1   **Dynamic selection of environmental variables to improve the prediction of aphid phenology: a**
2   **machine learning approach.**

3   Paul Holloway[1,2*], Daniel Kudenko[1,3], and James R. Bell[4]

4   [1]Department of Computer Science, University of York, Deramore Lane, York, YO10 5GH, UK

5   [2]Department of Geography, University College Cork, Cork, Ireland

6   [3]Saint Petersburg National Research Academic University of the Russian Academy of Science, Russia

7   [4]Rothamsted Insect Survey, Biointeractions and Crop Protection, Rothamsted Research, AL5 2JQ, UK

8   paul.holloway@ucc.ie   orcid.org/0000-0002-9773-4531

9   daniel.kudenko@york.ac.uk   orcid.org/0000-0003-3359-3255

10 james.bell@rothamsted.ac.uk   orcid.org/0000-0003-1640-6120

11

12 **Abstract**

13 Insect pests now pose a greater threat to crop production given the recent emergence of insecticide
14 resistance, the removal of effective compounds from the market (e.g. neonicotinoids) and the
15 changing climate that promotes successful overwintering and earlier migration of pests. As
16 surveillance tools, predictive models are important to mitigate against pest outbreaks. Currently
17 they provide decision support on species emergence, distribution, and migration patterns and their
18 use effectively gives growers more time to take strategic crop interventions such as delayed sowing
19 or targeted insecticide use. Existing techniques may have met their optimal usefulness, particularly
20 in complex systems and changing climates. Machine learning (ML) arguably is an advance over
21 current capabilities because it has the potential to efficiently identify the most informative time-
22 windows whilst simultaneously improving species predictions. In doing so, ML is likely to advance
23 the length of any integrated pest management opportunity when growers can intervene. As an
24 example, we studied the migration of 51 species of aphids, which include some of the most
25 economically important pests worldwide. We used a combination of entropy and C5.0 boosted
26 decision trees to identify the most informative time windows to link meteorological variables to
27 aphid migration patterns across the UK. Decision trees significantly improved the accuracy of first
28 flight prediction by 20% compared to general additive models; further, meteorological variables that
29 were selected by entropy significantly improved the accuracy by a further 3-5% compared to expert
30 derived variables. Coarser (e.g. monthly) weather variables resulted in similar accuracies to finer
31 (e.g. daily) variables but the most accurate model included multiple temporal resolutions with
32 different period lengths. This combined resolution model alone highlights the ability of machine
33 learning to accurately predict complex relationships between species and their meteorological
34 drivers, largely beyond the experience of experts in the field. Finally, we identified the potential of
35 these models to predict long-term first flight patterns in which machine learning attained equally
36 high predictive ability as shorter-term forecasts. Whilst machine learning is a statistical advance, it is
37 not necessarily a panacea: experts will be needed to underpin results with a mechanistic
38 understanding, thus avoiding spurious relationships. The results of this study should provide
39 researchers with an automated methodology to derive and select the most appropriate
40 environmental variables when predicting ecological phenomena, while simultaneously improving the
41 accuracy of such models.

42

Introduction:

The role of meteorological variables in identifying the drivers of ecological phenomena is well established (Gough et al. 1994; Awmack *et al.* 1997; Zhou *et al.* 1997; Harrington *et al.* 2001; Bale *et al.* 2002; Lobo et al. 2002; Awmack *et al.* 2004; Cocu *et al.* 2005; Westgarth-Smith *et al.* 2007; Lima *et al.* 2008; Estay *et al.* 2009; Sheppard *et al.* 2016; Thackeray *et al.* 2016); however, the use of basic or incorrectly identified weather signals can lead to unreliable predictions, and subsequently inappropriately timed management strategies (van de Pol 2016). Selecting the 'best' meteorological variables that are indicative of the ecological phenomena under study is therefore critical. Despite this importance, in a recent meta-analysis, van de Pol *et al.* (2016) found that variables were often selected based on narrow hypotheses founded on previous studies (66%), with little thought given to what other meteorological variables affect the phenomena of interest (86% only used a single weather variable), over what time period (62% did not refine the time window), or how these variables should be represented (55% only considered the arithmetic mean). Furthermore, 28% gave no justification for the choice of meteorological variable chosen. While many studies obviously do give considerable thought to the choice of meteorological variables, this is not always explicitly reported by authors, and moreover the issues identified by van de Pol *et al.* (2016) are indicative of a potentially broader issue in predictive ecological modelling.

Aphids are a major pest of global importance, causing substantial damage to a wide variety of commercial crops in agriculture, forestry, and horticulture. Aphids cause feeding damage and transmit plant viruses to hosts. For example, the worldwide distributed peach-potato aphid *Myzus persicae* is widely polyphagous feeding on over 40 plant families (CABI 2017) and transmits over 100 plant viruses mediated by its highly adaptive and plastic life cycle (Bass *et al.* 2014). The need to better understand the emergence, distribution, and migration patterns of such serious pests remains an on-going challenge for growers. Ecological indicators (such as first flight day) are an important tool for understanding aphid phenology in terms of the forthcoming season, and by understanding the environmental drivers responsible for aphid migration, predictions can be made. This provides land managers, farmers (small and large scale), forestry officials, and governments with vital decision support on species emergence, distribution, and migration patterns that would reduce the prophylactic use of insecticides.

Aphids have a low developmental temperature threshold of approximately 4°C, and above that continue to develop at a rapid rate (estimated generation time of 120 degree days) assuming that the temperatures do not exceed the optimum development threshold of approximately 25°C (Harrington *et al.* 2007). Once adult, the temperature thresholds for initiating first flight are considered to range from 11°C to 16°C for different aphid species (Irwin *et al.* 2007). In a recent study, Bell *et al.* (2015) corroborated that harsher winters (measured using the North Atlantic Oscillation – NAO) resulted in later first flight dates, while an increase in accumulated degree days (ADD) above 16°C in April and May had a linear relationship with earlier first flight dates for common species in the UK. While the importance of the host plant condition (Awmack and Leather 2002) and the emigration from host plants due to critical population size (Dixon *et al.* 1968) are important determinants for first flight initiation, the spatial scale of the meteorological drivers used in predictive entomological and ecological studies arguably supersede these biotic interactions (Stoner and Joern 2004; Wisz et al. 2013; Miller and Holloway 2015).

Although the importance of temperature and NAO in understanding and predicting aphid flight dates cannot be understated, the derivation of these variables is subject to a number of conceptual and methodological uncertainties. In particular, the effect of the temporal scale used in variable selection and how to select the most informative parameter needs to be considered. The temporal

89 extent (i.e. the overall time-period) and temporal resolution (i.e. the frequency of data collation,
90 hourly, daily etc) utilised for generating environmental variables will have important consequences
91 for any inferences made from resulting models.

92 For both annual and perennial species, the use of long-term averages can mask extreme
93 meteorological events that are important in determining specific indicators such as emergence,
94 migration, or death. Studies have subsequently begun to explore the 'window' of time over which
95 environmental variables are generated. For example, Thackeray *et al.* (2016) investigated the
96 differences in the seasonal periods within which climate had the most positive and negative
97 correlations with phenology of a large number of terrestrial and marine UK species, that included
98 aphid first flights. Thackeray *et al.*'s (2016) climate sensitivity profile approach improved the
99 understanding of long-term changes in phenological responses that are a consequence of climatic
100 changes. Similarly, van de Pol *et al.* (2016) introduced climwin, an R package that uses the Akaike
101 Information Criterion (AIC) to compare models fit using different predictor windows (Bailey and van
102 de Pol 2016). Studies have therefore begun to adopt a more flexible methodology in defining the
103 temporal extent used to generate the environmental variables that describe the physiological
104 tolerances of insect species (e.g. Cocu *et al.* 2005; Thackeray *et al.* 2016) as well as a large number of
105 other organisms (e.g. Reside *et al.* 2010; Price *et al.* 2013; Gillings *et al.* 2015; Selwood *et al.* 2015;
106 Fancourt *et al.* 2015; Holloway *et al.* 2016); however, there remains a need for research to identify
107 ecologically meaningful environmental time windows.

108 Like many organisms, environmental conditions drive each aphid life stage and these accumulate
109 over a period to determine when first flight will occur (Harrington *et al.* 2007). However, there is a
110 trade-off between data-volume and information that would otherwise make models slow to run and
111 unwieldy. For example, daily data provides a highly detailed, but possibly noisy account of the
112 temperature preceding the first-flight, while monthly data provides a more smoothed
113 representation of the preceding conditions but loses nuances, such as warm weather spikes, that
114 may have profound implications for migration to begin. It is unknown whether coarsening the
115 resolution significantly reduces the accuracy of predictive models, or whether daily data will result in
116 an over-fitted model. In certain instances, a combined resolution model may be more informative
117 and capture the relevant drivers at differing scales.

118 Machine Learning (ML) is a tool, which could resolve variable selection when modelling ecological
119 indicators across a large number of species with potentially differing meteorological drivers.
120 Applications of ML in ecological modelling are diverse, and due to their ability to model complex,
121 nonlinear ecological relationships have exhibited greater explanatory and predictive ability than
122 conventional, parametric approaches (Fielding 1999; Olden *et al.* 2008). ML has been utilized across
123 an array of ecological disciplines to identify migration patterns of species (Guilford *et al.* 2009),
124 quantify species richness (Knudby *et al.* 2010), automatically classify bird calls (Acevedo *et al.* 2009),
125 and predict habitat suitability (Franklin 2009).

126 Here we will use a machine learning approach to inform and predict aphid migration patterns using a
127 suite of meteorological variables. We focus on three main research questions: 1) does the modelling
128 approach influence the accuracy of predictions? 2) does data representation and variable choice in
129 predictive models affect the accuracy of the first flight indicator? and 3) does temporal scale, in
130 terms of a) extent and b) resolution affect first flight predictions?

131 **Methodology**

132 *Data Collection*

133 In the UK, the Rothamsted Insect Survey (RIS) has a network of suction-traps that continuously
134 measure the areal density of flying aphids (currently 16 traps in 2017), and provides daily records
135 during the main aphid flying season (Harrington et al. 2007; Bell *et al.* 2015). Data from 17 suction
136 traps that supplied 10,715 first flight dates for 55 aphid species were obtained from the RIS, from
137 1980 to 2010. In order to remove any issues of sample size or bias, we removed four species that
138 had less than 30 observations in the series, resulting in a total of 51 species for analysis. We also
139 removed observations from January as we were unable to distinguish between genuine first flight
140 dates and those that were a construct of the new Julian calendar year (e.g. a first flight day of 1
141 suggests the species did not initiate flight on January 1, but was rather already in the air on
142 December 31). First flights were converted to a binary Julian day series. Due to the continuous
143 monitoring of the suction traps, any date before first flight was recorded has to be associated with
144 no flight at the location of the suction trap. Therefore, for each first flight (FF) observation, we
145 generated a spatially explicit no flight (NF) counterpart, which occurred within 7-105 days prior to
146 the FF day (figure based on expert opinion). This resulted in 21,228 binary observations (10,614 FF :
147 10,614 NF) for use as response data in the analysis.

148 Daily temperature (mean, maximum and minimum) and pressure data was obtained from the Dark
149 Sky API (https://darksky.net/poweredby/) from 1979 to 2010, and daily North Atlantic Oscillation
150 (NAO - the difference in atmospheric pressure at sea level between Iceland and the Azores) data was
151 obtained from the National Weather Service (http://www.cpc.ncep.noaa.gov/) for the same period.
152 Accumulated degree days (ADD) were generated at different temperature thresholds using both
153 mean and maximum temperature ranging from 11°C to 16°C and measures of winter harshness were
154 calculated using NAO, pressure, and mean, maximum, and minimum winter temperatures. We used
155 a variety of dynamic temporal extents to calculate both ADD and winter variables. For ADD, we
156 calculated the temporal extent immediately preceding a FF or NF observation, including 7-, 14-, 21-,
157 28-, 60-, 90-, 120-, 180-, and 364-days. For example, for a FF observation recorded on May 28, a 7-
158 day extent would calculate ADD on the temperatures recorded from May 20 to May 27, while a 14-
159 day extent would calculate ADD on the temperatures recorded from May 13 to May 27. Similarly,
160 we calculated winter harshness across a number of dynamic temporal extents, including 6-2, 6-3, 6-
161 4, 6-5, 5-2, 5-3, and 5-4 months prior. For example, for the FF observation recorded on May 28, a 6-2
162 month dynamic temporal extent would calculate the winter variable (e.g. NAO) from November 27
163 of the previous year to March 27. The use of this methodology allows for dynamism in selecting the
164 time windows over which the variables are derived.

165 To explore the effect of the resolution (granularity) on results and subsequent predictions, we used
166 the daily data to calculate temperature data at three different resolutions: daily, weekly and
167 monthly. We then implemented these new variables separately in the machine learning
168 methodology. We also employed a mixed resolution model, which consisted of daily observations for
169 a two-month extent immediately preceding FF or NF, weekly data back until six months, and then
170 monthly data for the remainder of the year. Baseline variables were defined as those deemed the
171 most accurate and informative by Bell *et al.* (2015). For FF, these were ADD16 across an extent of 60
172 days and NAO across an extent of 6-2 months prior. It should be noted that Bell *et al.* (2015) used
173 these variables to predict Julian day of FF and not a binomial delineation of FF or NF; however, the
174 importance of these variables in determining FF should allow for comparison.

175 **Data Analysis – Machine Learning**

176 *Variable Selection: Entropy*

Variable selection is an important way to improve the performance of ML techniques. The goal is to automatically identify the most informative variables in terms of predicting the response variable. The most informative variables are then used as predictor variables in the ML process and subsequently in the final model, while the remaining variables are ignored. The entropy measure is a popular and efficient way to select variables by measuring how well (or badly) a predictor variable distributes the training data into partitions with respect to the response variable (FF or NF) values. High entropy means that the resulting partitions tend to be 'impure' (i.e. have a uniform distribution of the training examples with respect to the response variable values). For continuous predictor variables, entropy can be used to discretise the values into subintervals to maximize the purity of the resulting partitions of training examples (i.e. minimize the entropy values). Entropy for a binary classification with classes *a* and *b* (corresponding to the target variable values) is defined as:

$$Entropy(S) = -p(a) * \log(p(a)) - p(b) * \log(p(b))$$ Equation 1.

S denotes a partition of the training examples, and *p(x)* is the proportion of training examples of class x in partition *S*. Entropy is maximal when *p(a)* = *p(b)* = 0.5, and minimal when *p(a)* = 0 or *p(b)* = 0. The information gain of a variable is the decrease in entropy caused by splitting the training data according to its values. We implemented entropy using the FSelector package (Romanski and Kotthoff 2016) in R 3.3.1 (R Core Team 2016).

Model Selection: Decision Trees and General Additive Models

We used the C5.0 decision tree algorithm to test the predictive performance of the meteorological drivers of aphid flight. The C5.0 algorithm is considered the industry standard for producing decision trees due to its ability to perform comparably well on a variety of problems (Lantz 2013). C5.0 creates a branched tree, that identifies the value of a predictor variable that yields the largest information gain for the splitting the response variable (in this case FF or NF). The generation of the tree begins by creating a root node, chooses a variable to test at the current node, and recursively creates child nodes for each of the corresponding variable values. The training set is split accordingly, and the process continues until there is no further significant information gain. The advantage of C5.0 over previous implementations is that it builds smaller and more efficient trees. Another major advantage of decision tree learning is the readability of the output (i.e. a prediction hypothesis in form of a decision tree). Here we used the C50 package (Kuhn *et al.* 2015) in R 3.3.1 (R Core Team 2016) to create C5.0 boosted decision trees to identify meteorological drivers of aphid flight.

To illustrate the ability of decision tree learning to predict first flight, we compared the decision tree models with general additive models (gams), a commonly applied statistical model that has been used to explore similar ecological questions. Gams advance from general linear models by assuming that the functions are additive and the components are smoothed (Guisan *et al*. 2002). We define the expected value of the response value E(Y) as:

$$g\big(E(Y)\big) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n) + \varepsilon$$ Equation 2.

where g describes the link function, β_0 is the intercept term, f describes the smoothing function used on predictor variable x, and finally we assume that the error term, ε is constant across observations. We generated gams from the same data as outlined above, meaning we used a binomial link function to specify the distribution of the response variable and we specified the smoothing of the environmental variables using thin plate regression splines. Again, we undertook analysis in R 3.3.1 (R Core Team 2016) in the mgcv package (Wood 2011), with further details outlined in Supplementary Information 1.

221 *Model Evaluation*

222 Six accuracy metrics were utilised to evaluate the predictions. The area under the curve (AUC) is
223 calculated by summing the area under the receiver operating curve (ROC) plot, a graph of the false-
224 positive error rate on the x-axis plotted against the true positive rate on the y-axis. Values range
225 from 0 to 1, with a perfect classification recorded as 1, while 0.5 suggests a classifier that is no better
226 than random. Sensitivity measures the proportion of correctly predicted first flights and specificity
227 measures the proportion of correctly predicted no flights. Proportion correctly classified (PCC)
228 measures the total number of correctly predicted responses. The true skill statistic (TSS) measures
229 the combined sum of sensitivity and specificity. Kappa is a metric of categorical agreement that
230 incorporates the differences between the observed agreement and chance agreement, with a value
231 of 1 suggesting complete agreement. See Franklin (2009) for a discussion of these metrics in an
232 ecological modelling context. We evaluated our models using a 10-fold cross-validation technique.

233 **Results**

234 *Modelling Approach*

235 Implementation of a machine learning methodology resulted in significantly higher recorded first
236 flight (FF) accuracies in five of the six evaluation metrics considered when compared with a gam
237 implementation (measured to an $\alpha <= 0.01$ using paired sample t-tests - Figure 1). The differences in
238 mean accuracy ranged from an increase of 0.05 for AUC to 0.21 for sensitivity, with recorded
239 specificity 0.10 higher for the gam methodology. The higher sensitivity and lower specificity scores
240 suggest that the machine learning methodologies are generating a smaller number of false positives,
241 yet the gam methodology is predicting a larger number of false negatives and this is severely and
242 substantially affecting the predictive models and in turn their accuracy. This is also seen in the
243 evaluation metrics that take into account the correct prediction of both FF and NF, with decision
244 trees recording higher PCC, TSS, and Kappa scores than gam implementations.

245 *Data Representation (Variable Selection)*

246 The use of entropy identified the variation in real-world variables that appear to best delineate first
247 flight observations from no flight observations. Table 1 identifies the information gain across the
248 different accumulated degree days (ADD) and winter variables for *M. persicae,* an aphid of global
249 importance to agriculture*.* It can be seen that the information gain varies substantially across the
250 different temperature thresholds and temporal extents. The ADD values calculated from the
251 maximum temperature resulted in a higher information gain (or a purer split), than the use of the
252 mean temperature (measured to an $\alpha <= 0.01$ using paired sample t-tests). Furthermore, the use of
253 entropy to select one ADD and one winter metric resulted in an increase in accuracy for over 80% of
254 species (Table 2). When the proportion of correctly classified (PCC) observations were compared for
255 a decision tree implementation using entropy selected variables against baseline selected variables,
256 we observed an increase in the average accuracy of 0.027 (or ~3%). Again, this increase was
257 significant at an $\alpha <= 0.01$ when these values were compared using a paired sample *t*-test. The
258 accuracy of entropy selected variables were consistently more accurate than decision trees fit on
259 baseline variables across the different accuracy metrics including specificity.

260 When the frequency distribution of environmental variables with the highest information gain for
261 each species were analysed, we saw differences in both the thresholds (e.g. ADD value) and
262 temporal extents (e.g. number of days) reported, with certain patterns emerging when species were
263 analysed based on phenology (Figure 2). In general, the lower ADD thresholds resulted in higher
264 information gain (Figure 2a), with 20 of the 51 species best predicted by a threshold of 11°C. A

dynamic temporal window of 60 days immediately prior to FF was the most informative temporal extent for both host alternating and non-host alternating species (Figure 2b). The average maximum temperature reported higher information gains across the winter metrics, with neither NAO nor pressure resulting in the highest gain for any species (Figure 2c). Shorter dynamic temporal winter metrics were most informative, with longer periods of time prior to an observation indicative of a larger gain (Figure 2d).

Machine Learning and Entropy

Introducing temperature variables calculated at various temporal resolutions further increased the accuracy when compared to baseline and entropy models (Figure 3). When models fit with four different temperature resolutions were compared, we observed that the use of daily data (Figure 3a) did not result in an over-fitting of the models, reporting equally high accuracies when compared with other implementations (Figure 3e), although it did record a number of lower outliers that most likely resulted from noisy data. Similarly, the use of monthly data (Figure 3c) did not result in a smoothing of the models, reporting a higher number of more accurate models than the use of daily data (Figure 3e). The use of weekly data (Figure 3b) resulted in a higher frequency of PCC scores for non-host alternating species compared to other temporal resolutions (Figure 3e), while for host-alternating species the temperature data consisting of various resolutions (Figure 3d) recorded the highest frequency of PCC scores (Figure 3e). This model incorporated daily data for the two months dynamically preceding the observations, weekly data back until six months, and then monthly data for the remainder of the year, suggesting that the use of different scales improves the predictive ability of decision trees fit for host-alternating species of aphids.

Figure 4 illustrates the mixed temporal resolution (Figure 3d) decision tree for *M. persicae* which recorded the highest PCC score across the different models. Here we can see the benefit of incorporating variables at multiple resolutions. The root node splits the response data on the mean temperature for the twelfth month prior to the observation, with a temperature value of 9.4°C. Alternatively we see nodes (4, 5, 16, and 17) split the response data on recent daily temperature values. This identifies the hierarchical nature of the first flight phenomena, with both shorter term (less than a week) and longer-term (a year prior) temperature variables constantly deemed important in determining differentiation among flight or no flight. Many of the splits in the decision tree identify temperatures at these different temporal resolutions for which no-flight was recorded. Node 3 identifies that if the maximum temperature for the eleventh month prior to an observation was <=20.2°C, then no flight was recorded (with a pure node of 145 observations). This suggests that conditions during the previous aphid flight season restrict whether aphids will be recorded there in the next season. Similarly, node 8 illustrates that if the minimum temperature for 32 days before the observation was <=6.7°C, then no flight occurred (with a pure node of 65 observations).

Discussion

There has been a recent resurgence in the importance of selecting appropriate environmental variables when attempting to explain or predict ecological phenomena (van de Pol *et al.* 2016). In addressing this, the aims of our study were twofold. Firstly, we identified the power of machine learning (specifically decision tree learning) to address complex, hierarchical ecological questions, and illustrated how this method can be used to attain highly accurate models and identify previously unknown features of ecological importance. Secondly, we utilised this methodology to explore the ecological indicators for UK aphids, and improved upon existing prediction techniques. The results of this study should provide researchers with an automated methodology to derive and select the most

309   appropriate environmental variables when predicting ecological phenomena, while simultaneously
310   improving the accuracy of such models.

311   We identified from expert-opinion and well-established aphid literature (Harrington *et al.* 2007; Bell
312   *et al.* 2015) that a measure of spring development (ADD16 in the previous 60 days) and a measure of
313   winter harshness (mean NAO from 6-2 months' prior) are strong predictors of aphid first flight.
314   While we acknowledge that there are most likely other meteorological variables influencing these
315   complex processes, our aim was to utilise machine learning to refine the temporal scales associated
316   with these predictor variables. Here we used entropy to identify the most informative variables
317   (ADD thresholds and winter variables) and subsequent temporal extents (spring and winter) for 51
318   UK aphid species. Decision tree models fit on entropy derived variables resulted in significantly
319   higher accuracies compared to models fit on baseline variables (Table 2). NAO and pressure provided
320   little input into our predictive models, and consistently provided less information gain than
321   measures of winter temperature. While NAO incorporates various aspects of daily weather (e.g.
322   wind, precipitation, temperature), as well as being important for aphid population dynamics (e.g.
323   Westgarth-Smith *et al.* 2007), it does not take into account spatial variation. NAO provides one daily
324   value for the entire country and subsequently has been used to model ecological indicators at
325   coarser spatiotemporal resolutions than we investigated here. The use of temperature as a winter
326   metric provided both spatial and temporal variation among observations of aphid first flight,
327   resulting in more accurate predictions compared to the NAO (Table 2).

328   With approximately 4400 known species of aphid (Harrington *et al.* 2007), the temporal scales used
329   to generate variables to predict first flight are likely to be highly species or even clone specific.
330   Differences among aphid life cycles between species and clones (i.e. genotypes) are likely to
331   influence such decisions. Similarly, different species may respond to different thresholds in weather
332   patterns. The use of 'events' (e.g. heat wave, drought) or 'episodes' (e.g. degree day calculations)
333   have been widely used in entomological (and other ecological) research, and have been found to
334   improve interpretations of ecological phenomena (Bateman *et al.* 2012; Bell *et al.* 2015; Selwood *et*
335   *al.* 2015). Despite this, these methods of variable derivation are still subject to the scale
336   uncertainties associated with the temporal extent used.

337   The use of shorter temporal extents when generating ADD increased the accuracy for certain
338   species, in particular host-alternating species (Table 2; Figure 2). For example, *Hyalopterus pruni*
339   produces wings in the summer and migrates from *Prunus* trees to grasses. Our results suggest that
340   shorter-term changes in increasing temperature are better predictors than longer-term extents.
341   Similarly, for non-host alternating tree aphids *Myzocallis castanicola, Betulaphis quadrituberculata,*
342   and *Elatobium abietinum,* and the cereal aphid *Sitobion avenae* that respond to overcrowding or
343   senescing of host plants, a shorter temporal extent of a month coupled with higher temperature
344   thresholds were better predictors, possibly representing spikes in temperature which could cause
345   population booms or increased stress to plants (Dixon and Glen 1971; Watt and Dixon 1981). From
346   these results, we may infer that host-alternating species respond by taking first flight based on
347   short-term changes in meteorological variables, while non-host alternating species respond by
348   taking first flight once the population on the host-plant exceeds a certain number, which could be
349   represented by a meteorological variable indicative of egg development or generation time. The
350   variation in the ADD threshold that best predicted first flight was not unanticipated due to the
351   phenology of the 51 species. However, we did not expect initiation of first flight to be predicted at
352   11°C for 40% of the species (Table 2). The use of data-driven variable selection improves the
353   accuracy of these final predictive models, and illustrates the potential for such methodologies to be
354   utilised widely when selecting environmental variables. The results from this research should help

355　foster discussion on variable derivation in entomological research, but are also widely applicable to
356　any predictive (or explanatory) ecological research that uses ecological indicators.

357　An advantage of decision tree models is that they have the ability to handle more variables than
358　commonly applied parametric approaches, in part due to the fact that they are not as sensitive to
359　issues relating to multi-collinearity. Our results suggest that suitable information pertaining to the
360　prediction of aphid first flight can be ascertained from both voluminous daily data and smoothed
361　monthly data (Figure 3). This has important practical consequences for implementations of these
362　predictive models; projecting species-environment relationships into future space and time at a
363　monthly resolution would reduce the resources and cost of such an endeavour compared to using
364　daily data. Moreover, as ecological phenomena are often influenced by drivers operating across
365　multiple temporal scales, the use of variables generated at different temporal resolutions allowed
366　for both longer- term conditions that enforce an overarching influence and shorter-term variations
367　that describe finer-scale patterns to be included in the model (Figure 4). It should be noted that the
368　use of a large number of meteorological variables within such models could result in the
369　identification of an incorrect spuriously correlated variable that has nothing to do with aphid
370　lifecycles. However, the high accuracy obtained from these models in predicting aphid first flight and
371　the use of 10-fold cross-validation to control for such concerns suggests that our models do well in
372　identifying the meteorological drivers of this phenomena. One caveat to the study is that ML
373　methods require a rich data set for the training of the algorithm and thus this technique may be
374　limited to longer-term studies.

375　Finally, to illustrate the applicability of decision trees to mitigate against long-term future pest
376　outbreaks by strategic crop interventions, we implemented our methodology with filters (Figure 5a).
377　In every instance our objective was the same, to predict FF or NF for day x; however, we applied
378　filters (ranging from 28 days to 364 days) to the range of environmental days used in the generation
379　of predictor variables. This resulted in a new set of predictor variables to use when estimating FF or
380　NF at day x. The accuracy of our models decreased as the size of the filter increased towards a year;
381　however, this difference was minimal (0.0286 difference in AUC - Figure 5b). The decision tree
382　methodology discriminates between FF and NF for a variety of meteorological variables across a year
383　period, choosing among days, weeks, or months when thresholds are representative of FF or NF.
384　Therefore, the differences in FF predictions when fit on filtered environmental variables are
385　negligible when a dynamic time window is applied. Subsequently, the high accuracies obtained
386　should allow such methods to generate long-term predictions and mitigate against further crop
387　losses.

388　**Acknowledgements**

394　**References**

395　Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J., Aide, T.M., 2009.
396　Automated classification of bird and amphibian calls using machine learning: A comparison of
397　methods. *Ecological Informatics 4*, 206-214.

Awmack, C.S., Leather, S.R. 2002. Host plant quality and fecundity in herbivorous insects. *Annual Review of Entomology* 47, 817-844.

Awmack, C.S., Woodcock, C., Harrington, R. 1997. Climate change may increase vulnerability of aphids to natural enemies. *Ecological Entomology* 22, 366-368.

Awmack, C.S., Harrington, R., Lindroth, R.L. 2004. Aphid individual performance may not predict population responses to elevated CO2 or O3. *Global Change Biology* 10, 1414-1423.

Bailey, L.D., van de Pol, M. 2016. climwin: An R Toolbox for Climate Window Analysis. *PLoS ONE* 11, e0167980.

Bale, J.S., Masters, G.J., Hodkinson, I.D., Awmack, C., Bezemer, T.M., Brown, V.K., Butterfield, J., Buse, A., Coulson, J.C., Farrar, J., Good, J.E., 2002. Herbivory in global climate change research: direct effects of rising temperature on insect herbivores. *Global Change Biology*, *8*, 1-16.

Bass, C., Puinean, A.M., Zimmer, C.T., Denholm, I., Field, L.M., Foster, S.P., Gutbrod, O., Nauen, R., Slater, R., Williamson, M.S., 2014. The evolution of insecticide resistance in the peach potato aphid, Myzus persicae. *Insect biochemistry and molecular biology 51*, 41-51.

Bateman, B.L., VanDerWal, J., Johnson, C.N. 2012. Nice weather for bettongs: using weather events, not climate means, in species distribution models. *Ecography* 35, 306-314.

Bell, J.R., Alderson, L., Izera, D., Kruger, T., Parker, S., Pickup, J., Shortall, C.R., Taylor, M.S., Verrier, P., Harrington, R., 2015. Long-term phenological trends, species accumulation rates, aphid traits and climate: five decades of change in migrating aphids. *Journal of Animal Ecology 84*, 21-34.

CABI. 2017. *Myzus persicae* (green peach aphid). – Available at http://www.cabi.org/isc/datasheet/35642.

Cocu, N., Harrington, R., Rounsevell, M.D.A., Worner, S.P., Hulle, M., 2005. Geographical location, climate and land use influences on the phenology and numbers of the aphid, Myzus persicae, in Europe. *Journal of Biogeography 32*, 615-632.

Dixon, A.F.G., Burns, M.D., Wangboonkong, S. 1968. Migration in aphids: response to current adversity. *Nature* 220, 1337-1338.

Dixon, A.F.G., Glen, D.M. 1971. Morph determination in the bird-cherry oat aphid, Rhopalosiphum padi (L). *Annals of Applied Biology* 68, 11-21.

Estay, S.A., Lima, M., Harrington, R. 2009. Climate mediated exogenous forcing and synchrony in populations of the oak aphid in the UK. *Oikos* 118, 175-182.

Fancourt, B.A., Bateman, B.L., VanDerWal, J., Nicol, S.C., Hawkins, C.E., Jones, M.E., Johnson, C.N. 2015. Testing the role of climate change in species decline: is the eastern quoll a victim of a change in the weather?. *PloS one 10*, p.e0129420.

Fielding, A. 1999. *Machine learning methods for ecological applications*. Springer Science & Business Media.

Franklin, J. 2009. *Mapping species distributions: spatial inference and prediction*. Cambridge: Cambridge University Press.

Gillings, S., Balmer, D.E., Fuller, R.J. 2015. Directionality of recent bird distribution shifts and climate change in Great Britain. *Global Change Biology* 21, 2155-2168.

Gough, L., Grace, J.B., Taylor, K.L. 1994. The relationship between species richness and community biomass: The importance of environmental variables. *Oikos* 70, 271-279.

439 Guilford, T., Meade, J., Willis, J., Phillips, R.A., Boyle, D., Roberts, S., Collett, M., Freeman, R. and
440 Perrins, C.M., 2009. Migration and stopover in a small pelagic seabird, the Manx shearwater Puffinus
441 puffinus: insights from machine learning. *Proceedings of the Royal Society of London B: Biological*
442 *Sciences*, pp.rspb-2008.

443 Guisan, A., Edwards Jnr, T.C., Hastie, T. 2002. Generalized linear and generalized additive models in
444 studies of species distributions: setting the scene. *Ecological Modelling* 157, 89-100.

445 Harrington, R., Fleming, R.A., Woiwod, I.P. 2001. Climate change impacts on insect management and
446 conservation in temperate regions: can they be predicted?. *Agricultural and Forest Entomology* 3,
447 233-240.

448 Harrington, R., Clark, S.J., Welham, S.J., Verrier, P.J., Denholm, C.H., Hulle, M., Maurice, D.,
449 Rounsevell, M.D., Cocu, N., 2007. Environmental change and the phenology of European
450 aphids. *Global Change Biology 13*, 1550-1564.

451 Holloway, P., Miller, J.A., Gillings, S. 2016. Incorporating movement in species distribution models:
452 how do simulations of dispersal affect the accuracy and uncertainty of projections?. *International*
453 *Journal of Geographic Information Science* 30, 2050-2074.

454 Irwin, M.E., Kampmeier, G.E., Weisser, W.W., 2007. Aphid movement: process and
455 consequences. *Aphids as crop pests*, pp.153-186. CAB International.

456 Knudby, A., LeDrew, E., Brenning, A. 2010. Predictive mapping of reef fish species richness, diversity
457 and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. *Remote Sensing of*
458 *Environment* 114, 1230-1241.

459 Kuhn, M., Weston, S., Coulter, N., Quinlan, R. 2015. C50: C5.0 Decision Trees and Rule-Based Models.
460 - R package version 0.1.0-24. https://CRAN.R-project.org/package=C50

461 Lantz, B. 2013. Machine Learning with R. - Packt Publishing Ltd.

462 Lima, M., Harrington, R., Saldaña, S., Estay, S., 2008. Non-linear feedback processes and a latitudinal
463 gradient in the climatic effects determine green spruce aphid outbreaks in the UK. *Oikos 117*, 951-
464 959.

465 Lobo, J.M., Lumerat, J.P., Jay-Robert, P. 2002. Modelling the species richness distribution of French
466 dung beetles (Coleoptera, Scarabaeidae) and delimiting the predictive capacity of different groups of
467 explanatory variables. *Global Ecology and Biology* 11, 265-277.

468 Miller, J.A., Holloway, P. 2015. Incorporating movement in species distribution models. *Progress in*
469 *Physical Geography* 39, 837-849.

470 Olden, J.D., Lawler, J.J., Poff, N.L. 2008. Machine learning methods without tears: a primer for
471 ecologists. *The Quarterly review of biology* 83, 171-193.

472 Price, B., McAlpine, C.A., Kutt, A.S., Ward, D., Phinn, S.R., Ludwig, J.A. 2013. Disentangling how
473 landscape spatial and temporal heterogeneity affects Savanna birds. *PLoS One 8*, p.e74333.

474 R Core Team (2016). R: A language and environment for statistical computing. - R Foundation for
475 statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

476 Reside, A.E., VanDerWal, J.J., Kutt, A.S., Perkins, G.C. 2010. Weather, not climate, defines
477 distributions of vagile bird species. *PLoS One* 5, e13569.

478 Romanski, P., Kotthoff, L. 2016. FSelector: Selecting Attributes. - R package version 0.21.
479 https://CRAN.R-project.org/package=FSelector

480 Selwood, K.E., Thomson, J.R., Clarke, R.H., McGeoch, M.A., Mac Nally, R., 2015. Resistance and
481 resilience of terrestrial birds in drying climates: do floodplains provide drought refugia?. *Global*
482 *ecology and biogeography 24*, 838-848.

483 Sheppard, L.W., Bell, J.R., Harrington, R., Reuman, D.C. 2016. Changes in large-scale climate alter
484 spatial synchrony of aphid pests. *Nature Climate Change 6,* 610-613.

485 Stoner, K.J., Joern, A. 2004. Landscape vs. local habitat scale influences to insect communities from
486 tallgrass prairie remnants. *Ecological Applications* 14: 1306-1320.

487 Thackeray, S.J., Henrys, P.A., Hemming, D., Bell, J.R., Botham, M.S., Burthe, S., Helaouet, P., Johns,
488 D.G., Jones, I.D., Leech, D.I., Mackay, E.B. 2016. Phenological sensitivity to climate across taxa and
489 trophic levels. *Nature* 535, 241-245.

490 van de Pol, M., Bailey, L.D., McLean, N., Rijsdijk, L., Lawson, C.R., Brouwer, L. 2016. Identifying the
491 best climatic predictors in ecology and evolution. *Methods in Ecology and Evolution* 7, 1246-1257.

492 Watt, A.D., Dixon, A.F.G. 1981. The role of cereal growth stages and crowding in the induction of
493 alatae in Sitoban avenae. *Ecological Entomology* 6, 441-447.

494 Westgarth-Smith, A. R., Leroy, S.A., Collins, P.E., Harrington, R. 2007. Temporal variations in English
495 populations of a forest insect pest, the green spruce aphid (Elatobium abietinum), associated with
496 the North Atlantic Oscillation and global warming. *Quaternary International* 173, 153-160.

497 Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F.,
498 Forchhammer, M.C., Grytnes, J.A., Guisan, A., Heikkinen, R.K., 2013. The role of biotic interactions in
499 shaping distributions and realised assemblages of species: implications for species distribution
500 modelling. *Biological Reviews* 88, 15-30.Wood, S.N. 2011. Fast stable restricted maximum likelihood
501 and marginal likelihood estimation of semiparametric generalized models. *Journal of the Royal*
502 *Statistical Society (B)* 73, 3-36.

503 Zhou, X., Perry, J.N., Woiwod, I.P., Harrington, R., Bale, J.S., Clark, S.J., 1997. Temperature change
504 and complex dynamics. *Oecologia* 112, 543-550.

505

Tables

Table 1. Information gain for different accumulated degree day (ADD) temperature thresholds calculated at a range of temporal extents for first flight of *Myzus persicae* for both maximum temperature and mean temperature, and for winter harshness measures. Maximum information gain for each variable depicted by grid border.

	Days Prior	ADD11	ADD12	ADD13	ADD14	ADD15	ADD16
Max Temp	*7*	0.140	0.143	0.144	0.158	0.152	0.135
	14	0.147	0.191	0.181	0.155	0.147	0.143
	21	0.172	0.181	0.180	0.156	0.151	0.158
	28	0.155	0.162	0.167	0.201	0.206	0.174
	60	0.167	0.170	0.174	0.185	0.178	0.186
	90	0.131	0.133	0.142	0.166	0.161	0.170
	120	0.070	0.086	0.091	0.101	0.117	0.123
	180	0.038	0.041	0.042	0.043	0.043	0.044
	364	0.000	0.000	0.036	0.041	0.035	0.034
	Days Prior	**ADD11**	**ADD12**	**ADD13**	**ADD14**	**ADD15**	**ADD16**
Mean Temp	*7*	0.150	0.112	0.100	0.084	0.062	0.048
	14	0.176	0.132	0.115	0.091	0.069	0.048
	21	0.153	0.142	0.129	0.108	0.074	0.057
	28	0.185	0.152	0.141	0.112	0.073	0.057
	60	0.146	0.146	0.129	0.111	0.082	0.062
	90	0.098	0.106	0.096	0.094	0.060	0.050
	120	0.057	0.054	0.053	0.053	0.000	0.033
	180	0.063	0.065	0.063	0.067	0.057	0.048
	364	0.044	0.060	0.052	0.050	0.000	0.000
	Months Prior	**NAO**	**Pressure**	**MeanTemp**	**MaxTemp**	**MinTemp**	
Winter	*6-2*	0.000	0.027	0.131	0.127	0.155	
	6-3	0.000	0.025	0.164	0.173	0.139	
	6-4	0.000	0.000	0.146	0.179	0.186	
	6-5	0.000	0.000	0.177	0.179	0.186	
	5-2	0.000	0.062	0.110	0.094	0.114	
	5-3	0.000	0.034	0.183	0.119	0.126	
	5-4	0.000	0.000	0.195	0.163	0.145	

511

512   *North Atlantic Oscillation (NAO), Accumulated Degree Day (ADD).

14

513

Table 2. Accuracy (proportion correctly classified - PCC) of decision tree models fit on baseline
variables (accumulated degree days - ADD above 16°C for 60 days prior and North Atlantic
Oscillation - NAO for 6-2 months prior) and entropy selected accumulated degree days (ADD) and
winter variables. *depicts ADD calculated from mean temperature and no annotation depicts ADD
calculated from maximum temperature. Days (d) and Months (m) depicted within extent.
Monoecious and hetereocious alternations refer to non-host and host alternating aphid phenology
respectively. Some of the biggest differences between model accuracies were recorded for
Betulaphis quadrituberculata, Capitophorus similis and Rhopalosiphum maidis.

Species	Alternation	Baseline	Entropy	ADD Entropy Variable		Winter Entropy Variable	
				ADD	Extent	Winter	Extent
Acyrthosiphon pisum	monoecious	0.842	0.832	15	60d	Mean	6-5m
Anoecia corni	heteroecious	0.831	0.814	12*	60d	Max	6-5m
Aulacorthum solani	monoecious	0.918	0.871	16	60d	Mean	5-4m
Betulaphis quadrituberculata	monoecious	0.767	0.942	13	21d	Mean	6-5m
Brachycaudus helichrysi	heteroecious	0.861	0.889	14	60d	Mean	6-5m
Brevicoryne brassicae	monoecious	0.816	0.804	11	21d	Min	6-5m
Capitophorus hippophaes	monoecious	0.818	0.832	13	60d	Mean	6-5m
Capitophorus similis	heteroecious	0.767	0.880	11	60d	Max	6-5m
Cavariella aegopodii	heteroecious	0.803	0.850	12	60d	Min	6-5m
Cavariella archangelicae	heteroecious	0.860	0.785	12*	21d	Max	6-5m
Cavariella pastinacae	heteroecious	0.887	0.907	13	60d	Mean	6-4m
Cavariella theobaldi	heteroecious	0.847	0.875	15	90d	Mean	6-4m
Ceruraphis eriophori	heteroecious	0.794	0.846	14	90d	Mean	6-5m
Cryptomyzus galeopsidis	heteroecious	0.852	0.836	11*	60d	Max	6-5m
Drepanosiphum platanoidis	monoecious	0.852	0.912	11	60d	Max	6-5m
Elatobium abietinum	monoecious	0.841	0.872	14	28d	Max	6-5m
Eriosoma patchiae	heteroecious	0.765	0.752	11	120d	Mean	5-2m
Eriosoma ulmi	heteroecious	0.888	0.915	11	60d	Max	6-5m
Eucallipterus tiliae	monoecious	0.832	0.839	11*	90d	Max	6-5m
Euceraphis punctipennis	monoecious	0.859	0.862	13	60d	Mean	6-5m
Hyadaphis foeniculi	heteroecious	0.919	0.926	12*	60d	Min	6-5m
Hyalopterus pruni	heteroecious	0.871	0.908	11*	28d	Mean	6-5m
Hyperomyzus lactucae	heteroecious	0.857	0.882	13	60d	Max	6-5m
Macrosiphum euphorbiae	heteroecious	0.812	0.851	11	28d	Max	6-4m
Macrosiphum rosae	heteroecious	0.650	0.713	13	7d	Min	6-5m
Metopolophium dirhodum	heteroecious	0.754	0.800	11	28d	Mean	6-5m
Metopolophium festucae	monoecious	0.803	0.818	15	60d	Max	6-5m
Microlophium carnosum	monoecious	0.840	0.892	11	60d	Mean	6-5m
Myzocallis castanicola	monoecious	0.903	0.908	16	28d	Max	6-5m
Myzocallis coryli	monoecious	0.911	0.912	13*	60d	Min	6-5m
Myzus ascalonicus	monoecious	0.838	0.887	14*	180d	Max	6-4m
Myzus persicae	heteroecious	0.807	0.801	15	28d	Mean	5-4m
Myzus cerasi	heteroecious	0.854	0.872	11*	60d	Max	6-5m
Myzus lythri	heteroecious	0.825	0.925	12	60d	Min	6-4m
Ovatus crataegarius	heteroecious	0.913	0.923	12	60d	Max	6-5m

Periphyllus testudinaceus	monoecious	0.851	0.918	11	60d	Max	6-3m
Phorodon humuli	heteroecious	0.879	0.938	13	90d	Mean	6-4m
Phyllaphis fagi	monoecious	0.850	0.902	15	60d	Max	6-5m
Pterocallis alni	monoecious	0.840	0.866	11*	90d	Mean	6-5m
Rhopalosiphum insertum	heteroecious	0.852	0.885	11	60d	Max	6-5m
Rhopalosiphum maidis	heteroecious	0.676	0.739	12	14d	Max	6-5m
Rhopalosiphum padi	heteroecious	0.790	0.803	15	28d	Max	6-5m
Sitobion avenae	monoecious	0.822	0.837	13	28d	Max	6-5m
Sitobion fragariae	heteroecious	0.836	0.862	14	60d	Max	6-5m
Tetraneura ulmi	heteroecious	0.862	0.911	14*	90d	Max	6-5m
Thecabius affinis	heteroecious	0.815	0.848	11	120d	Mean	6-5m
Thelaxes dryophila	monoecious	0.914	0.944	12*	60d	Min	6-5m
Tuberculatus annulatus	monoecious	0.888	0.924	11*	60d	Max	6-5m
Tuberculatus borealis	monoecious	0.894	0.927	11*	60d	Mean	6-5m
Utamphorophora humboldti	heteroecious	0.827	0.779	15	21d	Max	5-4m
Wahlgreniella arbuti	heteroecious	0.779	0.823	11	60d	Min	6-4m

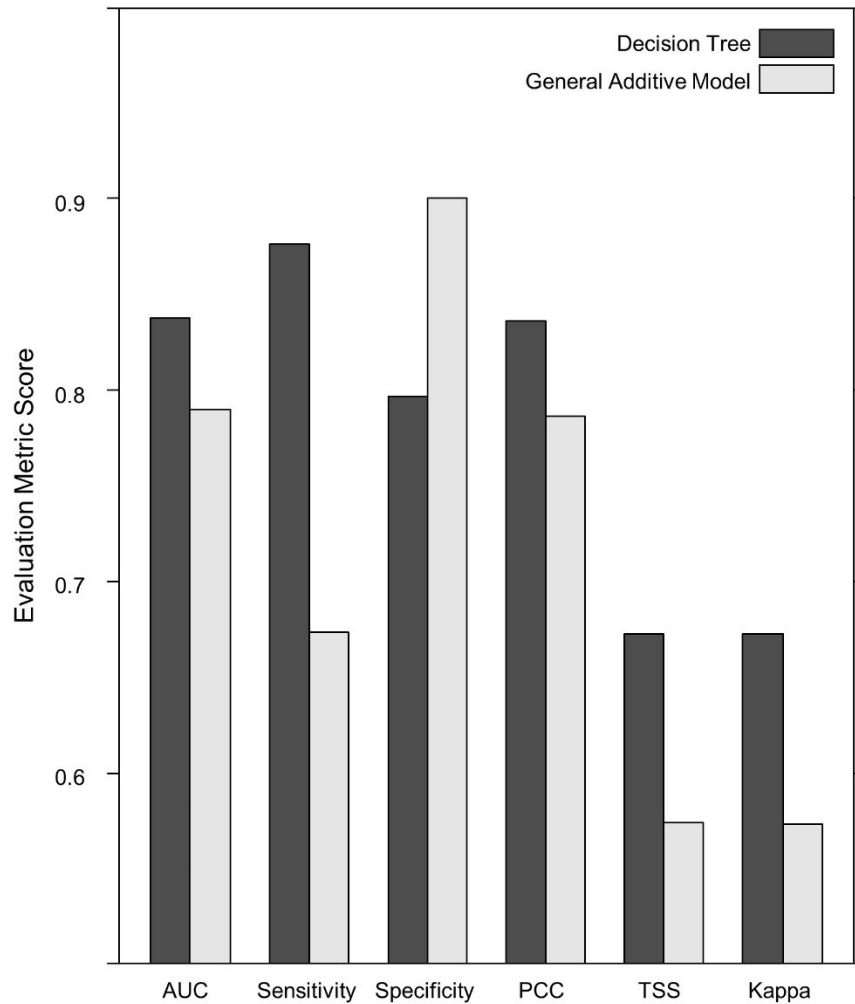| **Average**		
combined	0.836	0.863
heteroecious	0.824	0.851
monoecious	0.852	0.881

522

523

524

526 **Figures**

527

528 Figure 1. The average accuracy score of the predictive models measured across the 51 aphid species
529 when decision tree models fit with baseline variables were compared with general additive models
530 fit with the same baseline variables. Baseline variables were North Atlantic Oscillation (NAO) for the
531 dynamic 6-2 months prior to a first flight observation and accumulated degree days above 16°C
532 (ADD16) for 60 days prior a first flight observation. On average, the decision tree model recorded
533 significantly higher evaluation scores in five out of the six metrics (measured to an $\alpha <= 0.01$ using
534 paired sample t-tests). Accuracy statistics included Area Under the Curve (AUC), Sensitivity,
535 Specificity, Proportion Correctly Classified (PCC), True Skill Statistic (TSS), and Kappa. Readers are
536 directed to Franklin (2009) for a discussion on these evaluation metrics.

537

538
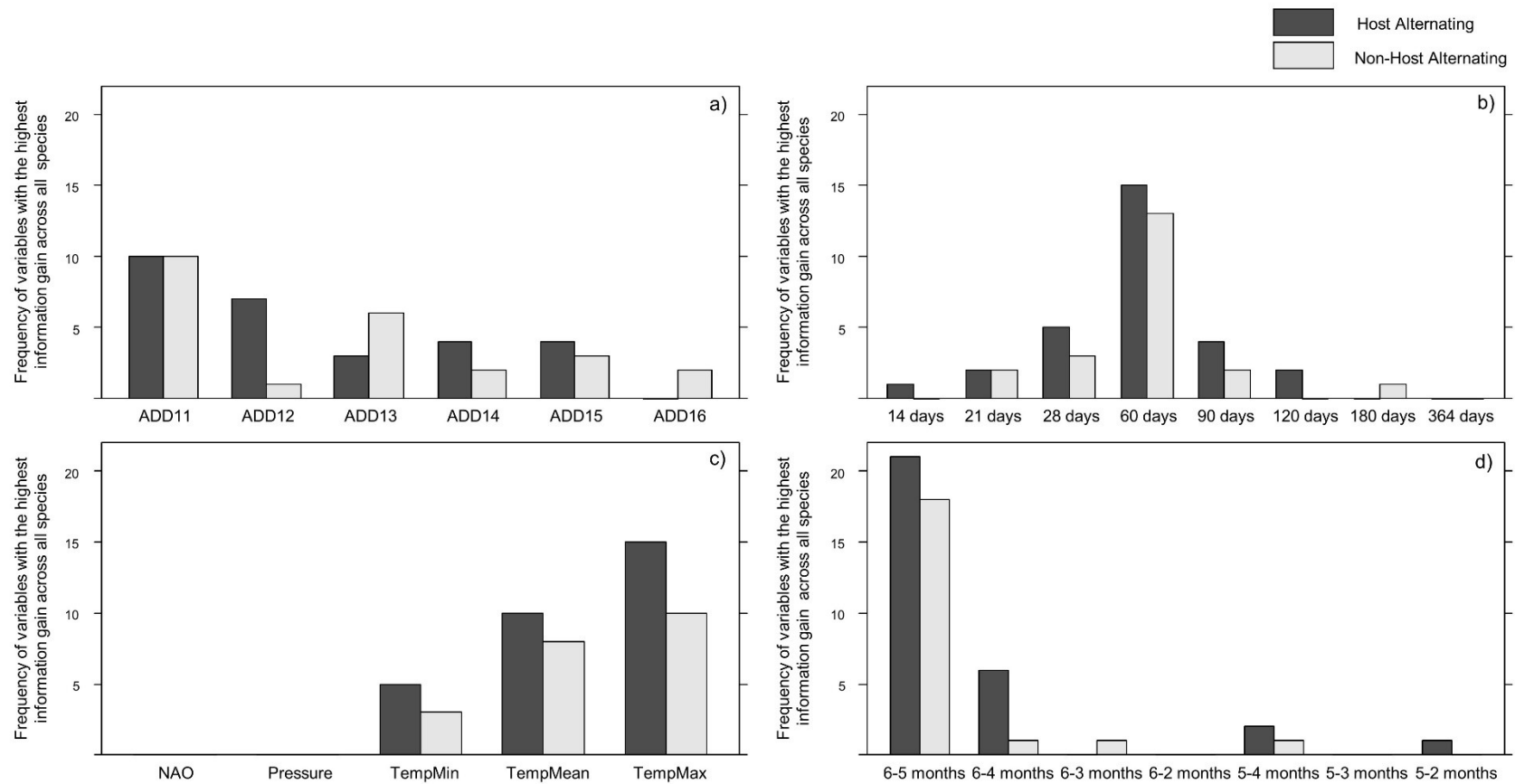
Figure 2: Frequency distribution of the variables attaining the highest information gain measured using entropy for each of the 51 species. The variables
representative of spring development were a) accumulated degree days (ADD) above certain temperature thresholds, and these were calculated at a
number of different b) dynamic temporal extents. The variables representative of winter harshness were c) North Atlantic Oscillation (NAO), Pressure, and
Minimum, Mean, and Maximum Temperature, and these were calculated at a number of different d) dynamic temporal extents.
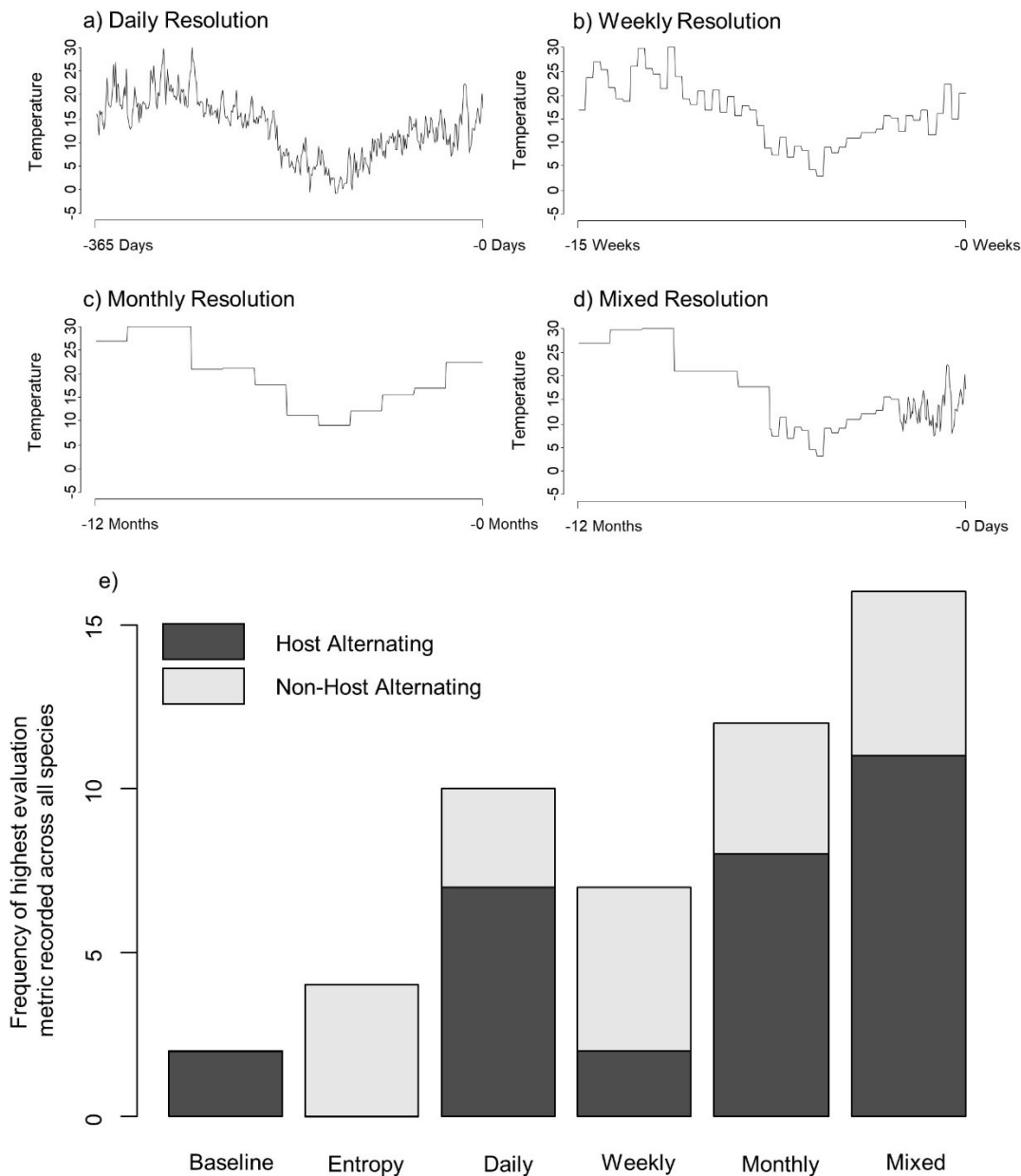
Figure 3: Illustration of the different resolutions that were used within the decision tree models to predict aphid first flight. Resolutions include a) fine-scale daily data, b) intermediate-scale weekly data, c) coarse-scale monthly data, and d) a mix of resolutions, including daily data going back 60 days, weekly data back until six months, and monthly data for the remainder of the year. These resolutions were used to fit decision tree models, and e) shows the frequency of a decision tree model to produce the highest accuracy (Proportion Correctly Classified – PCC) among models fit using these different resolutions, as well as both baseline variables (North Atlantic Oscillation – NAO at 6-2 months and Accumulated Degree Days – ADD above 16°C) and entropy selected variables for the 51 aphid species.

555

556
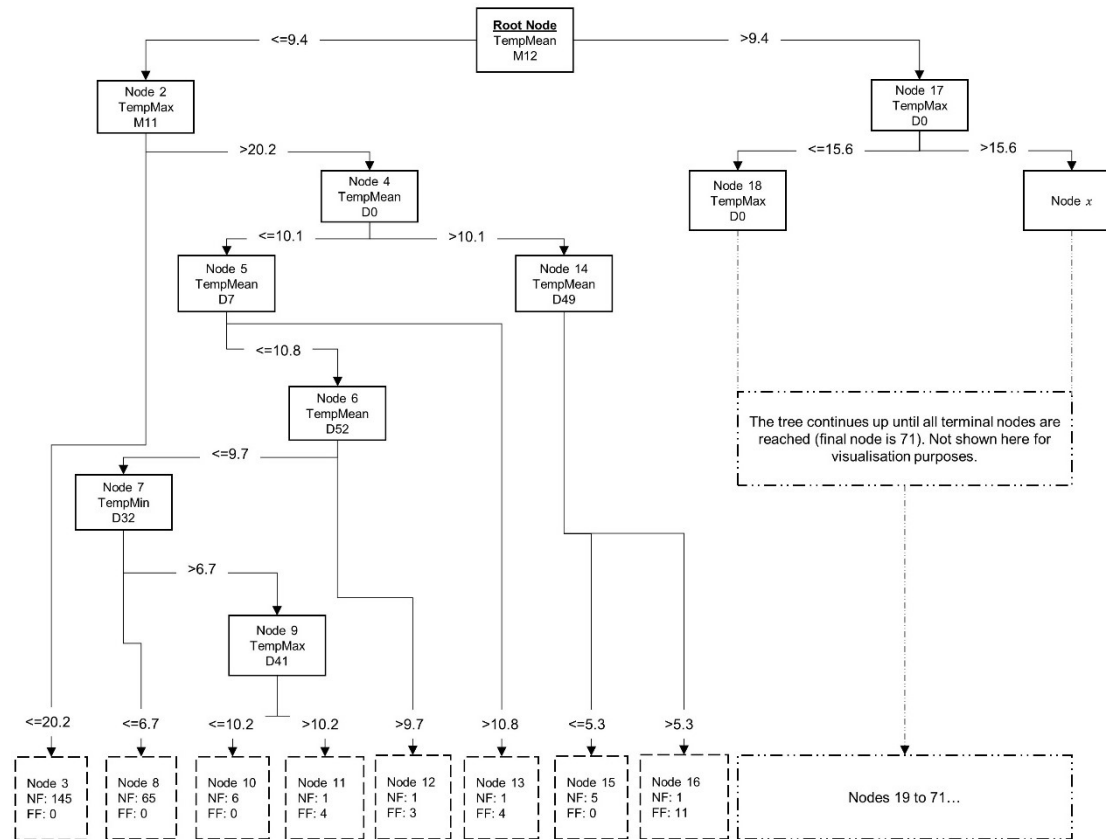
Figure 4: A decision tree for *Myzus persicae* fit using temperature data of mixed resolutions. The aim of a decision tree is to classify the response data correctly as either first flight (FF) or no flight (NF). The tree begins with the root node, that separates training data based on whether the mean temperature for the twelfth month prior to an observation was <= or > 9.4°C. For values with a mean temperature <= 9.4°C, the tree moves to the second node, which splits the response data based on whether the maximum temperature for the eleventh month prior was <= or > 20.2°C. Node 3 is a terminal node, whereby if the maximum temperature recorded was <=20.2°C, then the observation is classified as no flight. Based on our data, this was a pure node, with 145 observations classified as NF and 0 as FF. Temporal resolutions contained within the model are daily (D), weekly (W), and monthly (M). The tree continues in such a manner until all terminal nodes are reached (final node is 71). We only illustrate the tree up to split 18 in order to highlight the main decisions and the ability of decision tree modelling to predict hierarchical ecological systems.
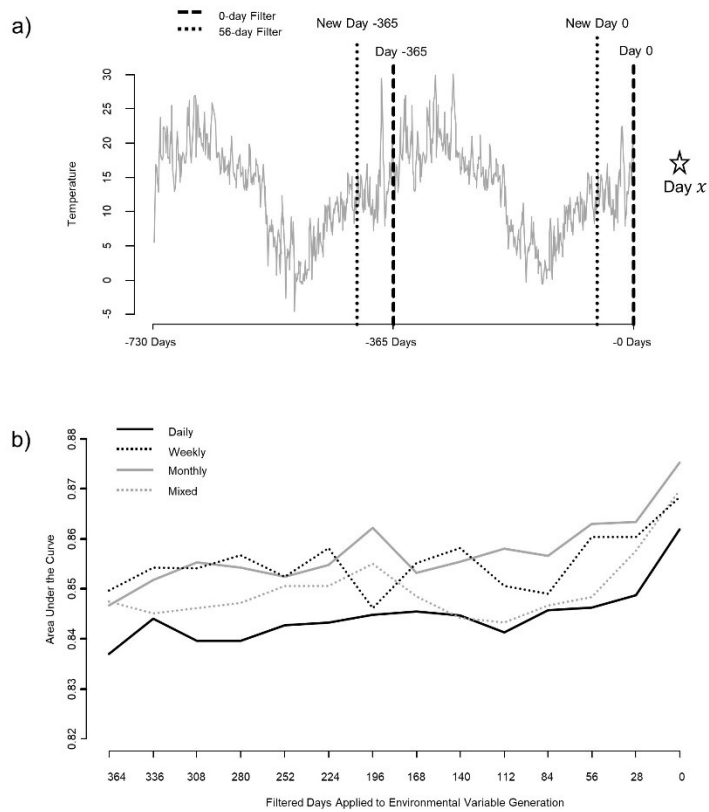
Figure 5: Illustration and results of the application of filters to the environmental variables in predicting aphid first flight. a) A two-year period of daily temperature data preceding a first flight or no flight recording at day 0, with a 0-day (red) and 56-day (blue) filter applied to the temperature data used to generated predictor variables. All models fit on these new environmental variables are predicting first flight or no flight at day x. b) the average Area Under the Curve (AUC) score of decision tree models fit on temperature data (at four different resolutions) with the filters applied for the 51 species. As the number of filtered days applied to the environmental variable generation increases, the recorded AUC value subsequently decreases; however this decrease is minimal. The largest recorded difference in AUC between a 0-day and a 364-day filter is 0.0286 for the monthly resolution decision tree.