A multi-farm global to local expert-informed machine learning system for strawberry yield forecasting

Matthew Beddows[a], Georgios Leontidis[a,*]

[a]*School of Natural and Computing Sciences & Interdisciplinary Centre for Data and AI, University of Aberdeen, AB24 3UE, Aberdeen, United Kingdom*

Abstract

The importance of forecasting crop yields in agriculture cannot be overstated. The effects of yield forecasting are observed in all aspects of the supply chain from staffing, supplier demand, food waste and other business decisions. However, the process is often inaccurate and far from perfect. This paper explores the potential of using expert forecasts to enhance the crop yield predictions of our global-to-local machine learning system. Additionally, it investigates the ERA5 climate model's viability as an alternative data source for crop yield forecasting in the absence of on-farm weather data. We find that by combining both the expert forecasts and the ERA5 climate model with the machine learning model, we can – in most cases – get better forecasts that outperform the growers' pre-season forecasts and the machine learning-only models. Our expert-informed model attains yield forecasts for 4 weeks ahead with an average RMSE of 0.0855 across all plots and an RMSE of 0.0872 with ERA5 climate data included.

Keywords: Machine Learning, XGBoost, Strawberry Yield Forecasting, Time Series

1. Introduction

The Agriculture sector is key to the UK economy, utilising 71% of the UK's total land area [1]. Globally, the agricultural sector has faced substantial disruption due to shifting geopolitics and the impact of COVID-19 on labor availability. In the UK these challenges have been exacerbated by the complexities of transitioning policies to restructure the industry post Brexit [2]. One such policy is the UK transitioning away from the European Union's Common Agricultural Policy (CAP), which compensated growers based upon the amount of land that they farmed[1] .

Amidst these evolving challenges, the ability to accurately forecast crop yields, both before and during the growing season, emerges as a critical tool for agricultural resilience and decision-making[3].

*Corresponding author
Email address: georgios.leontidis@abdn.ac.uk (Georgios Leontidis)

And these forecasts are among the most valuable pieces of information the grower could be provided with [4]. Increasing the accuracy of these forecasts means we can reduce the business risks the growers take[5].

It is well known that there will be several major fruit waves throughout a growing season. However, it is when the waves begin, which is the difficult part to predict [6]. This difficulty in forecasting is further compounded in tasks like predicting strawberry yields and prices, which are influenced by a myriad of complex factors. Variables such as weather, soil conditions, and irrigation play crucial roles in determining the yield. The inherent uncertainty of these factors adds layers of complexity to the forecasting process, making it a challenging task [7]. Growers create their forecasts based upon previous experience and seasonal conditions, and then used this as a guide to construct management decisions[8]. We propose a dynamic method that is able to use this expertise along side an ML solution. This research is also important as even as of 2020 there still is a significant need to develop ML techniques for fresh produce, including Strawberries [9]. Our system utilises a global to local method where we train a single model on data from various farms, and then use this single model to make individual predictions for all of the farms and their respective plots. This paper delves into the the intersection of machine learning and crop yield forecasting and investigates the integration of growers expert knowledge with machine learning techniques. We examine how embedding growers' seasoned wisdom into our model is able to enhance the precision and reliability of our predictions.

This paper also aims to address a critical challenge in the realm of agricultural data management. Specifically, we investigate the usage of the ERA5 climate model as an alternative data source for crop yield forecasting when weather data was not captured at the farm. A crop's yield largely depends on the weather conditions during the growing season [10]. Growers frequently rely on weather data to inform decisions about agricultural practices, such as planting, irrigation, and pest control. However, we have observed that many growers often utilize weather data on an ad-hoc basis without retaining it for future reference. This practice can lead to the under-utilization of valuable historical weather information, which could provide insights into crop trends and inform more resilient and sustainable agricultural strategies.

Addressing the issue of growers not retaining their weather data is not a quick fix, the change would involve infrastructural and behavioral changes. Even with growers adopting better data management practices moving forward, the challenge remains for historical data, which is currently unavailable. To tackle this issue comprehensively, we explore the potential of the ERA5 climate model as a valuable resource for providing historical weather data that can be incorporated into our dataset. By investigating the suitability of ERA5, we aim to provide a method for utilising historical yield data where temperature

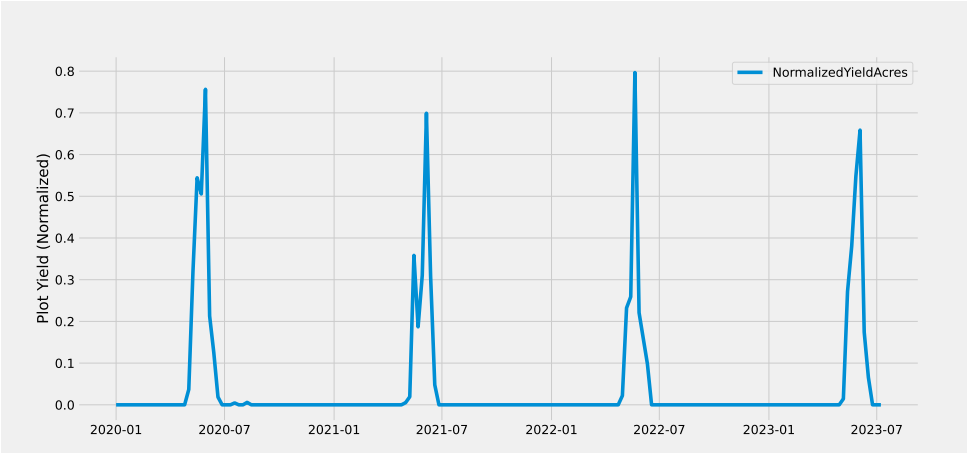


Figure 1: 4 Years of normalized yields from Farm 1 Plot 4.

data was not recorded.

Moreover, our study leverages real-world data provided by Angus Soft Fruits, a leading supplier of berries to UK and European retailers, enhancing the practical relevance of our findings. The valuable data they have provided has been instrumental in training our neural network (detailed in Section 2) and in evaluating their performance (in Section 4). The literature suggests that machine learning approaches are highly effective for yield forecasting, and our experiments confirm this. Notably, we incorporate XGBoost [11] into our end-to-end framework and benchmark against the growers' own forecasts.

2. Related work

The agri-food sector has experienced notable advancements through the integration of machine learning and data-driven approaches, leading to a vast array of applications in this broad field. These technologies are paving the way for agriculture to evolve into a data-driven, intelligent, agile, and autonomous connected system of systems [12, 13, 14, 15, 16, 17]. The sector has already seen the benefits of machine learning in a variety of different topics including pest prediction and prevention [18].

Machine learning within agri-food has also had success with yield forecasting [19]. However, there are fewer examples of this when we restrict the success to only strawberries, though even then we are able to find successful applications [20, 7, 21].

Recent advancements in agricultural forecasting for forecasting strawberry yields have shown promising results through the application of deep learning models. Notably, some studies have enhanced their predictive accuracy by incorporating satellite imagery and detailed soil parameter data [7].

However, these methods often rely on the availability of extensive environmental data and clear imagery for analysis. In our specific context, our data setup lacks the necessary sensors to collect detailed soil data, and also does not have the required historical data if we were to install them. Furthermore, our crops are housed within poly-tunnels, which poses a unique challenge as they obstruct the view of satellite cameras, rendering satellite imagery ineffective for monitoring the crops within.

In the realm of strawberry yield forecasting, transformers have been successfully applied to predict yield under varying conditions and settings [20, 21]. However their research utilized comprehensive datasets, including detailed irrigation information from tabletop systems, extensive environmental data from weather stations, and frequent yield quality reports from strawberry picking teams. These rich data sources, which were precise and high resolution are in contrast to much less detailed real world data we have access too. Even with this, they mentioned having issues with data availability. Despite this advantage in their data quality, we explored the use of Transformers in our research. However, as explained further in Section 3, they were not suitable for our task due to the lower resolution and quality of our data.

Other research has demonstrated the effectiveness of using unmanned aerial vehicles (UAVs) with mounted cameras for predicting strawberry yields and dry biomass, such as in the work presented by Zheng et al. in [22]. However, this approach may not be feasible for the farms under our consideration in Scotland, where strawberries are cultivated in polytunnels to adapt to the colder climate. This is in contrast to the open-field cultivation practices common in Florida's warmer environment.

3. Materials and methods

3.1. Angus Soft Fruits data

We collaborated with Angus Soft Fruits, a company that generously supplied us with both current and historical data on their soft-fruit crop yields from farms all across Scotland and England. Additionally, they provided pre-season and weekly forecasts. The pre-season is a forecast all of the growers make for every plot on their farm, they calculate both what yield they expect from each plot, and when they expect them. The weekly forecasts are just the pre-season forecast updated weekly, this is useful as it can account for changes in weather, any issues with crops or any other unexpected change during the year.

3.1.1. Pre-season forecast

For every year from 2020, we had a document called the "pre-season forecast". This document was compiled from the individual pre-seasons each farm submitted to the company. This document contained

the expected yield of each plot each week on the farm as well as other important information such as the date the crop was planted.

3.1.2. Weekly forecasts

Similarly to the pre-season forecast each farm also creates weekly updates from the growers. These are individual documents for each farm, for each week. These contain the growers updated forecasts. Often these forecasts are less accurate than the forecasts at the beginning of the year. Management noted that this is often because growers will often exaggerate the effects of positive/negative effects on their crops. As we only had the weekly forecasts for 2023 we utilised this data more as a method of reviewing the effectiveness of our predictions over time rather than implementing it as a feature.

3.2. Data wrangling

Though we had data ranging back to 2011 we elected to use data from 2020 onwards as we had the matching pre-season and weekly for these years. This provides use three years for training (2020,2021 and 2022) and one year for testing (2023). In the dataset we compiled from this data we focused only on strawberries and utilised the date, received yield, farm name, plot name, plot acres, annual predicted tonnes, strawberry variety, tunnel type, plant age and the growers prediction for the week. The model was trained and tested on the plots of one specific strawberry variety across 6 farms. The plots all very in size and shape, containing different amounts of poly-tunnels.

The size of the time-step was something that was very important when using XGBoost. The strawberry harvests, which typically occur twice a week, although the frequency can vary. On days without harvests, there was no data points, leading to irregular time-steps in our dataset. To address this, we modified the dataset by recording a yield of 0kg for days without harvests. This allowed us to maintain consistent daily records. We then aggregated this data on a weekly basis, providing a weekly yield for each plot at each farm throughout each year.

Another challenge we encountered was dealing with missing or incorrect data entries. It was common to find dates and values that were inaccurately recorded. Correcting these errors was a necessary step, as some level of data inconsistency is often inevitable in real-world scenarios. From the pre-season, we had a planted date for most crops. Any crops where this data was left blank we'd check the previous years to see if there was a planted data we could use, if not we would just use the current year. We could then use this to create an age for each plant. This was important as the age of a crop will have an effect on it's yield.

As we were using XGBoost, we adapted how we'd have to window the data in comparison to a Deep Learning network such as an LSTM or time-series Transformers which utilise a window size. We had to manually craft the input features to incorporate historical data. To emulate predicting yields 4 weeks ahead, we included the yields (Y) from 5, 6, 7, and 8 weeks prior from each specific plot as features in the model's input (X). This was applied not only to the yield data and the growers' forecasts but also to the historical temperature values which we pulled from the ERA5 system.

Our dataset contained categorical data, including variables like tunnel type, farm name, and plot name. To make this data compatible with our model, it was necessary to convert these categories into numerical form. We accomplished this using the LabelEncoder from the sklearn library. This method assigns a unique numerical value, starting from 0, to each category.

After finalizing the dataset, we normalised all of the numerical values to a range between 0 and 1. Although this normalization step did not significantly affect the model's performance, it was crucial for maintaining the data privacy of the farms involved. To do this we used a simple Min-Max scalar from sklearn shown below:

$$x'_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{1}$$

3.2.1. Datasets for comparison

With our data cleaned data we created four different datasets for our comparison:

1. The base dataset (machine learning model)
2. The data set with the growers forecast added (expert-informed model)
3. The data set with the added Era5 weather data added (machine learning model)
4. The data set with both the growers forecast and the Era5 weather data added (expert-informed model)

3.3. ERA5 data

As we had no weather data available and required an alternative we decided to use the "ERA5-Land Hourly - ECMWF Climate Reanalysis" dataset. The "Fifth Generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis" or ERA5 as it is more commonly known, represents a high-resolution and comprehensive dataset of various atmospheric variables providing historical weather data on a global scale. The dataset encompasses a wide array of variables including temperature, precipitation, wind speed, and more, with data available from 1979 to the present day [23].

To access the ERA5 data we built a python script pulling the data from Google Earth Engine, specifically from the "ERA5-Land Hourly - ECMWF Climate Reanalysis" dataset, converting the data to Celsius and then to a weekly mean temperature, finally, we would be matching up the data to the location of the farm.

To access the ERA5 data we developed a Python script to extract data from the Google Earth Engine. We specifically used the "ERA5-Land Hourly - ECMWF Climate Reanalysis" dataset. The script would take in the locations of all of the Angus Soft Fruits Farms and pull the relevant data for each location. Following this the script would then convert the temperature data from Kelvin to Celsius and then calculate the weekly average temperatures, this was to match the resolution of our harvest data, and the growers forecasts.

Due to the lack of available weather data, we opted for the "ERA5-Land Hourly - ECMWF Climate Reanalysis" dataset as an alternative. To effectively utilize the ERA5 data, we developed a Python script that retrieved the information from Google Earth Engine, converted the data to Celsius and then to a weekly mean temperature, finally it would match up the data to the location of the farm. This would become another feature in our model.

While we've emphasized the significance of weather data in our model, Angus Soft Fruits has been highly responsive by initiating weather sensor trials across a variety of polytunnels to gather data. However, a persistent challenge remains: historical data. To effectively train our model, we must have consistent historical weather data, spanning the past four years (The time of our dataset). In our quest for a reliable data source encompassing consistent historical and current data, we explored other options such the MET office. Unfortunately, historical data availability from this source was severely limited in terms of locations. We also examined alternative systems like MODIS; however, we encountered considerable inconsistencies when comparing the data trends of this satellite/sensor data to on-site weather stations in Scotland.

3.3.1. Comparison to Edinburgh Airport

To test the viability of the data we extracted from ERA5 we would need to compare it to some an on the ground weather station. There are public weather stations at most airports with years of historical data. We compare the data from Edinburgh Airport to ERA5 data pulled at that location.

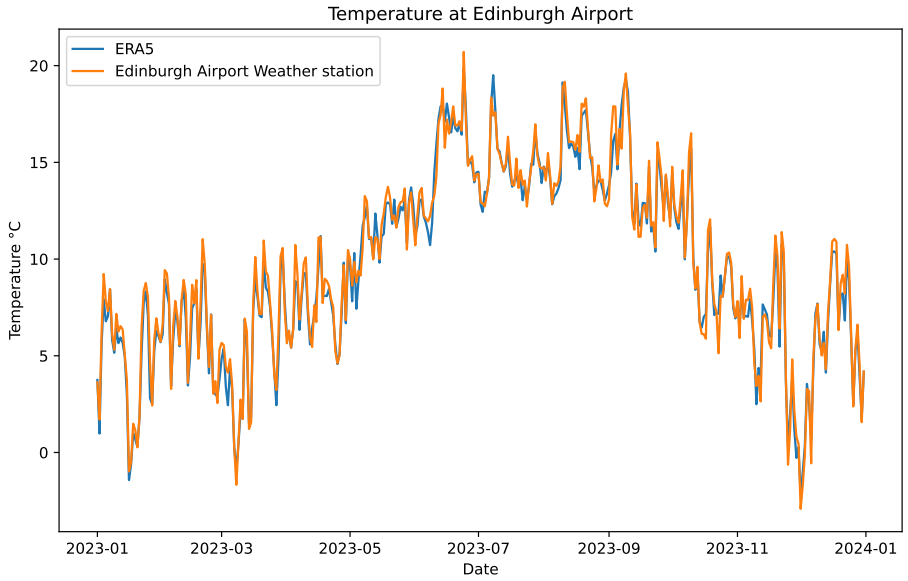


Figure 2: Temperature at Edinburgh Airport extracted from the ERA5 climate model and the local weather station.

As illustrated in Figure 2, the temperature trends observed from the ERA5 reanalysis dataset are remarkably consistent with the data recorded at the Edinburgh airport weather station. This correlation is noteworthy, considering the comprehensive and diverse sources from which ERA5 assimilates its data.

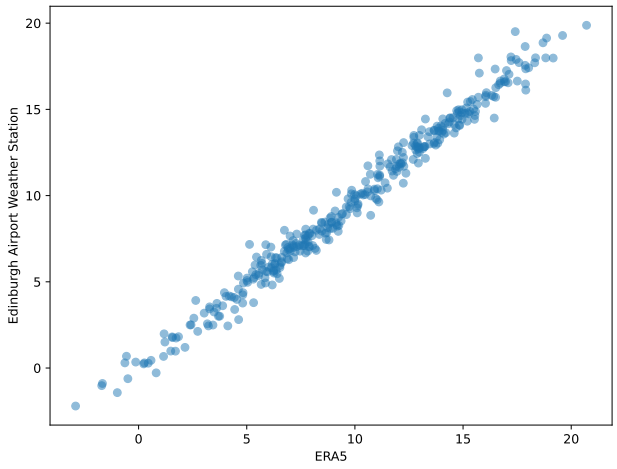


Figure 3: ERA5 data plotted against weather station data.

When performing a Pearson Correlation Coefficient, we got a score of 0.991. This indicates a very strong positive linear relationship between the two datasets.

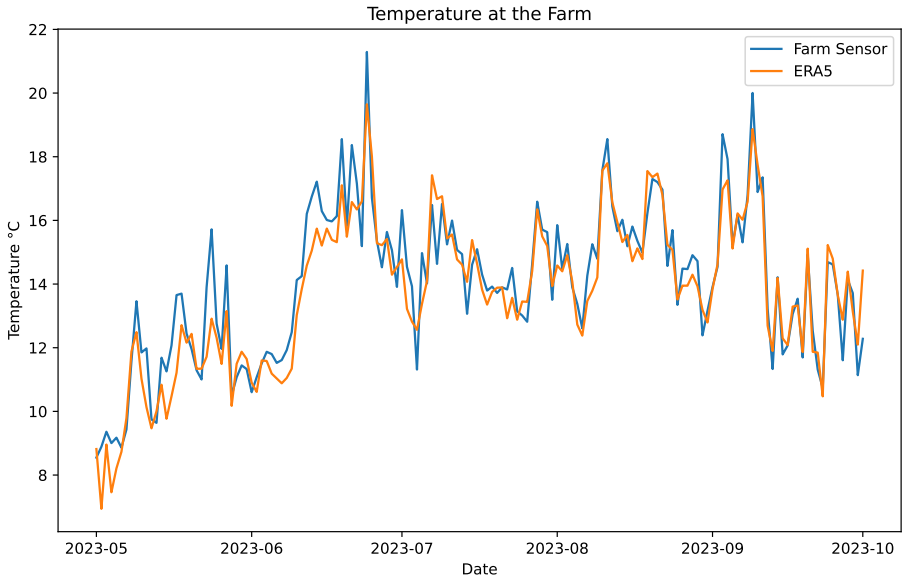


Figure 4: Temperature at an Angus Soft Fruits Farm.

3.3.2. Comparison to a farm weather station

The correlation between the ERA5 data and the Edinburgh Airport weather station is notable. However, considering that ERA5 assimilates data from a wide range of sources, including potentially public weather stations like that at Edinburgh Airport, it's important to evaluate the reliability of ERA5 data against a more independent source. This concern led us to compare the ERA5 data with measurements from a private weather station located at an Angus Soft Fruits farm.

The Pearson Correlation Coefficient between the ERA5 dataset and the farm's weather data was 0.938. This strong correlation underscores the reliability of the ERA5 dataset in reflecting actual weather conditions, even when compared to independent sources that are most certainly not being fed into the model. These results further reinforce the applicability of the ERA5 dataset for use in agriculture, as accurate local data is essential.

3.4. Models

In our research, we conducted a thorough comparison of various iterations of our XGBoost model with the Growers' pre-season and mid-season forecasts. This comparison was motivated by the consistently strong performance of the XGBoost model in preliminary studies. Our focus was to evaluate how these iterations of the XGBoost model performed in contrast to the currently relied on Growers forecasts.

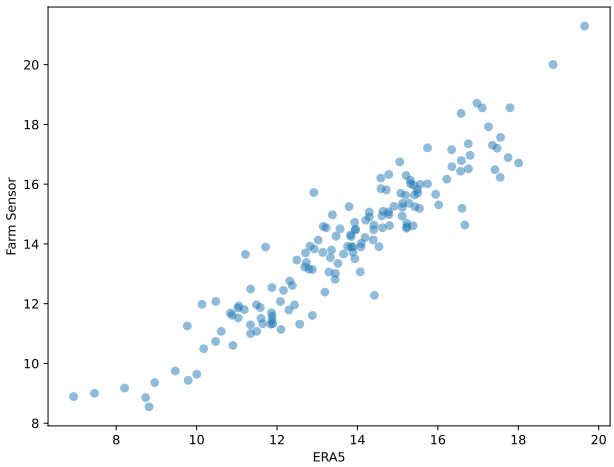


Figure 5: ERA5 data plotted against weather station data at an Angus Soft Fruits farm.

This analysis aimed to explore the potential of combining XGBoost the Growers manual predictions in enhancing predictive accuracy in the agricultural sector.

3.4.1. XGBoost

The XGBoost algorithm is a highly scalable end-to-end tree boosting system used in machine learning for classification and regression tasks [11]. The algorithm is renowned not only for its precision and flexibility, but also its automatic handling of missing values [24].

XGBoost stands out as one of the most widely adopted implementations of gradient-boosting decision trees (GBDT) due to both its robustness and effectiveness. Gradient trees are formed one by one, each addressing the errors of its predecessor. It employs gradient boosting to aggregate predictions from all trees, assigning greater weight to the more accurate ones, and ultimately combines these predictions for a final decision [25]. This is shown in Figure 6.

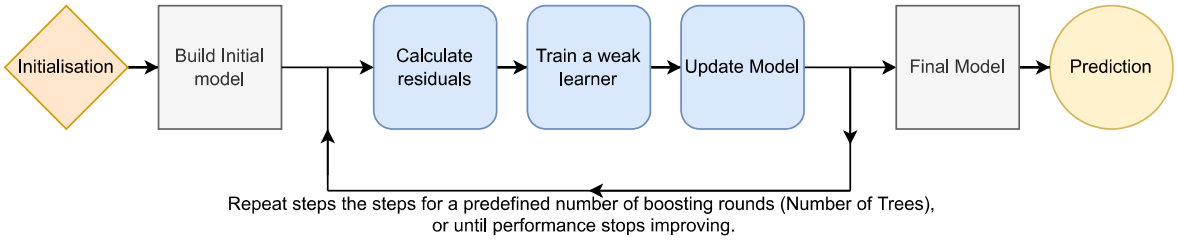


Figure 6: A simple visualisation of the XGBoost process.

XGBoost has often exhibited its capacity to surpass other models, including regular gradient-boosted

decision trees, ARIMA, Prophet, and LSTMs [26, 27].

3.4.2. Transformers

Transformers models, such as iTransformer and Autoformer, have achieved notable success in the field by adapting the transformer architecture for time series data. These models demonstrate the effective application of transformers in handling of time series data. Other research papers, especially those with access to more extensive data sets, have also reported success using transformer models, specifically in the field of strawberries for yield forecasting. For example the study we referenced earlier [20].

We tried to use transformers however the trends of the data were not followed, the data we have was too low resolution and messy. All of the predictions being made by the model mean values with day to day variance everyday. This was due to the model struggling to learn from the small dataset we had.

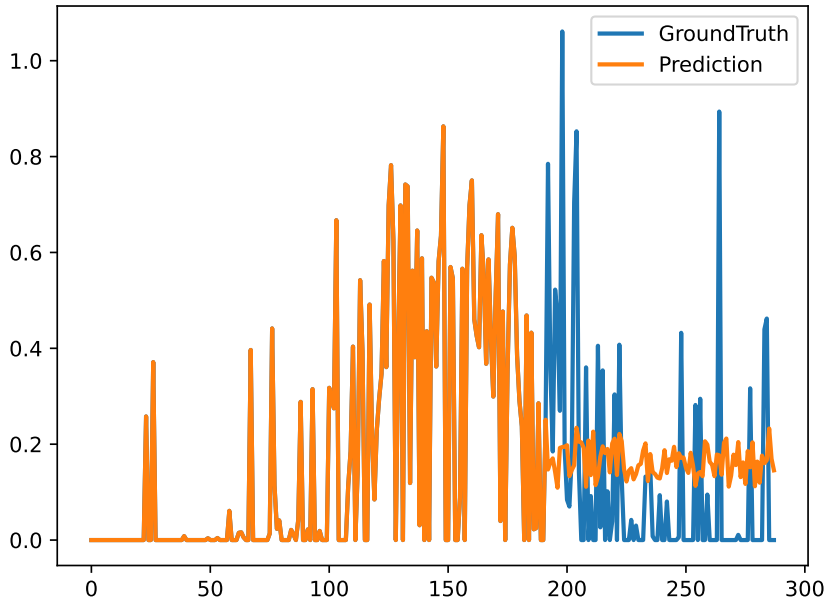


Figure 7: Prediction from the iTransformer model.

The both of the transformers model kept under-fitting and was not learning the patterns in the data. To address this, we made the model more complex. We did this by adding more layers, training for more epochs, increasing the learning rate and decreasing the batch size. This was all in an attempt to increase the amount the model was learning, we increased the complexity so much that we were deliberately pushing the boundaries towards potential over-fitting.

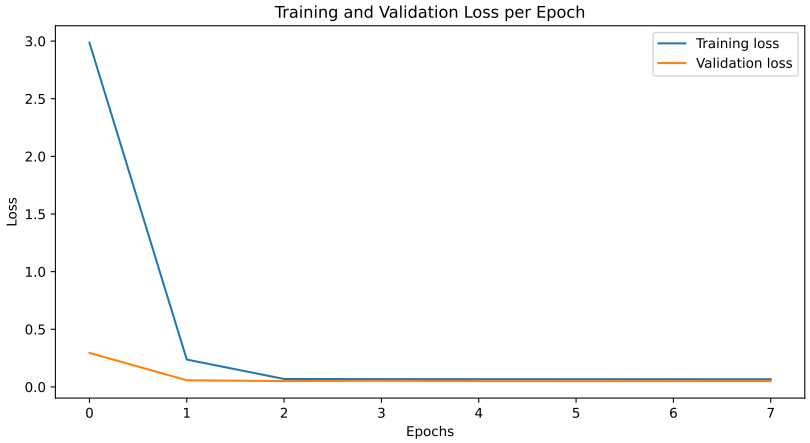


Figure 8: The loss from the iTransformer model plotted.

As we can see in figure 8, the model still did not over-fit. The predictions we see in figure 7 are useless to the growers, a similar harvest value every day over the course of months is not realistic and as such we decided not to work with transformers without further investigation about utilising these models for lower data time-series predictions.

3.4.3. Forecasting framework

Figure 9 illustrates the workflow of the application. Our system is a multi-farm global-to-local model, it is one model trained on the data from many farms across the UK which then makes predictions for each individual farm plot. Initially, the Dataframe Builder script is launched, which imports various CSV files containing historical yields from the growers' database, pre-season documents from previous years as well as the current year, and any mid-season forecasts available for the current year.

Additionally, the model incorporates data from the ERA5 climate model. This is processed by our Weather data tool script we created to pull the data from Google's Earth Engine. The Data-frame Builder will then process and window this data so that it can then be fed into XGBoost model to generate predictions.

4. Results

4.1. Model variations

To compare the different methods we created 4 predictions for every farm plot. We had the base model utilising current crop information and historical yields, the expert-informed model which added

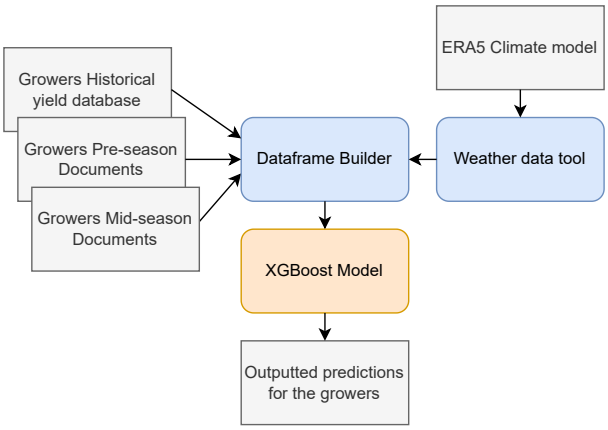
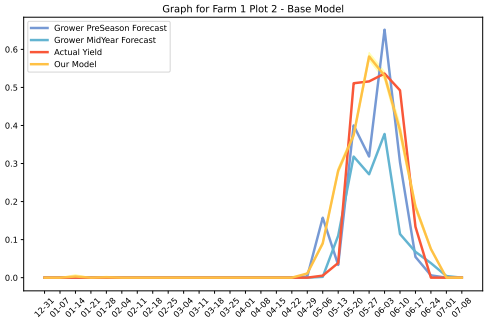
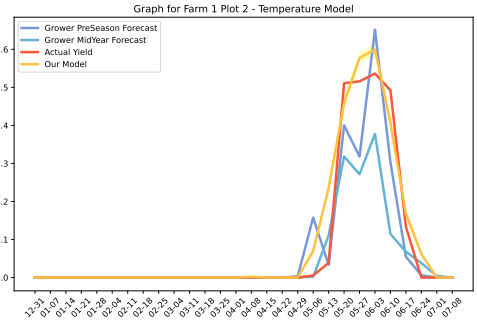


Figure 9: End-to-end forecasting framework.

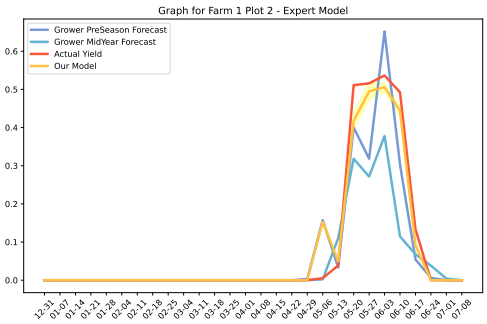
the Growers forecasts, the Sensor model which added temperature data from the ERA5 system and the expert-informed + Sensor system which added data from both the Growers forecasts as well as temperature data from the ERA5 system.



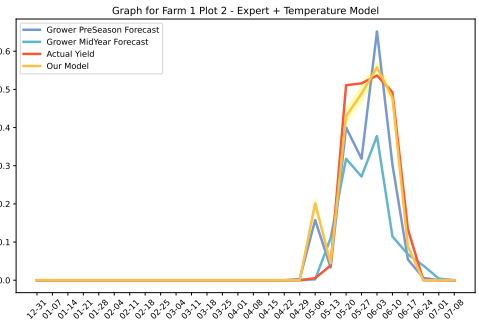
(a) Base Model



(b) Temperature Model



(c) Expert-informed Model



(d) Expert-informed Model + Temperature

Figure 10: Farm 1 Plot 2.

Farm 1 Plot 2 results for the 2023 predictions can be seen above in figure 10. Graphs for all the other farms and plots can be found attached in the appendix.

Table 1: Combined model variations - MAE and RMSE.

Farm	Plot	Base		Expert		Climate		Expert + Climate	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1	1	0.1239 ± 0.0026	0.0542 ± 0.0013	0.1097 ± 0.0024	0.0486 ± 0.0014	0.1340 ± 0.0018	0.0562 ± 0.0010	**0.1068** ± 0.0026	**0.0452** ± 0.0016
	2	0.0624 ± 0.0029	0.0281 ± 0.0015	**0.0360** ± 0.0018	0.0127 ± 0.0009	0.0462 ± 0.0016	0.0211 ± 0.0010	0.0387 ± 0.0019	**0.0126** ± 0.0011
	3	0.1670 ± 0.0012	0.0628 ± 0.0010	0.1601 ± 0.0018	**0.0537** ± 0.0012	0.1694 ± 0.0029	0.0653 ± 0.0014	**0.1565** ± 0.0028	0.0587 ± 0.0008
	4	0.0850 ± 0.0014	0.0366 ± 0.0007	0.0800 ± 0.0029	0.0331 ± 0.0016	**0.0724** ± 0.0018	**0.0310** ± 0.0009	0.0824 ± 0.0024	0.0351 ± 0.0010
	5	0.0601 ± 0.0024	0.0243 ± 0.0013	0.0536 ± 0.0019	0.0226 ± 0.0008	0.0635 ± 0.0043	0.0238 ± 0.0018	**0.0454** ± 0.0011	**0.0201** ± 0.0005
2	1	0.0436 ± 0.0023	0.0211 ± 0.0010	0.0478 ± 0.0045	0.0186 ± 0.0018	**0.0311** ± 0.0010	**0.0147** ± 0.0011	0.0365 ± 0.0024	0.0150 ± 0.0012
	2	0.0938 ± 0.0017	0.0412 ± 0.0008	0.0985 ± 0.0020	0.0448 ± 0.0010	**0.0774** ± 0.0018	**0.0370** ± 0.0010	0.0961 ± 0.0030	0.0442 ± 0.0015
3	1	0.1176 ± 0.0018	0.0314 ± 0.0009	**0.1154** ± 0.0019	0.0327 ± 0.0007	0.1184 ± 0.0017	**0.0313** ± 0.0008	0.1180 ± 0.0026	0.0326 ± 0.0006
	2	**0.0922** ± 0.0052	0.0452 ± 0.0027	0.1125 ± 0.0025	0.0472 ± 0.0010	0.0953 ± 0.0029	**0.0447** ± 0.0013	0.1267 ± 0.0027	0.0541 ± 0.0010
	3	0.0716 ± 0.0026	0.0348 ± 0.0009	**0.0650** ± 0.0017	**0.0238** ± 0.0006	0.0730 ± 0.0030	0.0322 ± 0.0015	0.0822 ± 0.0029	0.0318 ± 0.0013
4	1	0.0645 ± 0.0029	0.0259 ± 0.0015	0.0364 ± 0.0021	0.0149 ± 0.0009	0.0426 ± 0.0038	0.0175 ± 0.0018	**0.0337** ± 0.0022	**0.0126** ± 0.0013
5	1	0.1342 ± 0.0020	0.0620 ± 0.0010	0.1244 ± 0.0022	0.0516 ± 0.0018	0.1453 ± 0.0048	0.0649 ± 0.0021	**0.1233** ± 0.0010	**0.0497** ± 0.0009
6	1	0.1042 ± 0.0023	0.0455 ± 0.0010	**0.0718** ± 0.0017	**0.0302** ± 0.0008	0.0934 ± 0.0019	0.0350 ± 0.0011	0.0868 ± 0.0026	0.0323 ± 0.0010

4.2. Base model vs expert-informed model

The expert-informed model is generally more accurate and precise in its predictions than the Base model due to its consistently lower RMSE and MAE values across multiple farm plots. However, this comes with slightly more variability in performance as indicated by the higher standard deviations. The expert-informed model outperforms the Base model with lower average RMSE (0.0855 vs. 0.0939) and MAE (0.0334 vs. 0.0395), indicating greater accuracy and precision, albeit with slightly higher variability in performance as reflected by the standard deviations. A one-way ANOVA test with an Alpha of 0.05 was used, and confirmed that the improvement from the base model to the expert-informed model was significant, $P < 0.001$.

Table 2: Average RMSE and MAE values for the different models across all farms (ML: Machine Learning).

Method	**Average RMSE**	**Average MAE**
Base ML	0.0939	0.0395
Expert-informed ML	**0.0855**	**0.0334**
Climate ERA5 plus ML	0.0894	0.0365
Expert-informed ML plus Climate ERA5	0.0872	0.0342

4.3. Climate model data

The pattern that emerges from the analysis of the models indicates that while the expert-informed model generally provides superior performance compared to the Base model, As can be seen in Table2 the integration of Sensor data often leads to further improvements in average RMSE (0.0939 vs. 0.0894) and MAE (0.0395 vs. 0.0365). The expert-informed + Sensor model frequently achieves the best results, underscoring the value of combining expert analytical capabilities with sensor-derived data. The Sensor model alone also shows strong performance in specific instances, suggesting its utility in certain conditions. Overall, the data suggests a nuanced approach to model selection, where the choice of the model may depend on the specific characteristics and requirements of each farm plot. A one-way ANOVA test with an Alpha of 0.05 was used, and confirmed that the improvement from the base model to the model utilising climate data from ERA5 was significant, $P < 0.001$.

4.4. Expert-informed + climate model data

Combining the approaches presents us with a new model however as can be seen in Table 2, although superior to the base model average RMSE (0.0872 vs. 0.0939) and MAE (0.0342 vs. 0.0395), and the climate model data average RMSE (0.0872 vs. 0.0894) and MAE (0.0342 vs. 0.0365) on average the model is still beaten by the expert-informed model average RMSE (0.0872 vs. 0.0855) and MAE (0.0342 vs. 0.0334). Though these results are very similar, the improvement in the results just using the expert-informed method is consistent enough to be statistically significant when a one-way ANOVA was performed with an alpha of 0.05, leading to a p-value of 0.001.

4.5. Comparisons with grower forecasts

Table 3: Prediction comparisons.

Farm	Plot	Grower Pre-season		Grower Mid-season		Expert + Climate	
		RMSE	MAE	RMSE	MAE	RMSE	MAE
1	1	**0.0989**	**0.0355**	0.1082	0.0401	0.1068 ± 0.0026	0.0452 ± 0.0016
	2	0.0681	0.0307	0.0992	0.0413	**0.0387** ± 0.0019	**0.0126** ± 0.0011
	3	0.1756	0.0626	0.2133	0.0819	**0.1565** ± 0.0028	**0.0587** ± 0.0008
	4	**0.0646**	**0.0241**	0.1071	0.0329	0.0824 ± 0.0024	0.0351 ± 0.0010
	5	0.0765	0.0314	0.2925	0.1345	**0.0454** ± 0.0011	**0.0201** ± 0.0005
2	1	0.0426	0.0184	0.0809	0.0306	**0.0365** ± 0.0024	**0.0150** ± 0.0012
	2	0.1102	0.0471	0.1207	0.0459	**0.0961** ± 0.0030	**0.0442** ± 0.0015
3	1	**0.1029**	0.0327	0.1226	0.0337	0.1180 ± 0.0026	**0.0326** ± 0.0006
	2	0.1574	0.0704	**0.1169**	0.0545	0.1267 ± 0.0027	**0.0541** ± 0.0010
	3	0.1153	0.0456	0.1320	0.0440	**0.0822** ± 0.0029	**0.0318** ± 0.0013
4	1	0.0364	0.0146	0.0631	0.0277	**0.0337** ± 0.0022	**0.0126** ± 0.0013
5	1	0.1451	0.0674	0.1410	0.0665	**0.1233** ± 0.0010	**0.0497** ± 0.0009
6	1	0.1169	0.0549	0.1055	0.0404	**0.0868** ± 0.0026	**0.0323** ± 0.0010

In evaluating the effectiveness of our model we decided to compare it against the performance of the growers own forecasts. For our comparison, we used our expert-informed + Climate model data model. We chose this model as although the Average MAE and RMSE scores were slightly higher with this model compared to our expert-informed Model, this model had the most complete dataset, and plot-by-plot this model was the best performing with individual RMSE and MAE scores. We compared against both their pre-season forecast and one of their mid-year forecasts from May. Upon calculating the average values for each method, it was found that our model (expert-informed + sensor) demonstrated the highest accuracy with the lowest RMSE and MAE values. Specifically, this method showed an average RMSE of 0.0872 and an average MAE of 0.0342, outperforming both the Growers' pre-season (RMSE: 0.1008, MAE: 0.0412) and mid-season forecasts (RMSE: 0.1310, MAE:0.0519).

Table 4: Average RMSE and MAE values for each method (ML: Machine Learning).

Method	**Average RMSE**	**Average MAE**
Pre-season Grower	0.1008	0.0412
Mid-season Grower	0.1310	0.0519
Expert-informed ML plus Climate ERA5	**0.0872**	**0.0342**

5. Discussion

Our research underscores the value of integrating growers' forecasts into machine learning-based crop forecasting models. This approach effectively bridges traditional agricultural knowledge with advanced computational techniques, yielding better yield forecasts. This is particularly important when dealing with the inherent complexities of real-world agricultural data, this hybrid method(expert-informed model) demonstrates its strength.

We evaluated the performance of the expert-informed model, which incorporates the Growers pre-season forecast, in comparison to the model without any of this information. Our analysis revealed that the expert-informed model demonstrates superior accuracy in yield prediction. This is evidenced by its consistently lower Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values when compared across multiple farm plots. Specifically, the expert-informed model achieved an average RMSE of 0.0855, in contrast to the Base model's 0.0939, and an average MAE of 0.0334 compared to the Base model's 0.0395. These results indicate a notable improvement in the models' accuracy. The inclusion of growers' insights, derived from years of experience and deep understanding of their lands, complements the data-driven aspects of our models, which is important when handling the messy, small real-world data where other more advanced techniques such as Transformers proved not to be effective for timeseries forecasting. This hybrid strategy, therefore, represents a promising direction for future research in agricultural forecasting, utilising this practical, ground-level perspective provided by the growers.

Though we have access to the data of entire farms across the company, it would seem that data is still a limiting factor. Real-word data can often provide different challenges, we struggled to build a dataset with a strong independent variable. Going forward the company we worked with has already made an effort to more detailed thorough data after seeing the possibilities of ML while understanding the limitations of their current data collection.

Though these changes will have a drastic positive effect in the future and open the door the various other models and methods, in the meantime with the current historical training data we have we must find a way to make it work, this is where the importance of utilising data from both other expert data

sources and the ERA5 model come into play.

A significant aspect of our research focused on evaluating the potential of using historical ERA5 data as a feature in our predictive models, particularly as an alternative for instances where growers may not have recorded their own temperature data. This investigation stemmed from the need to provide a robust solution for growers who might lack local temperature recordings, a challenge we encountered in agricultural data collection.

Our findings indicate that incorporating temperature values from ERA5 data does indeed increase the accuracy of the predictive models, even with the crops being cultivated within polytunnels. This is particularly noteworthy in the realm of agriculture, where precise temperature data is often crucial for accurate yield forecasts. The improved model performance with ERA5 data integration demonstrates that even in the absence of locally recorded data, growers worldwide can still leverage machine learning techniques utilising weather data to make informed forecasts and decisions.

Looking forward, our research will evolve to try and adapt to the microclimatic conditions of the poly-tunnels that growers in Scotland utilise. Specifically, we plan to utilise sensor data gathered from within the poly-tunnels and analyze how these trends correlate with the ERA5 temperature data. The goal is to develop a model capable of using ERA5 data to infer corresponding internal poly-tunnel temperatures.

By achieving this, we anticipate a further improvement in our models' predictions. This advancement could be a game-changer for growers, enabling them to utilize predictive modelling effectively, even in scenarios where they lack extensive historical weather data collection. This research not only broadens the applicability of our model but also aligns with the broader objective of making machine learning a universally accessible tool in agriculture.

6. Conclusions

In this paper, we proposed an expert-informed global-to-local model designed for strawberry yield forecasting. The model incorporates real-world expert-generated data and achieves more precise predictions – in most cases – than the experts' forecasts as well as a machine learning model solely based on historical yield records, temperature readings from the ERA5 climate model, and various categorical variables. In addition, we observed that in scenarios where expert data is unavailable, integrating temperature data from the ERA5 climate model significantly enhances the accuracy of the forecasts, suggesting the need to consider this in future forecasting systems. Finally, we believe our system can form the basis for future developments in this area that will leverage already available historical data from farms for

developing accurate forecasting models that can support the growers' decision-making process.

CRediT authorship contribution statement

Matthew Beddows: Conceptualisation, Methodology, Software, Validation, Formal Analysis, Investigation, Data curation, Writing – original draft preparation, Visualization. Georgios Leontidis: Conceptualisation, Methodology, Investigation, Resources, Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was funded by the Data Lab, Angus Soft Fruits and a School of Natural and Computing Sciences PhD studentship.

References

[1] D. C. Rose, F. Shortland, J. Hall, P. Hurley, R. Little, C. Nye, M. Lobley, The impact of covid-19 on farmers' mental health: a case study of the uk, Journal of agromedicine 28 (3) (2023) 346–364.

[2] D. C. Rose, M. Bhattacharya, Adoption of autonomous robots in the soft fruit sector: Grower perspectives in the uk, Smart Agricultural Technology 3 (2023) 100118.

[3] P. Filippi, E. J. Jones, N. S. Wimalathunge, P. D. Somarathna, L. E. Pozza, S. U. Ugbaje, T. G. Jephcott, S. E. Paterson, B. M. Whelan, T. F. Bishop, An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning, Precision Agriculture 20 (2019) 1015–1029.

[4] B. Basso, L. Liu, Seasonal crop yield forecast: Methods, applications, and accuracies, advances in agronomy 154 (2019) 201–255.

[5] N. Kantanantha, N. Serban, P. Griffin, Yield and price forecasting for stochastic crop decision planning, Journal of agricultural, biological, and environmental statistics 15 (2010) 362–380.

[6] S. J. MacKenzie, C. K. Chandler, A method to predict weekly strawberry fruit yields from extended season production systems, Agronomy journal 101 (2) (2009) 278–287.

[7] M. Chaudhary, M. S. Gastli, L. Nassar, F. Karray, Deep learning approaches for forecasting strawberry yields and prices using satellite images and station-based soil parameters, arXiv preprint arXiv:2102.09024 (2021).

[8] O. Barrero, S. Ouazaa, C. I. Jaramillo-Barrios, M. Quevedo, N. Chaali, S. Jaramillo, I. Beltran, O. Montenegro, Rice yield prediction using on-farm data sets and machine learning, in: International conference on smart Information & communication Technologies, Springer, 2019, pp. 422–430.

[9] F. Jafari, K. Ponnambalam, J. Mousavi, F. Karray, Yield forecast of california strawberry: Time-series models vs. ml tools, in: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2020, pp. 3594–3598.

[10] S. Nonhebel, The importance of weather data in crop growth simulation models and assessment of climatic change effects, Wageningen University and Research, 1993.

[11] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[12] M. Lezoche, J. E. Hernandez, M. d. M. E. A. Díaz, H. Panetto, J. Kacprzyk, Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture, Computers in industry 117 (2020) 103187.

[13] A. Li, M. Markovic, P. Edwards, G. Leontidis, Model pruning enables localized and efficient federated learning for yield forecasting and data sharing, Expert Systems with Applications 242 (2024) 122847.

[14] P. Sheoran, P. Kamboj, A. Kumar, A. Kumar, R. K. Singh, A. Barman, K. Prajapat, S. Mandal, D. J. Yousuf, B. Narjary, et al., Matching n supply for yield maximization in salt–affected wheat agri–food systems: On-farm participatory assessment and validation, Science of The Total Environment 875 (2023) 162573.

[15] M. Thota, S. Kollias, M. Swainson, G. Leontidis, Multi-source domain adaptation for quality control in retail food packaging, Computers in Industry 123 (2020) 103293.

[16] A. Clarke, D. Yates, C. Blanchard, M. Islam, R. Ford, S. Rehman, R. Walsh, The effect of dataset construction and data pre-processing on the extreme gradient boosting algorithm applied to head rice yield prediction in australia, Computers and Electronics in Agriculture 219 (2024) 108716.

[17] A. Durrant, M. Markovic, D. Matthews, D. May, J. Enright, G. Leontidis, The role of cross-silo federated learning in facilitating data sharing in the agri-food sector, Computers and Electronics in Agriculture 193 (2022) 106648.

[18] S. Lee, C. M. Yun, A deep learning model for predicting risks of crop pests and diseases from sequential environmental data, Plant Methods 19 (1) (2023) 145.

[19] B. Alhnaity, S. Pearson, G. Leontidis, S. Kollias, Using deep learning to predict plant growth and yield in greenhouse environments, in: International Symposium on Advanced Technologies and Management for Innovative Greenhouses: GreenSys2019 1296, 2019, pp. 425–432.

[20] G. Onoufriou, M. Hanheide, G. Leontidis, Premonition net, a multi-timeline transformer network architecture towards strawberry tabletop yield forecasting, Computers and Electronics in Agriculture 208 (2023) 107784.

[21] M. A. Lee, A. Monteiro, A. Barclay, J. Marcar, M. Miteva-Neagu, J. Parker, A framework for predicting soft-fruit yields and phenology using embedded, networked microsensors, coupled weather models and machine-learning techniques, Computers and Electronics in Agriculture 168 (2020) 105103.

[22] C. Zheng, A. Abd-Elrahman, V. Whitaker, C. Dalid, Prediction of strawberry dry biomass from uav multispectral imagery using multiple machine learning methods, Remote Sensing 14 (18) (2022) 4511.

[23] R. Urraca, T. Huld, A. Gracia-Amillo, F. J. Martinez-de Pison, F. Kaspar, A. Sanz-Garcia, Evaluation of global horizontal irradiance estimates from era5 and cosmo-rea6 reanalyses using ground and satellite-based data, Solar Energy 164 (2018) 339–354.

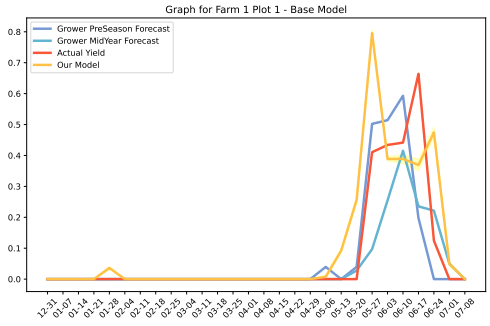
[24] L. Zhang, W. Bian, W. Qu, L. Tuo, Y. Wang, Time series forecast of sales volume based on xgboost, in: Journal of Physics: Conference Series, Vol. 1873, IOP Publishing, 2021, p. 012067.

[25] R. Potts, R. Hackney, G. Leontidis, Tabular machine learning methods for predicting gas turbine emissions, Machine Learning and Knowledge Extraction 5 (3) (2023) 1055–1075.

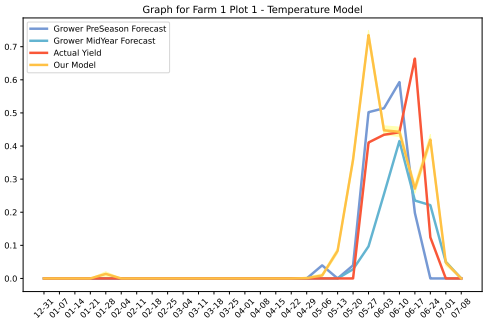
[26] J. Luo, Z. Zhang, Y. Fu, F. Rao, Time series prediction of covid-19 transmission in america using lstm and xgboost algorithms, Results in Physics 27 (2021) 104462.

[27] M. Abdurohman, A. G. Putrada, Forecasting model for lighting electricity load with a limited dataset using xgboost, Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control (2023).

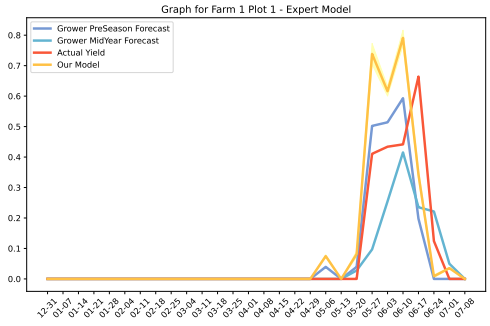
Appendix A.



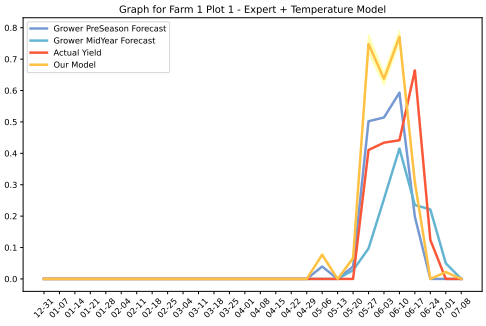
(a) Base Model



(b) Temperature Model

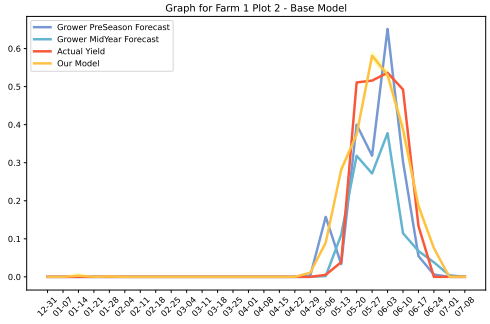


(c) Expert-informed Model

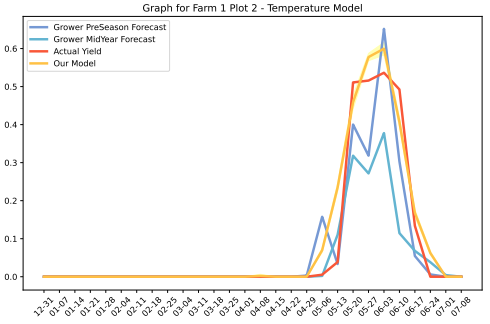


(d) Expert-informed Model + Temperature

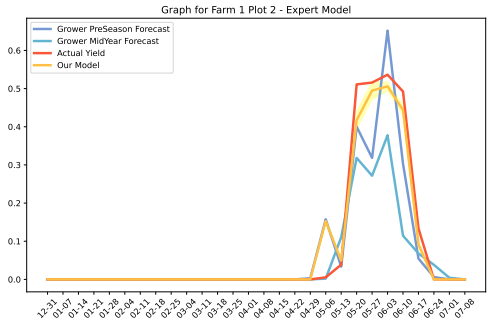
Figure A.11: Farm 1 Plot 1



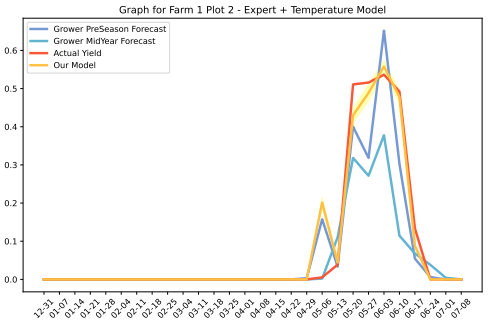
(a) Base Model



(b) Temperature Model

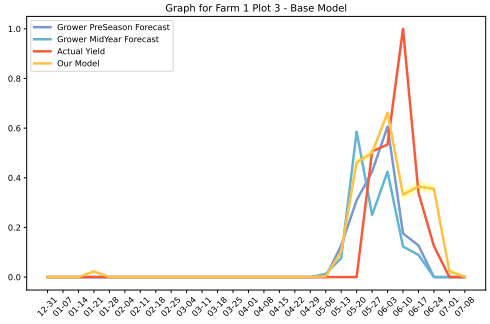


(c) Expert-informed Model

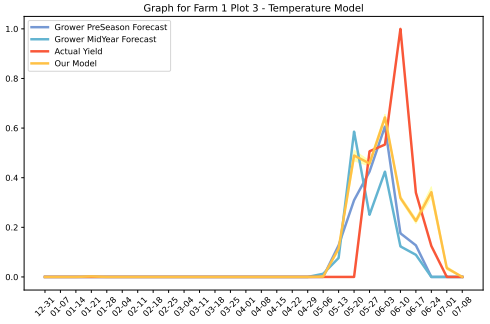


(d) Expert-informed Model + Temperature

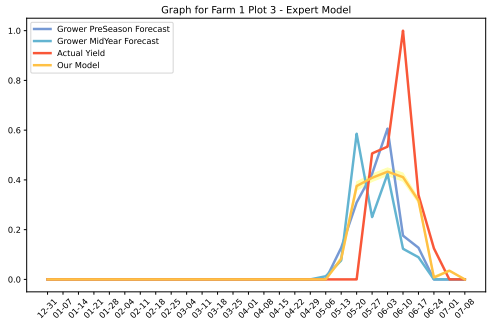
Figure A.12: Farm 1 Plot 2



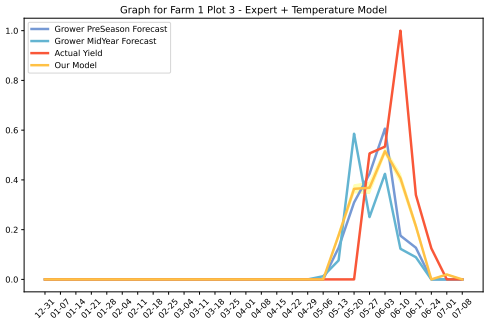
(a) Base Model



(b) Temperature Model

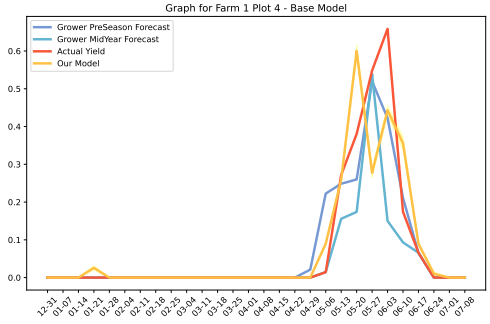


(c) Expert-informed Model

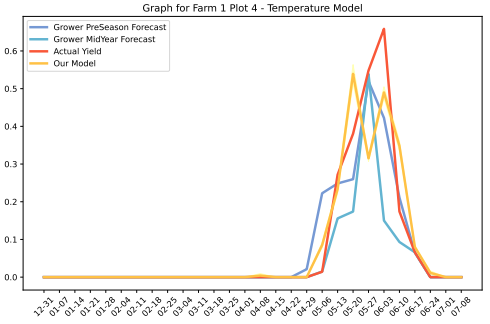


(d) Expert-informed Model + Temperature

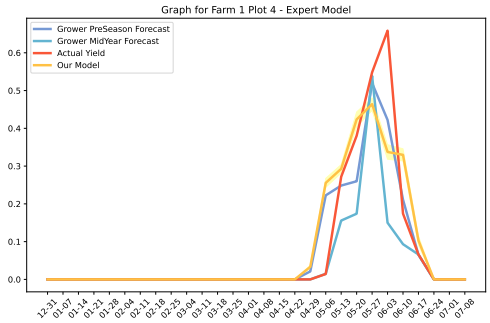
Figure A.13: Farm 1 Plot 3



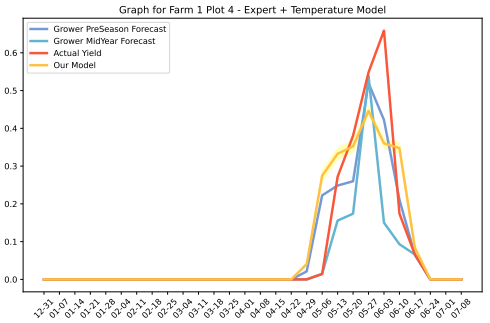
(a) Base Model



(b) Temperature Model

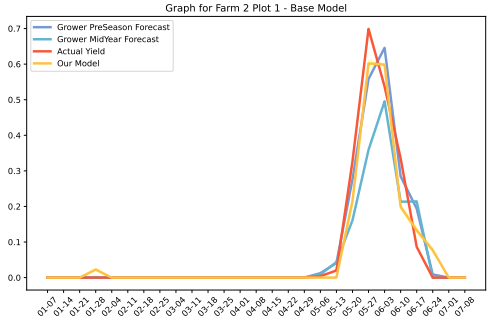


(c) Expert-informed Model

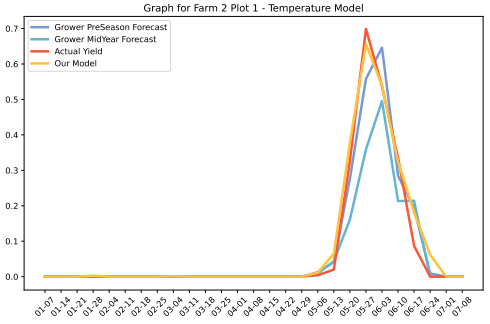


(d) Expert-informed Model + Temperature

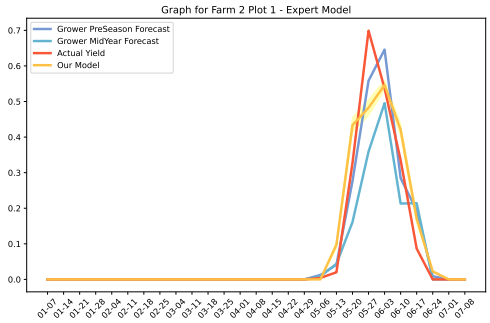
Figure A.14: Farm 1 Plot 4



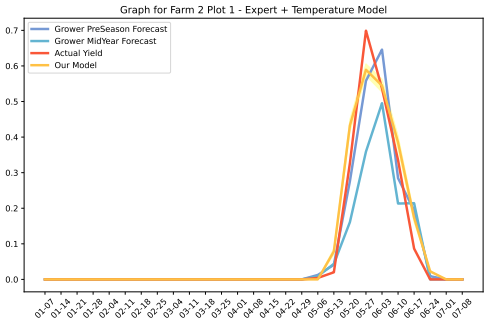
(a) Base Model



(b) Temperature Model

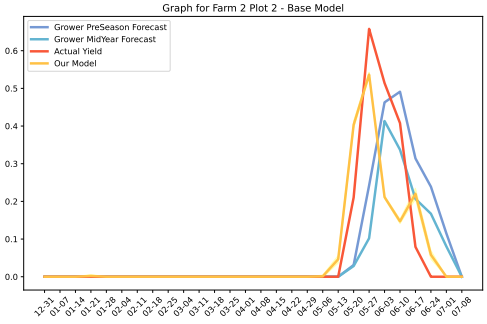


(c) Expert-informed Model

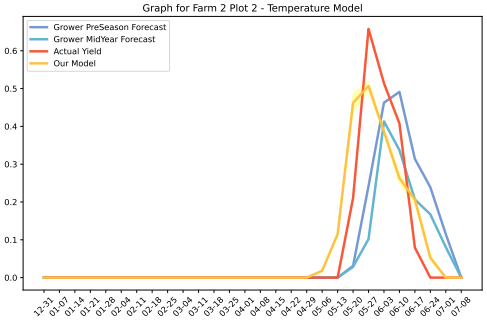


(d) Expert-informed Model + Temperature

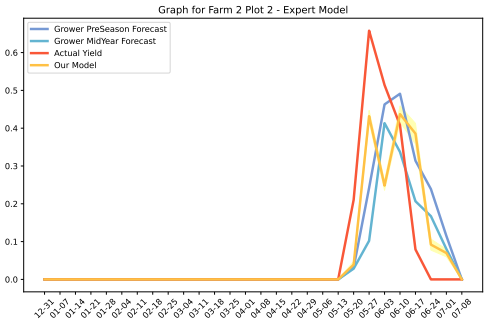
Figure A.15: Farm 2 Plot 1



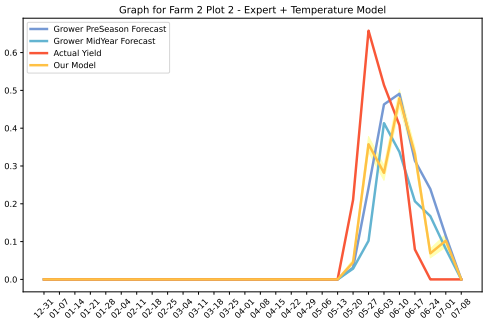
(a) Base Model



(b) Temperature Model

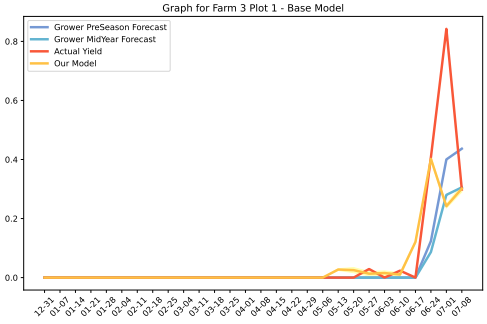


(c) Expert-informed Model

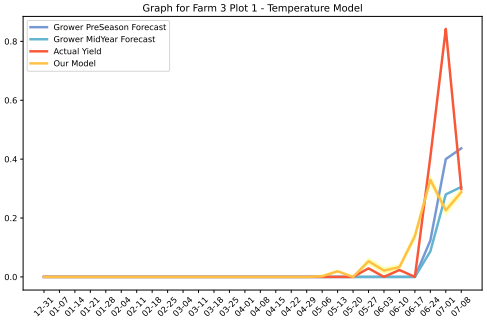


(d) Expert-informed Model + Temperature

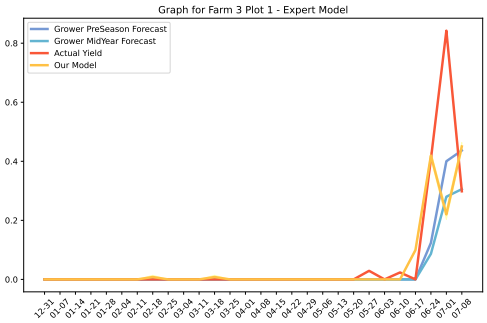
Figure A.16: Farm 2 Plot 2



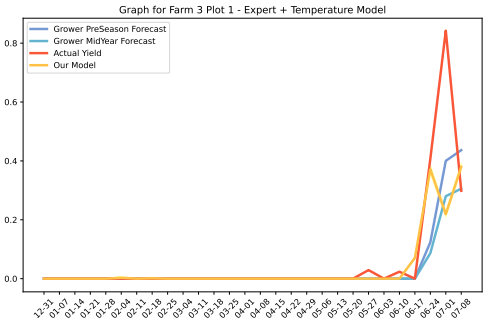
(a) Base Model



(b) Temperature Model

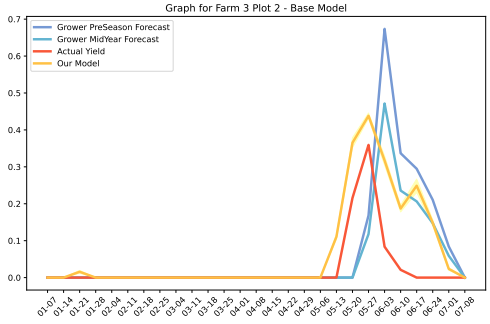


(c) Expert-informed Model

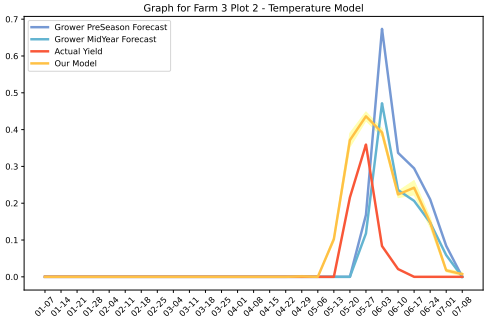


(d) Expert-informed Model + Temperature

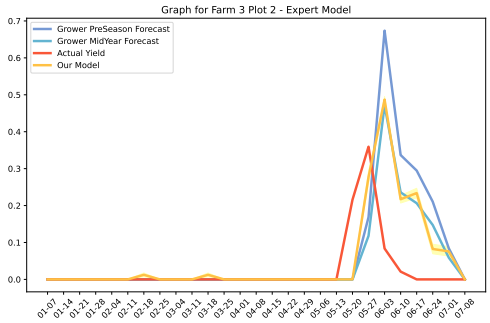
Figure A.17: Farm 3 Plot 1



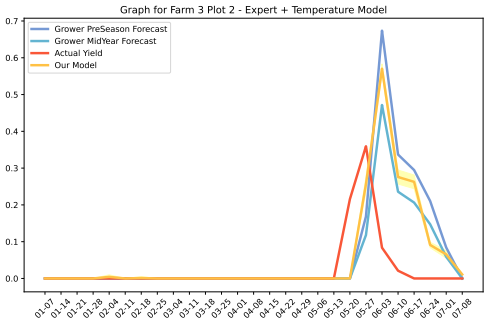
(a) Base Model



(b) Temperature Model

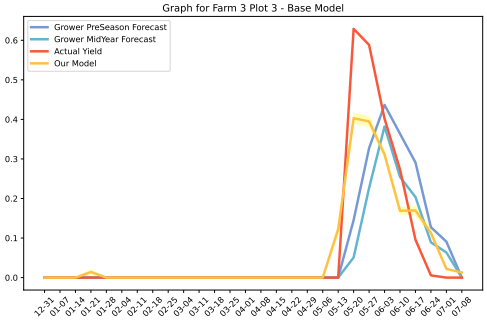


(c) Expert-informed Model

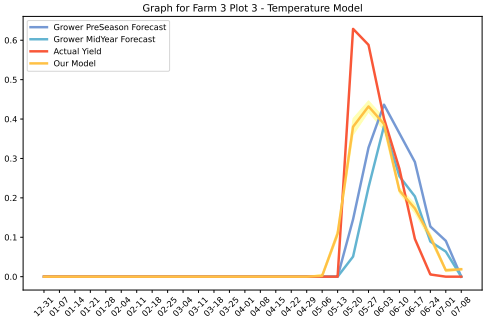


(d) Expert-informed Model + Temperature

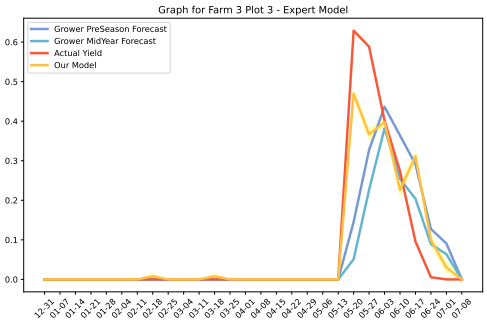
Figure A.18: Farm 3 Plot 2



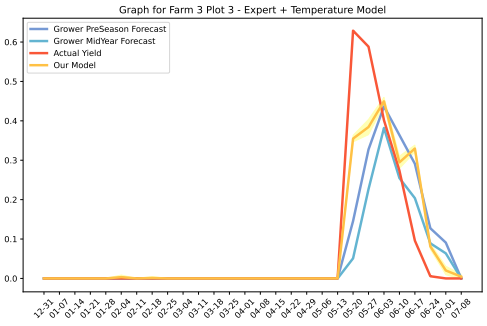
(a) Base Model



(b) Temperature Model

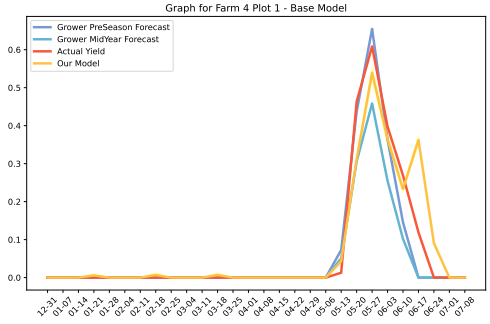


(c) Expert-informed Model

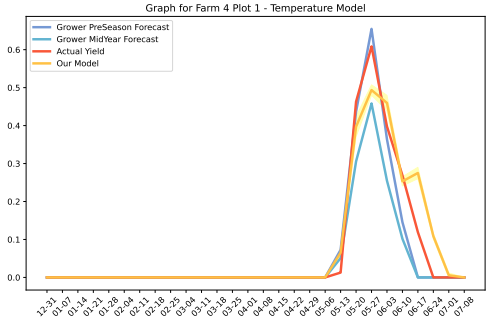


(d) Expert-informed Model + Temperature

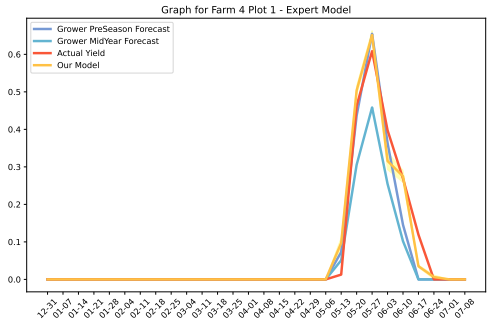
Figure A.19: Farm 3 Plot 3



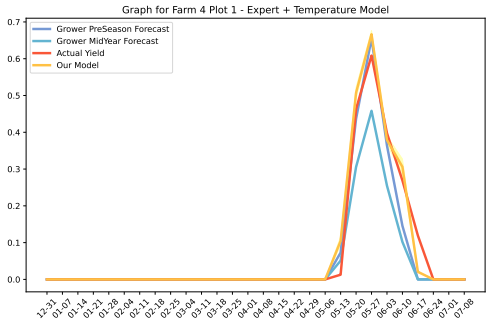
(a) Base Model



(b) Temperature Model

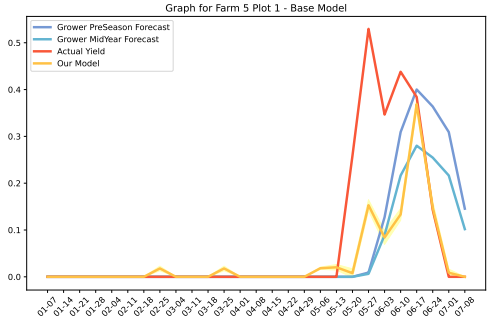


(c) Expert-informed Model

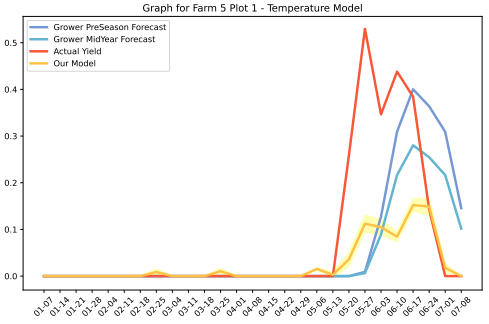


(d) Expert-informed Model + Temperature

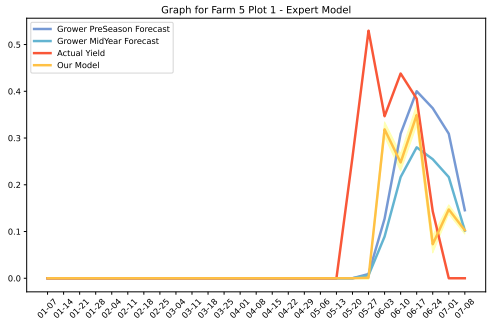
Figure A.20: Farm 4 Plot 1



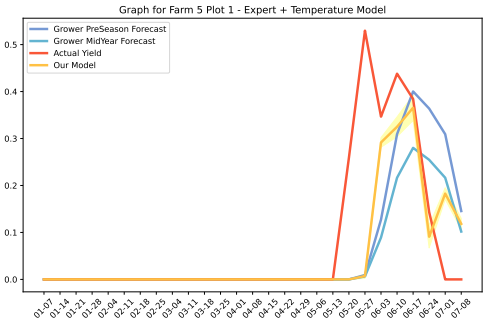
(a) Base Model



(b) Temperature Model

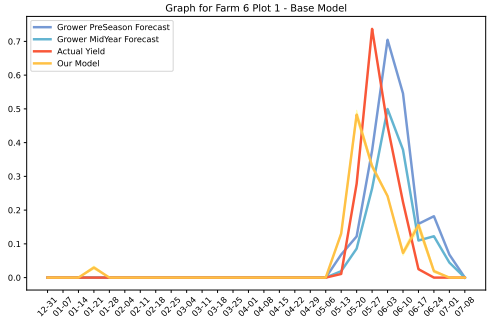


(c) Expert-informed Model

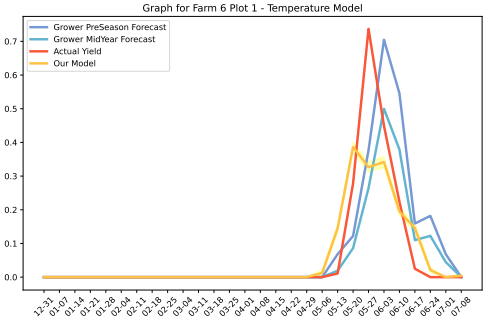


(d) Expert-informed Model + Temperature

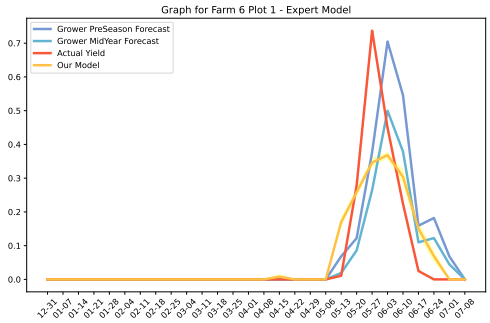
Figure A.21: Farm 5 Plot 1



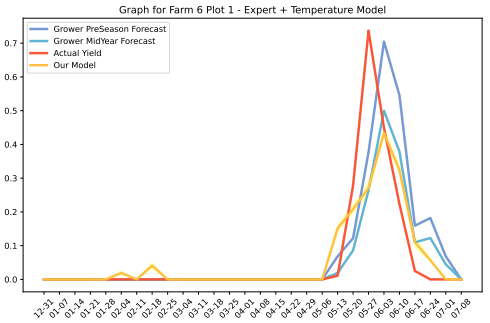
(a) Base Model



(b) Temperature Model



(c) Expert-informed Model



(d) Expert-informed Model + Temperature

Figure A.22: Farm 6 Plot 1