation of response strategies.**

***Index Terms*—Kiwifruit, Extreme Climate Events, Anomaly Detection, Complex Relationships**

I. Introduction

Climate change is triggering notable shifts in global weather patterns resulting in an increased frequency and intensity of adverse climate events. Rising global temperatures have altered atmospheric conditions intensifying the occurrences of heatwave, drought, and heavy precipitation events. Recent studies, such as [1] and [2], highlighted that even modest global temperature rises can substantially amplify these extreme events. Specifically, differences between warming scenarios of 1.5°C and 2.0°C above pre-industrial levels, as projected by the IPCC [3], have shown significant variability in projected consequences. Historical climate records provide further evidence to the occurrence of these relations documenting a clear increase in temperature extremes and heavy rainfall episodes [4]. As such climatic shifts intensify, having a comprehensive understanding of these extreme events is imperative for the development of effective mitigation and adaptation strategies.

The kiwifruit industry has already been experiencing significant disruptions due to these extreme events. Rising temperatures, altered rainfall patterns, and increased frequency of extreme weather events have posed substantial challenges to the farmers. Warmer autumn and winter conditions are reducing winter chilling hours — something crucial for kiwifruit budbreak and flower production which directly impacts both quantity and quality of yield — has prompted the migration of cultivation practices toward higher latitudes. Furthermore, events such as hailstorms, severe winds, and intense rainfall are becoming increasingly common resulting in notable financial losses as documented by [5]. Prolonged drought conditions and elevated evapotranspiration further heighten concerns around water availability prompting the need for improved irrigation practices and climate-resilient orchard management. As climate extremes grow more severe, proactive adaptation strategies will be essential for maintaining long-term viability and resilience within kiwifruit production.

A critical step towards assessing the impact of extreme events on kiwifruit yields involves accurately identifying and quantifying their occurrences, where national weather services leverage standardised indices such as SPI or percentage thresholds to detect such events. While such detection methods may not directly affect farmers' short-term day-to-day decisions, they are critical for analysis on modelling long-term yield risks and developing adaptation strategies. In this context, anomaly detection techniques play a key role though each of them offers distinct benefits but also limitations. Functional outlier detection methods, discussed in [6], utilise graphical tools such as rainbow plots to visualise anomalies yet often falter in high-dimensional scenarios and rely heavily on expert interpretation. Alternatively, multivariate anomaly detection approaches, which consider interactions among environmen-tal variables, typically enhance accuracy though necessitate extensive pre-processing to handle seasonal and correlated effects [7]. Density-based clustering techniques like DBSCAN can effectively detect grouped anomalies but exhibit high sensitivity to parameter selection and may under-perform with irregular temporal patterns [8]. Compared to these methods, Isolation Forest (IF) provides a computationally efficient and scalable alternative, adept at handling large and high-dimensional datasets due to its recursive partitioning approach [9].

In this study, we examine the complex relationship be-tween extreme climate events and kiwifruit yield variability in New Zealand, recognising the limitations of climate data alone in explaining observed yield outcomes. Our findings suggest that although climate anomalies significantly influence agricultural productivity, their effects are often modulated by farm-level management practices. This underscores the need for a nuanced interpretation of climate-yield relationships. Additionally, we investigate how different extreme events, such as drought, extreme rainfall, heatwave, and frost, vary in their impacts on yields revealing that these effects are neither uniform nor predictable across all conditions. These four events were selected due to their significant impacts on crop productivity and their prevalence as recorded by climate monitoring institutions in New Zealand. Finally, we explore alignments and discrepancies between climate anomalies and yield data, identifying instances where severe climate conditions do not consistently correspond to reduced yields, and vice versa. These findings highlight the complexities inherent in extreme event's impact assessments and emphasise the importance of integrating climate data with local management and environmental factors for a robust understanding of yield variability.

The remainder of this paper is structured as follows. Section II reviews relevant studies on anomaly and outlier detection methodologies with specific attention to agricultural and climate-related applications. Section III describes the dataset used, including data sources, granularity, temporal scope, and detailed definitions of extreme events according to New Zealand's national weather service, the National Institute of Water and Atmospheric Research (NIWA). The methodol-ogy of how we employ Isolation Forest and pre-processing techniques in this investigation is presented in Section IV. Results highlighting the alignment of detected extreme events with observed yield variations across farms are discussed in Section V. Section VI provides an in-depth discussion on the implications of our findings, limitations of current approaches, variability in the impact of extreme events, and proposed strategies for improvement. Finally, Section VIII concludes the paper by summarising our key contributions and insights derived from this study.

II. Related Work

Anomaly and outlier detection methods have been extensively adopted across diverse domains, including electricity monitoring [10], equipment fault diagnosis [11], and medical signal analysis [12], primarily due to their efficacy in identifying irregularities that signal potential system failures. In environmental sciences, such methodologies play a crucial role in detecting anomalous climate events. Flach et al. [7] compared multiple dimensionality reduction and anomaly detection algorithms, such as k-nearest neighbours mean distance, kernel density estimation, and recurrence analysis, and demonstrated their effectiveness in identifying environmental anomalies using Earth observation data. Mahmoud and Gan [13] further advanced this domain by employing non-parametric change-point detection integrated with GIS-based decision support systems, assessing urbanisation and climate change impacts on flood risks in Egypt to produce detailed flood suscepti-bility maps. Expanding the methodological scope, Hael and Yuan [6] proposed a functional outlier detection approach employing functional bag-plots and highest-density region box-plots, successfully identifying extreme rainfall anomalies in Yemen over a 21-year dataset. More recently, Zhang et al. [14] introduced an approach based on Support Vector Data Description (SVDD), utilising multi-class classification and multi-label recognition techniques to effectively discriminate normal patterns from extreme events, including typhoons and earthquakes.

In agricultural applications, the adoptions of anomaly detection have been specifically tailored to address cultivation practices and crop health monitoring. Peichl et al. [15] utilised random forest models to identify yield anomalies in winter wheat attributable to soil moisture variability and extreme weather events, highlighting soil moisture as having a more significant impact compared to meteorological variables, particularly when combined with spatial clustering techniques. Similarly, Mouret et al. [16] effectively employed Isolation Forests to analyse Sentinel-1 and Sentinel-2 time series data to detect anomalous crop development at the parcel level. Castillo-Villamor et al. [17] developed an Earth Observation-based Anomaly Detection (EOAD) system that automated the thresholding of optical vegetation indices derived from Sentinel-2 and PlanetScope imagery achieving high accuracy in identifying yield-limiting anomalies in rice fields. Subsequently, Moso et al. [18] proposed an ensemble-based anomaly detection approach designed for smart agriculture data streams, successfully validating their approach on GPS logs from combine harvesters and crop datasets to enhance farm efficiency. Mujkic et al. [19] extended anomaly detection methodologies further into precision agriculture by leveraging convolutional autoencoders, particularly semi-supervised variants utilising max-margin loss, to detect unknown obstacles in autonomous agricultural vehicles. Most recently, Sjulga˚rd et al. [20] analysed historical temperature and precipitation anomalies across Sweden, highlighting significant yield losses resulting from extreme summer weather events, particularly affecting spring-sown crops due to their shorter growing periods and sensitivity to soil-texture interactions.

While these studies significantly contribute to the field of anomaly detection in environmental and agricultural contexts, they underscore the ongoing need for methods that can integrate advanced machine learning with explainable and interpretable frameworks in the real world. Addressing this gap could enhance stakeholder understanding and decision-making capability, providing practical solutions tailored for real-world agricultural and environmental management scenarios.

III. Data

The datasets employed in this study comprise detailed climate and kiwifruit yield records from New Zealand, se-lected to facilitate an analysis of climate impacts on regional orchard productivity. Specifically, the kiwifruit yield data were sourced from The Yield Technology Solutions[1] (now Yamaha Agriculture), covering a total of 3,437 farms, thus enabling high-resolution, farm-level analysis. Concurrently,

[1]https://www.theyield.com/

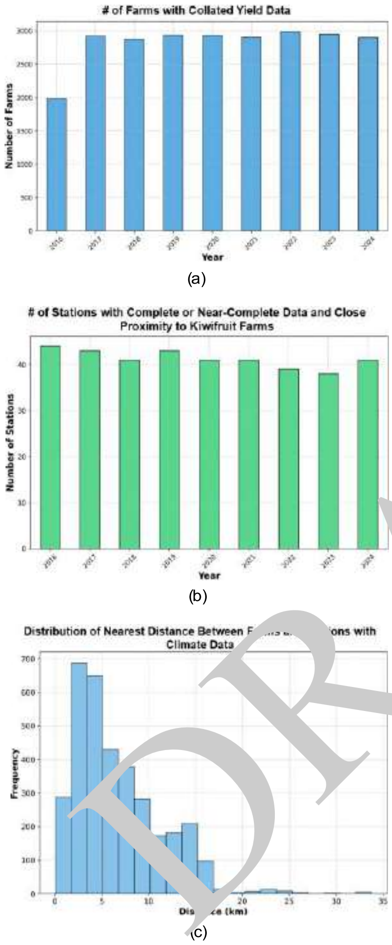


Fig. 1. Spatiotemporal statistics on the climate data and yield data

climate observations were obtained through the NIWA National Climate Database (CliFlo database)[2], providing daily meteorological records from a network of 318 climate stations strategically distributed nationwide. To accurately integrate climate observations with farm-specific yield data, spatial matching procedures were applied, whereby each farm was

[2]https://niwa.co.nz/climate-and-weather/national-climate-database

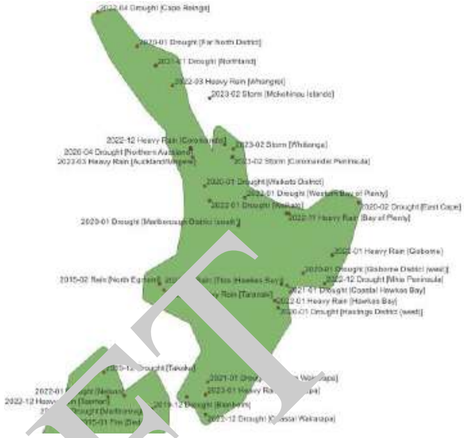


Fig. 2. Samples of extreme events summarised from NIWA seasonal summary documents.

matched to the nearest available climate stations, typically within a 10 km radius, with very few cases extending beyond 15 km (as shown in Figure 1c). Temporally, the climate dataset spans approximately twelve years, from December 2008 to January 2021, although the completeness of records varies between individual stations. In contrast, kiwifruit yield data are available from 2016 to 2024, ensuring sufficient temporal overlap for comprehensive analysis of climate-yield relationships (see Figure 1a and 1b). The climate variables include daily maximum and minimum temperatures, solar radiation, precipitation, and wind parameters. However, the availability of these features varies considerably between stations, and inconsistencies in temporal coverage across the network pose analytical challenges, particularly for multivariate outlier detection methods.

Extreme events data were collected from multiple authoritative sources. Large-scale extreme events, including droughts and anomalies in temperature and rainfall, were extracted from NIWA seasonal summaries [21], which provide detailed regional documentation and enable precise identification and categorisation of event severity (see Figure 2). Severity assessments were established using rule-based categorisations relative to long-term climatic baselines. However, frost events, due to their localised nature and inconsistent reporting, are less comprehensively covered in NIWA summaries. To address this gap, supplementary frost event data were obtained from Zespri monthly orchard reports [22], although these records typically lack quantifiable severity information. Consequently, frost events were represented using a binary classification to indicate presence or absence.

Collectively, the integration of these multi-source datasets aims to enhance the detection and characterisation of extreme weather events from daily climate data, thereby facilitating a robust assessment of their impacts on kiwifruit yields across regional orchards.

IV. Methodology

Figure 3 presents our methodological workflow employed in this study. The analysis began with the acquisition of three aforementioned data sources. Each dataset underwent a structured pre-processing pipeline to ensure data integrity and consistency. The pre-processing phase commenced with a comprehensive data cleaning procedure, addressing inconsistencies, missing values (found to be only less than 10 days in between), and potential errors. Specifically, missing values were handled using forward filling, where each missing entry was replaced with the last observed value. Following data cleaning, all datasets were spatio-temporally aligned by matching observations to the nearest weather stations and synchronising data across overlapping years.

Subsequently, the processed datasets advanced to the anomaly detection phase. Isolation Forest (IF) was employed as the primary anomaly detection method due to its computational efficiency and its ability to effectively isolate anomalies through recursive partitioning. Unlike density- or distance-based approaches, IF leverages random feature-space partitioning to isolate anomalous points efficiently. Mathematically, IF constructs multiple isolation trees through recursive binary partitioning. Given a dataset $X = \{x_1, x_2, ..., x_n\}$, each tree partitions the feature space iteratively until all instances are isolated in leaf nodes. The anomaly score $s(x)$ of an instance x is computed as:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where $E(h(x))$ represents the average path length from the root node to the node containing x across all isolation trees, and $c(n)$ is the expected path length of unsuccessful searches in a binary search tree, expressed as:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n},$$

where $H(i)$ denotes the harmonic number, defined as $H(i) = \ln(i) + 0.5772156649$ (Euler-Mascheroni constant). Instances exhibiting shorter average path lengths were classified as anomalies, as they were more readily isolated. The contamination hyperparameter of IF, which determines the proportion of data points identified as anomalies, was tuned between 0.005 and 0.05 through an iterative trial-and-error approach for each climate feature independently. The detected anomalies were subsequently validated against historical observational records to assess the model's precision and its capability in identifying significant deviations in climate data.

Finally, in the analysis phase, the results of the anomaly detection process were interpreted using instance-based explanations, incorporating domain expertise to contextualise and substantiate the findings. Additionally, the impact of extreme weather events on kiwifruit yield was quantitatively assessed by computing the reduction in yield relative to the average of preceding years.

V. Results

This section presents our findings from the integrated analysis of climate and agricultural data, highlighting the impacts of key climatic events on kiwifruit yields across northern New Zealand. Using anomaly detection and yield variation assessments, we examine how drought, heatwave, extreme rainfall, and frost influence crop production at a regional scale. The results provide insights into the extent of yield fluctuations under different climatic conditions, revealing both vulnerabilities to various extreme events and adaptive responses across farms and kiwifruit varieties: i) Green/Hayward (HW) variety, and ii) Gold (GA) variety. Note that the farm identifiers (4-digit numbers) were replaced by unique capital letters in this paper due to business confidentiality. The farms were selected due to their proximity to the named weather stations in northern New Zealand.

A. Drought

The integrated climate and agricultural data presented in Figure 4 illustrate critical relationships between drought conditions, climate anomalies, and kiwifruit yield fluctuations across northern New Zealand. Several moderate to severe drought episodes occurred from 2015 to 2024, notably aligning with the significant drought event of 2020 reported by NIWA. Anomaly detection results for relative humidity and rainfall at station-2006 further validate these climatic irregularities. Correspondingly, annual kiwifruit yields exhibited substantial variability across analysed farms during the drought year. Yield losses ranged from approximately -28% (increasing yield) to over 44% compared to the previous 5-year average, with farm D (HW variety) experiencing the most severe reduction of 44%, followed closely by farm E (GA variety) at approximately 41%. Conversely, farm D (GA variety) and farm B (GA variety) showed a yield increase of about 29% and 21% respectively, reflecting variability even within the same location across different crop types. Farms cultivating the HW variety, such as farm A and D, also displayed significant yield reductions (around 29% and 34%, respectively). These findings highlight considerable variability in kiwifruit yields in response to drought conditions, emphasising the necessity for farm and variety-specific adaptation strategies.

B. Heatwave

The integrated climate and agricultural data presented in Figure 5 illustrate critical relationships between the 2022 heatwave event, climate anomalies, and kiwifruit yield fluctuations across northern New Zealand. Anomaly detection results at station-37656 clearly identify significant tempera-ture deviations (Tmax and Tmin), confirming the presence of an exceptional heatwave, aligned with NIWA's report of record warmth from January to June 2022. Rainfall anomalies were comparatively less frequent during the same period. Interestingly, the yield data from selected farms indicate a generally positive influence from the heatwave, contrasting notably with drought-induced yield reductions observed previously. Compared to the 5-year window average from 2017

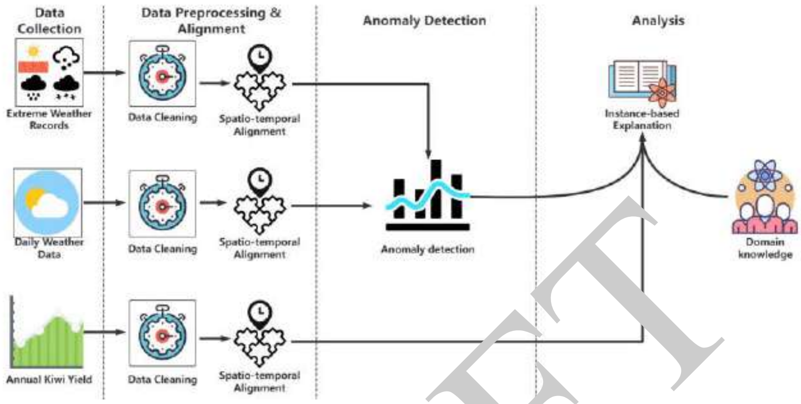


Fig. 3. Overview of the methodology workflow. The approach integrates diverse data sources, pre-processing steps, IsolationForest-based anomaly detection, and instance-based explanations supported by domain knowledge to evaluate the impact of extreme weather events on annual kiwifruit yield.

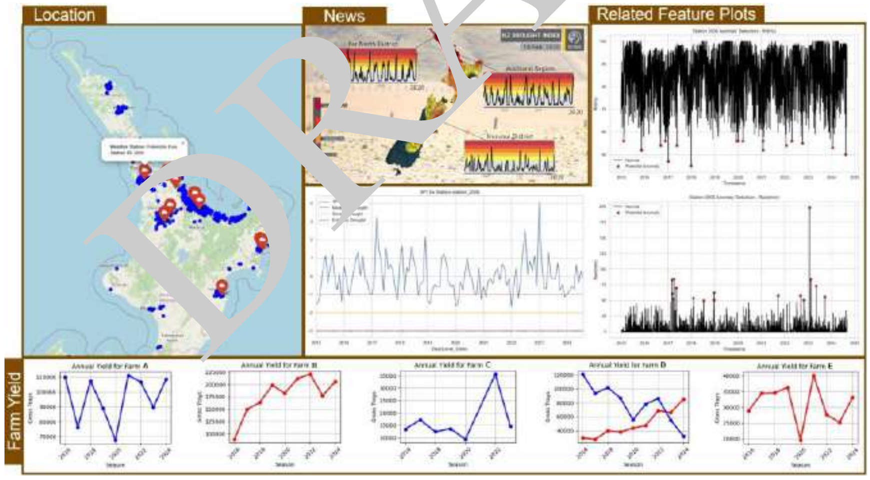


Fig. 4. Weather station (red cloud icons) and kiwifruit farm locations (blue dots) in northern New Zealand, alongside climate anomalies, drought indicators, and annual kiwifruit yields. The panels illustrate drought severity (NZ Drought Index, February 2020), Standardised Precipitation Index (SPI) drought intensity (station-2006), anomalies detected in relative humidity and rainfall (red points), and yield patterns for two kiwifruit crop types (GA in red; HW in blue) from five selected farms.

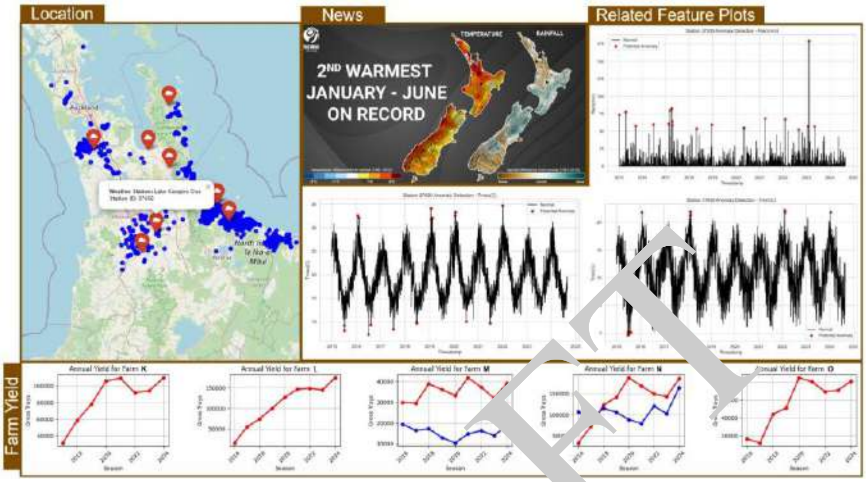


Fig. 5. Locations of weather stations (red cloud icons) and kiwifruit farms (blue dots) in northern New Zealand, alongside climate anomalies related to temperature and rainfall, with anomaly detection results at station-37656 for maximum temperature (Tmax), minimum temperature (Tmin), and rainfall. Annual kiwifruit yields for two crop types (GA in red; HW in blue) from five farms are also illustrated.

to 2021, all five selected farms demonstrated yield increment ranging from approximately 3% to 47%, notably farm H (GA) with an 3% improvement, farm G (GA) demonstrated a remarkable yield increase of 47%. Additionally, farm H (HW) and farm I (HW) also experienced modest yield improvements of approximately 13% and 25%, respectively. These positive responses highlight outstanding resilience for both kiwifruit varieties under heatwave conditions, reinforcing the various impact of extreme events on kiwifruit industry.

C. Rainfall

The integrated climate and agricultural data presented in Figure 6 illustrate significant impacts of extreme rainfall events on kiwifruit yields across northern New Zealand, particularly during the summer of 2022-2023. Anomaly detection results for rainfall at station-1520 confirm frequent high-intensity rainfall events, supported by the Standardised Precipitation Index (SPI)[3] peaks and NIWA's reports identifying record or near-record wet conditions at multiple locations. Correspondingly, yields across the analysed farms significantly declined in 2023, with losses ranging from approximately 3% to as high as 38%. Farm K and L (both HW variety) faced the most severe reduction of 38% and 30% respectively, while farm M (GA variety) experienced the least yield decrease (approximately 3%). Notably, both kiwifruit varieties, i.e., GA and HW, showed significant vulnerabilities, although variability within and between varieties persisted (GA: 21.27%; HW: 22.83%). These results clearly underscore the severe impact of extreme rainfall events on kiwifruit production, highlighting considerable risks and the critical need for farm-level adaptation and robust event management strategies.

D. Frost

The integrated climate and agricultural data presented in Figure 7 illustrate substantial kiwifruit yield impacts resulting from frost events in early 2023 across northern New Zealand. Frost, particularly late-spring frost, tends to be localised and often overlooked by government climate agencies due to its small spatial scale, making detection from standard anomaly results challenging. Relative humidity (RH), minimum temperature (Tmin), and wind speed (GustSpd) are critical indicators included here due to their direct relevance to frost formation: frost typically occurs under conditions of low minimum temperatures, calm winds, and high relative humidity. Although anomaly detection at station-1615 identifies sporadic anomalies in these variables, the precise timing and localised nature of frost mean these broader-scale detections may not fully capture its true intensity and frequency. Nonetheless, yield data confirm significant agricultural impacts, with yield reductions ranging from approximately 7% to over 42%. Farm P (26.17%), farm Q (29.67%) and farm R (33.66%) experienced the most severe yield losses on both GA and HW, emphasising kiwifruit vulnerability to frost. Conversely, farm S (GA) experienced the smallest reduction, approximately 7%. These substantial yet variable yield impacts underscore frost as

[3]https://niwa.co.nz/nz-drought-indicator-products-and-information/drought-indicator-maps/standardised-precipitation-index-spi

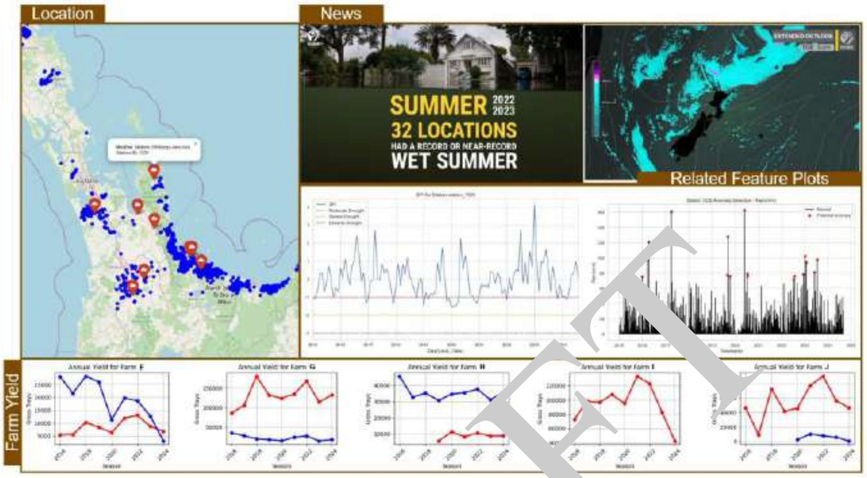


Fig. 6. Locations of weather stations (red cloud icons) and kiwifruit farms (blue dots) in northern New Zealand, alongside evidence of record-breaking rainfall events during the summer of 2022-2023. SPI values and rainfall anomaly detection at station-1520 validate the extreme intensity of these events. Kiwifruit yield fluctuations for two crop types (GA in red; HW in blue) from five farms illustrate yield changes associated with extreme rainfall.

a critical, though frequently underestimated, threat to kiwifruit productivity. A comparative analysis of yield losses between GA (13.33%) and HW (35.22%), conducted on five randomly selected farms, indicates that GA exhibits superior resilience to frost conditions.

VI. Discussion and Future Directions

The findings presented above reveal significant variability in the effects of extreme events, namely drought, extreme rainfall, heatwave, and frost, on kiwifruit yield in northern New Zealand, as shown in Figure 9. Among these, frost events exhibited the most severe impact, with average yield losses of approximately 27% across the five analysed farms. This was followed by extreme rainfall at 23%, drought at 20%, and heatwave, which had the least adverse effect with average benefit around 15%. The pronounced impact of frost highlights a major but often under-recognised threat to kiwifruit production, likely attributable to the localised and transient nature of frost events, which may lead to their under-reporting. On the other hand, when yield changes were disaggregated by kiwifruit variety, further patterns emerged. Both GA and HW varieties experienced substantial reductions during drought conditions, though intra-group variability was evident. On average, HW kiwifruit suffered greater losses (33%) compared to GA (5%). Heatwave, however, produced mixed outcomes: GA farms even reported yield increases (up to 31%), while HW farms showed around average yield with previous years. Extreme rainfall in early 2023 resulted in significant yield reductions for both types – approximately 22% for GA and 25% for HW. Similarly, frost caused major losses across both varieties, with average reductions of 25% for GA and 29% for HW, underscoring the widespread vulnerability of kiwifruit to frost and excessive rainfall, regardless of cultivar.

While anomaly detection techniques effectively identified broader climatic anomalies such as drought and extreme rainfall, they showed limitations in capturing the highly localised and transient nature of frost events. The challenge in detect-ing frost through standard anomaly indicators (e.g., relative humidity, minimum temperature, wind speed) highlights the limitations of generalised anomaly detection methodologies and points to the need for finer-scale, farm-specific monitor-ing systems. Apart from the limitation from the algorithm itself, the limited record of frost events in official reports also obstacle the research on frost in New Zealand. Based on the existing reports from kiwifruit industry [23], they suggest significant variability in frost impacts among growers. The reported variability in yield reduction aligns with the observed wide range in farm-level data within this study. Such variability may contribute to the absence of comprehensive governmental reporting on frost impacts.

Additionally, yield analyses revealed several counterexamples where farms experienced limited impacts despite exposure to extreme climatic events. Figure 8 illustrates orchards that exhibited relatively minor yield reductions. Based on empiri-cal observations, identifying such counterexamples was more feasible for drought events than for extreme rainfall or frost, suggesting that drought impacts may be more amenable to mitigation through human intervention. In contrast, orchards

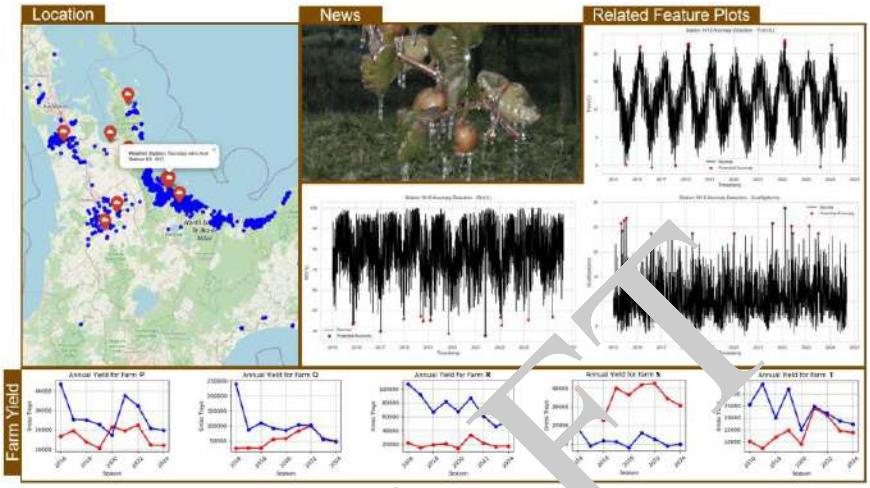


Fig. 7. Locations of weather stations (red cloud icons) and kiwifruit farm... dots) in northern New Ze..., accompanied by evidence of frost occurrence in early 2023 (source: Zespri). Anomaly detection results at station-1615 ... ...idity (RH), minim...m temperature (Tmin), and wind speed (GustSpd) provide further context. Yield data from five farms for two kiwifruit crop ty...es (GA ... in blue) demonstrate impacts associated with the frost event.

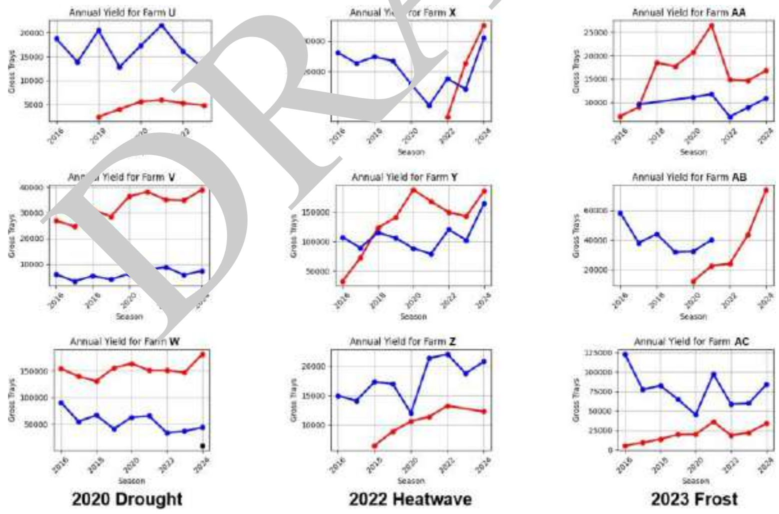


Fig. 8. Counterexamples of kiwifruit yield for 2020's drought, 2022's heatwave, and 2023's frost. For extreme rainfall, we find zero counterexample (GA in red; HW in blue).

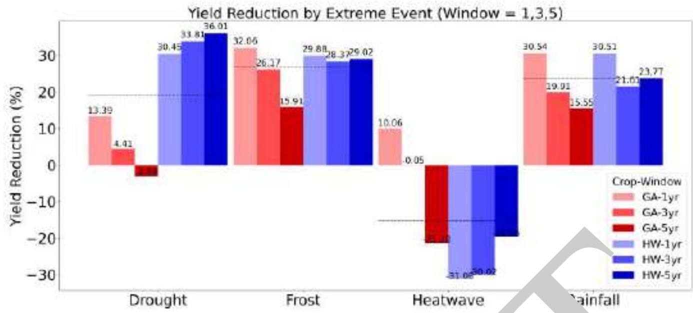


Fig. 9. Yield reduction percentages under different window averaging sizes. Each group consists of six bars, with the horizontal dashed lines indicating the mean yield reduction within each group.

appear to be less resilient to intense rainfall and frost, which are more difficult to manage effectively at the farm level. These insights also underscore the importance of incorporating farm-specific management variables (such as canopy architec-ture, irrigation strategies, and frost protection techniques) into future modelling efforts aimed at explaining or forecasting yield variation. Moreover, an important but under-explored dimension is the additive effect of multiple extreme events occurring within a single season. Situations where drought, intense rainfall, and frost co-occur — either sequentially or in combination — may compound stress on crop systems in non-linear ways. Understanding these compound events is crucial for both modelling and intervention, as their combined impact may not be adequately captured by analysing each of them in isolation. To address spatial variability in yield outcomes across different stations and farms, we also recommend the application of ensemble-based analyses. Integrating data from multiple weather stations and farms can help smooth out lo-calised anomalies and counterexamples, leading to more robust and generalisable inferences. Ultimately, such approaches will enhance the accuracy of future yield predictions and support the development of more effective and site-specific adaptation strategies.

VII. Acknowledgements

This research was supported by the Food Agility Coop-erative Research Centre under the project Yield Prediction Explainability & Climate Adaptiveness (Project Code: FA128), a collaboration between Food Agility CRC Limited, The Yield Technology Solutions Pty Ltd (Yamaha Agriculture), and the Data Science Institute at the University of Technology Sydney. The authors gratefully acknowledge this support, which made the research and its dissemination possible.

VIII. Conclusion

This study examined the impact of extreme climate events on New Zealand's kiwifruit yield using anomaly detection techniques. Our findings highlight significant yield reductions due to extreme rainfall and frost, while heatwave showed mixed effects. The results underscore the complexity of climate-yield interactions. Additionally, limitations in detect-ing localised events like frost suggest the need for finer-scale meteorological data and improved detection methods. Future work should integrate multi-farm data and agronomic variables to enhance anomaly detection and yield prediction supporting more resilient agricultural practices.

References

[1] H. Chen and J. Sun, "Projected changes in climate extremes in china in a 1.5 c warmer world," *International Journal of Climatology*, vol. 38, no. 9, pp. 3607–3617, 2018.

[2] C. F. Gaita´n, "Machine learning applications for agricultural impacts under extreme events," in *Climate extremes and their implications for impact and risk assessment*. Elsevier, 2020, pp. 119–138.

[3] V. Masson-Delmotte, P. Zhai, H.-O. Po¨rtner, D. Roberts, J. Skea, P. Shukla, A. Pirani, W. Moufouma-Okia, C. Pe´an, R. Pidcock, S. Connors, J. Matthews, Y. Chen, X. Zhou, M. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. e. Waterfield, *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways*. Intergovernmental Panel on Climate Change, 2018. [Online]. Available: https://www.ipcc.ch/sr15/

[4] J. Powell and S. Reinhard, "Measuring the effects of extreme weather events on yields," *Weather and Climate extremes*, vol. 12, pp. 69–79, 2016.

[5] Zespri, "Annual reports," https://www.zespri.com/en-NZ/publications/annualreports, accessed: 2025-01-30.

[6] M. A. Hael, Y. Yuan *et al.*, "Identifying extreme rainfall events using functional outliers detection methods," *Journal of Data Analysis and Information Processing*, vol. 8, no. 04, p. 282, 2020.

[7] M. Flach, F. Gans, A. Brenning, J. Denzler, M. Reichstein, E. Rodner, S. Bathiany, P. Bodesheim, Y. Guanche, S. Sippel *et al.*, "Multivariate anomaly detection for earth observations: a comparison of algorithms and feature extraction techniques," *Earth System Dynamics*, vol. 8, no. 3, pp. 677–696, 2017.

[8] M. E. Aslan and S. Onut, "Detection of outliers and extreme events of ground level particulate matter using dbscan algorithm with local parameters," *Water, Air, & Soil Pollution*, vol. 233, no. 6, p. 203, 2022.

[9] A. Ba˘ra, A. G. Va˘duva, and S.-V. Oprea, "Anomaly detection in weather phenomena: News and numerical data-driven insights into the climate change in romania's historical regions," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 134, 2024.

[10] S.-V. Oprea, A. Baˆra, F. C. Puican, and I. C. Radu, "Anomaly detection with machine learning algorithms and big data in electricity consumption," *Sustainability*, vol. 13, no. 19, p. 10963, 2021.

[11] A. Purarjomandlangrudi, A. H. Ghapanchi, and M. Esmalifalak, "A data mining approach for fault diagnosis: An application of anomaly detection algorithm," *Measurement*, vol. 55, pp. 343–352, 2014.

[12] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study," *Medical image analysis*, vol. 69, p. 101952, 2021.

[13] S. H. Mahmoud and T. Y. Gan, "Urbanization and climate change implications in flood risk management: Developing an efficient decision support system for flood susceptibility mapping," *Science of the Total Environment*, vol. 636, pp. 152–167, 2018.

[14] Y. Zhang, X. Wang, Z. Ding, Y. Du, and Y. Xia, "Anomaly detection of sensor faults and extreme events based on support vector data description," *Structural Control and Health Monitoring*, vol. 29, no. 10, p. e3047, 2022.

[15] M. Peichl, S. Thober, L. Samaniego, B. Hansju¨rgens, and A. Marx, "Machine-learning methods to assess the effects of a non-linear damage spectrum taking into account soil moisture on winter wheat yields in germany," *Hydrology and Earth System Sciences*, vol. 25, no. 12, pp. 6523–6545, 2021.

[16] F. Mouret, M. Albughdadi, S. Duthoit, D. Kouame´, G. Rieu, and J.-Y. Tourneret, "Outlier detection at the parcel-level in wheat and rapeseed crops using multispectral and sar time series," *Remote sensing*, vol. 13, no. 5, p. 956, 2021.

[17] L. Castillo-Villamor, A. Hardy, P. Bunting, W. Llanos-Peralta, M. Zamora, Y. Rodriguez, and D. A. Gomez-Latorre, "The earth observation-based anomaly detection (eoad) system: A simple, scalable approach to mapping in-field and farm-scale anomalies using widely available satellite imagery," *International journal of applied earth observation and geoinformation*, vol. 104, p. 102535, 2021.

[18] J. C. Moso, S. Cormier, C. de Runz, H. Fouchal, and J. M. Wandeto, "Anomaly detection on data streams for smart agriculture," *Agriculture*, vol. 11, no. 11, p. 1083, 2021.

[19] E. Mujkic, M. P. Philipsen, T. B. Moeslund, M. P. Christiansen, and O. Ravn, "Anomaly detection for agricultural vehicles using autoencoders," *Sensors*, vol. 22, no. 10, p. 3608, 2022.

[20] H. Sjulga˚rd, T. Keller, G. Garland, and T. Colombi, "Relationships between weather and yield anomalies vary with crop type and latitude in sweden," *Agricultural Systems*, vol. 211, p. 103757, 2023.

[21] T. N. I. of Water and A. R. L. (NIWA), "Seasonal climate summaries from summer 2001 to the present," https://niwa.co.nz/climate-and-weather/seasonal, accessed: 2025-01-30.

[22] Zespri, "It's what inside that counts," https://www.zespri.com/content/zespri/nz/en-NZ/publications/kiwiflier/2025.html, accessed: 2025-01-30.

[23] ——, "Kiwiflier oct 22," https://www.zespri.com/content/dam/zespri/nz/publications/Kiwiflier/kiwiflier-2022/KF-440-October-2022.pdf, accessed: 2025-01-30.

36.2 Multi-Hazard Early Warning Systems for Agriculture

Multi-Hazard Early Warning Systems for Agriculture with Featural-Temporal Explanations

Boyuan Zheng [a], Victor W. Chu [a], Fang Chen [a]

[a]Data Science Institute, University of Technology Sydney, 15 Broadway, Ultimo, Sydney, 2007, NSW, Australia

Abstract

Climate extremes present escalating risks to agriculture intensifying the need for reliable multi-hazard early warning systems (EWS). The situation is evolving due to climate change and hence such systems should have the intelligent to continue to learn from recent climate behaviours. However, traditional single-hazard forecasting methods fall short in capturing complex interactions among concurrent climatic events. To address this deficiency, in this paper, we combine sequential deep learning models and advanced Explainable Artificial Intelligence (XAI) techniques to introduce a multi-hazard forecasting framework for agriculture. In our experiments, we utilize meteorological data from four prominent agricultural regions in the United States (between 2010 and 2023) to validate the predictive accuracy of our framework on multiple severe event types, which are extreme cold, floods, frost, hail, heatwaves, and heavy rainfall, with tailored models for each area. The framework uniquely integrates attention mechanisms with TimeSHAP (a recurrent XAI explainer for time series) to provide comprehensive temporal explanations revealing not only which climatic features are influential but precisely when their impacts occur. Our results demonstrate strong predictive accuracy, particularly with the BiLSTM architecture, and highlight the system's capacity to inform nuanced, proactive risk management strategies. This research significantly advances the explainability and applicability of multi-hazard EWS, fostering interdisciplinary trust and effective decision-making process for climate risk management in the agricultural industry.

Keywords: Multi-hazard, Early warning systems, Explainable artificial intelligence, Temporal explainability

1. Introduction

Agriculture is known to be vulnerable to climate change, which has already impacted crop production in a measurable manner (Schlenker and Roberts, 2009; Lobell and Di Tommaso, 2025; Betts et al., 2018). In particular, extreme weather events are amplifying such problem and pose escalating threats to agricultural productivity and economic stability, especially as climate change drives an increase in the frequency and intensity of climate extremes worldwide (Lesk et al., 2016; Chen and Sun, 2018). Hazards, such as extreme cold, floods, frost, hail, heatwaves, and heavy rainfall, can devastate crop yields and supply chains incurring massive financial losses (Liu and Basso, 2020; Schmitt et al., 2022). In recent decades, climate-related disasters have cost hundreds of billions in damages, and recent years have seen a disproportionate share of these losses as extreme events become widespread (Powell and Reinhard, 2016). Early warning systems (EWS) are therefore becoming necessities to the agricultural sector and broader economy, providing advance hazard forecasts that enable farmers, communities, and decision-makers to take proactive measures to mitigate adverse climate impacts (Basher, 2006; Reichstein et al., 2025; Camps-Valls et al., 2025).

However, surging climate and weather extremes increasingly involve multiple, concurrent, or cascading hazards, which pose unprecedented challenges to traditional single-focus EWS (Brett et al., 2025). An intense weather episode may trigger various threats in tandem (e.g., heavy rain leading to floods and crop disease, or a late spring frost following an early warm spell), stressing the need for integrated monitoring of diverse perils. Recognizing this complexity, the international community has emphasized the development of multi-hazard early warning systems as a priority. For example, the United Nations and World Meteorological Organization launched the "Early Warnings for All" initiative[1] in 2022 to expand EWS coverage for all major hazards (WMO, 2022). A unified multi-hazard EWS offers the versatility of consolidating forecasts for different extremes within a single framework, which is especially valuable for agricultural risk management. By simultaneously predicting multiple hazard types (rather than operating separate siloed warning tools for each peril), a unified approach can provide more comprehensive situational awareness and efficient resource allocation for disaster preparedness (Zhang et al., 2023b). In practice, state-of-the-art forecasting methods have mostly remained hazard-specific — for instance, specialized data-driven models exist for individual perils such as floods, ~~droughts, or hailstorms~~ (Beillouin et al., 2020; Lobell et al.,

[1]https://www.un.org/en/climatechange/early-warnings-for-all

2011) — yet comparatively few attempts have been made to develop integrated multi-hazard systems (Hrast Essenfelder et al., 2025; Reichstein et al., 2025). This negligence highlights the need for an early warning solution capable of handling a suite of hazards in tandem, thereby improving the coordination of protective actions across different threat scenarios.

On the other hand, the agricultural domain is inherently interdisciplinary: meteorologists, agronomists, and policy-makers must collaborate and develop trust on common platforms. While black-box machine learning models are powerful forecasters, they often lack of transparency impeding their adoption in high-stakes decision environments (Zhou et al., 2021; Arrieta et al., 2020). Explainable Artificial Intelligence (XAI) techniques seek to address this issue by illuminating the reasoning behind model predictions, thereby enhancing users' trust, accountability, and uptake of AI-driven tools (Arrieta et al., 2020). In the context of hazard forecasting, XAI can bolster confidence in an EWS by clarifying why a model predicts an extreme event. For example, Budimir et al. (2025) identified the atmospheric signals or precursor conditions that led to alerts in his recent work to provide explanation. Such transparency is crucial for interdisciplinary trust: stakeholders from different fields can verify that the model's behavior aligns with domain knowledge and can more readily act on its warnings. Modern studies have accordingly begun integrating XAI into climate risk tools, with notable success in hazard-specific applications (e.g., AI models augmented with explainability to detect wildfires, landslides, or floods) (Hrast Essenfelder et al., 2025; Cilli et al., 2022). Nevertheless, most current XAI applications in time-series hazard forecasting tend to focus on feature-level explanations, emphasizing which input features (e.g., temperature, precipitation indices) most influenced a prediction (Hrast Essenfelder et al., 2025; Cilli et al., 2022). These feature attributions, while informative, do not capture the temporal development of extreme events — that is, they tell what factors mattered but not when. For sequential models that ingest meteorological time series, understanding the timing and evolution of predictive signals is vital. For instance, decision-makers may wish to know whether an impending drought warning was driven by anomalous rainfall patterns in preceding months or by a sudden recent weather shift. However, typical feature importance analyses (e.g., SHAP value rankings of features) lack a temporal dimension and thus cannot reveal how early-season conditions versus immediate precursors contributed to the hazard forecast. This limitation in existing XAI research motivates the pursuit of richer, temporal explanations that elucidate the sequence of events leading to a forecast.

To address these deficiencies, we develop a unified multi-hazard early warning system for agriculture that not only forecasts multiple extreme events but also provides explainable temporal insights into each prediction as shown in Figure 1. Our framework leverages sequential deep learning models with an attention mechanism, to learn patterns from a decade of weather station data (2010-2023) from the United States and to provide probabilistic forecasts for six hazard types: extreme cold, floods, frost, hail, heat, and heavy rainfall. The sequential models allows the system to capture complex temporal dependencies in meteorological data. Moreover, we embed explainability at the core of the system. The attention mechanism provides an intrinsic form of temporal explanation, by weighting the importance of past time steps when making each prediction. In addition, we adapt TimeSHAP (Bento et al., 2021) — a model-agnostic XAI method extending SHAP (Lundberg and Lee, 2017) to sequential data — to generate post-hoc explanations that quantify the contribution of each time step (and each feature) to a given warning. Time-SHAP computes feature-level and timestep-level attribu-tions for recurrent models, thereby identifying not only which features but also which time points or periods were most influential in the model's decision. By combining attention-based explainability with TimeSHAP analyses, our system yields robust temporal explanations for every hazard forecast, helping to reveal the seasonal and short-term precursors that drive extreme agricultural events. We posit that this unified, explainable multi-hazard EWS will enhance transparency and interdisciplinary trust in the model outputs, ultimately improving the system's effectiveness as a decision-support tool for climate resilience and agricultural risk management.

The structure of this paper is organized as follows: Section 1 introduces the motivation for developing a unified, explainable multi-hazard early warning system in agriculture. Section 2 reviews relevant literature on early warning systems and XAI for climate hazard forecasting. Section 3 describes the data sources, preprocessing steps, and selection of agricultural regions used in this study. Sec-tion 4 outlines the sequential modeling framework, train-ing mechanism, and the integration of explainability techniques. Section 5 presents forecasting outcomes and explainability analyses, including both local and global temporal explanations. Finally, Section 6 concludes the study with a comprehensive summary.

2. Related Work

Research on early warning systems (EWS) has gained renewed momentum since the launch of the United Nations' "Early Warnings for All" initiative in 2022 (WMO, 2022). In this context, Abdalzaher et al. (2023) proposed an architecture that integrates Internet-of-Things sensors with machine learning techniques to enhance earthquake early warning systems in urban environments. Their framework aims to process real-time seismic signals more efficiently, thereby improving alert responsiveness in smart cities. Yet, the effectiveness of EWS is not solely determined by technological advances. As shown by Shah et al. (2023), social and institutional challenges — such as limited trust in alerts, poor dissemination practices,

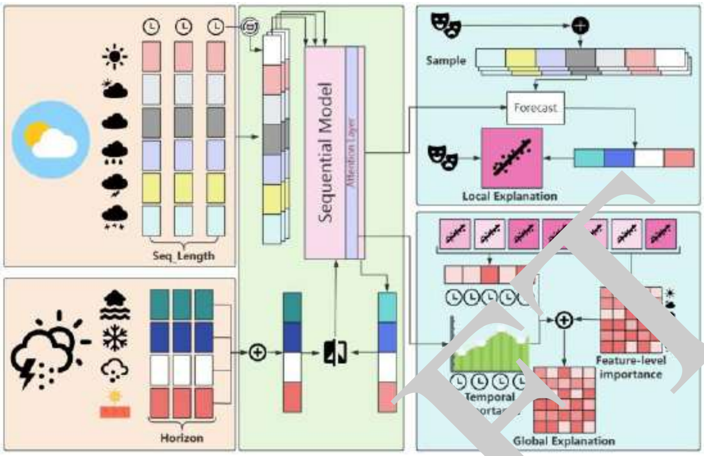


Figure 1: Graphical representation of the unified multi-hazard forecasting framework for agriculture. The diagram illustrates data integration from meteorological sources, sequential model architecture with attention mechanisms, and the dual-layered explainability approach employing TimeSHAP for both local and global hazard explanations.

and the marginalization of vulnerable groups — can undermine the utility of flood warnings in regions like Pakistan. Their findings underscore the importance of NGO-led communication strategies to fill these gaps. Moving into 2025, attention has increasingly turned to multi-hazard and integrated systems. For instance, Reichstein et al. (2025) presented a framework that couples AI-driven climate forecasting with impact assessments, advocating for user-focused design and transparent model governance to address complex climate risks more proactively. Complementing this direction, Rokhideh et al. (2025) assessed EWS progress within the Sendai Framework, identifying structural limitations such as fragmented hazard coverage, insufficient funding in the Global South, and inconsistent terminology across platforms. They call for more inclusive and coordinated systems that merge top-down policies with local engagement. In agriculture, De Clercq et al. (2025) advanced an operations research-based model for anticipatory action, using optimization methods to activate timely interventions — such as input distribution or early financial assistance — based on predictive agricultural hazard signals. This approach offers an alternative to rigid threshold-based warnings and has shown improved responsiveness in farming contexts.

In parallel with advances in early warning systems, the field has witnessed significant progress in explainable artificial intelligence (XAI), particularly in the context of time-series modeling. Nayebi et al. (2023) proposed WindowSHAP, a method that segments time series to efficiently estimate Shapley values, thereby identifying key temporal regions driving a model's output while keeping computational demands manageable. In a related effort, Zheng et al. (2024) addressed sequential decision-making by introducing R2RISE, a technique that highlights influential frames in imitation learning tasks, offering insights into how policies respond to specific temporal contexts. Within meteorology, XAI methods have begun to play a more prominent role in enhancing the explainability of hazard prediction systems. For example, Cilli et al. (2022) applied an explainable random forest model to identify wildfire risk in Southern Europe, leveraging feature attribution to uncover major contributing factors — such as drought indicators and vegetation health — and to delineate regions of elevated concern. Addressing a broader range of climate hazards, Hrast Essenfelder et al. (2025) designed an XAI framework guided by domain expertise, where models trained on historical agro-climatic data produce probabilistic warnings for extreme events like heatwaves and droughts, accompanied by explainable justifications that align with expert reasoning. Despite these advancements, concerns have been raised about the limits of post hoc explainability. O'Loughlin et al. (2025) argue for a shift toward inherently transparent modeling approaches, drawing inspiration from the modular structure of physics-based climate simulations, where each component can be systematically evaluated. They suggest that embedding explainability directly into the model architecture offers deeper and more actionable understanding than retrospective explanations. Complementing these perspectives, Pathania and Gupta (2025) developed a transformer-based drought forecasting model tailored for India, incorporating attention mechanisms to reveal which spatial and temporal inputs are most critical to the prediction. Their work demonstrates that high-performing deep learning models can still yield explainable outputs, thereby facilitating informed decision-making in climate-sensitive domains. Collectively, this body of work reflects a growing emphasis on not only achieving accurate multi-hazard forecasts but also ensuring their outputs are understandable and actionable for stakeholders in agriculture and environmental planning.

3. Data

For this study, we select climate data from four counties of the United States: i) Sonoma (California), ii) Kent (Michigan), iii) Adams (Pennsylvania), and iv) Yakima (Washington). They are widely recognized for their significant contributions to regional agriculture. These locations are among the leading fruit-producing areas in their respective states referring to FoodData Central. Daily weather observations at the county level, spanning from 2010 to 2023, were collected from nearby NOAA weather stations using the NCEI Climate Data Online (CDO) platform. Station selection prioritized feature completeness and temporal coverage. The meteorological dataset includes features related to temperature (e.g., daily highs and lows), precipitation, snowfall, and sunlight exposure. To capture weather-related hazards, we additionally sourced historical records of extreme events from NOAA's Storm Events Database covering the same period. This dataset includes event type, severity, and duration. We excluded events of light or moderate impact based on associated economic losses, agricultural damage, and casualties, retaining only six categories of severe events: Extreme Cold, Flood, Frost, Hail, Heat, and Extreme Rain. Droughts were excluded due to their infrequent reporting in the selected regions.

Prior to modeling, several preprocessing steps were applied to prepare the data. Meteorological features with substantial missing values were discarded, and event records were aggregated to the county level to ensure spatial consistency with the climate data. The six severe event types were converted into binary indicators through one-hot encoding, denoting event occurrence on a daily basis. The weather and event datasets were then temporally aligned and merged. A 14-day forecasting window was adopted: for each observation day, the number of severe events occurring within the following 14 days was summed by type to construct multi-label prediction targets. Lastly, all continuous meteorological features were standardized using a zero-mean, unit-variance transformation to facilitate model training.

4. Methodology

The forecasting framework presented in this research integrates several sequential models enhanced by attention mechanisms to simultaneously predict multiple agricultural hazards (as shown in Figure 1). Our approach employs a many-to-one structure, outputting predictions representing the anticipated occurrence of extreme events within a 90-day forecasting horizon. Daily meteorological inputs, gathered from weather stations distributed across the United States, are formatted as two-dimensional ar-rays, reflecting time steps and climate features, and expand into three-dimensional arrays when batch processing is used. Model performance and stability are rigorously assessed across various sequence lengths and stride sizes to optimize temporal coverage and resolution, thereby enhancing predictive precision.

We employ the Poisson Negative Log Likelihood (NLL) as loss function in model training. It was chosen for its suitability in modeling the nature of our target variables — discrete event counts such as the number of frost, hail, or drought occurrences over a given period. Unlike standard regression losses that assume continuous and normally distributed targets, Poisson NLL more accurately captures the sparsity and stochasticity typical of agricultural hazard data (Nelder, 1974). For optimization, we use the AdamW algorithm with a learning rate of 1×10^{-3}. AdamW is selected for its ability to decouple weight decay from the gradient update, offering more effective regularization (Parikh et al., 2014). This is particularly advantageous given the diversity of regional climates and the sequential nature of the input data, where overfitting is a common concern. Its stable convergence characteristics and generalization performance further support its application in our deep learning setup.

Several model architectures are evaluated, including Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Transformer-based methods. Attention mechanisms, specifically Bahdanau attention layers, are integrated into the LSTM and BiLSTM models to effectively capture critical temporal dependencies within input sequences (please refer to Zhang et al. (2023a) for details). Mathematically, given a sequence $X = x_1, x_2, \ldots, x_T$, attention-weighted representations are computed through $\alpha_t = \text{softmax}(\text{score}(h_t, s_T))$ and aggregated as $c = \sum_{t=1}^{T} \alpha_t h_t$, where h_t denotes the hidden state at time t, s_T the sequence's final state, and α_t the associated attention weights. The Transformer model incorporates multi-head self-attention mechanisms and position-wise feed-forward networks, deliberately configured with compact embedding dimensions and fewer transformer layers to optimize computational load and predictive accuracy.

Explainability of our framework is ensured through both local and global explanatory methodologies. At the local scale, TimeSHAP is employed to determine temporal feature significance using Shapley values, identifying how each feature at specific time points contributes to

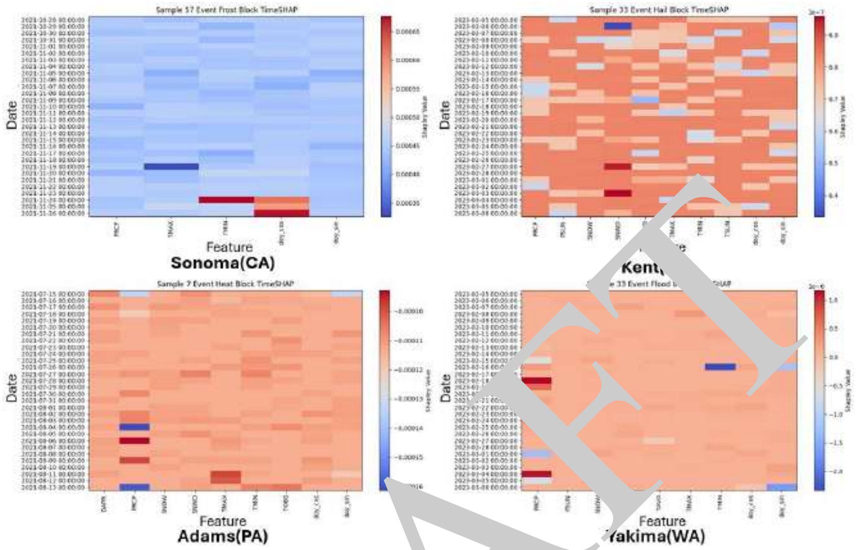


Figure 2: Local explainability heatmaps depicting the temporal and spatial variability in feature significance for distinct extreme weather events across selected agricultural counties. The color intensity indicates the magnitude of feature influence, reflecting the meteorological dynamics specific to each hazard type and location.

individual predictions. This method provides nuanced, instance-specific insights into model decision-making processes. Recognizing the limitations of directly aggregating local Shapley values due to their variability, we propose a novel global explanatory method. First, local TimeSHAP results generate two-dimensional importance matrices which are indexed temporally and by feature. Second, critical temporal segments are identified using either the magnitude of Shapley values or attention weights derived from the forecasting step. Finally, global feature relevance is calculated by summing these temporal and feature-specific importance measures. This global matrix offers a clearer view of which features tend to matter most, and when, across different hazard types and regions. For example, it can reveal emerging patterns such as rising temperatures in early spring being a consistent precursor to frost risk. Such insights support more informed, proactive responses by highlighting not just what the model has learned, but how its reasoning aligns with agronomic understanding. By integrating both instance-level specificity and population-level regularities, the dual-layered approach makes the model's decision process more transparent and actionable for stakeholders to evaluate predictions and agricultural risks.

5. Result

The forecasting performance of sequential models across the four selected agricultural counties is presented in Table 1. Given regional differences in meteorological data availability, we implemented region-specific models instead of a single universal approach. The results indicate particularly strong forecasting accuracy in Pennsylvania, with the BiLSTM model achieving the lowest errors (MAE: **0.0201**, RMSE: **0.0704**). In contrast, higher forecasting errors emerged in Washington, largely due to difficulties accurately predicting Flood events (MAE 0.30) and Heat events (MAE 0.10). Nonetheless, the sequential models demonstrated robust predictive capabilities overall, with BiLSTM models frequently outperforming Transformer models. Based on these outcomes, we selected the optimal model for each region to inform subsequent explainability analyses.

Figure 2 visualizes local explainability analyses for individual extreme weather events, clearly illustrating both temporal and spatial variations. The differing magnitudes within the heatmaps reflect shifts in the temporal dynamics and significance of meteorological features throughout event periods. Furthermore, distinct feature prominence across regions highlights unique local climatic con-

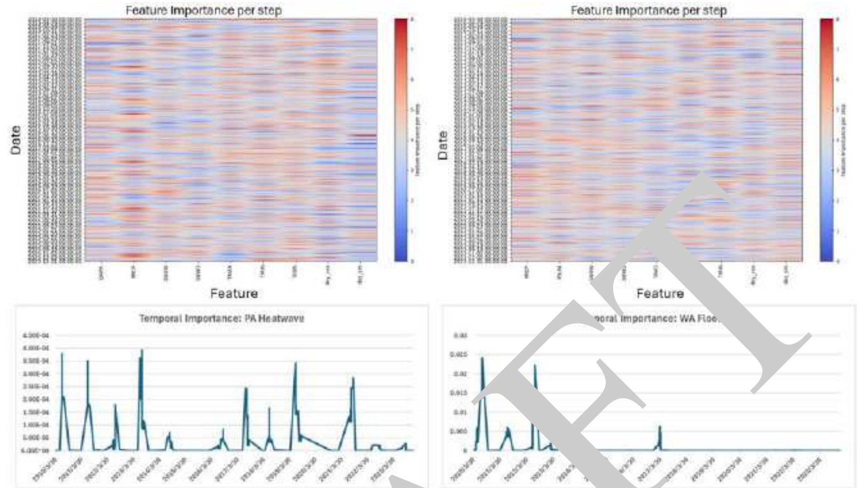


Figure 3: Aggregated global explainability analysis for California and Michigan, highlighting distinct seasonal patterns and the dominant climatic features underpinning frost and hail event predictions. The analysis emphasizes temperature and precipitation dynamics, aligning closely with recognized meteorological processes.

Table 1: Average model performance over all selected extreme events on unseen samples

	LSTM		BiLSTM		Transformer	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Sonoma(CA)	0.0562	0.1642	**0.0442**	0.1649	0.0608	**0.1629**
Kent(MI)	0.0365	0.1697	**0.0313**	**0.1692**	0.0360	0.1701
Adams(PA)	0.0261	0.0711	**0.0201**	**0.0704**	0.0308	0.0716
Yakima(WA)	0.0758	0.2901	0.0683	0.2941	**0.0652**	**0.2791**

ditions influencing extreme event profiles. For instance, Sonoma County (California) experiences pronounced temporal shifts in minimum (TMIN) and maximum (TMAX) temperatures during frost events, directly corresponding to frost formation mechanisms. In Kent County (Michigan), hail events exhibit variable importance for snow depth (SWND), suggesting its potential influence on near-surface thermal dynamics. Adams County (Pennsylvania) prominently features daytime maximum temperatures and precipitation (PRCP) during heat events, capturing the agricultural impact of prolonged dry spells. In Yakima County (Washington), flood events underscore precipitation and soil moisture levels, consistent with the region's susceptibility to sustained rainfall and flooding. These distinctions align well with established meteorological principles, offering clear guidance for targeted agricultural risk mitigation.

Figure 3 provides aggregated global explanations for California and Michigan, elucidating how regional climate patterns influence frost and hail events. In California, a pronounced seasonal rhythm emerges, with meteorological feature importance peaking sharply during colder months, coinciding precisely with typical frost conditions. Minimum and maximum temperatures, along with precipitation, emerge consistently as key predictive features. These findings resonate with known meteorological processes, notably nocturnal radiative cooling and saturation condi-tions critical for frost formation. In Michigan, however, the analysis highlights different seasonal periodicities associated with hail events. Here, precipitation and snow depth emerge as the dominant indicators. Heavy precipitation marks strong convective storm activity, which is vital for hail development. Additionally, snow depth potentially affects surface temperature gradients, influencing boundary-layer conditions conducive to storm formation during seasonal transitions. This alignment of model interpretations with recognized meteorological dynamics underscores their credibility.

Combining local and global explanations, alongside with feature-level and temporal perspectives, provides a comprehensive framework for interpreting extreme event forecasts. This integrated approach is crucial for unified multi-hazard forecasting, as it captures the intricate relation-ships of meteorological features across varying temporal scales and spatial contexts. By addressing both immediate and longer-term climatic influences, enhanced explainability significantly aids stakeholder decision-making, enabling precise interventions, efficient resource management, and improved disaster risk mitigation strategies tailored to diverse agricultural environments.

6. Conclusion

This study introduces a unified multi-agricultural hazard forecasting system that effectively integrates deep learning models with robust temporal explainability techniques. Our approach demonstrates high predictive accuracy for multiple concurrent climate hazards across diverse agricultural contexts. By combining attention mechanisms and TimeSHAP, the proposed system not only pinpoints key predictive features but also reveals critical timing in the buildup to hazardous events. These insights empower stakeholders, such as farmers and policymakers, to implement timely and targeted mitigation strategies. Ultimately, the integration of explainability within the forecasting model fosters greater transparency and stakeholder confidence, enhancing practical resilience to climate-driven agricultural risks.

References

Abdalzaher, M.S., Elsayed, H.A., Fouda, M.M., Salim, M.M., 2023. Employing machine learning and iot for earthquake early warning system in smart cities. Energies 16, 495.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion 58, 82–115.

Basher, R., 2006. Global early warning systems for natural hazards: systematic and people-centred. Philosophical transactions of the royal society a: mathematical, physical and engineering sciences 364, 2167–2182.

Beillouin, D., Schauberger, B., Bastos, A., Ciais, P., Makowski, D., 2020. Impact of extreme weather conditions on european crop production in 2018. Philosophical Transactions of the Royal Society B 375, 20190510.

Bento, J., Saleiro, P., Cruz, A.F., Figueiredo, M.A., Bizarro, P., 2021. Timeshap: Explaining recurrent models through sequence perturbations, in: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp. 2565–2573.

Betts, R.A., Alfieri, L., Bradshaw, C., Caesar, J., Feyen, insecurity projected at 1.5 c and 2 c global warming with a higher-resolution global climate model. Philosophi-cal Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376, 20160452.

Brett, L., Bloomfield, H.C., Bradley, A., Calvet, T., Champion, A., De Angeli, S., de Ruiter, M.C., Guerreiro, S.B., Hillier, J., Jaroszweski, D., et al., 2025. Science–policy–practice insights for compound and multi-hazard risks. Meteorological Applications 32, e70043.

Budimir, M., Trogrlić, R.Š., Almeida, C., Arestegui, M., Vásquez, O.C., Cisneros, A., Iriarte, M.C., Dia, A., Lizon, L., Madueño, G., et al., 2025. Opportunities and challenges for people-centered multi-hazard early warning systems: Perspectives from the global south. iScience 28.

Camps-Valls, G., Fernández-Torres, M.Á., Cohrs, K.H., Höhl, A., Castelletti, A., Pacal, A., Robin, C., Martinuzzi, F., Papoutsis, I., Prapas, I., et al., 2025. Artifi-cial intelligence for modeling and understanding extreme weather and climate events. Nature Communications 16, 1919.

Chen, H., Sun, J., 2018. Projected changes in climate extremes in china in a 1.5 c warmer world. International Journal of Climatology 38, 3607–3617.

Cilli, R., Elia, M., D'Este, M., Giannico, V., Amoroso, N., Lombardi, A., Pantaleo, E., Monaco, A., Sanesi, G., Tangaro, S., et al., 2022. Explainable artificial intelligence (xai) detects wildfire occurrence in the mediterranean countries of southern europe. Scientific reports 12, 16349.

Climate Data Online, N.C.f.E.I., . Climate data online. https://www.ncei.noaa.gov/cdo-web/datasets. Accessed: 2025-05-30.

De Clercq, D., Xu, L., de Ruiter, M., van den Homberg, M., van der Velde, M., Hall, J., Jaegermyer, J., Mahdi, A., 2025. Towards optimal anticipatory action: Maximizing the effectiveness of agricultural early warning systems with operations research. International Journal of Disaster Risk Reduction , 105249.

FoodData Central, U.D.o.A., . Fooddata central. https://fdc.nal.usda.gov/. Accessed: 2025-05-30.

Hrast Essenfelder, A., Toreti, A., Seguini, L., 2025. Expert-driven explainable artificial intelligence models can detect multiple climate hazards relevant for agriculture. Communications Earth & Environment 6, 207.

L., Friedlingstein, P., Gohar, L., Koutroulis, A., Lewis, K., Morfopoulos, C., et al., 2018. Changes in climate extremes, fresh water availability and vulnerability to food

Lesk, C., Rowhani, P., Ramankutty, N., 2016. Influence of extreme weather disasters on global crop production. Nature 529, 84–87.

Liu, L., Basso, B., 2020. Impacts of climate variability and adaptation strategies on crop yields and soil organic carbon in the us midwest. PloS one 15, e0225433.

Lobell, D.B., Di Tommaso, S., 2025. A half-century of climate change in major agricultural regions: Trends, impacts, and surprises. Proceedings of the National Academy of Sciences 122, e2502789122.

Lobell, D.B., Schlenker, W., Costa-Roberts, J., 2011. Climate trends and global crop production since 1980. Science 333, 616–620.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

Nayebi, A., Tipirneni, S., Reddy, C.K., Foreman, B., Subbian, V., 2023. Windowshap: An efficient framework for explaining time-series classifiers based on shapley values. Journal of biomedical informatics 144, 104438.

Nelder, J., 1974. Log linear models for contingency tables: a generalization of classical least squares. Journal of the Royal Statistical Society: Series C (Applied Statistics) 23, 323–329.

O'Loughlin, R.J., Li, D., Neale, R., O'Brien, T.A., 2025. Moving beyond post hoc explainable artificial intelligence: a perspective paper on lessons learned from dynamical climate modeling. Geoscientific Model Development 18, 787–802.

Parikh, N., Boyd, S., et al., 2014. Proximal algorithms. Foundations and trends® in Optimization 1, 127–239.

Pathania, A., Gupta, V., 2025. Interpretable transformer model for national scale drought forecasting: Attention-driven insights across india. Environmental Modelling & Software 187, 106394.

Powell, J., Reinhard, S., 2016. Measuring the effects of extreme weather events on yields. Weather and Climate extremes 12, 69–79.

Reichstein, M., Benson, V., Blunk, J., Camps-Valls, G., Creutzig, F., Fearnley, C.J., Han, B., Kornhuber, K., Rahaman, N., Schölkopf, B., et al., 2025. Early warning of complex climate risk with integrated artificial intelligence. Nature Communications 16, 2564.

Rokhideh, M., Fearnley, C., Budimir, M., 2025. Multi-hazard early warning systems in the sendai framework for disaster risk reduction: Achievements, gaps, and future directions. International Journal of Disaster Risk Science , 1–14.

Schlenker, W., Roberts, M.J., 2009. Nonlinear temperature effects indicate severe damages to us crop yields under climate change. Proceedings of the National Academy of sciences 106, 15594–15598.

Schmitt, J., Offermann, F., Söder, M., Frühauf, C., Finger, R., 2022. Extreme weather events cause significant crop yield losses at the farm level in german agriculture. Food Policy 112, 102359.

Shah, A.A., Ullah, A., Khan, N.A., Khan, A., Tariq, M.A.U.R., Xu, C., 2023. Community social barriers to non-technical aspects of flood early warning systems and ngo-led interventions: The case of pakistan. Frontiers in Earth Science 11, 1068721.

Storm Events Database, N.C.f.E.I., . Storm events database. https://www.ncdc.noaa.gov/stormevents/. Accessed: 2025-05-30.

WMO, 2022. Early warnings for all. https://library.wmo.int/viewer/58209. Accessed: 2025-05-30.

Zhang, A., Lipton, Z.C., Li, M., Smola, A.J., 2023a. Dive into deep learning. Cambridge University Press.

Zhang, T., Wang, D., Lu, Y., 2023b. Machine learning-enabled regional multi-hazards risk assessment considering social vulnerability. Scientific reports 13, 13405.

Zheng, B., Zhou, J., Liu, C., Li, Y., Chen, F., 2024. Explaining imitation learning through frames. IEEE Intelligent Systems .

Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A., 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics 10, 593.

References

[1] Victor, B., He, Z., Nibali, A.: A systematic review of the use of deep learning in satellite imagery for agriculture. arXiv preprint arXiv:2210.01272 (2022)

[2] Arrieta, A.B., D´ıaz-Rodr´ıguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garc´ıa, S., Gil-López, S., Molina, D., Benjamins, R., *et al.*: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)

[3] Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5), 593 (2021)

[4] Paudel, D., Wit, A., Boogaard, H., Marcos, D., Osinga, S., Athanasiadis, I.N.: Interpretability of deep learning models for crop yield forecasting. Computers and Electronics in Agriculture **206**, 107663 (2023)

[5] SHAP. Accessed: 2024-12-22 (2024). https://www.biaodianfu.com/shap.html

[6] Extremes, C.: El Niñ o's Impact on Australia's Weather and Climate. https://climateextremes.org.au/wp-content/uploads/El-Nin%CC%83os-Impact-on-Australias-Weather-and-Climate-ARC-Centre-of-Excellence-for-Climate-Extremes.pdf. Accessed: 2025-01-30

[7] Keele, S., et al.: Guidelines for performing systematic literature reviews in software engineering. Technical report, ver. 2.3 ebse technical report. ebse (2007)

[8] He, L., Fang, W., Zhao, G., Wu, Z., Fu, L., Li, R., Majeed, Y., Dhupia, J.: Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods. Computers and electronics in agriculture **195**, 106812 (2022)

[9] Mateo-Sanchis, A., Adsuara, J.E., Piles, M., Munoz-Mar´ı, J., Perez-Suay, A., Camps-Valls, G.: Interpretable long short-term memory networks for crop yield estimation. IEEE Geoscience and Remote Sensing Letters **20**, 1–5 (2023)

[10] Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., *et al.*: Integrating satellite and climate data to predict wheat yield in australia using machine learning approaches. Agricultural and forest meteorology **274**, 144–159 (2019)

[11] Sihi, D., Dari, B., Kuruvila, A.P., Jha, G., Basu, K.: Explainable machine learning approach quantified the long-term (1981–2015) impact of climate and soil properties on yields of major agricultural crops across conus. Frontiers in Sustainable Food Systems **6**, 847892 (2022)

[12] Nayak, H.S., Silva, J.V., Parihar, C.M., Krupnik, T.J., Sena, D.R., Kakraliya, S.K., Jat, H.S., Sidhu, H.S., Sharma, P.C., Jat, M.L., *et al.*: Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in northwest india. Field Crops Research **287**, 108640 (2022)

[13] Kern, A., Barcza, Z., Marjanovi´c, H., Árendás, T., Fodor, N., Bónis, P., Bognár, P., Lichtenberger, J.: Statistical modelling of crop yield in central europe using climate data and remote sensing vegetation indices. Agricultural and forest meteorology **260**, 300–320 (2018)

[14] Wolanin, A., Mateo-Garc´ıa, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Liangzhi, Y., Guanter, L.: Estimating and understanding crop yields with explainable deep learning in the indian wheat belt. Environmental research letters **15**(2), 024019 (2020)

[15] Paudel, D., Boogaard, H., Wit, A., Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I.N.: Machine learning for large-scale crop yield forecasting. Agricultural Systems **187**, 103016 (2021)

[16] Newman, S.J., Furbank, R.T.: Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data. Nature Plants **7**(10), 1354–1363 (2021)

[17] Hu, T., Zhang, X., Bohrer, G., Liu, Y., Zhou, Y., Martin, J., Li, Y., Zhao, K.: Crop yield prediction via explainable ai and interpretable machine learning: Dan-gers of black box models for evaluating climate change impacts on crop yield. Agricultural and Forest Meteorology **336**, 109458 (2023)

[18] Mariadass, D.A., Moung, E.G., Sufian, M.M., Farzamnia, A.: Extreme gradient boosting (xgboost) regressor and shapley additive explanation for crop yield prediction in agriculture. In: 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 219–224 (2022). IEEE

[19] Khaki, S., Wang, L., Archontoulis, S.V.: A cnn-rnn framework for crop yield prediction. Frontiers in Plant Science **10**, 492736 (2020)

[20] Khaki, S., Wang, L.: Crop yield prediction using deep neural networks. Frontiers in plant science **10**, 452963 (2019)

[21] Ryo, M.: Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. Artificial Intelligence in Agriculture **6**, 257–265 (2022)

[22] Huber, F., Engler, H., Kicherer, A., Herzog, K., Töpfer, R., Steinhage, V.: Grouping shapley value feature importances of random forests for explainable yield prediction. In: Intelligent Systems Conference, pp. 210–228 (2023). Springer

[23] Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S.,

Singh, A.K.: Crop yield prediction integrating genotype and weather variables using deep learning. Plos one **16**(6), 0252402 (2021)

[24] Celik, M.F., Isik, M.S., Taskin, G., Erten, E., Camps-Valls, G.: Explainable artificial intelligence for cotton yield prediction with multisource data. IEEE Geoscience and Remote Sensing Letters (2023)

[25] Jones, E.J., Bishop, T.F., Malone, B.P., Hulme, P.J., Whelan, B.M., Filippi, P.: Identifying causes of crop yield variability with interpretive machine learning. Computers and Electronics in Agriculture **192**, 106632 (2022)

[26] Torres-Tello, J., Ko, S.-B.: Interpretability of artificial intelligence models that use data fusion to predict yield in aeroponics. Journal of Ambient Intelligence and Humanized Computing **14**(4), 3331–3342 (2023)

[27] Duan, Z., Zheng, C., Zhao, S., Feyissa, T., Merga, T., Jiang, Y., Zhang, W.: Cold climate during bud break and flowering and excessive nutrient inputs limit apple yields in hebei province, china. Horticulturae **8**(12), 1131 (2022)

[28] Santos, J.A., Malheiro, A.C., Karremann, M.K., Pinto, J.G.: Statistical mod-elling of grapevine yield in the port wine region under present and future climate conditions. International Journal of Biometeorology **55**, 119–131 (2011)

[29] Knowling, M.J., Walker, R.R., Pellegrino, A., Edwards, E.J., Westra, S., Collins, C., Ostendorf, B., Bennett, B.: Generalized water production relations through process-based modeling: A viticulture example. Agricultural Water Management **280**, 108225 (2023)

[30] Gurbuz, I.B., Ozkan, G., Er, S.: Exploring kiwi fruit producers' climate change perceptions. Applied Fruit Science, 1–9 (2024)

[31] Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N.H., Islam, N.: A systematic literature review on crop yield prediction with deep learning and remote sensing. Remote Sensing **14**(9), 1990 (2022)

[32] Minoli, S., Müller, C., Elliott, J., Ruane, A.C., Jägermeyr, J., Zabel, F., Dury, M., Folberth, C., Fran¸cois, L., Hank, T., *et al.*: Global response patterns of major rainfed crops to adaptation by maintaining current growing periods and irrigation. Earth's Future **7**(12), 1464–1480 (2019)

[33] Farooq, M.S., Khaskheli, M.A., Uzair, M., Xu, Y., Wattoo, F.M., Rehman, O.U., Amatus, G., Fatima, H., Khan, S.A., Fiaz, S., *et al.*: Inquiring the inter-relationships amongst grain-filling, grain-yield, and grain-quality of japonica rice at high latitudes of china. Frontiers in Genetics **13**, 988256 (2022)

[34] Australia, V.: SA WINEGRAPE CRUSH SURVEY. https://vinehealth.com.au/news/surveys/sa-winegrape-crush-survey/. [Online; accessed 8-May-2024] (2023)

[35] Van Klompenburg, T., Kassahun, A., Catal, C.: Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture **177**, 105709 (2020)

[36] Zespri: Zespri Climate Risk Opportunities Report. https://www.zespri.com/content/dam/zespri/nz/sustainability/Zespri-Climate-Risk-Opportunities-(TCFD)-Report.pdf. [Online; accessed 9-May-2024] (2021)

[37] Zespri: Zespri Climate Change Adaptation Plan Summary. https://www.zespri.com/content/dam/zespri/nz/sustainability/Zespri-Climate-Change-Adaptation-Plan-Summary-2022.pdf. [Online; accessed 9-May-2024] (2022)

[38] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89 (2018). IEEE

[39] Chhetri, T.R., Hohenegger, A., Fensel, A., Kasali, M.A., Adekunle, A.A.: Towards improving prediction accuracy and user-level explainability using deep learn-ing and knowledge graphs: A study on cassava disease. Expert Systems with Applications **233**, 120955 (2023)

[40] Ahmed, M.S., Tazwar, M.T., Khan, H., Roy, S., Iqbal, J., Rabiul Alam, M.G., Hassan, M.R., Hassan, M.M.: Yield response of different rice ecotypes to meteorological, agro-chemical, and soil physiographic factors for interpretable precision agriculture using extreme gradient boosting and support vector regression. Complexity **2022**, 1–20 (2022)

[41] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

[42] Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R., D´ıaz-Rodr´ıguez, N.: Explainable artificial intelligence (xai) on timeseries data: A survey. arXiv preprint arXiv:2104.00950 (2021)

[43] USApple: Notes from the Orchard: Bloom Season in Michigan. https://usapple.org/news-resources/notes-from-the-orchard-bloom-season-in-michigan. [Online; accessed 9-May-2024] (2019)

[44] Lee, I.B., Jung, D.H., Kang, S.B., Hong, S.S., Yi, P.H., Jeong, S.T., Park, J.M.: Changes in growth, fruit quality, and leaf characteristics of apple tree (malus domestica borkh.'fuji') grown under elevated co2 and temperature conditions. (2023)

[45] Gallardo, R.K., Grant, K., Brown, D.J., McFerson, J.R., Lewis, K.M., Einhorn, T., Sazo, M.M.: Perceptions of precision agriculture technologies in the us fresh

apple industry. HortTechnology **29**(2), 151–162 (2019)

[46] Cusick, D., News, E.: How Climate Change Hurt this Year's Apple Har-vest. https://www.scientificamerican.com/article/how-climate-change-hurt-this-years-apple-harvest. [Online; accessed 15-May-2024] (2021)

[47] Li, M., Guo, J., Xu, C., Lei, Y., Li, J.: Identifying climatic factors and circulation indices related to apple yield variation in main production areas of china. Global ecology and conservation **16**, 00478 (2018)

[48] USApple: Industry Outlook 2023 report. https://usaa.memberclicks.net/assets/2023/Outlook2023/USAPPLE-OutlookReport-2023.pdf. [Online; accessed 15-May-2024] (2023)

[49] Wilton, J.: Summer management for quality. https://apal.org.au/summer-management-for-quality/. [Online; accessed 9-May-2024] (2018)

[50] Harris, R., Remenyi, T., Hayman, P., Thomas, D., Risbey, J., Petrie, P., Thatcher, M., Bindoff, N.: Australia's wine future: Adapting to short-term climate variability and long-term climate change. Final report to Wine Australia, Antarctic Climate and Ecosystems Cooperative Research Centre. Hobart, Tasmania (2019)

[51] Mosedale, J.R., Abernethy, K.E., Smart, R.E., Wilson, R.J., Maclean, I.M.: Climate change impacts and adaptive strategies: lessons from the grapevine. Global change biology **22**(11), 3814–3828 (2016)

[52] Australia, W.: National vintage report. https://www.wineaustralia.com/market-insights/national-vintage-report. [Online; accessed 8-May-2024] (2023)

[53] Zespri: WHERE ARE KIWI FRUIT GROWN. https://www.zespri.com/en-US/blogdetail/where-are-kiwifruit-grown#:~:text=Kiwifruit%20are%20grown%20commercially%20all,tart%20from%20a%20green%20kiwifruit. [Online; accessed 9-May-2024] (2024)

[54] Cradock-Henry, N.A.: New zealand kiwifruit growers' vulnerability to climate and other stressors. Regional Environmental Change **17**(1), 245–259 (2017)

[55] Zespri: Zespri Annual report. https://canopy.zespri.com/content/dam/new-canopy/nz/en/documents/public/about/annual-report/Annual-Report-2022-23.pdf. [Online; accessed 20-May-2024] (2023)

[56] Tait, A., Paul, V., Sood, A., Mowat, A.: Potential impact of climate change on hayward kiwifruit production viability in new zealand. New Zealand journal of crop and horticultural science **46**(3), 175–197 (2018)

[57] Kenny, G.J., Porteous, A.S.: Adapting to climate change in the kiwifruit industry. Ministry of Agriculture and Forestry (2008)

[58] westpac: Kiwifruit-Climate-Change-Fact-Sheet. https://www.westpac.co.nz/assets/Business/agribusiness/documents/Kiwifruit-Climate-Change-Fact-Sheet-Westpac-NZ.pdf. [Online; accessed 20-May-2024] (2022)

[59] Hernández, G., Craig, R., Tanner, D.: Assessment of alternative bud break enhancers for commercial kiwifruit production of 'zesy002'(gold3). Advances in Plant Dormancy, 289–300 (2015)

[60] Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V.K., Tanwar, S., Sharma, G., Bokoro, P.N., Sharma, R.: Explainable ai for healthcare 5.0: opportunities and challenges. IEEE Access **10**, 84486–84517 (2022)

[61] Černevičienė, J., Kabaˇsinskas, A.: Explainable artificial intelligence (xai) in finance: A systematic literature review. Artificial Intelligence Review **57**(8), 216 (2024)

[62] Bommer, P.L., Kretschmer, M., Hedström, A., Bareeva, D., Höhne, M.M.-C.: Finding the right xai method—a guide for the evaluation and ranking of explainable ai methods in climate science. Artificial Intelligence for the Earth Systems **3**(3), 230074 (2024)

[63] Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems **32**(11), 4793–4813 (2020)

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

[65] Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and computing **14**, 199–222 (2004)

[66] Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems **33**(12), 6999–7019 (2021)

[67] Reynolds, D.A., et al.: Gaussian mixture models. Encyclopedia of biometrics **741**(659-663) (2009)

[68] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)

[69] Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach,

CA, USA, pp. 4765–4774 (2017)

[70] Aas, K., Jullum, M., Løland, A.: Explaining individual predictions when fea-tures are dependent: More accurate approximations to shapley values. Artificial Intelligence **298**, 103502 (2021)

[71] Molnar, C., König, G., Bischl, B., Casalicchio, G.: Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. Data Mining and Knowledge Discovery **38**(5), 2903–2941 (2024)

[72] Mukherji, A.: Climate change 2023 synthesis report (2023)

[73] Lobell, D.B., Schlenker, W., Costa-Roberts, J.: Climate trends and global crop production since 1980. Science **333**(6042), 616–620 (2011)

[74] Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 4765–4774 (2017)

[75] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)

[76] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998) https://doi.org/10.1109/5.726791

[77] Han, H., Li, Y., Zhu, X.: Convolutional neural network learning for generic data classification. Inf. Sci. **477**, 448–465 (2019) https://doi.org/10.1016/J.INS.2018.10.053

[78] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

[79] Bar-Joseph, Z., Gifford, D.K., Jaakkola, T.S.: Fast optimal leaf ordering for hierarchical clustering. Bioinformatics **17**(suppl 1), 22–29 (2001)

[80] Ge, T., Luo, X., Wang, Y., Sedlmair, M., Cheng, Z., Zhao, Y., Liu, X., Deussen, O., Chen, B.: Optimally ordered orthogonal neighbor joining trees for hierarchical cluster analysis. IEEE Transactions on Visualization and Computer Graphics **30**(8), 5034–5046 (2023)

[81] Kukal, M.S., Irmak, S.: Climate-driven crop yield and yield variability and climate change impacts on the us great plains agricultural production. Scientific reports

8(1), 1–18 (2018)

[82] Zespri: Climate Risk Opportunities Report. https://www.zespri.com/content/dam/zespri/nz/sustainability/Zespri-Climate-Risk-Opportunities-(TCFD)-Report.pdf. Accessed: 2025-01-30

[83] Kaur, N., *et al.*: Projected climate change under different scenarios in central region of punjab, india. Journal of Agrometeorology **18**(1), 88 (2016)

[84] Chen, H., Sun, J.: Projected changes in climate extremes in china in a 1.5 c warmer world. International Journal of Climatology **38**(9), 3607–3617 (2018)

[85] Powell, J., Reinhard, S.: Measuring the effects of extreme weather events on yields. Weather and Climate extremes **12**, 69–79 (2016)

[86] Liu, L., Basso, B.: Impacts of climate variability and adaptation strategies on crop yields and soil organic carbon in the us midwest. PloS one **15**(1), 0225433 (2020)

[87] Crane-Droesch, A.: Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters **13**(11), 114003 (2018)

[88] Taylor, S.J., Letham, B.: Forecasting at scale. The American Statistician **72**(1), 37–45 (2018)

[89] Kaushik, I., Singh, S.M.: Seasonal arima model for forecasting of monthly rainfall and temperature. Journal of Environmental Research and Development **3**(2), 506–514 (2008)

[90] Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. Journal of econometrics **31**(3), 307–327 (1986)

[91] Ray, D.K., Gerber, J.S., MacDonald, G.K., West, P.C.: Climate variation explains a third of global crop yield variability. Nature communications **6**(1), 5989 (2015)

[92] Wang, T., Li, N., Li, Y., Lin, H., Yao, N., Chen, X., Li Liu, D., Yu, Q., Feng, H.: Impact of climate variability on grain yields of spring and summer maize. Computers and Electronics in Agriculture **199**, 107101 (2022)

[93] McGregor, S., Gallant, A., Rensch, P.: Quantifying ensos impact on australia's regional monthly rainfall risk. Geophysical Research Letters **51**(6), 2023–106298 (2024)

[94] Cleland, E.E., Chuine, I., Menzel, A., Mooney, H.A., Schwartz, M.D.: Shifting plant phenology in response to global change. Trends in ecology & evolution **22**(7), 357–365 (2007)

[95] Webb, L., Whetton, P., Bhend, J., Darbyshire, R., Briggs, P., Barlow, E.: Earlier wine-grape ripening driven by climatic warming and drying and management practices. Nature Climate Change **2**(4), 259–264 (2012)

[96] Rogiers, S.Y., Greer, D.H., Liu, Y., Baby, T., Xiao, Z.: Impact of climate change on grape berry ripening: An assessment of adaptation strategies for the australian vineyard. Frontiers in Plant Science **13**, 1094633 (2022)

[97] Hassan, T., Gulzar, R., Hamid, M., Ahmad, R., Waza, S.A., Khuroo, A.A.: Plant phenology shifts under climate warming: a systematic review of recent scientific literature. Environmental Monitoring and Assessment **196**(1), 36 (2024)