Crop Harvest Forecast via Agronomy-Informed Process Modelling and Predictive Monitoring

Jing Yang[1,2(✉)], Chun Ouyang[1,2], Güvenç Dik[1,2], Paul Corry[1,2], and Arthur H. M. ter Hofstede[1]

[1] Queensland University of Technology, Brisbane, QLD 4000, Australia
{roy.j.yang,c.ouyang,g.dik,p.corry,a.terhofstede}@qut.edu.au
[2] Food Agility Cooperative Research Centre, Sydney, NSW 2000, Australia

Abstract. Reliable and timely forecasts on crop harvest bring significant benefits to agri-food industries by providing valuable input to complex decisions on production planning. Useful predictions on crop harvest require continual effort by seasoned field agronomists. However, they are often scarce resources in the real-world. A feasible way to facilitate crop harvest forecast is through developing predictive models that can exploit data relevant to crop growth and automatically generate consistent predictions. To this end, this paper presents our design of a systematic and data-driven approach to supporting online forecasts on crop harvest. Underpinned by process modelling and predictive monitoring techniques, our approach can utilise crop-growth-related information from multiple data sources and progressively generate crop harvest predictions within the crop growing season. The approach has a flexible design informed by agronomic knowledge applicable to crop growth in general, and may be tailored to different crops and production scenarios. A case study with a local farming company using its real-life production data demonstrates the feasibility and efficacy of our approach.

Keywords: Crop forecast · Predictive process monitoring · Event logs · Process modelling · Knowledge discovery

1 Introduction

The resilience of agriculture and food production is one of the key challenges in today's world. Growers of fresh produce often have to make fast and complex decisions about production planning based on outcomes of crop harvest. Reliable and timely forecasts on crop harvest can bring significant benefits to agri-food industries, as they provide valuable input to support growers' decision-making, e.g., schedule for appropriate harvest dates to ensure the quality of yield [10]. However, forecasting crop harvest is complicated by the fact that (i) crop growth is affected by many factors, such as climate variability, geographical location, soil

X. Franch et al. (Eds.): CAiSE 2022, LNCS 13295, pp. 201–217, 2022.
https://doi.org/10.1007/978-3-031-07472-1_12

and water quality, and (ii) the impact of relevant factors on crop production is often dynamic, i.e., it changes along the crop growing process.

Effective forecasts on crop harvest concern various aspects and factors related to crop growth, and demand efforts by experienced agronomists working in the fields [2]. However, they are often scarce resources in the real-world. Recent advancement in digitalisation and Internet of Things (IoT) technologies enables the recording of various data relevant to crop production by information systems deployed in the agri-food sector [10] (e.g., farm management systems). It becomes an interesting yet challenging problem how to exploit such relevant and multi-sourced data to provide accurate and timely crop forecasts within the crop growing season automatically.

To this end, we propose a systematic and data-driven approach to supporting online forecasts on crop harvest. By modelling the process nature of crop growth, our approach can integrate static and dynamic crop-growth information from various data sources into one consistent data input and utilise it for reliable crop harvest predictions. Built upon process modelling and predictive monitoring [16] capabilities, the approach enables predictions to be generated automatically and progressively during the crop growing season. Furthermore, the approach has a flexible design informed by agronomic knowledge applicable to crop growth in general, and may be tailored to different crops and production scenarios. A case study with a local farming company using historical real-life crop production data demonstrates the feasibility and efficacy of the approach.

Our research contributes an effective approach to forecasting crop harvest based on process science and machine learning. The contribution is three-fold. Firstly, it presents a novel attempt to adapt a predictive process monitoring framework in the domain of agri-food production. Secondly, the proposal of using process modelling to integrate multi-sourced crop-growth data into one standardised, consistent data input in the form of event logs, is a potential contribution to the field of Information Systems Engineering. Last but not least, for the farming company engaged in the case study, our work lays the technical foundation for the company's capability building in data-driven crop forecast and contributes to improved production planning and resource deployment.

2 Background and Related Work

Forecasting crop harvest within crop growing seasons plays a vital role in crop production planning and decision-making. There are three types of approaches to crop forecasting [2]. Field survey is a traditional yet expensive way. Farm managers and farmers collect and assess information such as the number of pods and pod weight close to the harvest period and give estimation of the final yield [12]. Field surveys usually require trained operators to carry out data collection across multiple locations. A typical example concerns the Objective Yield surveys conducted by the US Department of Agriculture. Field survey data can be used to make forecasts, which often depend on agronomists' opinions about growing-season conditions (like weather events) and their expectations on

the final yield. As a result, field surveys may risk uncertainty and inconsistency due to their reliance on agronomists' expertise [2].

Crop simulation modelling is an effective way to deriving crop forecasts. Simulation models consist of mathematical equations to characterise plant development and growth processes, considering factors of genotypes, environment, management, and their interactions [13]. Historical data, such as actual observed weather and averaged weather data, as well as climate model outputs, can be used to establish parameters of a simulation model. Predictions on the end-of-season harvest outcomes are then generated by running the instantiated model. Since simulation models do not predict based on real-time observed data, they are prone to risks of unknown climate situations between forecast dates and harvest dates [2]. To overcome this issue, some research (e.g., [5]) studies how to calibrate simulation models at runtime within crop growing seasons.

Another promising way to approach crop harvest forecasting is through the use of machine learning techniques, which has received growing interest in recent years. Machine learning models, e.g., linear regression or neural network, are trained to fit crop data from historical seasons, and can then be applied to new crop data and make predictions for coming seasons. Crop data may cover various types of information, including weather (e.g., temperature and solar radiation), soil (e.g., soil type and nutrients), water (e.g., rainfall and humidity), and the crop itself (e.g., crop variety and plant weight) [15]. Recent advancement in sensor technologies has enabled possibilities to use data collected by dedicated proximal sensors (e.g., Internet of Things devices deployed on fields) and remote sensing platforms (e.g., satellites and Unmanned Aerial Vehicles or UAVs) [17] as inputs to machine learning techniques. These data are usually images, from which various vegetation indices and biophysical parameters can be derived and utilised for prediction and change analysis [8]. For example, remotely-sensed images acquired by UAVs are used to calculate the Normalized Difference Vegetation Index, contributing an important feature alongside weather variables for predicting pasture biomass development [4]. Notably, a recent systematic literature review on application of machine learning to crop yield prediction [15] highlights the need and challenge for future work to utilise data from different data sources to improve predictions.

To automatically generate reliable forecasts on crop harvest within crop growing seasons, it is vital to integrate multi-sourced data relevant to crop growth—data reflecting aspects of the plant nature, growing environment, and production management—and to be able to utilise such data and synchronise with dynamic changes during growing seasons. Event logs provide a flexible view that aggregates multi-dimensional, time-series data relevant to the same process and potentially from multiple data sources [1]. Crop growing seasons adhere to the plant's phenological nature [2] and can be captured using a process notation. This makes event logs a suitable choice for integrating multi-sourced data relevant to crop growth.

Predictive process monitoring techniques can be used to exploit event log data and make predictions on running processes with regard to performance,

outcomes, risks, or future states [11]. These techniques use machine learning algorithms with a process focus [16] and strengthen organisations' capabilities of making timely decisions about processes at runtime.

In this paper, we propose an approach to integrating multi-sourced, crop growth-related data into the form of event logs and using predictive process monitoring techniques for online predictions of crop harvest. In the context of research on process analytics [1] and predictive process monitoring [11], our approach represents an application of these techniques to address key problems in the agri-food domain.

3 Approach

In this section, we present our approach to supporting online predictions of crop harvest. Figure 1 depicts an overview of the approach, which takes as input multiple data sources related to crop growing and generates predictions related to crop growth and harvest (e.g., *days to harvest* and *yield*). The design of the approach is informed by the relevant agronomic knowledge and is built upon a benchmark predictive process monitoring workflow [16]. As such, it is capable of addressing the need of making timely crop predictions automatically and progressively during the crop growing season. There are two key components—data fusion guided by a crop growing process model, and an agronomy-informed process-aware (AIPA) predictive model for crop forecast.

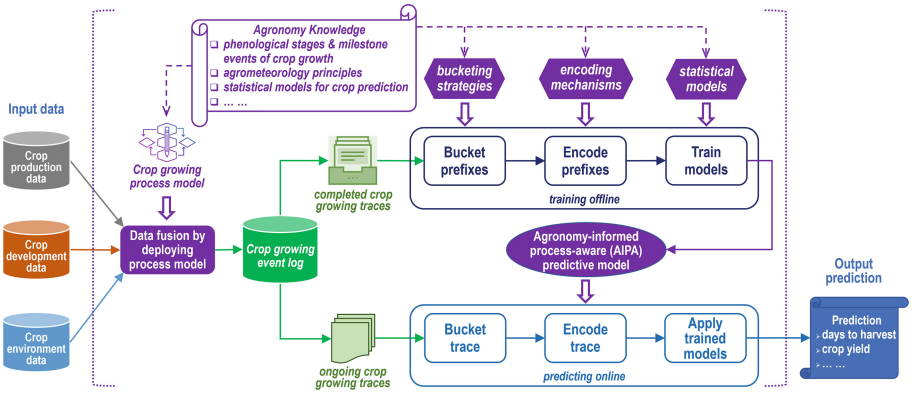


Fig. 1. An overview of the approach supporting online forecasts on crop harvest

3.1 Data Fusion

We consider three main data sources which provide useful information related to crop growing. *Crop production data* records information of crop production relevant to farm and supply chain management, such as production order, crop

plant variety, growing location, sowing date, and harvest date and yield (only available in historical data). *Crop development data* contains information about crop growth, including important events like sowing and flowering; and physical characteristics of growing plants like size of the plant, tiller number, etc. *Crop environment data* records information concerning the crop growing environment like soil and water quality and, in particular, *weather information* of the crop growing locations such as temperature, radiation, and rainfall.

We use rice production as an example to explain how data fusion is carried out, as illustrated in Fig. 2. While the three data sources are inherently different from each other, the purpose of data fusion is to integrate them into one crop growing dataset as the input for prediction.

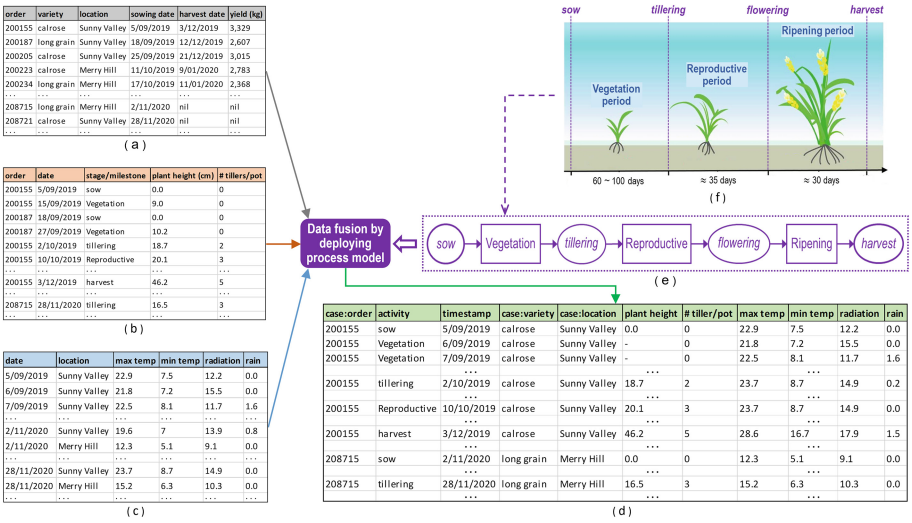


Fig. 2. Illustration of data fusion using a simple hypothetical example, where (a) rice production data, (b) rice plant development data and (c) weather data of rice growing locations are integrated into (d) rice growing event log data, by referring to (e) a rice growing process model (in a simplified view) of which the design is informed by (f) three main phenological growth stages and milestone events of rice crop (adapted from [7])

The key to enabling data fusion in our approach is the design of a crop growing process model which captures the phenological process of a given crop. Consider the example of rice crop in Fig. 2(f). The growth of a rice crop consists of three main phenological stages—Vegetation, Reproductive and Ripening, and two intermediate milestone events—*tillering* which indicates the transition from Vegetation to Reproductive growth, and *flowering* which marks the transition from Reproductive to Ripening growth [7]. Such agronomic knowledge can be used to inform the design of a rice growing process model. Figure 2(e) depicts a simplified view of such a model capturing rice growing activities in terms of

three main phenological growth stages and two milestone events between *sow* (the start event of rice growth) and *harvest* (the end event of rice growth).

Once a crop growing process model is established, it can be deployed to replay or monitor the process of crop growth by incorporating the input data from available data sources. As a result, a log of events recording the occurrences of crop growing activities can be generated, where data from the various sources is systematically and consistently integrated into relevant data attributes associated with individual events. For example, Fig. 2(d) shows a fragment of a rice growing event log generated by deploying the rice growing process model in (e) with the input data of rice production in (a), rice plant development in (b) and relevant weather information in (c).

Concepts and Notations. We define several key concepts and notations for describing crop growing event log data.

A *crop growing event* (denoted e) is an instance of a crop growing activity (including a start, end or intermediate milestone event), and has *attributes* carrying crop production, development, and environment data associated with the activity. Each crop growing event e has three mandatory attributes—production order (denoted o), crop growing activity (a) and timestamp (t), and is uniquely identified by a combination of the values carried by these attributes. Hence, e can be represented as a tuple (o, a, t). For example, in Fig. 2(d) each row corresponds to a rice growing event and each column to an attribute, and the rice growing event on the first row can be written as (200155, *sow*, 5/09/2019).

A *case-level attribute* is an attribute of which values remain identical across all crop growing events that belong to the same production order. By definition, production order is always a case-level attribute of a crop growing event. Revisiting the example in Fig. 2(d), each rice growing event has three case-level attributes—crop production order, variety, and location.

A *crop growing trace* (denoted σ) is a non-empty finite sequence of crop growing events that belong to the same production order. For each crop growing trace σ, the order of crop growing events in σ is determined by event timestamps. Let σ_o represent the crop growing trace of production order o, σ_o can be written as $[(o, a_1, t_1),\ldots,(o, a_n, t_n)]$ (where $t_1<\ldots<t_n$). Revisiting the example in Fig. 2(d), the following two rice growing traces can be observed:

- σ_{200155}: [(200155, *sow*, 5/09/2019), (200155, *Vegetation*, 6/09/2019), (200155, *Vegetation*, 7/09/2019), . . . , (200155, *harvest*, 3/12/2019)]
- σ_{208715}: [(208715, *sow*, 2/11/2020), . . . , (208715, *tillering*, 28/11/2020)]

A *crop growing event log* is a set of crop growing traces. There are: *completed crop growing traces*, which begin with the start event (e.g., *sow*) of a crop growing process and finish with the end event (e.g., *harvest*) of the process; and *ongoing crop growing traces*, which begin with the start event but finish with an event other than the end event. In the above example, σ_{200155} is a completed rice growing trace, and σ_{208715} is an ongoing rice growing trace.

3.2 AIPA Predictive Model

The availability of crop growing event log data (as output of data fusion) makes it possible to develop a predictive model to forecast crop harvest by exploiting predictive process monitoring capabilities. We adopt a benchmark predictive process monitoring workflow [16]. The main idea is to train a predictive model using historical data of completed crop growing traces, and then use the trained model to make predictions for ongoing crop growing traces (see Fig. 1). In particular, we focus on applying agronomic knowledge and domain expertise and support model explainability in the design of our approach.

Bucket Prefixes. The first step is to extract *prefixes* from completed traces and group them into *buckets* (or bins) according to certain criteria. Given a completed trace σ, a prefix of σ is defined as a sequence of the first l ($1 \leq l \leq |\sigma|$) events of σ. Hence, a completed trace σ can be used to extract $|\sigma|$ prefixes. These prefixes capture the history of crop growth related to the trace progressively, and are the input for feature encoding and model training. A *bucketing strategy* is used to specify the criteria for grouping the extracted prefixes into buckets. A typical example known as prefix-length-based bucketing is to group prefixes of the same length into the same bucket. For crop prediction, we propose to apply the relevant agronomic knowledge to the design of a bucketing strategy. For example, the growth of a crop plant consists of different phenological stages. Each stage is associated with a specific set of factors affecting crop harvest outcome. According to this, prefixes of crop growing traces can be grouped into buckets depending on which phenological stage each prefix belongs to.

Encode Prefixes. In the second step, prefixes in each bucket are encoded as feature vectors using an *encoding mechanism*. Our focus is on how to encode *event-specific* attributes, which change from event to event and are considered *dynamic* attributes of an event log. Although there exist various feature encoding techniques in data mining research, they are not necessarily suitable for crop prediction. For example, weather data attributes are typical event-specific attributes of a crop growing event log. Since weather plays an important role in crop growth, it has been studied in the field of agrometeorology, where specific measures and algorithms for aggregation of weather attributes are established with an emphasis on their impact on crop production. Hence, we propose that relevant agronomic knowledge such as agrometeorology principles should be used to guide the design of encoding mechanisms for crop prediction.

Train Models. In this step, feature vectors in each bucket are used to train a predictive model. While there are various machine learning techniques that one can choose from, we propose two key rationales for making a design decision. Firstly, since complex machine learning models developed to build advanced predictive capabilities are often used as a 'black-box', the recent body of literature in machine learning has emphasised the importance to apply models that are

transparent and explainable [14]. Secondly, among the models that are explainable by design, statistical models are a good choice as they have already been used to make crop predictions (see Sect. 2). A typical example is the use of statistical regressions for predicting crop yield, where the applicability of such a model is often driven by its simplicity and transparency [2]. Hence, we consider the use of statistical models for crop prediction in our approach.

At the end of this step, a predictive model that is agronomy-informed and process-aware (i.e., an AIPA predictive model) is generated. It comprises a set of buckets of prefixes, and encoded feature vectors and trained models associated with each of the buckets. The performance of an AIPA predictive model can be assessed using appropriate evaluation measures for the given predictive target.

Online Prediction. During the online phase, a trained AIPA predictive model is used to make predictions for ongoing crop growing traces. Given an ongoing trace, the correct bucket for the trace is firstly determined, then the feature encoder for the bucket is used to encode the trace data into a feature vector, and finally the trained model for the bucket is deployed to obtain a prediction.

4 Case Study

This section reports on a case study using real-life data provided by a farming company X in Australia to predict the harvest of a crop Y.[1]

4.1 Context

Crop Y is a common type of crop grown by company X. To increase market value of crop Y, the company wishes to develop a scientific solution that provides timely predictions about the crop's harvest date and yield, using historical crop production and meteorological data. While its current predictions rely heavily on the manual labour of field agronomists, company X expects the solution to automatically produce predictions on a weekly basis during the crop growing season. We applied our approach to address this need.

For this case study, company X provided us with historical crop production data recording the actual production orders (*orders* for short) of Y over the past five years. Each order in the dataset can be uniquely identified and records information on the growth period of a certain amount of crop Y. More specifically, an order record has four types of information, including (i) location and area of the production unit, (ii) variety of crop Y, (iii) key dates during the crop growth period, e.g., harvest dates, (iv) quantities of the order, i.e., the ordered quantity (corresponded to sales orders) and the delivered quantity (i.e., yield).

For crop environment data, we collected and used the public weather data released by the state government. For each local area in the state, the weather dataset records the daily maximum and minimum temperature, radiation, rainfall, evaporation, and vapour pressure.

[1] For confidentiality reasons, we cannot disclose the company's name and the specific crop considered in this case study.

4.2 Application of the Approach

Data Fusion and Preprocessing. We discussed with agronomists from company X and built a process model capturing the phenological growth process of crop Y. As shown in Fig. 3, the growing season of crop Y is divided into two stages marked by two milestone events, namely "PF" and "HA".

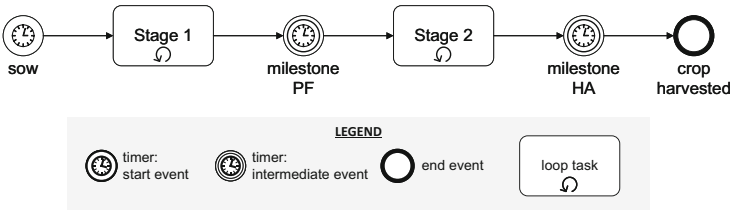


Fig. 3. Process model capturing the growth of crop Y using BPMN

We utilised the created crop growing process model to integrate the collected order data and weather data, and generated the crop growing event log for the given orders of crop Y. By matching the production unit locations in the order data against the geographical locations in the weather data, we were able to extract the daily weather observations during the crop growing season for every order. Then, the created process model was deployed to generate for each order a crop growing trace containing daily crop growing events. As a result, we obtained a crop growing event log with 348 completed traces consisting of 22,115 events. Table 1 shows a fragment of the event log.

Table 1. A fragment of the anonymised event log used in the case study

Order	Event/ subprocess	timestamp	Days to PF	Days to harvest	case: unit_loc	case: unit_area	case: var	case: order_qty	case: deliv_qty	radn	max_t	min_t	rain	evap	vp
12763	sow	2016-08-10	74	56	M1	4.2	VG	32320	38350	17.4	24	5.4	0	2.4	11.6
12763	Stage 1	2016-08-11	73	55	M1	4.2	VG	32320	38350	8.1	22.9	7.2	0	2.9	13
12763	Stage 1	2016-08-12	72	54	M1	4.2	VG	32320	38350	16	21.9	8.5	2.9	3	12.4
...															
12763	Stage 1	2016-10-05	0	18	M1	4.2	VG	32320	38350	25.8	26	7.1	0	6	9.7
12763	Stage 2	2016-10-06	n/a	17	M1	4.2	VG	32320	38350	25.5	28.3	6.6	0	5.4	7.2
12763	Stage 2	2016-10-07	n/a	16	M1	4.2	VG	32320	38350	26.3	29.8	5.8	0	5.2	10.5
...															
13687	sow	2019-10-14	60	45	M3	2	W	3900	5700	20.1	29.6	15.4	0.8	5.3	17.8
13687	Stage 1	2019-10-15	59	44	M3	2	W	3900	5700	22	34.3	16.3	0	7.5	20.4
...															

(i) Attributes derived from the order data: "days to PF", days from the recorded event to PF; "days to harvest", days from the recorded event to the harvest date; "case:unit_loc", location of the production unit; "case:unit_area", area of the production unit (hectare); "case:var", crop variety; "case:order_qty", ordered quantity (kg); "case:deliv_qty", delivered quantity (kg). (ii) Attributes derived from the weather data: "radn", solar radiation (MJ/m^2); "max_t", maximum temperature (°C); "min_t", minimum temperature (°C); "rain", rainfall (mm); "evap", evaporation (mm); "vp", vapour pressure (hPa). Attributes that start with "case" are case-level attributes.

In a real-life scenario, historical crop growing traces are used to train an AIPA predictive model offline, which is then used for making predictions about ongoing crop growing traces. In our evaluation, we simulated such a scenario by splitting the event log dataset into two subsets. We ordered all traces by their sowing dates. The first 75% was used for training the AIPA model (as "historical" traces), while the more recent 25% was used for testing the derived model (as "ongoing" traces).

Bucketing. We applied a bucketing strategy based on the crop's phenological growth stages. As mentioned, the growing season of crop Y is divided into two stages. Therefore, we used two buckets to group encoded prefixes based on whether milestone PF was reached.

Encoding. We consulted agronomists from company X to identify and model factors impacting the growth of crop Y. It was suggested that temperature and radiation are the most important weather attributes. Specifically, temperature can be characterised by two measures in agronomy, namely, growing degree days (GDD) [9] and heat stress days [6]. These measures can be derived[2] based on the daily maximum and minimum temperature in our dataset. For radiation, we applied numerical aggregation functions including *sum, mean, max, min* and *std* to encode features [16].

Furthermore, in terms of predicting harvest date, the agronomists advised that the length of stage 1 (from sowing to milestone PF) is a good indicator for estimating final harvest dates in practice. Therefore, for prefixes in the second bucket (at stage 2), we included the duration from sowing to milestone PF as an encoded feature. We set the field "days to harvest" as the prediction target.

For predicting yield, the agronomists suggested that the number of heat stress days occurred during the first five days after milestone PF may be a useful predictor besides the foregoing features. We included this for prefixes in the second bucket. We set the delivered quantity per unit area as the prediction target to eliminate the area difference across production units. This variable can be derived from dividing "case:delivered_qty" by "case:unit_area".

Statistical Model. In this case study, we employed ordinary linear regression models with least squares. The reasons are two-fold. First, applying a linear regression model requires minimal configuration, which helps avoid hyperparameter tuning often required by other more complex regression techniques. Second, the simplicity and transparency of a linear model helps us better communicate our prediction results with company X when explaining how predictions are obtained and identifying factors that have high impact on the predictions.

[2] Calculation of those meteorological measures was done based on an R-package `cropgrowdays` (https://gitlab.com/petebaker/cropgrowdays).

Model Evaluation. When selecting the evaluation metrics, we refer to the goals set by company X: (i) predicted harvest dates are expected to be within 2 days on either side of the actual harvest dates, and (ii) predicted yield (as delivered quantity per unit) values are expected to deviate no more than 35% (either side) from the actual delivered yield. Given these goals, we employed Mean Absolute Error (MAE) for evaluating harvest date prediction and Mean Absolute Percentage Error (MAPE) for yield prediction, respectively.

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

where n is the number of samples, y_i and $\hat{y}_i$ are the observed and predicted target values of the i-th sample [16]. Furthermore, we also included the adjusted R^2 score (i.e., coefficient of determination) considering the regression nature of the tasks.

We also considered earliness [16] in our evaluation, which refers to the earliest point of time when the prediction accuracy (in terms of a selected measure) satisfies a given goal. To enable earliness evaluation, we applied two filters on prefixes in both the training and testing data set: (i) we excluded the final event for all prefixes (i.e., event recording harvests), since predictions obtained on the day of harvest will be trivial, and (ii) we excluded prefixes with excessive length[3] to avoid unreliable results due to an unbalanced number of samples [16].

In our experiments, we conducted model evaluation in two ways. First, we performed an overall model evaluation by calculating prediction scores and errors for prefixes in the two buckets separately, i.e., stage 1 (pre-PF) and stage 2 (post-PF). In this way, we were able to obtain an overview on the efficacy of linear regression models using different features. Second, we performed a weekly-based evaluation to assess how our approach generates predictions within the crop growing season. We took all prefixes from both buckets and re-grouped them by prefix length based on weeks, e.g., prefixes with lengths ranged from 1 to 7 comprise the first group and those from 8 to 14 comprise the second group, etc. We then calculated the prediction errors for each weekly group and compared the results against the prediction goal.

4.3 Results Analysis and Discussion

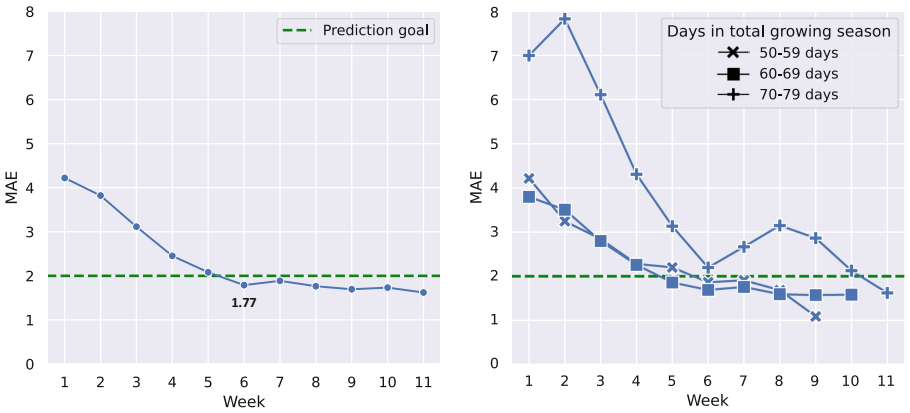
Overall Model Evaluation. Table 2 show the results of the overall model evaluation by prediction errors and scores. In predicting days to harvest, the average adjusted R^2 scores show that regression models built for both buckets have decent performance. Meanwhile, the MAE values are 2.77 and 1.85. The decrease in errors signifies that including the duration from sowing to milestone PF as a feature contributes to making better predictions, which aligns with the advice given by the agronomists of company X.

[3] Note that this does not reduce the size of the training or the testing set, in terms of the number of traces (orders).

Table 2. Results of overall model evaluation

Prediction task	Metric	Bucket 1: pre-PF	Bucket 2: post-PF
Days to harvest	Adj.R^2	0.93	0.85
	MAE	2.77	1.85
Yield	Adj.R^2	0.62	0.60
	MAPE	0.37	0.31

In predicting yield, we obtained adjusted R^2 scores of 0.62 and 0.60, respectively. The scores are lower compared to those in the previous task, but this is expected—in predicting yield, all prefixes of the same trace have an identical target value (i.e., the final unit quantity delivered), due to the lack of data tracking crop development in terms of its final yield. This limitation of data impeded the yield predictive model in capturing how weather conditions within the growing season may have impacted the final yield.



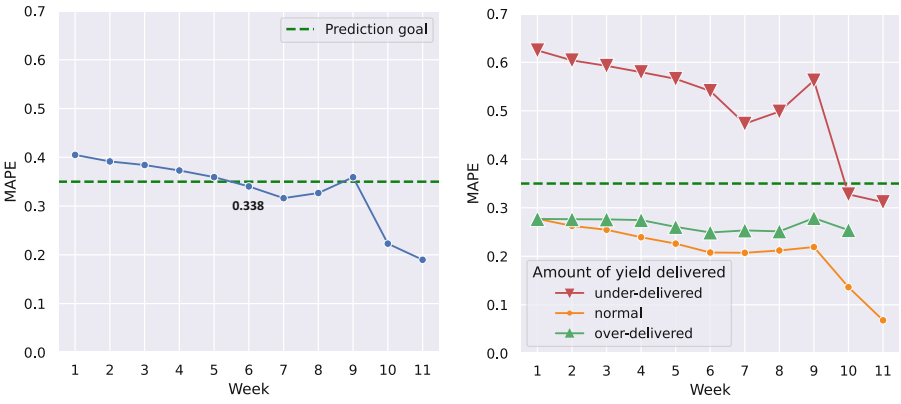
(a) all traces (b) per trace length cluster

Fig. 4. Weekly evaluation by Mean Absolute Error on predicting days to harvest

Weekly-Based Model Evaluation. In terms of predicting days to harvest, Fig. 4a shows the results of weekly evaluation by MAE over prefixes of all traces. It is clear that the prediction error decreases weekly as the crop growing season proceeds weekly. Also, the developed models show good performance in terms of earliness, since the prediction errors satisfy the goal set by company X as early as week 6 (when MAE $= 1.77 < 2$).

Furthermore, we unfolded the evaluation results to investigate how our approach performed on traces related to crop growing seasons of different lengths. We did this by classifying traces into three clusters based on the number of days from crop sowing to harvest: 50–59 days, 60–69 days, and 70–79 days,

respectively. Figure 4b shows the results. We can observe that the predictive models performed well on the 50-day and 60-day clusters, with a similar trend of progressively decreasing error as seen before. However, the longer traces (i.e., those in the 70-day cluster) seem to be the outliers. Even though the prediction error decreases in general, it has an unexpected increase after week 6, which is approximately the point that splits the two buckets by milestone PF. Also, the prediction error never reaches a satisfactory level until the last week (week 11), which would then have very little value in terms of earliness. A possible reason is that it is less common that the total growing season spans over 70 days. We confirmed this by examining the distribution of traces in our dataset. It was found that the longer traces comprise a small percentage of all traces, only around 15% in both the overall dataset and the subset used for model training. Therefore, the predictive models may likely have under-fitted data of the longer traces.



(a) all traces (b) per delivered quantity cluster

Fig. 5. Weekly evaluation by Mean Absolute Percentage Error on predicting yield

In terms of predicting yield, Fig. 5a show the results of weekly evaluation by MAPE over prefixes of all traces. At first, the prediction error decreases in general, albeit at a small scale, and meets the prediction goal in week 6 (MAPE = 33.8% < 35%). During week 8 to week 9, there is an unexpected increase, but the prediction error stays close to the accepted level. From week 10 onwards, the prediction error becomes much lower to around 20%. These observations provide further evidence to our previous conclusion that predicting yield with the current dataset is challenging due to the limitation of the data—yield predictions become much more accurate in terms of the actual final yield only when the crop growing season enters its later stages.

We further examined the yield prediction results. Again, traces were clustered but, in this case, by calculating the ratio of delivered quantity to ordered quantity (see data attributes in Table 1), which then represents the percentage of ordered quantity delivered. We considered three clusters: (i) "normal", if the

ratio is between 90% and 110%; (ii) "over-delivered", if the ratio is higher than 110%; and (iii) "under-delivered", if the ratio is lower than 90%. Figure 5b shows the unfolded results per cluster. Clearly, the predictive models performed well on the "normal" cluster, showing an expected pattern of both low and progressively decreasing error. Performance on the "over-delivered" cluster is acceptable and remains stable over time. The most interesting finding concerns the "under-delivered" cluster, where prediction error is consistently high in the early weeks and merely reaches the prediction goal from week 10 onwards. Consider that "under-delivered" traces accounted for 42% of the training set, the underperformance of the models is unlikely to be caused by insufficient data. We surmised that crop growth relevant to those "under-delivered" traces was impacted by unseen factors not captured by information in the current dataset. This speculation was later confirmed through our discussion with the agronomists reported below. The relatively high proportion of "under-delivered" traces also explains the increase of prediction error in week 7 as previously shown in Fig. 5a.

Summary. Experiment results show that our approach can provide reasonably accurate predictions about the harvest date and yield of crop Y during its growing season. Predictions generated from our approach become progressively more accurate, and can already provide insights usable by the company as early as week 6 from the date of sowing. We communicated our findings with company X and received positive feedback. Agronomists from the company expressed high interest in our solution in terms of how it integrates up-to-date weather observations and enables them to obtain prediction outcomes on a daily basis. Specifically, regarding our speculation of predicting yield, we learned from the discussion that the "delivered quantity" recorded in the dataset (which we specified as the target of yield prediction) does not fully represent the amount of fresh yield. Instead, the delivered quantity records the result from a quality control process which took place after crop harvest. Not surprisingly, data in the "under-delivered" clusters are related to more human intervention in quality control, compared to other clusters. Since the dataset we used does not cover any information about that post-processing step, it makes sense that the derived predictive models did not perform well on the "under-delivered" clusters.

5 Discussion

Our work reported in this paper paves new avenues to some interesting future work. For one, our approach has a flexible design which enables it to be generalised to potentially any field crop for which phenological events can be identified and captured by process modelling. Also, the use of a general predictive process monitoring workflow allows many machine learning techniques to be plugged in. For example: (i) when given data that embeds complex, non-linear patterns, advanced regression algorithms, e.g., Support Vector Machine (SVM), can be used to potentially improve the prediction accuracy [15]; and (ii) when encoding

prefixes, feature selection methods for sequence prediction [18] can be applied to complement the use of expert knowledge from agronomists.

With this flexibility, it is worthwhile to investigate guidelines on how to configure different steps of the approach according to factors like actual crops, ecosystems (soil, growing season, etc.), and input data characteristics [8].

The use of process modelling in our approach can transcend the purpose of integrating data from multiple, various sources to feed the predictive process monitoring workflow. Process models formally capturing the phenology of crop growth enable many existing process analytics tools to be applied "off-the-shelf". For example, historical crop growth can be visually analysed by replaying crop growing event logs [3], so that agronomists can leverage the collated data to examine how crop harvests were impacted by geographical locations, date-times, weather conditions, etc. We consider the use of process modelling an enabler of future deployment and application of process analytics over data integrated from various non-standard data sources—this can potentially be extended to domains other than forecasting crop harvest, making a contribution to the field of Information Systems Engineering.

Last but not least, an interesting direction concerns how to integrate this approach into the decision workflow of production planning for growers, specifically how it can synergise with the management of other business processes in crop production, e.g., delivery after crop harvest.

A limitation of our work is that the case study is subject to one crop grown in the local area, and the collected data contains only crop production data and crop environment data recording weather conditions. Therefore, an immediate next step is to extend the case study to other crops and include more relevant data. In particular, we are interested in exploring the use of crop development data, for example, as collected by agronomists through field visits or by dedicated sensors and remote sensing platforms [17]. Such ancillary data will provide precise information on crop growth monitoring and thus the opportunity to improve the efficacy of our approach. We also seek to improve the input data quality, e.g., to use production data that captures the fresh yield rather than the processed one.

6 Conclusion

Reliable and timely forecasts on crop harvest benefits decision-making in the agri-food industries and contribute to the resilience of agriculture and food production. In this paper, we present an approach that systematically utilises crop-growth data from multiple sources and generates online predictions on crop harvest automatically and progressively. Our approach offers a flexible solution for crop harvest predictions and can be tailored according to different crops (by redesigning the crop growing process model) and relevant agronomic knowledge (by altering the strategies for bucketing, encoding, and the statistical models). Meanwhile, our research findings contribute to the field of process science by making a novel and successful attempt to use process modelling and predictive monitoring techniques to address a key problem in the agri-food domain.

Acknowledgments. This work was supported by Food Agility CRC Ltd, funded under the Commonwealth Government CRC Program. We also received highly valuable input from A/Prof Miranda Mortlock, a specialist in agronomy and statistics, and Dr David Carey, a senior horticulturist from Queensland Department of Agriculture and Fisheries.

References

1. Van der Aalst, W.M.P.: Process Mining - Data Science in Action. Springer, second edn. (2016). https://doi.org/10.1007/978-3-662-49851-4, https://doi.org/10.1007/978-3-662-49851-4_1
2. Basso, B., Liu, L.: Chapter Four - Seasonal Crop Yield Forecast: Methods, Applications, and Accuracies. Advances in Agronomy. Academic Press, vol. 154, pp. 201–255 (2019)
3. De Leoni, M., Suriadi, S., ter Hofstede, A.H.M., van der Aalst, W.M.P.: Turning event logs into process movies: animating what has really happened. Softw. Syst. Model. **15**(3), 707–732 (2014). https://doi.org/10.1007/s10270-014-0432-2
4. De Rosa, D., et al.: Predicting pasture biomass using a statistical model and machine learning algorithm implemented with remotely sensed imagery. Comput. Electron. Agric. **180**, 105880 (2021)
5. Inoue, Y., Moran, M.S., Horie, T.: Analysis of spectral measurements in paddy field for predicting rice growth and yield based on a simple crop simulation model. Plant Prod. Sci. **1**(4), 269–279 (1998)
6. Kaushal, N., Bhandari, K., Siddique, K.H.M., Nayyar, H.: Food crops face rising temperatures: an overview of responses, adaptive mechanisms, and approaches to improve heat tolerance. Cogent Food Agric. **2**(1), 1134380 (2016)
7. Krishnan, P., Ramakrishnan, B., Reddy, K.R., Reddy, V.: Chapter Three - High-Temperature Effects on Rice Growth, Yield, and Grain Quality. Advances in Agronomy. Academic Press, vol. 111, pp. 87–206 (2011)
8. Lhermitte, S., Verbesselt, J., Verstraeten, W.W., Coppin, P.: A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. Remote Sens. Environ. **115**(12), 3129–3152 (2011)
9. McMaster, G.S., Wilhelm, W.: Growing degree-days: one equation, two interpretations. Agric. Forest Meteorol. **87**(4), 291–300 (1997)
10. Miranda, J., Ponce, P., Molina, A., Wright, P.K.: Sensing, smart and sustainable technologies for agri-food 4.0. Comput. Ind. **108**, 21–36 (2019)
11. Márquez-Chamorro, A.E., Resinas, M., Ruiz-Cortés, A.: Predictive monitoring of business processes: a survey. IEEE Trans. Serv. Comput. **11**(6), 962–977 (2018)
12. Nandram, B., Berg, E., Barboza, W.: A hierarchical Bayesian model for forecasting state-level corn yield. Environ. Ecol. Stat. **21**(3), 507–530 (2013). https://doi.org/10.1007/s10651-013-0266-z
13. Reynolds, M., et al.: Role of modelling in international crop research: overview and some case studies. Agronomy **8**(12), 291 (2018)
14. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)
15. Van Klompenburg, T., Kassahun, A., Catal, C.: Crop yield prediction using machine learning: a systematic literature review. Comput. Electron. Agric. **177**, 105709 (2020)

16. Verenich, I., Dumas, M., Rosa, M.L., Maggi, F.M., Teinemaa, I.: Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. ACM Trans. Intell. Syst. Technol. **10**(4), 34:1–34:34 (2019)
17. Weiss, M., Jacob, F., Duveiller, G.: Remote sensing for agricultural applications: a meta-review. Remote Sens. Environ. **236**, 111402 (2020)
18. Xing, Z., Pei, J., Keogh, E.: A brief survey on sequence classification. ACM SIGKDD Explor. Newslett. **12**(1), 40–48 (2010)