European Journal of Agronomy 160 (2024) 127319

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

Evaluating the performance of models predicting the flowering times of twenty-six apple cultivars in England

Haidee Tang [a,b,*], Xiaojun Zhai [b], Xiangming Xu [a]

[a] NIAB, East Malling, Kent ME19 6BJ, UK
[b] School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK

ARTICLE INFO

Keywords:
Apple
Flowering time
Model
Parameterization
StepChill
PhenoFlex

ABSTRACT

The timing of the transition between endodormancy and ecodormancy remains uncertain. However, with advancements in phenology modelling, we can now fit models which allow for variable transitions between chilling and forcing models. Previous studies have primarily focused on single-cultivar parameterisation, and few have explored multi-cultivar comparative modelling. In this paper, we address this gap by evaluating three parameterisation approaches based on the recently developed PhenoFlex framework using a large flowering time dataset of twenty-six apple cultivars collected at the same location in England. The three parameterisation approaches were: cultivar-specific, group-specific with the groups derived using the K-means algorithm on mean bloom and variation of bloom dates, and a common model (for all twenty-six cultivars). The three PhenoFlex models fitted to each of three groups of cultivars based on their flowering time and the common model fitted to all cultivars achieved similar predictive performance, better than predictions using the average bloom date of each cultivar. The best approach to apply would depend on the amount of data present. The common model works best with large number of cultivars with small datasets (~10 years), the mean flowering date grouped works best with medium numbers of datasets (~20 years) and the cultivar-specific model should only be used when each cultivar has at least 30 years of data, however, it is more biased, so it is likely to predict bloom dates later than the observed bloom dates. Finally, the PhenoFlex model was shown to perform better than the StepChill model, where no overlapping is allowed between chilling and heat models. The result of this study indicates that the PhenoFlex model can be used to determine apple flowering time at the species level.

1. Introduction

Apple tree dormancy is regulated by temperature (Heide and Prestrud, 2005; Rea and Eccel, 2006). Low temperatures slow plant processes so that young meristems such as leaf and flower buds are protected against harsh environments. There are two types of dormancies which affect the flowering period: endodormancy and ecodormancy. Endodormancy is the state when the trees stop their growth in response to chilling conditions. Following this, ecodormancy sets in, signifying that chilling requirements have been met but the plants are awaiting warm temperatures to resume growth. Specifically, the chilling requirement refers to the duration and intensity of chilling temperatures needed to complete endodormancy and likewise, the heat requirement refers to the duration and intensity of warm temperatures needed to overcome ecodormancy. Both the chilling and heat requirements are species-specific (Perry, 1971). Flowering signals the end of dormancy.

Apples (*Malus domestica* (Suckow) Borkh.) have been cultivated for several thousand years, dating back to 1000 BCE (Janick, 2005) and while only about 100 cultivars are grown commercially, there are over 7500 unique cultivars (Shultz, 2003). Predicting the timing of apple flowering time is important for variety selection and crop management. Phenology models aim to explain the realities of biology with mathematical representations. Many phenology models have been developed to predict flowering time but there is no consensus to the best phenology model for apple (Carsten et al., 2022; Chuine, 2000; Darbyshire et al., 2016; Landsberg, 1974). A complete flowering time model comprises of two sub-models: the first models chilling accumulation and the second models heat accumulation. Chilling models estimate completion of endodormancy whereas forcing models represent the hypothetical relationship between warm temperatures and plant development.

Chilling models include the Chilling Hours, Utah and Dynamic models. The Chilling Hours model is a long-established chilling model

* Corresponding author at: NIAB, East Malling, Kent ME19 6BJ, UK.
E-mail addresses: Haidee.Tang@niab.com, ht21074@essex.ac.uk (H. Tang).

https://doi.org/10.1016/j.eja.2024.127319
Received 12 February 2024; Received in revised form 13 August 2024; Accepted 13 August 2024
Available online 27 August 2024

and takes the cumulative number of hours below a certain temperature threshold (Luedeling and Brown, 2011). However, it does not consider the negative effects of chill accumulation by high temperatures unlike the Utah model and the Dynamic model. The Utah model calculates chilling accumulation by giving each hour a positive, negative, or no chilling value (Richardson, 1974). As the Utah model results in negative chilling units in tropical and subtropical climates, this model is only appropriate for temperate climates (Melke, 2015). The Dynamic model is the only known empirical chilling model that utilises bud break experiments to formulate the model, which is why this paper will focus on this chilling model (Campoy et al., 2011; Erez and Couvillon, 1987; Hauagge and Cummins, 1991; Shaltout and Unrath, 1983). The Dynamic model calculates chilling accumulation in two steps. The first step accumulates a pseudo product called precursor to the dormancy-breaking factor (PDBF) which is created in chilling temperatures but destroyed in warm temperatures. Both the formation and destruction follow Arrhenius law, describing the effect of temperature on the rate of chemical reactions in an exponential relationship. The PDBF fluctuating in response to temperature allows the model to adapt to warmer regions, since longer periods of warm temperatures will not result in negative chilling. Additional chilling activates the second step, the irreversible conversion of the PDBF to a chill portion after a critical amount of chilling is reached. Chill portions cannot be destroyed by heat.

The most widely used forcing model is the Growing Degree Hour (GDH) model (Anderson et al., 1985). The GDH model is the hourly variation of the growing degree days model, which takes the average temperature of the day rather than the hour (Anderson et al., 1985; Réaumur, 1735). A GDH model sums the temperatures suitable for growth (temperatures above a threshold temperature) and each degree is weighted based on its proximity to the optimal growing temperature. The date when the GDH units reach the heat requirement is predicted as the flowering date.

The boundary between endodormancy and ecodormancy is unclear due to a lack of measurable physiological changes. Thus, recent phenology models were developed to integrate chilling and forcing models into a single framework for predictive purposes. To ensure a flexible modelling approach, several forms of transitions from chilling to forcing have been studied, including (1) treating chilling and heat accumulations completely sequentially and separately (chilling must be completed before heat accumulation starts), (2) assuming a complete overlap (chilling and heat accumulation is simultaneous), and (3) assuming a partial overlap between the chilling and heat accumulation phases (heat accumulation can start before chilling is completed but not before the initiation of chilling).

The Parallel model assumes that chilling and forcing units accumulate at the same time (Landsberg, 1974). However, we know that a chill requirement must be met before forcing units can accrue because any response to heat before chilling satisfaction may lead to flowering in unsuitable conditions (Campoy et al., 2011; Erez, 2000; Petri and Leite, 2003; Pope et al., 2014). In contrast to parallel accumulations of chilling and heat, the Sequential and Unified models both assume independent phases for chilling and heat accumulation (Ashcroft et al., 1977; Chuine, 2000). The difference between the Sequential and Unified models is that the Sequential model uses user-chosen chilling and forcing models, whereas the Unified model chooses the chilling and forcing models based on the data. The shape of the adjustable function within the Unified model can follow the shape for Triangular Chilling, Chilling Days, Sigmoidal and Growing Degree Days models and anything in between for chilling and forcing. These models are commonly used in older combination models. The parameters determine the shape of the model, so it does not prematurely constrain the data to fit a particular model (Chuine, 2000). The StepChill model is a simpler variation of the Unified model. It showed similar predictive performance in a recent comparative study in predicting bud break in major forest tree species compared to the full Unified model despite the use of fewer parameters (Asse et al., 2020; Luedeling et al., 2021).

Later models were developed with the concept of an overlap existing between the chilling and heat phases and, moreover, they consider a beneficial effect on heat accumulation through additional chilling after chilling requirements have been reached (Erez and Couvillon, 1987; Guerriero et al., 1985; Naor et al., 2003). The Chill Overlap model attempts to model an overlap between the chilling and heat models. This model is based on the idea that once endodormancy is satisfied, additional exposure to chilling temperatures will reduce the heat requirement (Cannell and Smith, 1983; Darbyshire et al., 2016; Pope et al., 2014). The PhenoFlex model was developed in 2021 and was used to predict apple and pear flowering time (Luedeling et al., 2021). This model integrates the Dynamic and GDH models as the chilling and forcing models. The difference is that while the Chill Overlap model requires a preset overlap value, the PhenoFlex model allows for no overlap, various degrees of overlap or complete overlap between the chilling and forcing models.

Previously, apple flowering time of different cultivars has been predicted with a range of different approaches. The recent studies have used parameterisation to minimise the root mean square error (RMSE) to get the highest level of predictive accuracy. A typical sequential model combination is the Dynamic and GDH model combination. This combination resulted in fairly good predictions in validation datasets for Boskoop (4.2 days), Cox's Orange (5.7 days), Golden Delicious (5.12 days) and Jonagold (4.57 days) apples in Belgium (Drepper et al., 2020) but poor predictions for Crispps Pink apples in Australia (RMSE of 14.7 days) (Darbyshire et al., 2016). A chill overlap model was developed for Cripps Pink apples grown in Australia with a RMSE value of 5.9 days (Darbyshire et al., 2016). A StepChill model resulted in a RMSE of 7.68 days for Boskoop apple trees in Germany, but a PhenoFlex model improved these results, reducing the RMSE to 3.82 days (Luedeling et al., 2021). These few studies on a limited number of apple cultivars indicated that the overlapping modelling framework for predicting apple flowering time may be better than a completely sequential model and that apple cultivars may differ significantly in the exact model formulation as well as in parameter values within the same model formulation.

A study in apricots (Andreini et al., 2014) shows that while applying flowering data from many cultivars to a crop does not result in accurate predictions, consolidating cultivars based on their flowering group can improve model accuracy. Apple cultivars are classified into flowering groups (based on similar flowering times), but it would be interesting to see if a naïve clustering method can be used to classify cultivars into groups that are used to generate parameters suitable for the PhenoFlex model.

Previous research has focused on parameterisation for individual cultivars separately, thus in the present study we explore comparative modelling of multiple cultivars based on the PhenoFlex modelling framework. Specifically, the present study focuses on estimation of PhenoFlex model parameters for (1) individual cultivars separately (cultivar-specific), (2) cultivar groups derived from K-means clustering of mean flowering dates and their variation across years (mean flowering time), and (3) a common model fitted to all cultivars (common model). We evaluated the three approaches for their accuracies in predicting flowering dates. The cultivar-specific approach follows conventional protocols when fitting a phenology model. The common model approach is to explore whether a large dataset from combining all different apple cultivars improves model performance. The K-means approach is an intermediate between the cultivar-specific and common model approaches and is used to determine if cultivars can be grouped together using a naïve clustering method based on pattern similarities in their mean flowering dates. We hypothesise that the cultivar-specific approach will prove to be the most accurate in predicting flowering dates, followed by K-means grouped and the common approach. Finally, we use the best of the three approaches and compare it to the StepChill model to assess if the PhenoFlex model can achieve comparative results.

2. Materials and methods

2.1. 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, we can limit the environmental variation within our dataset and directly determine the variation between apple cultivars. The specific years are indicated in Table 1 and depicted in Fig. 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 we assume flowering behaviour is determined by the scion.

2.2. Temperature data

Temperature records were obtained from the East Malling weather station (51.2876°N, 0.4486°E, 33 m 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 and 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, 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).

2.3. Model formulation

2.3.1. 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 (s_1) 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

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)	Number of years (training)	Number of years (testing)	Standard deviation 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	8.30
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	8.90
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

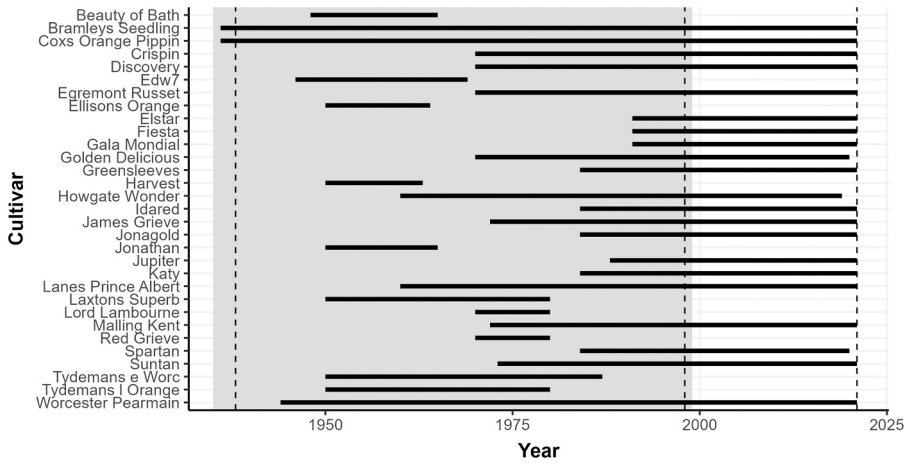

Fig. 1. Year range for each of the twenty-six cultivars, shown as black bars. The data highlighted in grey indicate simulated temperatures using latitude and maximum and minimum daily temperatures. The unhighlighted area represents temperature data from real recorded hourly temperatures, supplemented with simulated data where there are missing hourly values. The dashed lines indicate the specific years in which the three days and five hours of missing data occurs.

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 s_1. Large values of s_1 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 and 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 and 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).

2.3.2. StepChill model

The five-parameter StepChill model was fitted with the StepChill_Wrapper() function from the chillR package (Luedeling and Fernandez, 2022). The first parameter is the chilling threshold (T_c). Any temperature lower than T_c does not contribute to chilling function (CF); temperatures above T_c contributes 1 to the CF (Eq. 1).

$$CF(T) = \begin{cases} 0 & T \leq T_c \\ 1 & T > T_c \end{cases} \quad (1)$$

Chilling hours accumulate until it reaches the chilling requirement (C☆), the second parameter of this model. The heat (forcing) model then begins and is represented by Eq. 2. This function determines the amount of heat contributed by the temperature (x). In the simplified version of the Unified model, the StepChill function sets parameter *a* to 0. The third and fourth parameters, *b* and *c*, affect whether the sigmoidal curve starts positive and shifts to negative or *vice versa* and at what temperature this shift occurs. The fifth coefficient F☆ represents the heat requirement and budbreak occurs when the heat accumulated reaches F☆ (Chuine, 2000; Luedeling and Fernandez, 2022).

$$Heat(T) = \frac{1}{1 + e^{a(x-c)^2 + b(x-c)}} \quad (2)$$

2.4. 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 and 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 and 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 and Anderson, 2003), model efficiency (EFF) (Nash and 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 (Eq. 3).

$$AICc = 2k + nlog\left(\frac{RSS}{n}\right) + \frac{2k^2 + 2k}{n - k - 1} \quad (3)$$

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 (Eq. 4).

$$EFF = \frac{RSS}{\sum (t_i - \bar{t})^2} \quad (4)$$

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 our 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.

2.5. Comparative modelling

2.5.1. 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.

2.5.2. Comparing PhenoFlex and StepChill models using common parameters

In addition to the common PhenoFlex model, a common StepChill model was fitted to the pooled training data of all twenty-six cultivars. The common PhenoFlex and StepChill models were then assessed for the performance.

2.5.3. 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. Results

3.1. PhenoFlex models fitted to individual cultivars

Cultivar-specific parameter estimates for the PhenoFlex model (Table S1) 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 (Table S2 and Fig. 2A). 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 (Fig. 2B). 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.

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 (Table S2).

3.2. 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

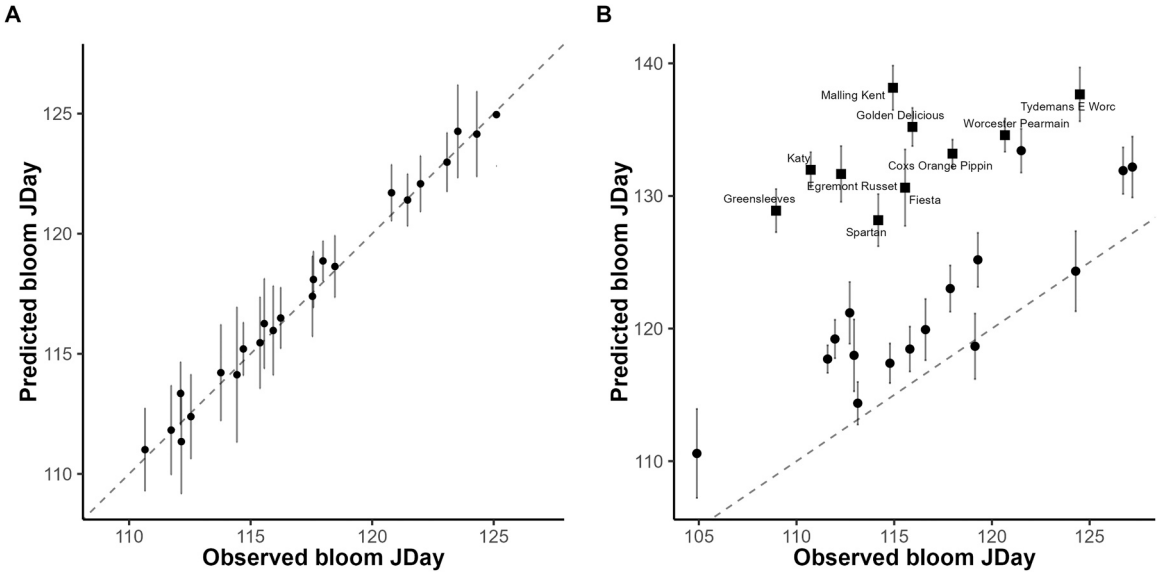

Fig. 2. A comparison of the mean observed and predicted bloom dates using cultivar-specific parameters on the PhenoFlex model on A) training data and B) test data. The dashed line represents the line of equality or the y = x relationship between the x and y coordinates. The square points in B represent cultivars which have RMSEs greater than 13 days.

scores (Fig. 3A), with three clusters identified using the silhouette method (Fig. 3B). 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 (Table 1). We applied K-means 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 (Table 1).

We fitted PhenoFlex models for three groups of cultivars identified via K-means clustering of the means and variations of flowering dates (Fig. 4A). 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 (Fig. 4). 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 (Table S2). 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.

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 (Fig. 6A). In the test dataset, the mean flowering clustered group outperformed both the cultivar-specific and common model approaches in all but Beauty of Bath (Fig. 6B).

3.3. 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 (Fig. 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 (Fig. 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

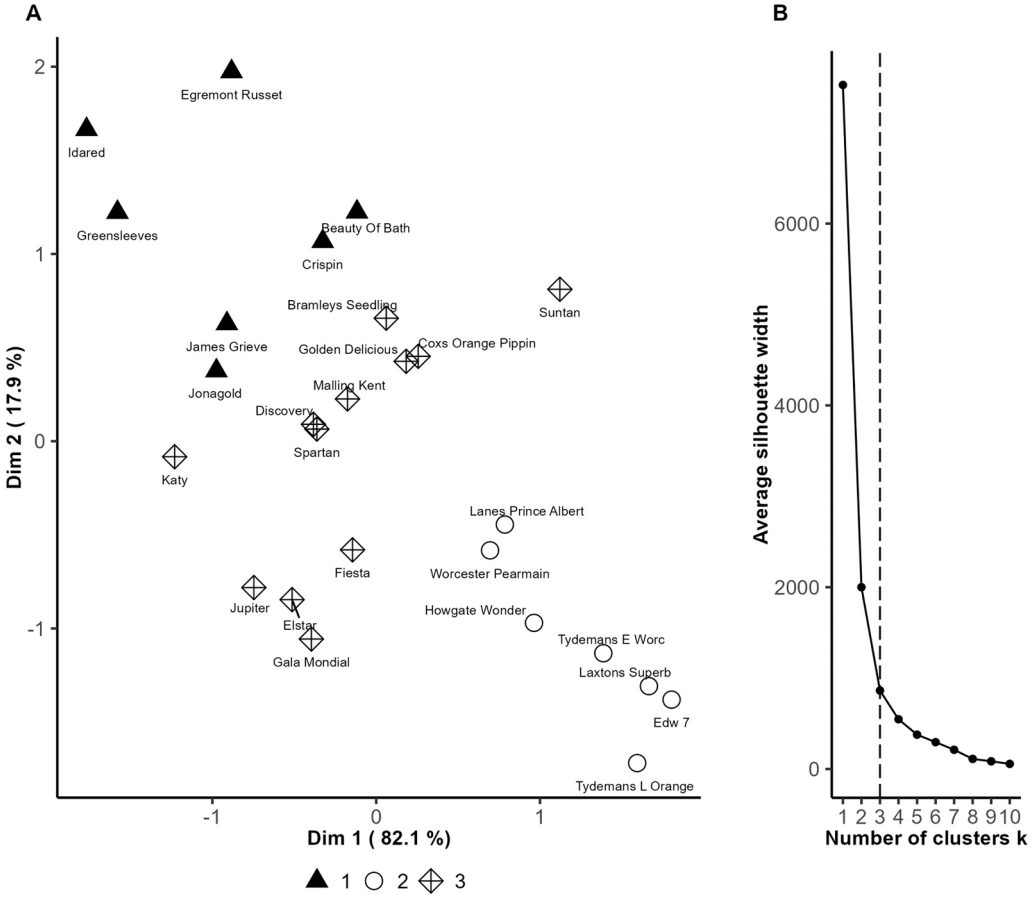

Fig. 3. A) K-means clustering presented in a PCA plot for mean flowering time and flowering variation across years for the twenty-six cultivars. B) Silhouette plot indicating three optimal clusters.

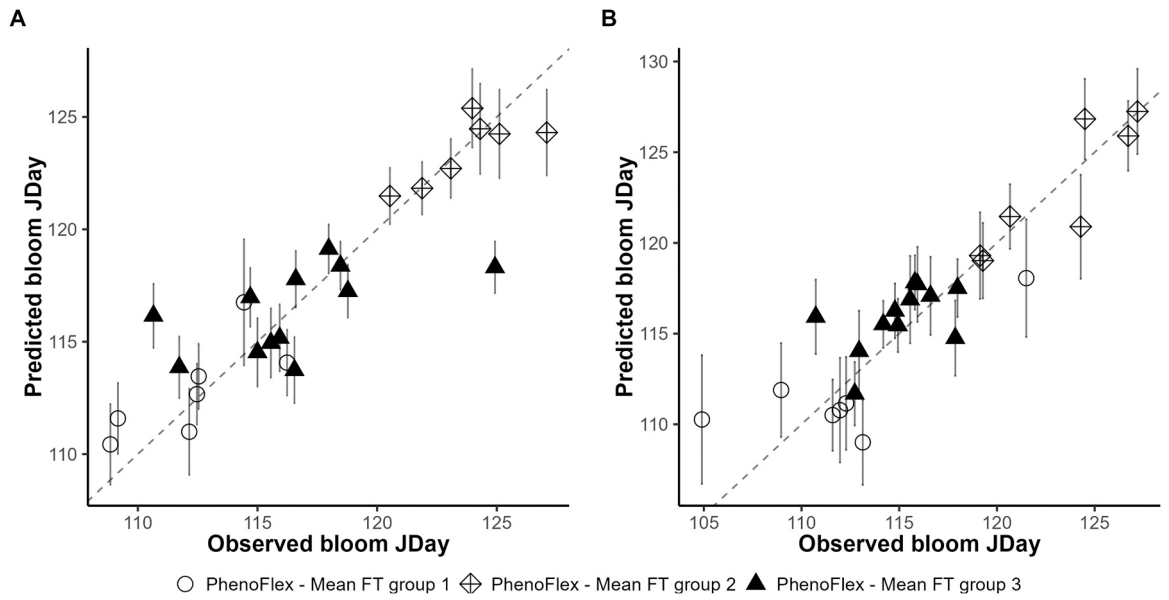

Fig. 4. A comparison of the observed and predicted bloom dates of cultivars grouped by K-means clustering on mean flowering and variation on the PhenoFlex model for A) training data and B) test data. The dashed line represents the line of equality or the y = x relationship between the x and y coordinates.

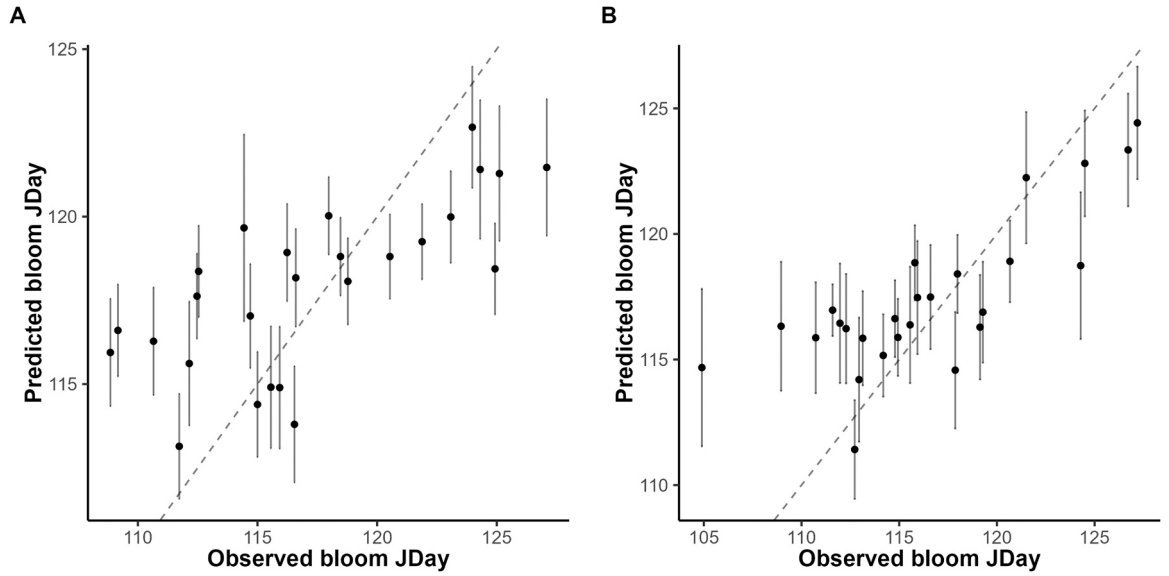

Fig. 5. A comparison of the observed and predicted bloom dates using common parameters on the PhenoFlex model on A) training data and B) test data. The dashed line represents the line of equality or the y = x relationship between the x and y coordinates.

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.

The cultivar-specific model was able to more accurately predict bloom dates better on the training data (Fig. 6A), but the common model predictions outperformed the cultivar-specific model predictions or were the same between model approach predictions in the test dataset (Fig. 6B). The common model approach was outperformed by the mean flowering cluster approach on Greensleeves and Laxton's Superb (Fig. 6B).

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 years of data, the mean bloom date cluster performs well since this method increases the number of data-points 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 (Table S3).

3.4. Comparison of the common PhenoFlex and StepChill models

Fitting the training the data with the StepChill model resulted in an average RMSE of 9.60 ± 0.34 days and an R^2 of 0.52 (Fig. 7A). The common StepChill model did not predict bloom dates well for the test data, resulting in a RMSE of 32.4 ± 0.46 days and R^2 of -30.86. All predicted flowering dates were later than the observed bloom date (Fig. 7B). Overall, the PhenoFlex model predicted flowering dates 82.59 % more accurately than the StepChill model. This agrees with the observed RPIQ values. The common StepChill model produced smaller RPIQ (0.29) on the test dataset compared to the common PhenoFlex model (Table S2).

AICc model selection was used to select between possible models which were trained on the same dataset and the Nash-Sutcliffe efficiency

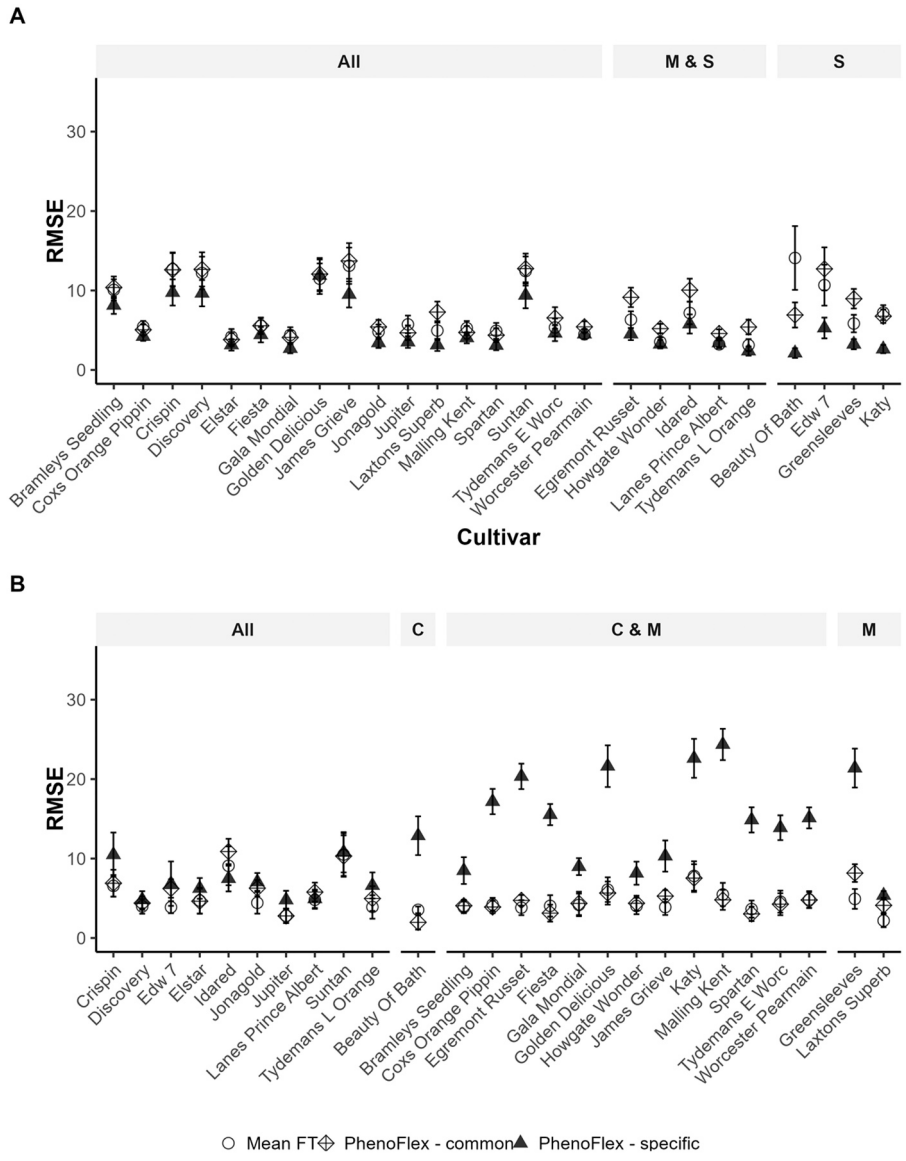

Fig. 6. Comparison of RMSE of the cultivar-specific, mean flowering grouped and common models for A) the training data and B) the test data. The graphs are separated by whether the specific (S), grouped by mean flowering date (M) or common (C) models perform better. When two of the approaches perform equally well, two letters are shows (M&S and C&M) or whether there is no difference between approaches (All).

of the model was used to quantify the predictive effectiveness of a model, with a good efficiency being close to 1. The common PhenoFlex model with twelve parameters was deemed to be better than the StepChill model with five parameters as PhenoFlex model AICc (3438.53 ± 5.29) was lower than StepChill model AICc (3611.21 ± 4.65), and the PhenoFlex model efficiency (0.27 ± 0.003) was higher than StepChill model efficiency (0.08 ± 0.004).

The common PhenoFlex model was more reliable as it was able to predict bloom dates for all the years, but the common StepChill model was unable to predict 2 datapoints from 1950 on the test dataset.

4. Discussion

In this study, we used a large collection of flowering dates of twenty-six apple cultivars at East Malling, Kent, England to fit PhenoFlex models to predict flowering time, and to compare the PhenoFlex and StepChill models that were fitted to the training data of all twenty-six cultivars.

Predictive performance of cultivar-specific approach is the worst of the three modelling approaches, with large RMSEs (13.8 ± 0.53 days). Most predictions were worse than taking the average bloom dates of each cultivar as the predicted bloom date. Flowering predictions for seven of twenty-six cultivars were better than taking the average flowering date. The groups identified by applying K-means on mean flowering dates and their variation were significantly better than using the cultivar-specific model. The model predictions were better than taking the average bloom dates for each cultivar. The common model performed as well as the second approach, with similar RMSEs and RPIQ.

We can speculate that the difference in model performance results from model overfitting on training data of smaller datasets, excessive model complexity and noisy training data, with more emphasis placed on the first two factors. Cultivar-specific models tended to do well when the dataset is large (approximately 30 years or more of training data per cultivar), the mean flowering date grouped model did well with at least 20 years of training data per cultivar and the common model does well with even less training data per cultivar. Small datasets can restrict the optimisation functions' performance, increasing the chance that the model parameters converge at a local minimum rather than the global minimum. In the present study, flowering data were only available for a

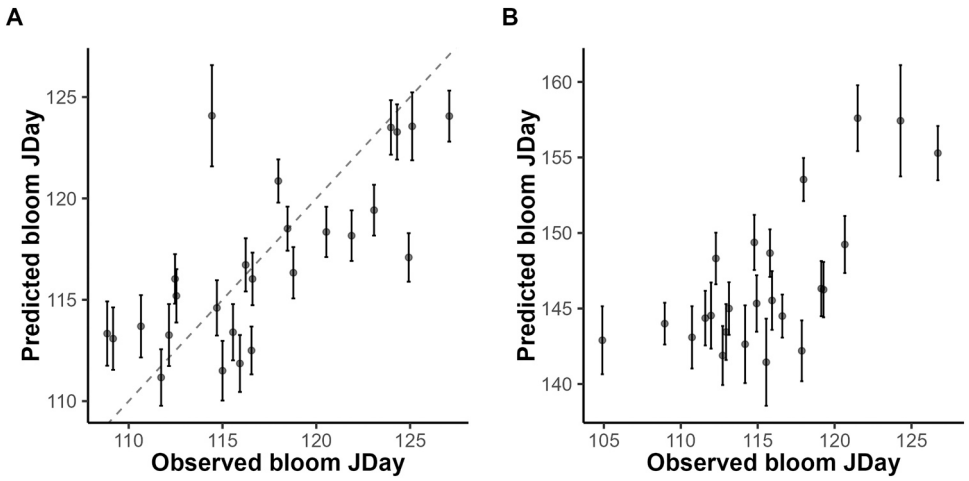

Fig. 7. A comparison of the observed and predicted bloom dates using common parameters on the StepChill model on A) training data and B) test data. The dashed line represents the line of equality or the y = x relationship between the x and y coordinates.

limited number of years for several cultivars and for some, flowering dates were only observed for a limited number of trees, thus limiting the genetic influence on flowering time. Another reason could be the large variability in flowering dates between years for some cultivars or variability in flowering dates between cultivars in the same year, which could make parameter convergence difficult in model optimisation. Problems associated with small datasets are further exacerbated when complex models are fitted.

The common PhenoFlex model performs as well as previous literature since its average RMSE (5.64 ± 0.30 days) is comparable to previous results predicting flowering dates for Boskoop (RMSE = 4.2 days), Cox's Orange (RMSE = 5.7 days), Golden Delicious (RMSE = 5.12 days), Jonagold apples (RMSE = 4.57 days) (Drepper et al., 2020) and Crispps Pink apples (RMSE = 14.7 days) (Darbyshire et al., 2016) using the Dynamic and GDH models sequentially. However, previous application of the PhenoFlex model on a single cultivar (Boskoop apples with RMSE = 3.82 days) (Luedeling et al., 2021) attained better predictive accuracy than our grouped and common models. As discussed above, the size of available data and number of individual trees are key factors affecting model predictive performance. A potential improvement for the common model would be to retain the CR and HR of the individual models but identify a common set of parameters for the other 10 parameters. In theory, this would be the best common model as previous studies agree that CR and HR are species specific (Perry, 1971).

Present modelling suggested that cultivars differ in the exact PhenoFlex model parameter estimates, although grouped by mean flowering date and variation of bloom dates and common PhenoFlex model led to more accurate predictions than the cultivar-specific model. Predictions given by both approaches were more or less balanced (namely varying around the observed dates). In contrast, cultivar-specific models predicted flowering later than the observed. Consistent large over-predictions (bias) may suggest model under-fitting. Small datasets may be the cause of poor predictive performance of cultivar-specific models. Under-fitting is more likely due to the fact that the data do not contain sufficient information on the generic feature due to large differences in the generic features between cultivars.

We grouped cultivars together based on their mean flowering date and variation of bloom dates. A previous study attempted a similar modelling approach with several phenology models to predict apricot flowering time, but they did not find good predictive results from modelling at the species level (common model approach) (Andreini et al., 2014). They found good results when split into smaller precocity groups for early, intermediate and late flowering, which is similar to what we observe in our current study. We speculate that our common model works well because our data originates from one site, thus reducing the generic variability and difference in response to different conditions.

Drepper et al. (2020) identified much larger chilling and heat requirements for Coxs Orange Pippin (CR = 79.6 and HR = 4430), Golden Delicious (CR = 59.84 and HR = 4980), Jonagold (CR = 60.66 and HR = 4980) and Boskoop (CR = 59.16 and HR = 4430). This agrees with the chilling (45) and heat (8500) requirements used by Darbyshire et al., (2016) on Boskoop apples. Using a combined chilling and heat model which can adjust for the overlap between the models appears to lower the chill and heat requirements for flowering by 2- to 3-fold. Luedeling et al., (2021) identified parameters very similar (CR = 25.4 ± 3.2 and HR = 348 ± 31) to the parameters identified in our common model (Table 2) as well as similar levels of overlap as shown by their s_1 value of 1 ± 22. The common model implements a large overlap between the chilling and heat models, as indicated by the smaller s_1 parameter (0.55 ± 0.08). The present model is more likely to have a greater level of overlap as the s_1 parameter estimate has a much smaller standard error than in Luedeling et al. (2021).

The common PhenoFlex model had much better predictive performance than the StepChill model. Although both models had similar goodness of fit for the training dataset, the StepChill model had much worse predictive performance, predicting flowering dates much later than observed. Moreover, for 1950, it failed to predict flowering time. The large bias in predictions indicated a model underfitting, suggesting that the model formulation does not capture much of the generic responses of apple flowering development in response to temperature.

In our study, we generated over 70 % of hourly temperature data using the recorded daily maximum and minimum temperatures following a sine curve for warming and logarithmic decay for cooling temperatures in our data prior to year 2000 as temperatures were only recorded as daily maximum and minimum values. In doing so, we must assume that temperatures more or less follow these trends for warming and cooling. Since our models are parameterised on our data, our models are only as good as our simulated data. In this current study, we are comparing the effectiveness of the three approaches in predicting flowering date, and since the hourly temperatures are the same for all approaches, this would unlikely impact the comparisons between models. Nevertheless, it would be informative to repeat this experiment with non-simulated hourly temperatures to understand the impact of using simulated hourly data on model fitting and performance.

5. Conclusion

The present research showed that the PhenoFlex model approach is an improved approach over the StepChill model to predict apple

Table 2
The best fit parameters of the PhenoFlex model for the K-means groups and Common (all cultivars). The standard errors are calculated on 99 bootstraps and the original parameters.

Model	y_c	z_c	s_1	T_u	E0	E1	A0	A1	T_f	T_c	T_b	slope
Common	25.36 ± 0.14	286.01 ± 9.34	0.55 ± 0.08	21.22 ± 0.29	3.82e+03 ± 4.33e−01	1.03e+04 ± 2.74e−01	8.51e+03 ± 6.70e+00	5.91e+13 ± 2.07e+08	−3.64 ± 1.73	27.27 ± 3.13	2.85 ± 0.06	46.87 ± 8.45
Group 1	22.6 ± 0.46	321.95 ± 13.09	0.55 ± 7.23	18.61 ± 0.55	3.82e+03 ± 1.22e+00	1.02e+04 ± 1.49e+00	1.025e+04 ± 2.73e+01	5.89e+13 ± 2.22e+08	0.82 ± 2.18	28.74 ± 3.52	3.53 ± 0.17	40.22 ± 8.78
Group 2	22.3 ± 0.23	355.14 ± 3.31	22.96 ± 8.04	20.24 ± 0.23	3.92e+03 ± 7.72e−01	1.04e+04 ± 5.08e−01	1.084e+04 ± 2.71e+01	5.94e+13 ± 2.08e+08	2.35 ± 1.69	27.15 ± 3.31	3.18 ± 0.15	52.1 ± 9.48
Group 3	28.44 ± 0.18	305.96 ± 4.19	0.76 ± 2.23	19.73 ± 0.15	3.82e+03 ± 1.57e−01	1.03e+04 ± 4.70e−01	9.996e+03 ± 1.25e+01	5.92e+13 ± 2.27e+08	−2.15 ± 1.99	29.7 ± 3.55	3.09 ± 0.15	31.01 ± 10.23

flowering development in relation to temperature, as concluded from higher EFF and lower AICc and RMSE values. Moreover, the common PhenoFlex model had no failed predictions. Contrary to our hypothesis, we find that the common PhenoFlex model and cultivars grouped by mean flowering time results in the best predictive accuracy and highest RPIQ compared to the cultivar-specific approach. Grouping cultivars by similar flowering dates can be used to adjust for low numbers of data in individual cultivars or grouping all data of a species from one region will yield better results than modelling each cultivar independently. Much of the poor model performance may be associated with small data sizes for fitting a complex model such as the PhenoFlex model with 12 parameters. In future, we should develop PhenoFlex models with a much-increased dataset through merging data from different locations as well as formulate specific CR and HR for each cultivar whilst having common values for the other 10 parameters.

Formatting of funding sources

The research is funded by the UK Biotechnology and Biological Sciences Research Council (grant number: BB/W510762/1)

CRediT authorship contribution statement

Xiangming Xu: Writing – review & editing, Supervision. **Haidee Tang:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Xiaojun Zhai:** Writing – review & editing, Supervision.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used AI image generator (deepai.org) in order to create the icons and images in the graphic abstract. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Haidee Tang reports financial support was provided by UK Biotechnology and Biological Sciences Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgements

Thank you to Mike Davies, Karen Jones and other NIAB staff who collected the flowering data at NIAB over the last eighty-five years. Without their work, this research would not be possible. I am also grateful to Jinya Su who offered guidance on this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.eja.2024.127319.

References

Anderson, J.L., Richardson, E.A., Kesner, C.D., 1985. Validation of chill unit and flower bud phenology models for 'Montmorency' sour cherry. Int. Symp. . Comput. Model. Fruit. Res. Orchard Manag. 184, 71–78. https://doi.org/10.17660/actahortic.1986.184.7.

Andreini, L., de Cortázar-Atauri, I.G., Chuine, I., Viti, R., Bartolini, S., Ruiz, D., Campoy, J.A., Legave, J.M., Audergon, J.-M., Bertuzzi, P., 2014. Understanding dormancy release in apricot flower buds (Prunus armeniaca L.) using several process-based phenological models. Agric. For. Meteorol. 184, 210–219. ⟨https://doi.org/10.1016/j.agrformet.2015.08.109⟩.

Ashcroft, G.L., Richardson, E.A., Seeley, S.D., 1977. A statistical method of determining chill unit and growing degree hour requirements for deciduous fruit trees. HortScience. https://doi.org/10.21273/hortsci.12.4.347.

Asse, D., Randin, C.F., Bonhomme, M., Delestrade, A., Chuine, I., 2020. Process-based models outcompete correlative models in projecting spring phenology of trees in a future warmer climate. Agric. Meteor. 285, 107931. https://doi.org/10.1016/j.agrformet.2020.107931.

Burnham, K.P., Anderson, D.R., 2003. Model selection and multimodel inference: A practical information-theoretic approach. Springer Science & Business Media, New York.

Campoy, J.A., Ruiz, D., Egea, J., 2011. Dormancy in temperate fruit trees in a global warming context: a review. Sci. Hortic. 130, 357–372. https://doi.org/10.1016/j.scienta.2011.07.011.

Cannell, M.G.R., Smith, R.I., 1983. Thermal time, chill days and prediction of budburst in Picea sitchensis. J. Appl. Ecol. 951–963. https://doi.org/10.2307/2403139.

Carsten, U., Luedeling, E., Schiffers, K., 2022. PhenoFlex vignette. https://cran.r-project.org/web/packages/chillR/vignettes/PhenoFlex.html (accessed 8.7.23).

Chuine, I., 2000. A unified model for budburst of trees. J. Theor. Biol. 207, 337–347. https://doi.org/10.1006/jtbi.2000.2178.

Darbyshire, R., Pope, K., Goodwin, I., 2016. An evaluation of the chill overlap model to predict flowering time in apple tree. Sci. Hortic. 198, 142–149. https://doi.org/10.1016/j.scienta.2015.11.032.

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, 73. https://doi.org/10.3390/agronomy10010073.

Erez, A., 2000. Bud dormancy; phenomenon, problems and solutions in the tropics and subtropics. in: Temperate Fruit Crops in Warm Climates. Springer, pp. 17–48. https://doi.org/10.1007/978-94-017-3215-4_2.

Erez, A., Couvillon, G.A., 1987. Characterization of the influence of moderate temperatures on rest completion in peach. J. Am. Soc. Hortic. Sci. (USA). https://doi.org/10.21273/jashs.112.4.677.

Guerriero, R., Indiogine, S.E.P., Scalabrelli, G., 1985. The effect of cyclic and constant temperatures in fulfilling the chilling requirement of two apricot cultivars. : VIII Int. Symp. . Apric. Cult. Decline 192, 41–48. https://doi.org/10.17660/actahortic.1986.192.9.

Hauagge, R., Cummins, J.N., 1991. Phenotypic variation of length of bud dormancy in apple cultivars and related Malus species. J. Am. Soc. Hortic. Sci. 116, 100–106. https://doi.org/10.21273/jashs.116.1.100.

Heide, O.M., Prestrud, A.K., 2005. Low temperature, but not photoperiod, controls growth cessation and dormancy induction and release in apple and pear. Tree Physiol. 25, 109–114. https://doi.org/10.1093/treephys/25.1.109.

Janick, J., 2005. The origins of fruits, fruit growing, and fruit breeding. Plant Breed. Rev. 25, 255–320.

Landsberg, J.J., 1974. Apple fruit bud development and growth; analysis and an empirical model. Ann. Bot. 38, 1013–1023. https://doi.org/10.1093/oxfordjournals.aob.a084891.

Luedeling, E., Brown, P.H., 2011. A global analysis of the comparability of winter chill models for fruit and nut trees. Int J. Biometeorol. 55, 411–421. https://doi.org/10.1007/s00484-010-0352-y.

Luedeling, E., Fernandez, E., 2022. chillR: Statistical Methods for Phenology Analysis in Temperate Fruit Trees.

Luedeling, E., Schiffers, K., Fohrmann, T., Urbach, C., 2021. PhenoFlex-an integrated model to predict spring phenology in temperate fruit trees. Agric. Meteor. 307, 108491 https://doi.org/10.1016/j.agrformet.2021.108491.

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. Nachr. Dtsch Pflanzenschutz 46, 141–153.

Melke, A., 2015. The physiology of chilling temperature requirements for dormancy release and bud-break in temperate fruit trees grown at mild winter tropical climate. J. Plant Stud. 4 https://doi.org/10.5539/jps.v4n2p110.

Naor, A., Flaishman, M., Stern, R., Moshe, A., Erez, A., 2003. Temperature effects on dormancy completion of vegetative buds in apple. J. Am. Soc. Hortic. Sci. 128, 636–641. https://doi.org/10.21273/jashs.128.5.0636.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — a discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Perry, T.O., 1971. Dormancy of Trees in Winter: Photoperiod is only one of the variables which interact to control leaf fall and other dormancy phenomena. Science 171 (1979), 29–36. https://doi.org/10.1126/science.171.3966.29.

Petri, J.L., Leite, G.B., 2003. Consequences of insufficient winter chilling on apple tree bud-break. : VII Int. Symp. . Temp. Zone Fruits Trop. Subtrop. 662, 53–60. https://doi.org/10.17660/actahortic.2004.662.4.

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. Agric. Meteor. 198 15–23. https://doi.org/10.1016/j.agrformet.2014.07.009.

Rea, R., Eccel, E., 2006. Phenological models for blooming of apple in a mountainous region. Int J. Biometeorol. 51, 1–16. https://doi.org/10.1007/s00484-006-0043-x.

Réaumur, R.A., 1735. Observations du thermomètre faites pendant l'année MDCCXXXV comparées a celles qui ont été faites sous la ligne a l'Isle-de-France, a Alger et en quelques-unes de nos Isles de l'Amérique. Mémoires De. l'Acad. émie R. Des. Sci. 1735, 545–576.

Richardson, E.A., 1974. A model for estimating the completion of rest for 'Redhaven' and 'Elberta' peach trees. HortScience 9, 331–332. https://doi.org/10.21273/hortsci.9.4.331.

Shaltout, A.D., Unrath, C.R., 1983. Rest completion prediction model for 'Starkrimson Delicious' apples. J. Am. Soc. Hortic. Sci. 108, 957–961. https://doi.org/10.21273/jashs.108.6.957.

Shultz, S., 2003. Apples. J. Agric. Food Inf. 5, 77–84. https://doi.org/10.1300/J108v05n04_08.