Final Report

October 2025

Student Project No. [SF/TF 170a]

Title: Using climatic and imaging data to predict apple phenology.

Haidee Tang

New Road, East Malling, Kent, ME19 6BJ

Supervisors:

Xiangming Xu, Xiaojun Zhai

Report No: [AHDB Use only]

This is the final report of a PhD project that ran from November 2021 to September 2025. The work was funded by AHDB BBSRC (BB/W510762/1).

CONTENTS

1. Industry Summary

This project focuses on using phenology models and deep learning models on hyperspectral images to predict the optimal harvest window. Conventional maturity assessments require a lot of time and manpower, yet they are destructive and can only estimate the harvest window within seven days of harvest at the orchard-level. A need for a quicker, non-destructive method that can achieve maturity estimates per fruit will enable apples to be picked at their optimum maturity. Picking fruit at their optimum maturity can improve shelf-life and the final fruit quality at market. The aims of the study were to:

- Evaluate the accuracy of the PhenoFlex phenology model and determine the best method of parameterisation when dealing with numerous cultivars.
- Determine the effect of flowering variation on fruit maturity, as the major shortcoming of phenology models is that they predict only the average flowering date. This neglects the effect of flowering variation on fruit maturity.
- Explore the use of hyperspectral imaging with deep learning models to predict apple maturity features. This results in real-time, non-destructive maturity assessments for each fruit.

The main contributions of our studies showed that phenology models can be applied at the species level, generalising the model across all apple cultivars. This simplifies the use of phenology models for growers or any person(s) wanting to determine the flowering date of apples. Flowering time can add up to 20% of variation to fruit maturity, with this effect being more pronounced in early flowering cultivars. It should be considered when estimating the harvest window. Knowing this allows growers to refine harvest date predictions. For the hyperspectral images, a large, diverse dataset collected across cultivars, seasons and countries was collected. This dataset is the largest collection of apple hyperspectral images and will be published as an available dataset for research use. This will allow other researchers to use the collected data to further similar apple research, or for general use to practice building deep learning models. Using this dataset, deep learning models predicting the levels of Brix and firmness showed a reasonable level of accuracy in indoor conditions. Moreover, key wavelengths for determining Brix and firmness were identified, allowing for simpler and cheaper camera systems to be built for apple maturity assessments, making imaging more accessible to growers. As a whole, we proposed a way to integrate phenology and imaging into a framework for harvest window predictions.

2. Introduction

Harvest maturity has been known to influence post-storage fruit quality. However, it is difficult to determine the optimal harvest maturity with conventional maturity assessments. Conventional apple maturity assessments rely on destructive sampling methods that are labour-intensive, time-consuming and often subjective, particularly for starch measurements. These methods typically provide forecasts within seven days before harvest and offer orchard-level insights that often fail to capture intra-orchard variability. As a result, growers face challenges in efficiently allocating resources and may harvest fruit that are either not completely developed or overripe.

In contrast, phenology models and hyperspectral imaging offer promising non-destructive alternatives to maturity predictions. Phenology models can predict flowering time and harvest windows based on temperature, while imaging can achieve objective, real-time assessments of apple traits (firmness and sugar content) without the need for destructive sampling.

The purpose of this study was to identify a non-destructive method of maturity prediction in apples more than seven days before harvest. By leveraging the well-defined phenological cycle — where flowering time can be predicted from dormancy, and the harvest date can be predicted from flowering time, primarily through temperature-driven models — phenology models provide early-season forecasts. Imaging techniques, applied closer to harvest, offer more precise assessments of fruit maturity. These approaches have the potential to improve orchard planning, enhance post-storage fruit quality and reduce food loss.

The key gap in using phenology models is the limited understanding of how accurately flowering time can be predicted across different apple cultivars. To address this, the PhenoFlex model was fitted on twenty-six apple cultivars grown at a single site in East Malling. The model was chosen as it is theoretically the most biologically accurate model currently available, enabling a flexible amount of overlap through parameter tuning for each cultivar. Additionally, a combined parameterisation approach, based on similarities in flowering times, and a generic parameterisation approach, using all cultivars in the study, were conducted to determine whether parameterisation was required at the cultivar, grouped-flowering-time or at the species (common apple) level.

The next section addresses the understudied impact of flowering time variation on fruit maturity; most harvest date predictions focus on the average flowering date, disregarding the effect of the spread of flowering time. The total sum of growth units was calculated for a collection of apples, using the temperature recorded from flowering until harvest. This was assessed its fruit maturity at harvest to determine the effect of flowering time on harvest. This analysis aimed to evaluate the limitations of phenology-based harvest predictions—particularly those relying on average flowering dates—and to determine whether a more targeted, image-based approach closer to harvest could offer improved accuracy for non-destructive maturity assessment.

The last section focused on directly assessing apples using imaging techniques as a non-destructive approach to predict apple maturity. An extensive dataset of hyperspectral images was collected from multiple regions, seasons, and cultivars, yielding the largest collection of apple hyperspectral images. Four different model architectures were tested to determine the best-performing model architecture. To optimise computational efficiency, without losing spatial information, input images were downscaled. Edge cropping was applied to isolate the key region of the apples, ensuring that the model focused only on the face of the apples. Cultivar information was embedded into the input layer to provide genetic information during training. Bayesian optimisation was employed to efficiently tune hyperparameters and reduce computational complexity. Shapley analysis was used to quantify the relative importance of spectral wavelengths and spatial regions within the images. Additionally, the necessity of multiple views of the fruit are required for accurate prediction was evaluated to assess the feasibility of simplified imaging setups for practical orchard use. Lastly, models were independently trained on seasonal data to observe seasonal-specificity in the data.
The key results from the studies are presented and evaluated for their applicability in developing a non-destructive framework for predicting apple maturity.

3. Materials and methods

3.1. Flowering data

The flowering time of twenty-six apple cultivars have been recorded from East Malling, United Kingdom, over the last eighty-five years. The flowering data collected for each cultivar ranged between eighteen to eighty-five years. Some records are shorter than others as some cultivars may have been recorded earlier than others, discontinued or are newer cultivars. Due to the apples being monitored from one site, the environmental variation can be limited within the dataset and directly determine the variation between apple cultivars. The specific years are indicated in Table 1. For each cultivar, their date of the first flower (BBCH 60, according to the BBCH-scale for fruit phenology ((Meier et al., 1994))), will be used for modelling purposes as the first flowers are less affected by environmental conditions other than temperature ((Darbyshire et al., 2016; Pope et al., 2014)). The flowering dates are either an average of four trees of the same cultivar or an individual tree, depending on the data availability. The number of trees are recorded in Table 1.The cultivars are grown on sixteen different rootstocks, which have not been used as a factor in the analysis as flowering behaviour is assumed to be determined by the scion.

Table 1: Summary of the range of flowering time data available from East Malling, the total number of years and number of datapoints used to train and test each model for each of the twenty-six cultivars. The table is split by K-means clustering on mean flowering dates and variation across years. The last column represents the standard deviation of flowering.

Cultivar	Starting year	Ending year	Total years	Tree (n)	Training years	Testing years	SD of flowering dates
Group 1							
Beauty of Bath	1948	1965	18	5	13	5	9.89
Crispin	1970	2021	51	5	36	15	9.75
Egremont Russet	1970	2021	52	5	36	16	10.53
Greensleeves	1984	2021	38	3	27	11	9.89
Idared	1984	2021	34	4	24	10	10.27
James Grieve	1972	2021	49	5	34	15	9.35
Jonagold	1984	2021	38	5	27	11	9.12
Group 2							
Edw7	1946	1969	24	13	17	7	7.27
Howgate Wonder	1960	2019	54	3	38	16	7.74
Lanes Prince Albert	1960	2021	62	3	43	19	8.3
Laxton's Superb	1950	1980	25	3	18	7	7.35
Tydeman's Early Worcester	1950	1987	31	5	22	9	7.55
Tydeman's Late Orange	1950	1980	26	5	18	8	6.85
Worcester Pearmain	1944	2021	70	11	49	21	8.16
Group 3							
Bramley's Seedling	1936	2021	81	9	57	24	8.9
Cox's Orange Pippin	1936	2021	85	17	59	26	9.19
Discovery	1970	2021	50	5	35	15	8.84
Elstar	1991	2021	29	1	20	9	7.88
Fiesta	1991	2021	31	2	22	9	8.16
Gala Mondial	1991	2021	30	2	21	9	7.64
Golden Delicious	1970	2020	51	4	36	15	9.17
Jupiter	1988	2021	34	1	24	10	7.95
Katy	1984	2021	37	2	26	11	8.67
Malling Kent	1972	2021	50	5	35	15	8.98
Spartan	1984	2020	36	2	25	11	8.82
Suntan	1973	2021	49	6	34	15	9.52

3.2. Flowering records

Apple flowers grow in clusters. Flowering records were done by tagging clusters of flowers with the date of bloom. The bloom date was noted when the majority (three of five flowers) of the cluster was fully open. Therefore, the flowering dates used in this study were when flowers on positions 2 and 3 were fully open, which usually occurred a day after the king bloom flower opened and a day before the flowers at positions 4 and 5 opened. A total of 1199 flower clusters were tagged between the 14 trees, made up of 12 to 85 clusters per tree.

3.3. Fruit Samples

Fruit of six apple cultivars were collected from Kent, United Kingdom and Hawke's Bay and Nelson, New Zealand in three harvest seasons. Fruit were harvested during the harvest seasons from NZ (February to April) in 2023 and 2024, and from UK (September to October) in 2024. In England, Cox, Braeburn, Fuji, Gala and Golden Delicious were picked from East Malling (51°17'07.0"N 0°27'13.2"E) and additional Gala, Braeburn, Cox and Jazz were sourced from orchards around Harbledown, UK (51°16'33.5"N 1°2'28.4"E). All cultivars, except Jazz, were picked from orchards at Plant and Food (PFR) Research Hawke's Bay (39°39'37.4"S 176°52'45.9"E). Additional Gala and Fuji were picked from a commercial orchard (39°36'11.5"S 176°49'20.9"E), Golden Delicious was picked from a local grower (39°36'54.2"S 176°50'52.1"E), and Braeburn and Fuji from PFR Nelson (41°06'49.7"S 172°59'04.3"E) in 2023 and 2024. All apples were picked in the morning, imaged, then the firmness, Brix and starch of each fruit were measured within 36 hours of harvest at ambient temperature. A total of 5756 apples were harvested.

3.4. Maturity measurements

The equipment reported for each maturity feature was used, as follows, in the UK and NZ, respectively. Firmness was measured using a fruit texture analyser (Llyod LRX, UK and GÜSS, South Africa) with an 11 mm diameter probe to a depth of 8 mm. Two regions (approximately perpendicular to each other) along the equatorial region of the fruit were punctured after peeling away the skin. The force at maximum depth was used in this study. Atago refractometers (portable benchtop palette series, model PR-32α and pocket PAL series, model PAL-1) were used to measure the apple juice collected from the puncture sites made during the firmness process. The refractometers were calibrated at the start of each sampling day using distilled water. Lastly, the starch percentage was measured by cutting the apples in half horizontally then a potassium iodine solution was applied to one half of the apple by dipping them in a 1% w/v iodine and 4% w/v potassium iodide solution or spraying with 1% w/v potassium iodide and 0.25% w/v iodine. The

apples were left to stain for 30 minutes (Per comms) before recording the staining percentage. The firmness readings from both sides of each fruit were averaged to create an average firmness reading.

3.5. Tree canopy zones

A single tree canopy was divided into 7 zones as a proxy for fruit exposure to light: north, south, east, west, upper, inner and lower (Table 4.1). The first 5 zones are regions on the outer areas of the trees with greater light exposure, whereas fruit from the inner and lower zones were mostly shaded by foliage during fruit development. The upper region consisted of fruit within the upper 25% of the tree. The fruit from the four cardinal directions were picked from the outer edge of the trees. Fruit picked from the inner and lower of the trees were located close to the trunk and within the lower 25% of the trees, respectively.

3.6. Temperature data

Temperature records were obtained from the East Malling weather station (51.2876°N, 0.4486°E, 33m above mean sea level), an official UK Meteorological Office Station. The orchards were located within 0.75 miles east and 0.31 miles north, 0.21 miles west, 0.18 miles south of the weather station. Fluctuations in hourly temperatures in a day follow typical patterns between maximum and minimum daily temperatures. These patterns can be modelled by a sine function for daytime warming and logarithmic decay for nighttime cooling, respective to a specific geographical latitude. A simulated approach following these patterns was used to generate missing hourly temperature values ((Luedeling et al., 2021; Luedeling & Fernandez, 2022)). Data from 1935 to 1999 are recorded as daily maximum and minimum temperatures so the simulated approach was used to generate hourly temperatures for all hourly observations between 1935 and 1999. These generated hourly temperatures were also used to fill in missing hourly datapoints from 2000 to 2021. Overall, 72.15% of the data was formed by the simulated approach, and 27.84% was of real data values. The remaining 0.011% is due to the inability to simulate hourly temperatures due to missing daily minimum and maximum temperatures. In total this accounts to three days and five hours of data which is unlikely to significantly affect model predictions. The hourly data from 2000 to 2021 was mostly complete. It consists of 99.8% of real temperature values, 0.24% in simulated data and a negligible amount in missing temperatures (28 hours).

3.7. Model Formulation

3.7.1. PhenoFlex Model

The PhenoFlex model, implemented in the PhenoFlex_GDHwrapper() function from the chillR package, integrates the framework from the Dynamic model and the GDH model (Luedeling and Fernandez, 2022). The PhenoFlex model (Luedeling et al., 2021) is fitted with twelve parameters,

with the parameters for the chilling requirement (y_c), the heat requirement (z_c) and slope ($s1$) linking the Dynamic and GDH models. The heat accumulated at any point in time (t) is calculated by the PhenoFlex model equation, incorporating the total heat accumulated so far (z) and a portion of the GDH function over the elapsed time and temperature (T) (Luedeling et al., 2021). *P y(y)* is a function following a sigmoidal pattern which determines the proportion or size (s) of heat that can be accumulated, as a function of the accumulated chill (y) (Luedeling et al., 2021). The inflection point is determined by the critical chilling threshold (y_c) and the slope of the transition is determined by the parameter *s1*. Large values of *s1* indicates lower levels of overlap and vice versa.

Six of the twelve PhenoFlex parameters are associated with the Dynamic model. The hypothetical process to form and destroy the precursor to the dormancy-breaking factor (PDBF) follows Arrhenius law. *E0* and *E1* represent the time-independent activation energy, and *A0* and *A1* refer to the amplitude of the function. *E0* and *A0* contribute to PDBF formation, while *E1* and *A1* are involved in PDBF destruction. When x reaches 1, a portion of the PDBF is converted to a stable chilling portion where it cannot be destroyed by warm temperatures (Erez & Couvillon, 1987). The pseudo-intermediate (x) is calculated as a function over time, where t is the new time and t_j is the level of x at time j. The portion converted is determined by a sigmoidal function with the inflection point at T_f and slope governed by the slope parameter (Erez & Couvillon, 1987).

Three parameters are associated with the GDH model. The contribution to heat accumulation is dictated by the optimal temperature (T_u), the upper temperature limit (T_c) and the lower temperature limit (T_b) (Anderson et al., 1985). The difference between optimal and lower temperatures are multiplied with a function which determines the effectiveness of GDH in driving the biological process under consideration (Anderson et al., 1985).

3.7.2. Linear Growing Degree Hours

The linear GDH, established by Anderson & Seeley (1992), assumes a linear relationship of growth with accumulative temperatures above a temperature (base) threshold. It has 3 parameters, T_b, T_c and T_u, representing the base, critical and optimal temperatures. Temperatures below the base do not count towards GDH units, nor temperatures exceeding the critical threshold. The contribution of each degree increase in temperature from the base linearly increases as temperature increases, up until the optimum temperature. Temperatures between the optimum and critical temperatures accumulate GDH units at the maximum rate. Thus, strictly speaking, this GDH is not linear but two lines joining at T_u.

$$GDH_{linear}(T) = \sum f_{linear_{GDH}}(T)$$

$$f_{linear_{GDH}} = \begin{cases} T_u - T_{b,} & if\ T_u \leq T < T_c \\ T - T_b, & if\ T_b < T < T_u \\ 0, & if\ T \leq T_b\ or\ T_c \leq T \end{cases}$$

In grid search, T_b spanned from 0°C to 10°C, T_u from 15°C to 25°C, and T_c from 30°C to 40°C. The base temperature originally proposed by Richardson (1975) was 4.5°C for peach trees. However, a recent study by Tang et al. (2024) found that the base temperatures of apple trees may be lower than 4.5°C. The search was extended from 0°C to 10°C to explore the best fitting base temperature. T_u was chosen based on the expected best growth conditions of most living organisms, and finally, the critical temperature was expected to range somewhere between 30 and 40°C

3.7.3. Non-linear Growing Degree Hours

The second is another well-established model growing degree hour model by Anderson et al. (1985) Opposed to the linear non-linear GDH model, this model assumes a non-linear accumulative relationship of growth with temperature. Each temperature increase from the base causes a non-linear increase in GDH up until the optimum temperature. Temperatures above the optimal gradually decrease in effectiveness in GDH accumulation.

$$GDH(T) = (T_u - T_b) * f_{non-linear_{GDH}}(T)$$

$$f_{non-linear_{GDH}} = \begin{cases} \frac{1}{2} * \left(1 + cos\left(\pi + \pi * \frac{T - T_b}{T_u - T_b}\right)\right), & if\ T_b < T < T_u \\ 1 + cos\left(\frac{\pi}{2} + \frac{\pi}{2} * \frac{T - T_u}{T_c - T_u}\right), & if\ T_u < T < T_c \\ 0, & otherwise \end{cases}$$

3.7.4. Thermodynamic model

The Thermodynamic model is a non-linear growth rate model based on the theory of enzyme activity rate variation in response to temperature changes (Wagner et al., 1984; Xu, 1996). The parameters for the Thermodynamic model are *B, C, TH* and *ρ*.

$$R(K) = \frac{\frac{\rho K}{298} exp\left[B(1 - \frac{298}{K})\right]}{1 + exp\left[C(1 - \frac{TH}{K}\right]}$$

In the grid search, the range of parameters *B* was 15 to 40°K, *C* was 5 to 30°K and *TH* was 290 to 300°K. The last variable in the equation, *ρ*, is a scaling factor and does not affect the correlation of the estimated growth unit from flowering to maturity with SPI, so it was fixed to 1.

3.8. Model optimisation and performance evaluation

The flowering data was split into a training dataset and a test dataset by randomly selecting 70% of the years for each cultivar and leaving the last 30% of the unselected data for the test dataset. This split was done for each cultivar then the split was maintained for subsequent model fitting and

comparisons between approaches. Specific models were fitted to the corresponding training data with a simulated annealing algorithm, wrapped in the phenologyFitter() function from the chillR package (Luedeling & Fernandez, 2022). The simulated annealing mechanism generates model parameters for the model chosen, then aims to reduce the residual sum of squares (RSS) by choosing a new set of model parameters. This process is repeated up to 1000 times or until there are no improvements after 250 iterations. The best fit model was then bootstrapped 99 times with the function bootstrap.phenologyFit() in the chillR package (Carsten et al., 2022; Luedeling & Fernandez, 2022). The standard errors for the parameters were calculated on the 99 bootstrap values and the original set of fitted parameter values. This process was repeated at least seven times using different starting parameters as the PhenoFlex model results can be sensitive to the initial parameters. The reported results were from the run with the smallest residual sum of squares (RSS).

Fitted models were evaluated with Akaike information criterion (AIC) (Burnham & Anderson, 2002), model efficiency (EFF) (Nash & Sutcliffe, 1970) and RMSE for both the training and test datasets. AIC is a measure of model goodness of fit that considers the number of parameters in the fitted model. The function AICc contains a penalty term adjusting for small sample sizes.

$$AICc = 2k + nlog(\frac{RSS}{n}) + \frac{2k^2 + 2k}{n - k - 1}$$

The number of parameters is represented by the letter k, the number of samples, n, and the residual sum of squares, RSS. AICc will be used to assess the fitted models and their parameters and determine the model that minimises information loss. AICc values are relative to each other, the smaller the AICc value, the better. The model efficiency (EFF) compares models that were fitted to the same training dataset. The efficiency is the ratio between the residual sum of squares and the squared sum of the differences between the observed values and the mean.

$$EFF = \frac{RSS}{\sum(t_i - \bar{t})^2}$$

Root mean squared error (RMSE) is a commonly used metric of prediction accuracy. AIC and EFF can only be used to compare between models fitted with the same dataset, while RMSE can be used to compare between models. This is why RMSE is used to evaluate the test dataset. Due to differences in the amount of flowering data for each cultivar, the Ratio of Performance to InterQuartile distance (RPIQ) was used to standardise the prediction errors against the variation of the observed flowering dates.

3.9. Comparative modelling

3.9.1. Comparing PhenoFlex models between apple cultivars

Firstly, PhenoFlex models were fitted to individual cultivars, and thus the model was fitted to twenty-six test datasets (one for each cultivar). Next, K-means clustering was applied on mean

flowering dates and their variation across years to determine flowering groups. K-means clustering is an unsupervised method which assigns each observation into a group based on their similarities with other observations in the same group. Principal Component Analysis (PCA) of standardised Z scores were used for interpretability and visualisation. PhenoFlex models were then fitted to each flowering group of cultivars as identified by the K-means clustering. Finally, a single PhenoFlex model was fitted to all twenty-six cultivars. Model performance, particularly for the test datasets, was then evaluated and compared among the three sets of models.

3.10. Assessment of the relative importance of experimental factors in fruit maturity

Logistic regression with starch proportion as the response variable was used to determine the effect of flowering time (as approximated by the estimated temperature-based growth unit), year, individual trees, and fruit position within the tree canopy on fruit maturity. In the GLM analysis, a binomial distribution was assumed for the residual errors. The deviance explained by each experimental factor was calculated by extracting the residual deviance from ANOVA tables calculated using Chi-square (synonymous to likelihood ratio) as the test function. Since this study focused on the temperature effect (flowering time) on fruit maturity, the accumulated growth models estimated by one of the three models was first added in GLM analysis of SPI for each cultivar. Then, year, tree, and canopy region were added sequentially. A nested model approach was used to test for statistical significance of the effect of specific factors on fruit development (SPI).

3.11. Obtaining fruit images and initial image processing

Images were taken using two different Specim IQ hyperspectral cameras (Specim Imaging Ltd., Oulu, Finland) using the same settings (Behmann et al., 2018); one was in New Zealand and the other was in the United Kingdom. The spectral range of the cameras was 400-1000 nm, with 204 spectral bands at 7 nm resolution. Two 750 W tungsten halogen lights (ARRILITE 750 Plus, ARRI, Germany) were each positioned approximately 1 m apart, forming a triangular setup with the photo shooting tent. This setup ensured little shadow was cast on the apples. The tent diffused the light to avoid overexposure caused by direct light on the waxy layer of apples. The camera was set in front of the tent. The camera set-up and calibration were followed according to the manufacturer's manual. The reflectance was generated by correcting with white and dark references to reduce background noise. Images were captured following appropriate integration times.

In a dark room, four images of each apple were taken by rotating the apples by 90 degrees on the horizontal axis to attain four equatorial images. These were processed at room temperature before the maturity assessments. Depending on the size of the fruits, each image contained between 2 to 8 apples arranged in two rows. A total of 3636 images were taken. The image metadata was inputted during the imaging process, ensuring the date, cultivar and fruit numbers were recorded.

Some apples may not have images from all four sides due to human errors during the collection process.

3.12. Image Processing

Since multiple apples were presented in the images, object detection was done using Segment Anything Model 2 (SAM2) (Ravi et al., 2024) on the RGB images generated by the hyperspectral camera. Occasionally, SAM2 would not detect all the apples within the image. To ensure that we retained the correct number of actual apples and those detected by the image detection AI, we counted the number of bounding boxes detected and matched it to the count of fruit noted in the metadata. When the numbers did not match for that image, the image and the maturity data for those apples were removed from the final dataset. 2.75% of the images were filtered out due to mismatches between the detected and actual number of apples in the image. Valid masks were applied onto the hyperspectral images so that any pixel coordinates for any wavelength outside of the segmented area is zeroed.

The apple images within the bounding box coordinates (apple and zeroed background) were divided into equal parts before locally averaging the pixels, forming either 30-by-30 or 50-by-50 pixel images. This method reduces the size of the tensors and retains spatial information. To further localise just the centre of the images, 5 pixels were removed from each edge of the generated images, thus forming 20-by-20 or 40-by-40 pixels images. Due to the curvature of apples and lower quality of pixel values around the edge of the fruit, removing the edges may improve model accuracy (Wang et al., 2022). The spectral data, 204 contiguous bands, made up the third dimension of the image. In addition to the hyperspectral data, each sample was labelled with its corresponding cultivar information using one hot encoding so that it is trainable by machine learning algorithms. The six cultivars in the study were encoded into a binary format and integrated into the last layer of the images, increasing the third dimension from 204 to 210 channels, unless otherwise stated.

The data was first split into their respective cultivars before randomly splitting into training, test and validation datasets at 80, 18 and 2%, respectively. The different training, test and validation datasets for each cultivar were rejoined after the random split to ensure all cultivars were equally represented in all three datasets. The images were further processed to test whether specific regions of the apple were important in model predictions. The images were processed as above, then they were split in half vertically and horizontally before the quadrants were swapped diagonally.

3.13. Deep Learning Models

The selection of these four model architectures was intended to systematically compare the performance differences of various deep learning strategies in processing hyperspectral data for the values of Brix, firmness, and starch prediction. Specifically, 2D-Convolutional Neural Networks

(CNN) excel at extracting local spatial features; 3D-CNNs can simultaneously capture joint features in both spatial and spectral domains; Hybrid CNN-Transformer models attempt to combine the local receptive field capabilities of CNNs with the global dependency modelling capabilities of Transformers; and Vision Transformer (ViT) models rely entirely on self-attention mechanisms for feature learning.

3.14. Hyperparameter Tuning

Bayesian optimisation is a method to optimise model hyperparameters (Wu et al., 2019). Bayesian optimisation was used to determine the hyperparameters only for the ViT model, as it demonstrated the strongest baseline performance among the four architectures tested. A combination of the number of layers (1-5), patch size (2-10), projection dimension (64-256), number of heads (1-5), multilayer perceptron (MLP) head units (small head: [128, 64], medium head: [256-128] or large head: [256-128-64]) and dropout rates (0.1-0.5) were tested. The larger the head, the more complex patterns can be learnt, but it may overfit and require more time to train. The final model was evaluated as follows.

3.15. Model Evaluation and Feature Selection

The three models return outputs for Brix, average firmness and starch percentage, independently. Root Mean Squared Error (RMSE) and the Coefficient of Determination (R2) were used to determine the performance of the model. Evaluation of important wavebands, apple regions of interest and cultivar was done using Shapley values. Shapley value is based on coalition theory to fairly split a prize pool of money, based on their contribution. The greater their contribution, the greater their winnings. We calculate Shapley values for each feature on our final machine learning models to determine the contribution of each waveband to each maturity feature prediction, as well as the significance of each region of the apple and cultivars.

Floating Point Operations (FLOPs) were used to measure the complexity of models. The lower the FLOPs value, the lower the computational cost. We used FLOPs to assess the difference between the Bayesian optimised models and the models trained on default parameters

3.16. Hardware and software environment

3.16.1. R version

The analysis was run on R version 4.3.2 (2023-10-31 ucrt). Model fitting and bootstrapping was run on a high throughput computer running R version 4.2.3 (2023-03-15)

3.16.2. Python

Models were trained and evaluated with Python 3.10.13, using Keras (v2.15.0), Pytorch (v2.1.0), scikit-learn (v1.4.2), Numpy (v1.26.4), Pandas (v2.2.3), OpenCV (v4.8.1) and Matplotlib (v3.8.2). All ML models were trained on a NVIDIA A100 (80 GB), 503 GB RAM system.

4. Results

4.1. 3.3.1 PhenoFlex models fitted to individual cultivars

Cultivar-specific parameter estimates for the PhenoFlex model were used to predict flowering dates for individual cultivars. The parameters were derived from running the model 10 times with different starting parameters. The parameters were selected from the runs with the lowest RSS for each cultivar. The average RSS was 1194.15 ± 103.69 for the specific model.

The PhenoFlex model fitted well to the individual training datasets of twenty-six cultivars, resulting in an average RMSE of 6.15 ± 0.22 days and an R^2 value of 0.99 (Figure 1A). A decline in model performance (RMSE 13.8 ± 0.53 days) was observed when the fitted model was used to predict flowering date on the test datasets (Figure 1B). The resulting R^2 value was negative (-3.93), indicating a poor model fit. Poor model performance was particularly apparent for ten cultivars: Cox's Orange Pippin, Egremont Russet, Fiesta, Golden Delicious, Greensleeves, Katy, Malling Kent, Spartan, Tydeman's Early Worcester, and Worcester Pearmain, with RMSEs above 13 days. When these ten cultivars were excluded, the R^2 value of the test data improved to 0.43.

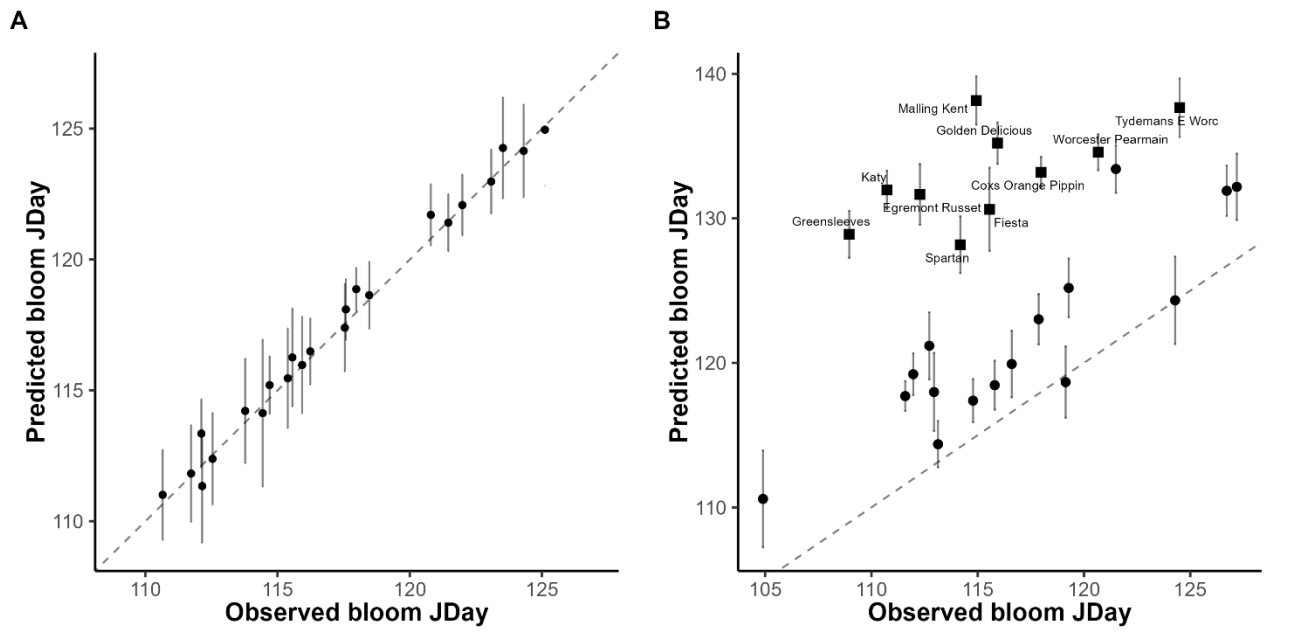

Figure 1. A comparison of the mean observed and predicted bloom dates using cultivar-specific parameters on the PhenoFlex model on A) training data and B) test data. The dashed line represents the line of equality or the y = x relationship between the x and y coordinates. The square points in B represent cultivars which have RMSEs greater than 13 days.

Linear regression with forwards and backwards selection were used to determine which of the twelve PhenoFlex parameters are correlated with high RMSEs. There were no significant correlations between high RMSE and any of the twelve parameters.

Of the twenty-six cultivars, seven cultivars – Egremont Russet (5.12 ± 1.26 days), Gala Mondial (4.54 ± 1.58 days), Howgate Wonder (4.49 ± 1.25 days), Jupiter (4.95 ± 1.34 days), Katy (6.42 ±

1.88 days), Lane's Prince Albert (5.44 ± 1.22 days) and Malling Kent (4.42 ± 1.16 days) – resulted in test data RMSEs smaller than the standard deviation of flowering dates between years. The RPIQ of the cultivar-specific PhenoFlex model was 1.64 for the training data but only 0.75 in the test data.

4.2. PhenoFlex models fitted to groups of cultivars as identified by mean flowering dates and variation across years

The cultivars separated well in the first two dimensions of the PCA scores (Figure 2A), with three clusters identified using the silhouette method (Figure 2B). Group one contained 7 cultivars with 280 flowering dates, group 2 contained 7 cultivars with 292 flowering dates and group 3 contained 12 cultivars with 563 flowering dates. K-means was applied on mean flowering dates and variation across years to divide the cultivars by their flowering behaviours. The variance of flowering patterns in group 1 were the highest at 9.48 days, followed by group 3 at 8.96 days and lastly group 2 at 5.36 days. The cultivars from group 1 contains the least genetic variability with a total of 32 trees, followed by group 2 with 43 trees and group 3 with 56 trees.

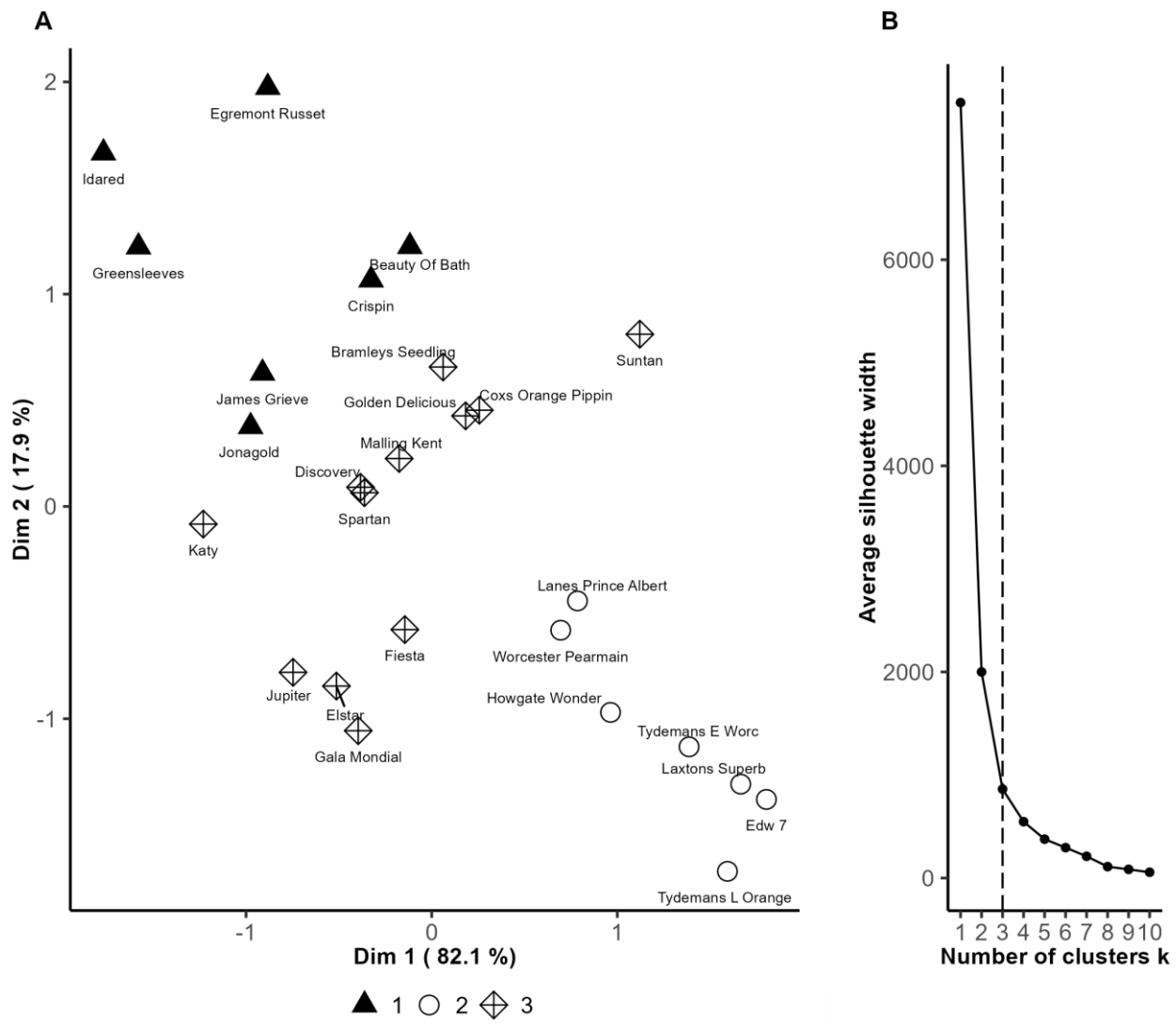

Figure 2. K-means clustering presented in a PCA plot for mean flowering time and flowering variation across years for the twenty-six cultivars. B) Silhouette plot indicating three optimal clusters.

We fitted PhenoFlex models for three groups of cultivars identified via K-means clustering of the means and variations of flowering dates (Figure 3). The model was run 10 times, each with different initial parameters. The average RSS identified from the 10 runs on the mean flowering time groups were 19507.29 ± 201.39, 5661.96 ± 111.61 and 2911.12 ± 209.45 for groups 1, 2 and 3, respectively. Group 1 consists mostly of cultivars which bloom earlier in the season, group 3 consists of cultivars blooming late in the season, and group 2 contains cultivars which bloom sometime in between. As the groups were split by their mean flowering dates, R^2 values are not that relevant, but were reported to be 0.51, 0.58 and 0.33 on the training data and 0.54, 0.74 and -0.06 for the test data for groups 1, 2 and 3, respectively. Evaluation against the test datasets showed the best model performance among the three model approaches. The RMSE for groups 1,2 and 3 were 9.68 ± 0.69, 4.98 ± 0.35 and 8.42 ± 0.43 days on the training data, respectively.

The RMSE remains consistent on the test dataset, at 5.46 ± 0.60, 4.34 ± 0.47 and 5.50 ± 0.42 days for groups 1, 2 and 3, respectively (Figure 3). The average RMSE for the three groups were all less than the standard deviation of the interannual flowering dates, so these model parameters identified were significantly better than taking the average flowering date of each cultivar. The mean flowering date approach is a significant improvement on the predictive accuracy of flowering dates compared to using cultivar-specific parameters. The RPIQ of the mean flowering time groups 1, 2 and 3 were 1.06, 1.60 and 1.11, respectively. Their RPIQs improved when applied to the test dataset (group1 = 1.75, group 2 = 1.86 and group 3 = 1.55). Overall, the mean flowering date clustered groups performed better than the cultivar-specific model.

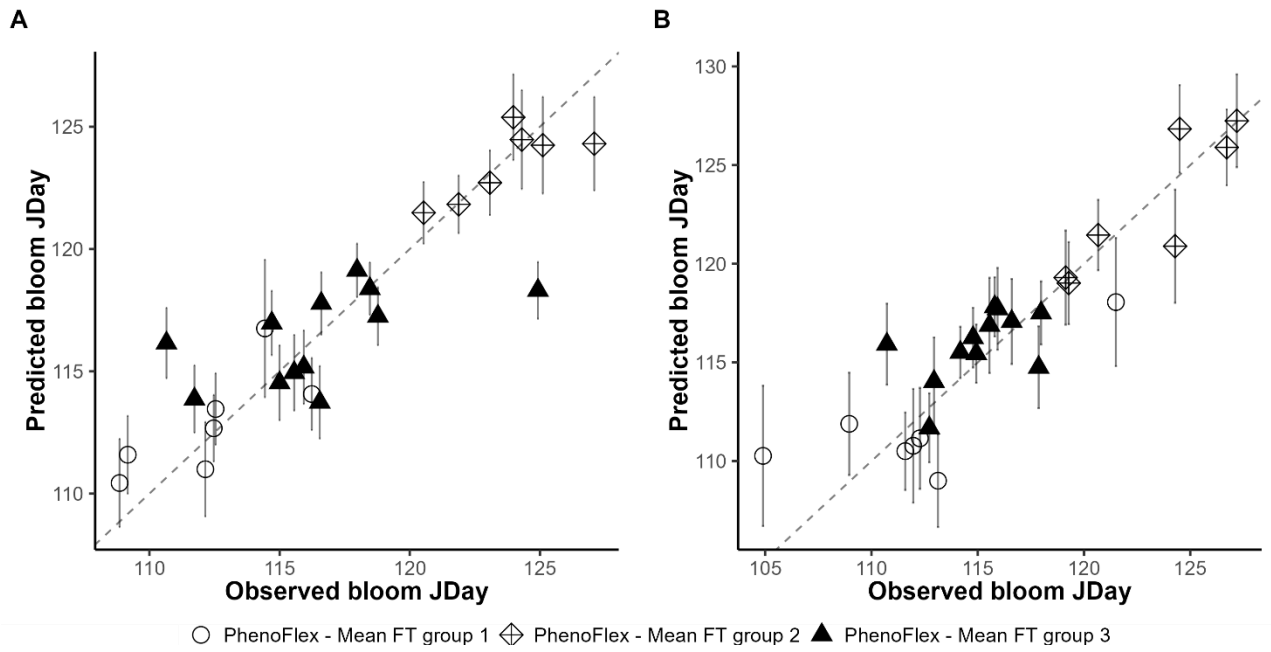

Figure 3. A comparison of the observed and predicted bloom dates of cultivars grouped by K-means clustering on mean flowering and variation on the PhenoFlex model for A) training data and B) test data. The dashed line represents the line of equality or the y = x relationship between the x and y coordinates.

The mean flowering clustered groups performs well but are outperformed by the cultivar-specific model predictions on four cultivars (Beauty of Bath, Edw7, Greensleeves and Katy) on the training data (Figure 4A). In the test dataset, the mean flowering clustered group outperformed both the cultivar-specific and common model approaches in all but Beauty of Bath (Figure 4B).

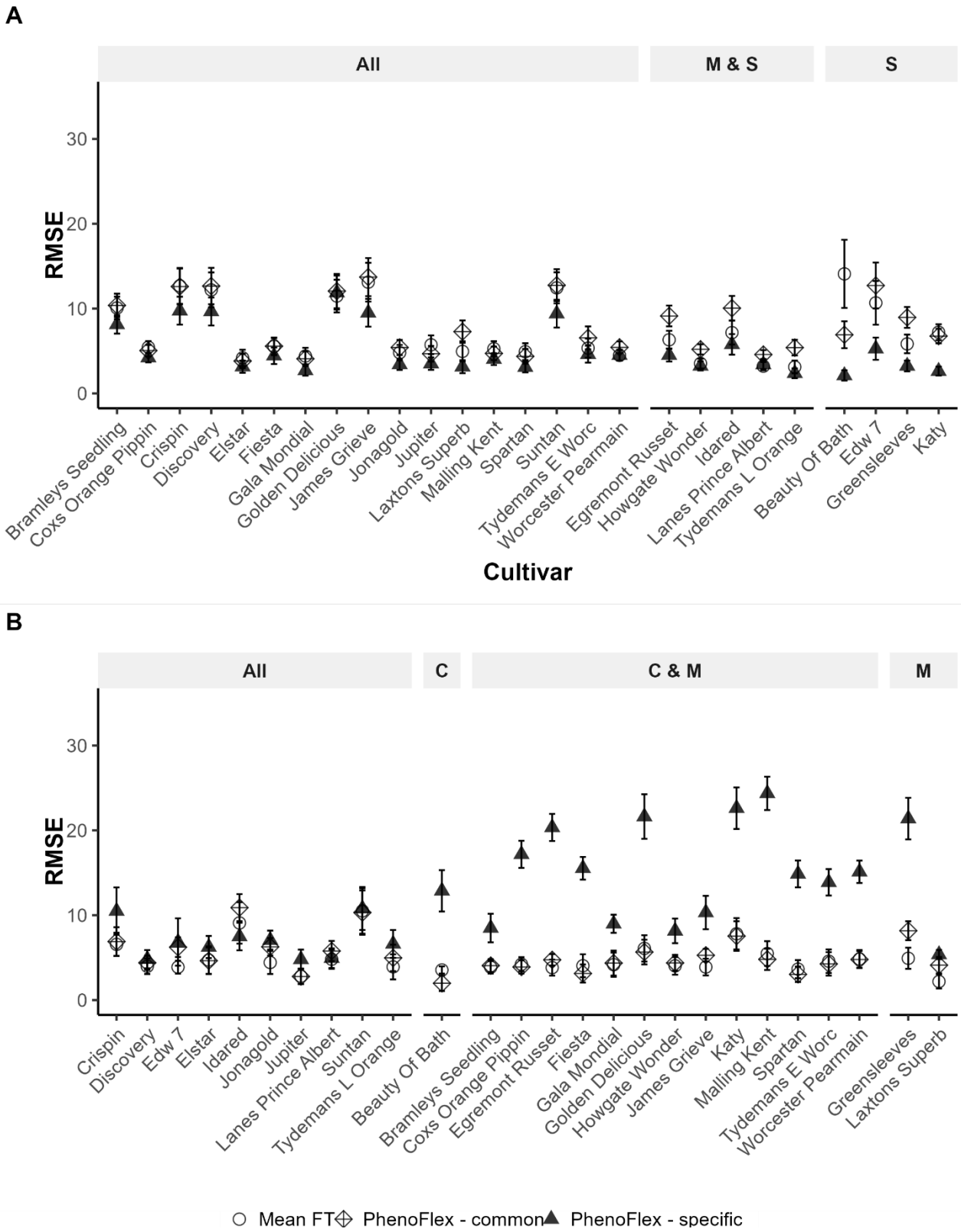

Figure 4. Comparison of RMSE of the cultivar-specific, mean flowering grouped and common models for A) the training data and B) the test data. The graphs are separated by whether the specific (S), grouped by mean flowering date (M) or common (C) models perform better. When two of the approaches perform equally well, two letters are shown (M&S and C&M) or whether there is no difference between approaches (All).

4.3. A common PhenoFlex model fitted to all cultivars

The common model was run 7 times with various initial parameters. The average RSS identified across the seven runs was 60431.79 ± 785.74. The standard error is higher than the previous approaches, likely because the model needs to allow for larger errors to fit a more generalised model with more cultivars. The common model performance on the training data was 8.62 ± 0.31 days, with an R^2 value of 0.44 (Figure 5). Its performance on the test data yielded a RMSE of 5.64 ± 0.30 days and R^2 of 0.53, which was better than cultivar-specific model performance. Unlike the specific model, which predicts bloom dates later than the observed bloom date, the common model predicts flowering time around the observed flowering date (Figure 5). The RPIQ observed for the common model was 1.16 for the training data and 1.67 for the test data. This RPIQ is on par with the RPIQ observed using clustered mean flowering dates. The common model produced smaller RMSE than the standard deviation of flowering time for all cultivars, suggesting that model predictions are better than taking the average bloom date for each cultivar.

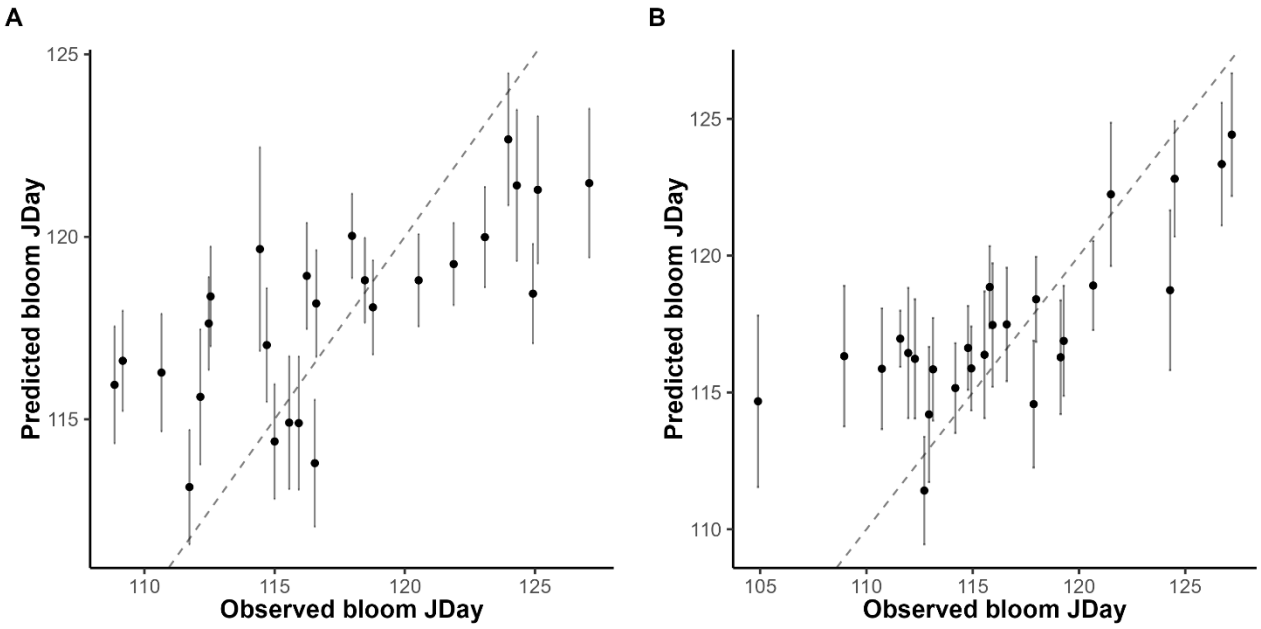

Figure 5. A comparison of the observed and predicted bloom dates using common parameters on the PhenoFlex model on A) training data and B) test data. The dashed line represents the line of equality or the y = x relationship between the x and y coordinates.

The cultivar-specific model was able to more accurately predict bloom dates better on the training data (Figure 4A), but the common model predictions outperformed the cultivar-specific model predictions or were the same between model approach predictions in the test dataset (Figure 4B). The common model approach was outperformed by the mean flowering cluster approach on Greensleeves and Laxton's Superb (Figure 4B).

When there are large numbers of data per cultivar (more than 30 years), the cultivar-specific approach does well in predicting bloom dates. When there are around 20 year of data, the mean bloom date cluster performs well since this method increases the number of datapoints by

including more cultivars. When there are only few numbers of years per cultivar (approximately 10 years), but many cultivars are present, it is better to apply the common approach.

4.4. Temperature-based fruit development

The correlation between calculated linear GDH and starch proportion had a maximum of 0.54 from Braeburn and a minimum of 0.22 from Fuji (Table 2). Golden Delicious and Fuji have the smallest correlation, have nearly identical parameters (Table 2) and hence temperature-growth rate relationship (A). They have the highest optimum and critical temperatures. Gala has the smallest effective temperature range between 9.1 – 30 °C, which may be compensated by the lowest optimal temperature (16.1 °C). Braeburn and Cox have similar temperature-growth rate relationship, but the parameters for Cox are 3 °C lower than Braeburn for all three parameters.

Table 2. Linear Growing Degree Hour model parameters estimated by the best correlation (Kendall's Tau) to Starch. Where multiple combinations result in the best correlation, the parameters presented are the closest to the median values. The errors represent the standard deviation of the best correlated parameters.

Cultivar	Best correlation	T_b	T_u	T_c
Braeburn	0.54	10.0 ± 0.00	19.0 ± 0.29	37.3 ± 1.86
Cox	0.41	7.8 ± 0.00	16.3 ± 0.00	34.4 ± 3.28
Fuji	0.22	10.0 ± 0.05	24.9 ± 0.12	39.2 ± 0.21
Gala	0.47	9.1 ± 0.11	16.1 ± 0.27	30.0 ± 0.63
Golden Delicious	0.33	10.0 ± 0.00	24.9 ± 0.06	39.6 ± 0.26

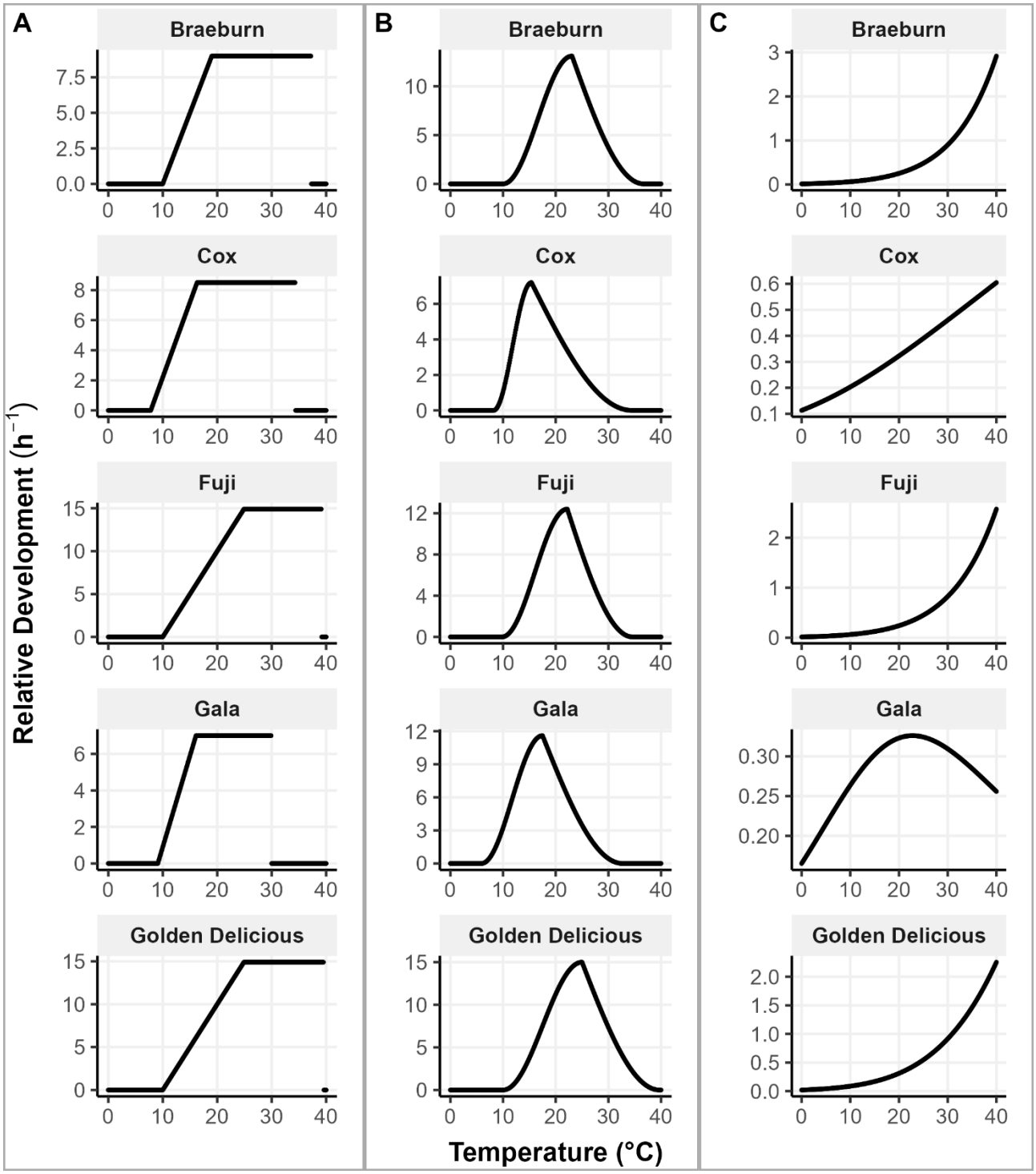

Figure 6. Growth rate of apple cultivars for hourly exposure to temperatures

For the non-linear GDH model, Braeburn has the highest correlation of 0.54. The least correlated is Fuji, with a weak negative correlation (Table 3). Fuji, Golden Delicious and Braeburn have similar temperature-growth rate relationship (Figure 6B). Their parameters vary slightly, with Fuji having lower temperature requirements and Golden Delicious being acceptive of high temperatures (Tc = 39.7°C). The critical temperature for Gala is the lowest of the five cultivars. Gala and Cox have low base temperature values (5.9 and 8.2°C, respectively).

The temperature-growth rate relationship as modelled by the Thermodynamic model for Braeburn, Fuji and Golden Delicious are similar (Table 4 and Figure 6C). They all follow an exponential pattern, suggesting their maximum growth rate has not been reached at 40°C. Similarly, Cox has also not reached its maximum growth rate 40°C, but its growth rate appears to be almost linearly related to temperature. In contrast, the maximum growth rate for Gala is at 22°C. The estimated relative growth rate at 22°C were similar for all cultivars.

Table 3. Non-linear Growing Degree Hour model parameters estimated by the best correlation (Kendall's Tau) to Starch. Where multiple combinations result in the best correlation, the parameters presented are the closest to the median values. The errors represent the standard deviation of the best correlated parameters.

Cultivar	Best correlation	T_b	T_u	T_c
Braeburn	0.54	10.0 ± 0.00	23.1 ± 0.28	36.7 ± 2.32
Cox	0.41	8.2 ± 0.28	15.4 ± 0.28	34.5 ± 1.71
Fuji	-0.17	9.8 ± 0.10	22.2 ± 0.49	34.6 ± 2.85
Gala	0.47	5.9 ± 0.53	17.5 ± 0.77	32.7 ± 1.66
Golden Delicious	0.30	10.0 ± 0.00	25.0 ± 0.00	39.7 ± 0.00

Table 4. Thermodynamic model parameters estimated by the best correlation (Kendall's Tau) to Starch. Where multiple combinations result in the best correlation, the parameters presented are the closest to the median values. The errors represent the standard deviation of the best correlated parameters.

Cultivar	Best correlation	B	C	TH
Braeburn	0.53	40.0 ± 0.19	7.9 ± 1.03	293.6 ± 3.34
Cox	0.41	18.8 ± 2.18	17.1 ± 5.06	290.0 ± 0.00
Fuji	0.23	39.8 ± 2.50	9.1 ± 5.30	290.0 ± 0.32
Gala	0.37	16.4 ± 0.52	29.9 ± 0.11	290.7 ± 0.93
Golden Delicious	0.34	40.0 ± 0.00	20.9 ± 0.00	300.0 ± 0.00

As expected, there is a negative correlation between the proportion of starch and accumulated growth unit for all three growth models —linear GDH (Figure 7A), non-linear GDH (Figure 7B) and Thermodynamic (Figure 7C) models. For Braeburn, Cox and Gala, the relationship follows a logistic shape as starch proportion does not change significantly in the early stages of development, but rapidly degrades after accumulation of certain growth units (Figure 7). The difference in the trend between years is consistent with the observed relationship of temperature accumulation with maturity: harvesting fruit appeared to be too early in 2023. For Golden Delicious, proportion of starch appears to decrease linearly with increasing accumulated growth units for all three models. The correlation between maturity and relative growth rates for Fuji apples were consistently the lowest (Table 2-Table 3Table 4), this is reflected by the weak trends observed for

Fuji (Figure 7). The calculated accumulated growth units were higher in 2023 than in 2022 for linear GDH and Thermodynamic models, but the opposite was true for the non-linear model (Figure 7). This change in the accumulated growth units does not occur in the other four cultivars.

4.5. Factors contributing to maturity variation

Table 5 shows the summary of deviance in the fruit maturity (SPI) attributable to individual factors for individual cultivars. Comparing the deviance explained by the accumulated growth units across the three temperature growth models, the linear GDH model is the most effective for Gala (14.71%), the non-linear GDH model suits Braeburn, explaining 3.37%, and the Thermodynamic model works best for Cox (19.68%) and Golden Delicious (6.76%). However, it should be noted that the deviance attributable to the accumulated growth units was very similar among the three models (Table 5). The effect of accumulated growth units is not always statistically significant; only the linear GDH model for Braeburn (2.88%) and Gala (14.71%), non-linear GDH model for Braeburn (3.37%) and Thermodynamic model for Golden Delicious (6.76%) were statistically significant. For Fuji, < 1% of deviance in proportion of starch was explained by accumulated growth units (Table 5).

Some of the differences between the two seasons are expected to be accounted for by the accumulated growth units. The year effect did not contribute much to the deviance in proportion of starch for Braeburn or Gala, but it did affect proportion of starch significantly for Fuji with the linear GDH model (2.17%) and Golden Delicious with the non-linear GDH (1.75%) and Thermodynamic models (1.84%).

Differences between individual trees did not significantly affect proportion of starch for Braeburn, Fuji and Gala. In contrast, for Cox's Orange Pippin, tree effects were highly significant for all growth models, contributing 6.71%, 6.75% and 7.55% of deviance in the linear GDH, non-linear GDH and Thermodynamic models, respectively. For Golden Delicious, tree effects were significant for the linear and non-linear GDH models (Table 5). The regions within the canopy contributed to less than 6% of the deviance in the observed proportion of starch, none of which was statistically significant.

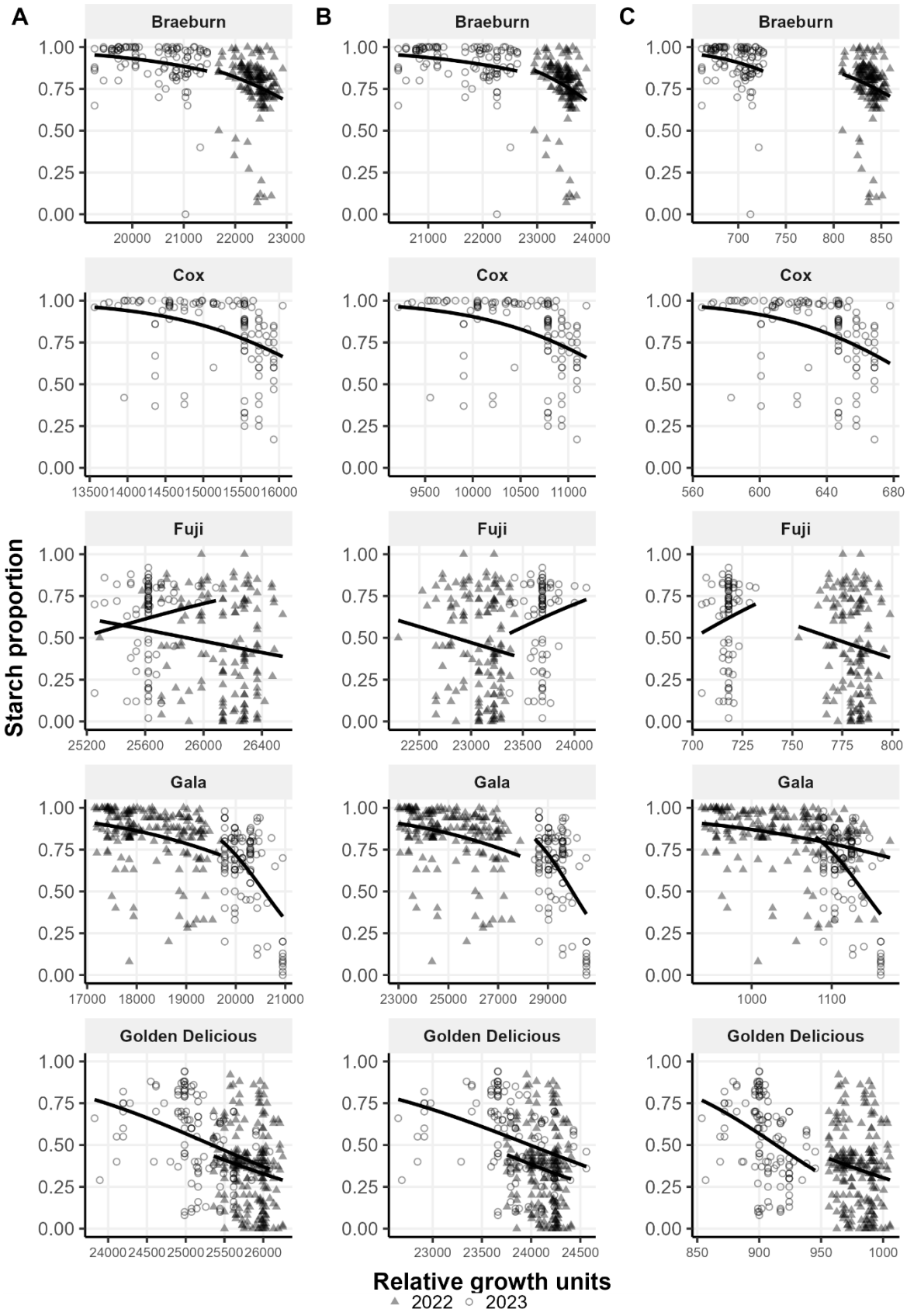

Figure 7. Proportion of starch (1 – immature and 0 – mature) against relative growth units calculated using A) linear Growing Degree Hours, B) non-linear Growing Degrees Hours and C) Thermodynamic model for each apple cultivar across 2 years. The trend lines show the trends of the values from 2022 and 2023 for each cultivar on each model.

Table 5. The percentage of the total variation explained by the linear GDH, non-linear GDH and Thermodynamic models for all cultivars. The values represent the percentage explained by each variable. Chi-squared test was used to determine the significance of each variable

Terms	Linear GDH	Non-linear GDH	Thermodynamic
Braeburn			
Relative growth units	2.88 ***	3.37 ***	2.18
Year	1.07	0.84	0.05
Tree	0.65	0.83	0.46
Region	1.95	1.93	2.10
Cox's Orange Pippin			
Relative growth units	18.44	18.70	19.68
Tree	6.71 ***	6.78 ***	7.55 ***
Region	4.82	4.88	5.61
Fuji			
Relative growth units	0.52	0.61	0.29
Year	2.17 ***	6.40	0.03
Tree	0.58	0.61	0.54
Region	2.90	2.86	3.16
Gala			
Relative growth units	14.71 ***	13.42	12.18
Year	0.64	1.53	4.88
Tree	0.33	0.49	0.66
Region	3.57	3.35	3.06
Golden Delicious			
Relative growth units	6.15	5.86	6.76 ***
Year	0.02	1.75 ***	1.84 ***
Tree	10.86 ***	10.61 ***	11.50
Region	2.09	1.99	2.30
The significance codes denote the p-value thresholds.			
*** $p < 0.001$			
** $p < 0.01$			
* $p < 0.05$			
. $p < 0.1$			

4.6. Hyperspectral Imaging

As this work has not been published yet, we provide a summary of results from the study.
We evaluated the state-of-the-art effective model and tested the effects of wavelengths, regions of interest, cultivar encoding, and seasonality on models predicting Brix, firmness and starch. We found that imaging is an effective method for predicting Brix and firmness, but it does not work well for starch. The best model performance is achieved by training Vision Transformer models for each maturity feature, on datasets from different seasons and geographic locations, training the models on more than one side of the fruit and encoding cultivar information in the training process. Reductions in computational processes can be achieved by reducing the wavelengths to the top 50% of most informative wavelengths (assessed with Shapeley analysis). We achieved RMSE of 0.76 and R^2 of 0.63 for firmness and RMSE = 0.91, R^2 = 0.75 for Brix. Starch model predictions did not perform well.
It was found that only half of the wavelengths between 400-1000 nm were required to attain the same level of Brix and firmness predictions. The best model predictions came from using all 4 sides of the fruit to train the models.

5. Discussion

This study demonstrates the practical potential of the combination of phenology models and hyperspectral imaging with ViT models for non-destructive assessment of apple maturity (Brix and firmness). By combining both systems, growers can plan harvests more effectively, using long-term forecasting over entire orchards with phenology models and real-time maturity assessment on individual fruit with hyperspectral imaging. Traditional maturity assessment methods are fundamental in establishing the optimal harvest window. However, assessments are destructive, labour-intensive and timeconsuming. In this thesis, we investigated the potential of phenology and hyperspectral imaging as predictive tools for harvest timing. Our findings suggest that phenology models and hyperspectral imaging offer a viable and efficient alternative to conventional methods.
In the first study, we evaluated the effectiveness of phenology models in predicting apple flowering time. We applied the PhenoFlex model to our extensive flowering data, collected across 85 years, on a range of different cultivars from East Malling, UK. The study showed that a common apple model at the species level was better at predicting flowering time for the twenty-six apple cultivars studied than using models trained on individual cultivars. The trained model could predict the flowering date within 5 days of harvest. Similar results can be found with models trained on groups of apple cultivars clustered by flowering time. The predictions in this approach depended on the internal consistency of each group, and results were within 5-6 days of the harvest window. Both approaches gave predictions comparable to previous flowering time prediction studies, with results ranging between 3-6 days (Darbyshire et al., 2016; Drepper et al., 2020; Luedeling et al., 2021).

The results from this study were slightly higher, since we are reporting the average RMSE for all twenty-six cultivars as opposed to individual cultivars. There is some risk in training models on a single cultivar; models trained on individual cultivars can yield polarising (ranging from precise to misleading) results. Our results show that generalised models — models trained on data either aggregated with a large number of cultivars or divided into smaller groups by flowering date are appropriate and more reliable for making flowering time predictions compared to single cultivar approaches.

Since flowering time can be predicted with reasonable accuracy, the harvest date can be forecasted using the average flowering date, a convention commonly used to estimate the harvest date of apples (Blanpied, 1982; Luton & Hamer, 1983). However, due to the use of the average flowering date in this process, harvest date predictions often overlook the variation in flowering time and its influence on fruit maturity. We recorded the flowering date and fruit maturity from five apple cultivars harvested over two seasons to determine how much flowering time influences the fruit maturity. The growth units were calculated from the flowering date to harvest date for each apple cluster. The results show that up to 20% of maturity variation is explained by flowering time variation, with this effect being more pronounced in early-flowering cultivars than late-flowering cultivars. This suggests that accounting for flowering time variability is crucial for improving the accuracy of harvest predictions, particularly for early-season cultivars. Therefore, phenology models that rely on modelling the climate data alone are insufficient for accurate harvest date predictions. Given the remaining variability in harvest timing predicted by phenology models, there is a need for methods that can assess fruit maturity more accurately, closer to the harvest date. Hyperspectral imaging was tested as a non-destructive method to evaluate fruit maturity. We imaged over 5000 apples over three seasons from both the United Kingdom and New Zealand and applied deep learning models to find patterns between the hyperspectral data and maturity data. Our results demonstrated that deep learning vision transformer models could reliably predict key maturity indicators — soluble sugar content and firmness — with prediction accuracies of 0.91°Brix and 0.79 kgF, respectively. Our results are consistent with previous studies predicting firmness with hyperspectral imaging (Çetin et al., 2022; Ekramirad et al., 2017), but our Brix model underperformed in comparison with previous models (Fan et al., 2020; Wang et al., 2022) suggesting that Brix does not require deep learning models for accurate Brix predictions. Despite less accurate results, our model can still predict well within reasonable errors. Overall, phenology models are valuable to ensure that seasonal weather conditions will satisfy the chilling and forcing requirements specific to each cultivar. Without sufficient environmental cues, the apples will not flower and therefore will not fruit. They can also approximate the harvest window for each cultivar. While phenology models provide useful estimates for the harvest window, they fail to capture the full variability induced by differences in flowering time. On the other hand, deep learning models trained on hyperspectral images can accurately determine the levels of Brix and firmness at the

individual fruit level. However, the deep learning model trained on hyperspectral images may not generalise well beyond the conditions of the current study. Therefore, integrating phenology models with hyperspectral imaging offers a promising alternative to traditional harvest prediction methods. By combining both systems, growers could plan harvests more effectively, using long-term forecasting over entire orchards with phenology models and real-time maturity assessment on individual fruit with hyperspectral imaging.

5.1. Applications

Phenology models, such as the PhenoFlex model, can be applied to forecast the harvest window. As temperature data is required as the model inputs, and growers are primarily interested in the coming season's harvest date, historical temperature data can be used to initially simulate temperatures to predict an estimated harvest date, then refine the predictions with actual temperatures experienced by the trees. According to the results from our studies, this will attain an estimated harvest window within ± 7 days of the actual harvest date. Hyperspectral imaging can be applied from the earliest point of the potential harvest day, tracking the development of Brix and firmness over time. By monitoring the progress of Brix and firmness, the optimal harvest maturity can be determined by observing the changes in Brix, which increases when ripe, and firmness, which decreases with maturity, for each fruit. Future applications may extend beyond apples and into other fruits.

5.2. Limitations and Future Research

Phenology models are dependent on temperature data. When simulated data is used, prediction accuracy can be compromised. This would impact regions where temperature is inconsistent or where there is limited coverage or lack of historical data to simulate with. The increasing unpredictability of the climate due to unusual weather events may introduce more frequent and extreme anomalies. However, previous phenology models will also be vunerable to unusual weather patterns. Training models with updated temperature data will help maintain model robustness. Apple varieties are constantly produced through breeding programs. While we present a large range of existing apple cultivars, our results may not generalise well to newer varieties. As a result, our findings may not fully translate to future varieties, limiting the long-term applicability of the results. The models should be continuously trained and recalibrated with the emergence of new cultivars.

As we captured apple images in a highly controlled environment, the feasibility of hyperspectral imaging will need to be tested in-field with the reduced wavebands identified in our study. The effects of variable lighting and occlusion by leaves or other fruit will affect the accuracy of model results. The distance of the apples from the camera will also affect model results. The apples captured for our study were aligned in a row, with small distances between the camera and each apple. In-field applications would typically scan the entire canopy height for fruit from approximately

1 m away, thus reducing the resolution of the pixels of each apple. Reduced pixel resolution may limit the accuracy and reliability of predictive models. Moreover, the distance of each apple from the camera will introduce inconsistencies, further influencing the model results. These factors will need to be investigated in the future to determine their impact on the model performance in outdoor settings. Further preprocessing of image data will be needed to standardise inputs and mitigate the variability.

6. References

Anderson, J. L., Richardson, E. A., & Kesner, C. D. (1985). Validation of chill unit and flower bud phenology models for'Montmorency'sour cherry. *I International Symposium on Computer Modelling in Fruit Research and Orchard Management 184*, 71–78. https://doi.org/https://doi.org/10.17660/actahortic.1986.184.7

Anderson, J. L., & Seeley, S. D. (1992). Modelling strategy in pomology: Development of the Utah models. *III International Symposium on Computer Modelling in Fruit Research and Orchard Management 313*, 297–306. https://doi.org/https://doi.org/10.17660/actahortic.1992.313.36

Behmann, J., Acebron, K., Emin, D., Bennertz, S., Matsubara, S., Thomas, S., Bohnenkamp, D., Kuska, M. T., Jussila, J., & Salo, H. (2018). Specim IQ: evaluation of a new, miniaturized handheld hyperspectral camera and its application for plant phenotyping and disease detection. *Sensors*, *18*(2), 441.

Blanpied, G. D. (1982). Observations of the Ripening and Harvest Indices at Commercial Harvest Dates of 'Delicious' Apple at its Extreme Northern Latitudes1. *HortScience*, *17*(5), 783–785.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.

Carsten, U., Luedeling, E., & Schiffers, K. (2022, August 17). *PhenoFlex*. https://cran.r-project.org/web/packages/chillR/vignettes/PhenoFlex.html

Çetin, N., Karaman, K., Kavuncuoğlu, E., Yıldırım, B., & Jahanbakhshi, A. (2022). Using hyperspectral imaging technology and machine learning algorithms for assessing internal quality parameters of apple fruits. *Chemometrics and Intelligent Laboratory Systems*, *230*, 104650.

Darbyshire, R., Pope, K., & Goodwin, I. (2016). An evaluation of the chill overlap model to predict flowering time in apple tree. *Scientia Horticulturae*, *198*, 142–149.

Drepper, B., Gobin, A., Remy, S., & Van Orshoven, J. (2020). Comparing apple and pear phenology and model performance: what seven decades of observations reveal. *Agronomy*, *10*(1), 73.

Ekramirad, N., Rady, A., Adedeji, A. A., & Alimardani, R. (2017). Application of hyperspectral imaging and acoustic emission techniques for apple quality prediction. *Transactions of the ASABE*, *60*(4), 1391–1401.

Erez, A., & Couvillon, G. A. (1987). Characterization of the influence of moderate temperatures on rest completion in peach. *Journal of the American Society for Horticultural Science (USA)*.

Fan, S., Wang, Q., Tian, X., Yang, G., Xia, Y., Li, J., & Huang, W. (2020). Non-destructive evaluation of soluble solids content of apples using a developed portable Vis/NIR device. *Biosystems Engineering*, *193*, 138–148.

Luedeling, E., & Fernandez, E. (2022). *chillR: Statistical Methods for Phenology Analysis in Temperate Fruit Trees*. https://CRAN.R-project.org/package=chillR

Luedeling, E., Schiffers, K., Fohrmann, T., & Urbach, C. (2021). PhenoFlex-an integrated model to predict spring phenology in temperate fruit trees. *Agricultural and Forest Meteorology*, *307*, 108491.

Luton, M. T., & Hamer, P. J. C. (1983). Predicting the optimum harvest dates for apples using temperature and full-bloom records. *Journal of Horticultural Science*, *58*(1), 37–44.

Meier, U., Graf, H., Hack, H., Heb, M., Kennel, W., Klose, R., Mappes, D., Seipp, D., StauB, R., & Streif, J. (1994). Phänologische Entwicklungsstadien des Kernobstes, des Steinobstes der Johannisbeere und der Erdbeere. *Nachrichtenblatt Des Deutschen Pflanzenschutzdienstes*, *46*, 141–153.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290.

Pope, K. S., Da Silva, D., Brown, P. H., & DeJong, T. M. (2014). A biologically based approach to modeling spring phenology in temperate deciduous trees. *Agricultural and Forest Meteorology*, *198*, 15–23.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., & Gustafson, L. (2024). Sam 2: Segment anything in images and videos. *ArXiv Preprint ArXiv:2408.00714*.

Richardson, E. A. (1975). Pheno-climatography of spring peach bud development. *HortScience*, *10*, 236–237.

Tang, H., Zhai, X., & Xu, X. (2024). Evaluating the performance of models predicting the flowering times of twenty-six apple cultivars in England. *European Journal of Agronomy*, *160*, 127319.

Wagner, T. L., Wu, H.-I., Sharpe, P. J. H., Schoolfield, R. M., & Coulson, R. N. (1984). Modeling insect development rates: a literature review and application of a biophysical model. *Annals of the Entomological Society of America*, *77*(2), 208–220.

Wang, F., Zhao, C., Yang, H., Jiang, H., Li, L., & Yang, G. (2022). Non-destructive and in-site estimation of apple quality and maturity by hyperspectral imaging. *Computers and Electronics in Agriculture*, *195*, 106843.

Xu, X. (1996). The effects of constant and fluctuating temperatures on the length of the incubation period of apple powdery mildew (Podosphaera leucotricha). *Plant Pathology*, *45*(5), 924–932.