

This is a repository copy of *Dynamic selection of environmental variables to improve the prediction of aphid phenology : a machine learning approach*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/127857/>

Version: Accepted Version

---

**Article:**

Holloway, Paul, Kudenko, Daniel [orcid.org/0000-0003-3359-3255](https://orcid.org/0000-0003-3359-3255) and James, Bell (2018) Dynamic selection of environmental variables to improve the prediction of aphid phenology : a machine learning approach. *Ecological Indicators*. pp. 512-521. ISSN 1470-160X

<https://doi.org/10.1016/j.ecolind.2017.10.032>

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

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

3 Paul Holloway<sup>1,2\*</sup>, Daniel Kudenko<sup>1,3</sup>, and James R. Bell<sup>4</sup>

4 <sup>1</sup>Department of Computer Science, University of York, Deramore Lane, York, YO10 5GH, UK

5 <sup>2</sup>Department of Geography, University College Cork, Cork, Ireland

6 <sup>3</sup>Saint Petersburg National Research Academic University of the Russian Academy of Science, Russia

7 <sup>4</sup>Rothamsted Insect Survey, Biointeractions and Crop Protection, Rothamsted Research, AL5 2JQ, UK

8 [paul.holloway@ucc.ie](mailto:paul.holloway@ucc.ie) [orcid.org/0000-0002-9773-4531](https://orcid.org/0000-0002-9773-4531)

9 [daniel.kudenko@york.ac.uk](mailto:daniel.kudenko@york.ac.uk) [orcid.org/0000-0003-3359-3255](https://orcid.org/0000-0003-3359-3255)

10 [james.bell@rothamsted.ac.uk](mailto:james.bell@rothamsted.ac.uk) [orcid.org/0000-0003-1640-6120](https://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 **Keywords:** Entropy, Scale; Weather; Decision Trees; Generalised Additive Models, GAM; First Flight

43 **Introduction:**

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

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

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

85 Although the importance of temperature and NAO in understanding and predicting aphid flight  
86 dates cannot be understated, the derivation of these variables is subject to a number of conceptual  
87 and methodological uncertainties. In particular, the effect of the temporal scale used in variable  
88 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***

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

$$188 \quad \text{Entropy}(S) = -p(a) * \log(p(a)) - p(b) * \log(p(b)) \quad \text{Equation 1.}$$

189  $S$  denotes a partition of the training examples, and  $p(x)$  is the proportion of training examples of  
190 class  $x$  in partition  $S$ . Entropy is maximal when  $p(a) = p(b) = 0.5$ , and minimal when  $p(a) = 0$  or  $p(b) =$   
191  $0$ . The information gain of a variable is the decrease in entropy caused by splitting the training data  
192 according to its values. We implemented entropy using the FSelector package (Romanski and  
193 Kotthoff 2016) in R 3.3.1 (R Core Team 2016).

#### 194 **Model Selection: Decision Trees and General Additive Models**

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

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

$$213 \quad g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_n(x_n) + \varepsilon \quad \text{Equation 2.}$$

214 where  $g$  describes the link function,  $\beta_0$  is the intercept term,  $f$  describes the smoothing function  
215 used on predictor variable  $x$ , and finally we assume that the error term,  $\varepsilon$  is constant across  
216 observations. We generated gams from the same data as outlined above, meaning we used a  
217 binomial link function to specify the distribution of the response variable and we specified the  
218 smoothing of the environmental variables using thin plate regression splines. Again, we undertook  
219 analysis in R 3.3.1 (R Core Team 2016) in the mgcv package (Wood 2011), with further details  
220 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  $\alpha \leq 0.01$  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  $\alpha \leq 0.01$  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  $\alpha \leq 0.01$  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



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

### 271 **Machine Learning and Entropy**

272 Introducing temperature variables calculated at various temporal resolutions further increased the  
273 accuracy when compared to baseline and entropy models (Figure 3). When models fit with four  
274 different temperature resolutions were compared, we observed that the use of daily data (Figure 3a)  
275 did not result in an over-fitting of the models, reporting equally high accuracies when compared with  
276 other implementations (Figure 3e), although it did record a number of lower outliers that most likely  
277 resulted from noisy data. Similarly, the use of monthly data (Figure 3c) did not result in a smoothing  
278 of the models, reporting a higher number of more accurate models than the use of daily data (Figure  
279 3e). The use of weekly data (Figure 3b) resulted in a higher frequency of PCC scores for non-host  
280 alternating species compared to other temporal resolutions (Figure 3e), while for host-alternating  
281 species the temperature data consisting of various resolutions (Figure 3d) recorded the highest  
282 frequency of PCC scores (Figure 3e). This model incorporated daily data for the two months  
283 dynamically preceding the observations, weekly data back until six months, and then monthly data  
284 for the remainder of the year, suggesting that the use of different scales improves the predictive  
285 ability of decision trees fit for host-alternating species of aphids.

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

### 300 **Discussion**

301 There has been a recent resurgence in the importance of selecting appropriate environmental  
302 variables when attempting to explain or predict ecological phenomena (van de Pol *et al.* 2016). In  
303 addressing this, the aims of our study were twofold. Firstly, we identified the power of machine  
304 learning (specifically decision tree learning) to address complex, hierarchical ecological questions,  
305 and illustrated how this method can be used to attain highly accurate models and identify previously  
306 unknown features of ecological importance. Secondly, we utilised this methodology to explore the  
307 ecological indicators for UK aphids, and improved upon existing prediction techniques. The results of  
308 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**

389 This research was jointly funded by BBSRC and InnovateUK under grant reference BB/M006980. We  
390 would like to thank Ian Bedford for initial comments on this research. We are also grateful to  
391 collaborators and staff who have contributed to the Rothamsted Insect Survey suction-trap network  
392 since the 1960s. We thank Jon Pickup and others in SASA for their use of data. The RIS is a National  
393 Capability funded by BBSRC.

### 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.

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

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

402 Awmack, C.S., Harrington, R., Lindroth, R.L. 2004. Aphid individual performance may not predict  
403 population responses to elevated CO<sub>2</sub> or O<sub>3</sub>. *Global Change Biology* 10, 1414-1423.

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

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

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

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

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

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

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

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

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

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

428 Fancourt, B.A., Bateman, B.L., VanDerWal, J., Nicol, S.C., Hawkins, C.E., Jones, M.E., Johnson, C.N.  
429 2015. Testing the role of climate change in species decline: is the eastern quoll a victim of a change  
430 in the weather?. *PLoS one* 10, p.e0129420.

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

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

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

437 Gough, L., Grace, J.B., Taylor, K.L. 1994. The relationship between species richness and community  
438 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., Rijdsdijk, 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 *Sitona 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

506 **Tables**

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

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

511

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

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

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



<i>Periphyllus testudinaceus</i>	monoecious	0.851	0.918	11	60d	Max	6-3m
<i>Phorodon humuli</i>	heteroecious	0.879	0.938	13	90d	Mean	6-4m
<i>Phyllaphis fagi</i>	monoecious	0.850	0.902	15	60d	Max	6-5m
<i>Pterocallis alni</i>	monoecious	0.840	0.866	11*	90d	Mean	6-5m
<i>Rhopalosiphum insertum</i>	heteroecious	0.852	0.885	11	60d	Max	6-5m
<i>Rhopalosiphum maidis</i>	heteroecious	0.676	0.739	12	14d	Max	6-5m
<i>Rhopalosiphum padi</i>	heteroecious	0.790	0.803	15	28d	Max	6-5m
<i>Sitobion avenae</i>	monoecious	0.822	0.837	13	28d	Max	6-5m
<i>Sitobion fragariae</i>	heteroecious	0.836	0.862	14	60d	Max	6-5m
<i>Tetraneura ulmi</i>	heteroecious	0.862	0.911	14*	90d	Max	6-5m
<i>Thecabius affinis</i>	heteroecious	0.815	0.848	11	120d	Mean	6-5m
<i>Thelaxes dryophila</i>	monoecious	0.914	0.944	12*	60d	Min	6-5m
<i>Tuberculatus annulatus</i>	monoecious	0.888	0.924	11*	60d	Max	6-5m
<i>Tuberculatus borealis</i>	monoecious	0.894	0.927	11*	60d	Mean	6-5m
<i>Utamphorophora humboldti</i>	heteroecious	0.827	0.779	15	21d	Max	5-4m
<i>Wahlgreniella arbuti</i>	heteroecious	0.779	0.823	11	60d	Min	6-4m
<b>Average</b>							
	combined	0.836	0.863				
	heteroecious	0.824	0.851				
	monoecious	0.852	0.881				

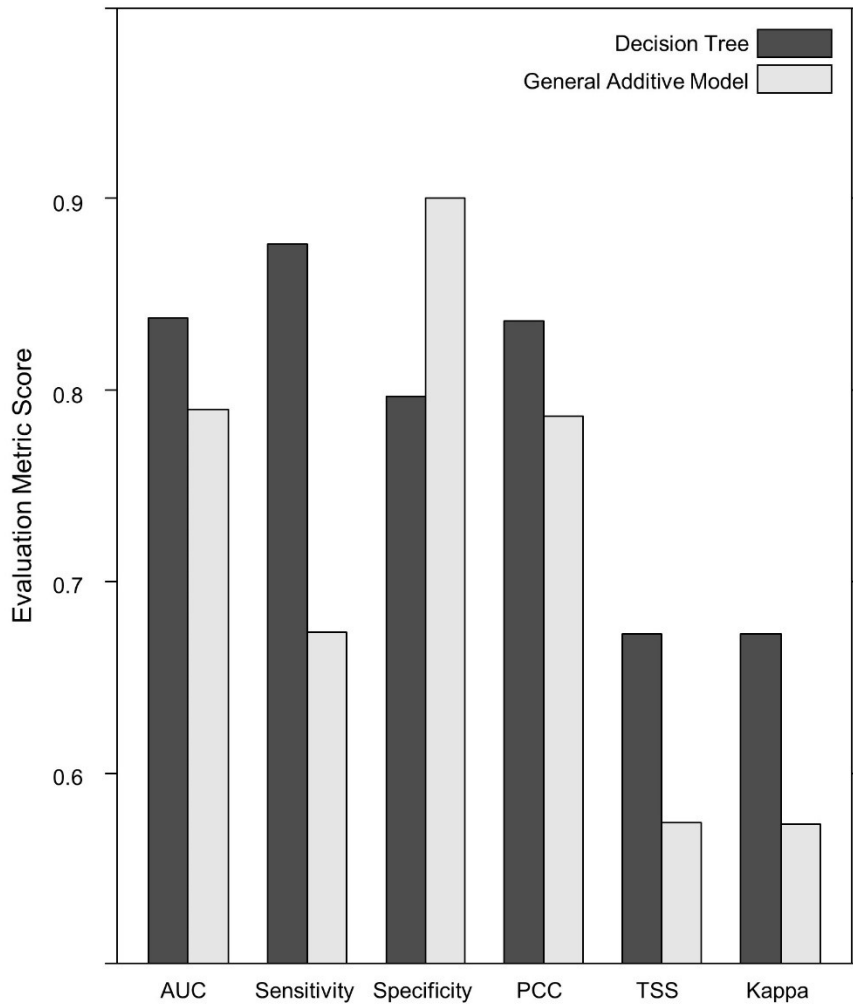
522

523

524

525

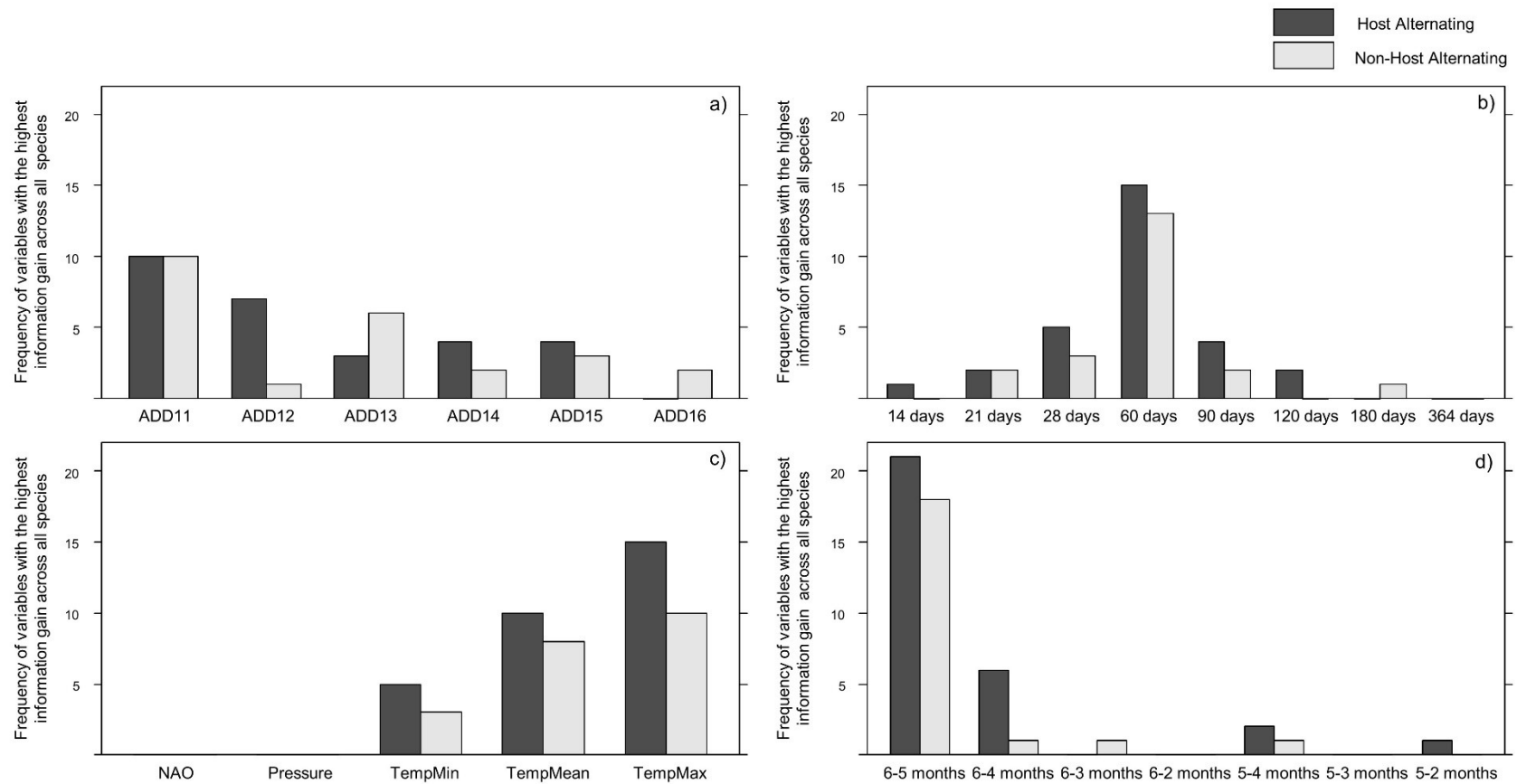
526 **Figures**



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  $\alpha \leq 0.01$  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

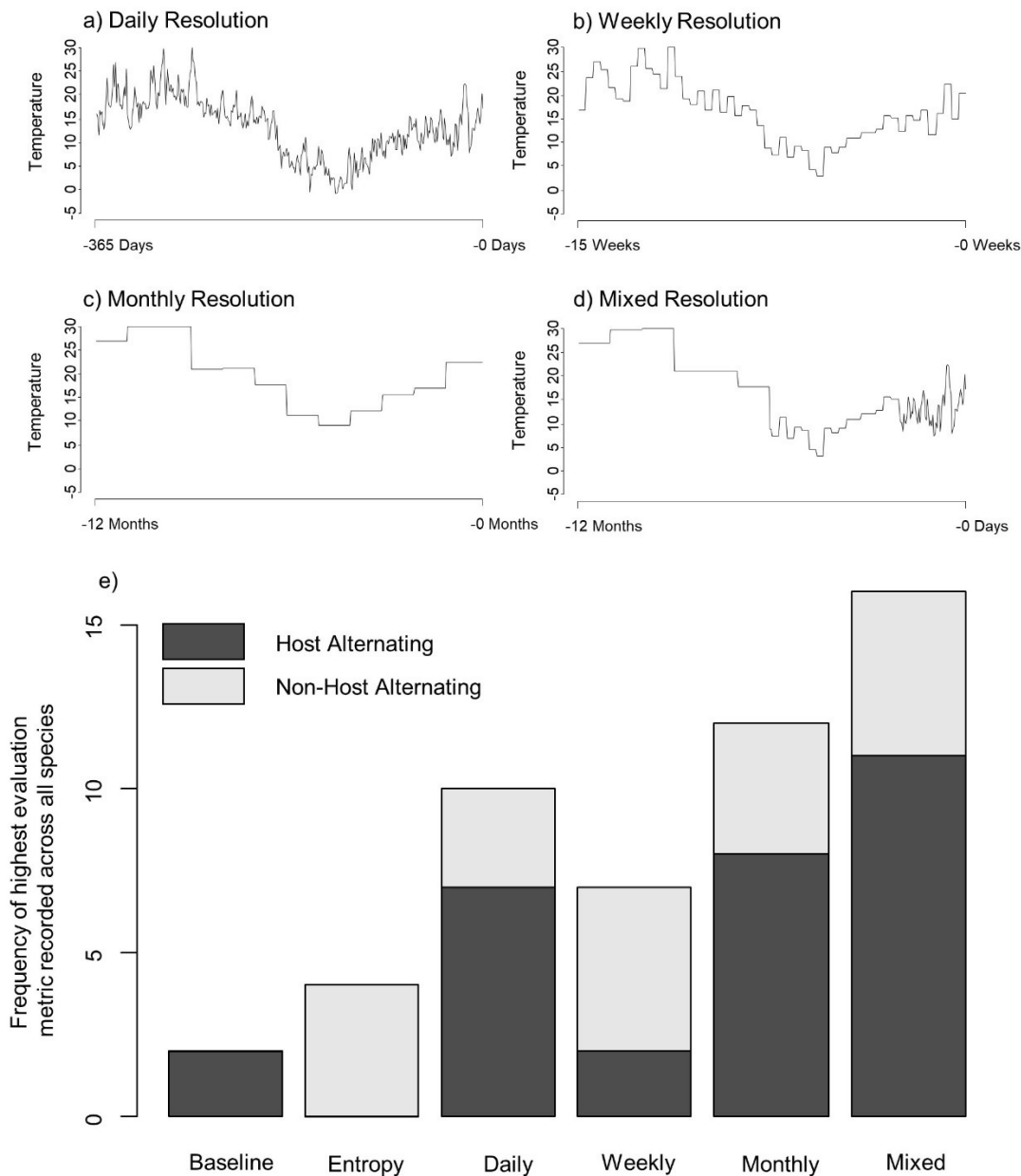


538

539 Figure 2: Frequency distribution of the variables attaining the highest information gain measured using entropy for each of the 51 species. The variables  
 540 representative of spring development were a) accumulated degree days (ADD) above certain temperature thresholds, and these were calculated at a  
 541 number of different b) dynamic temporal extents. The variables representative of winter harshness were c) North Atlantic Oscillation (NAO), Pressure, and  
 542 Minimum, Mean, and Maximum Temperature, and these were calculated at a number of different d) dynamic temporal extents.

543

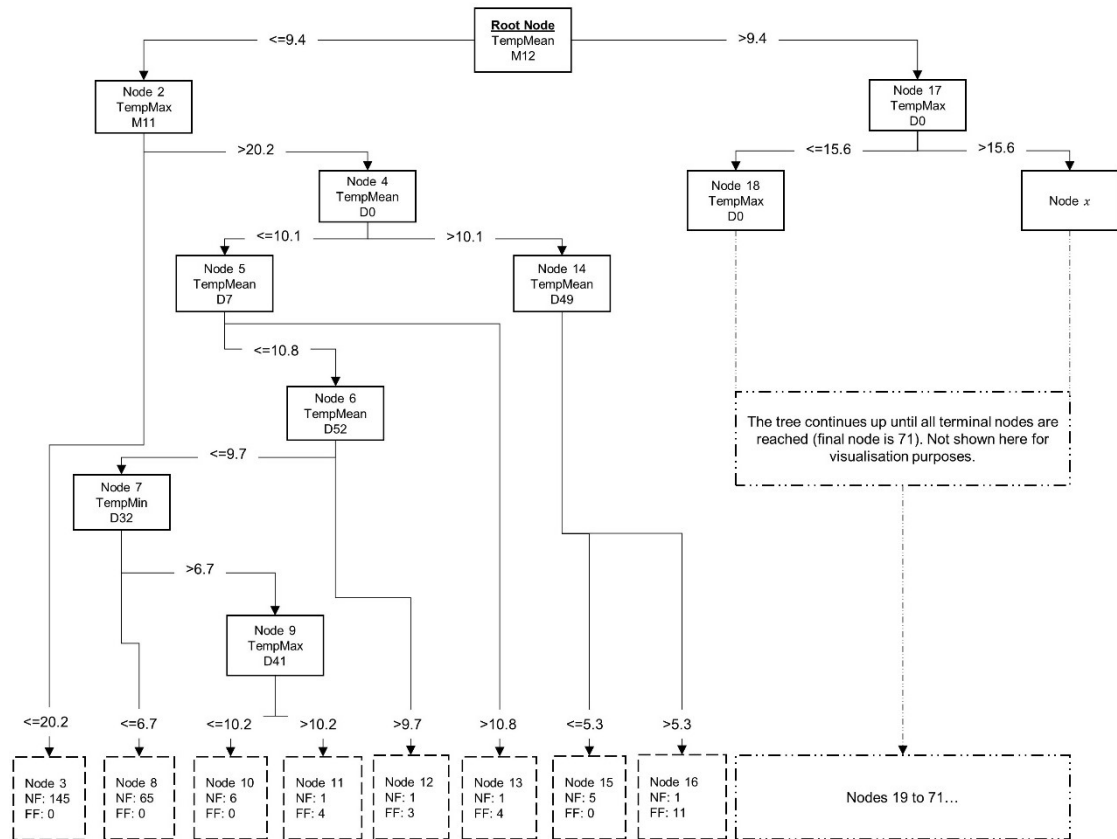
544



545

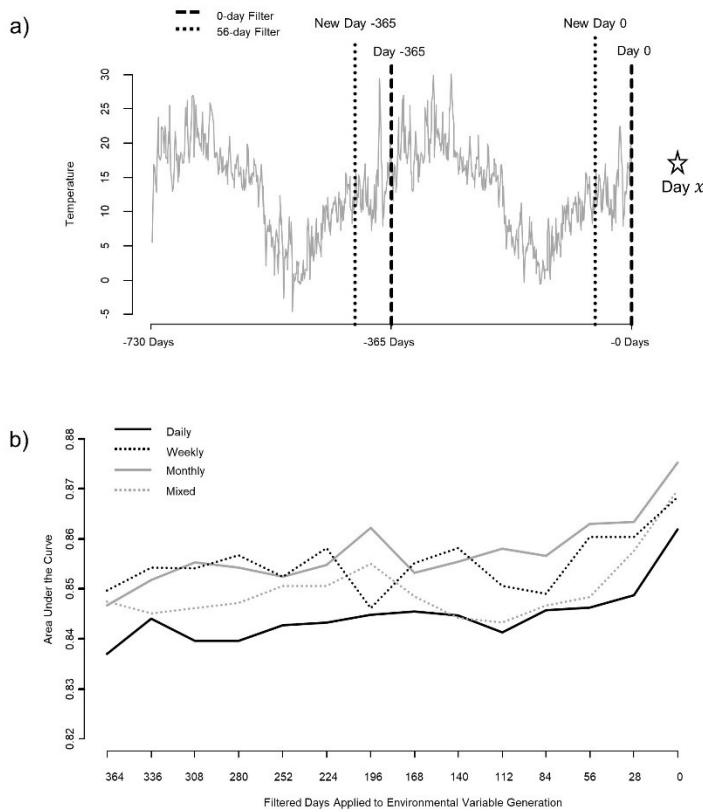
546 Figure 3: Illustration of the different resolutions that were used within the decision tree models to  
547 predict aphid first flight. Resolutions include a) fine-scale daily data, b) intermediate-scale weekly  
548 data, c) coarse-scale monthly data, and d) a mix of resolutions, including daily data going back 60  
549 days, weekly data back until six months, and monthly data for the remainder of the year. These  
550 resolutions were used to fit decision tree models, and e) shows the frequency of a decision tree  
551 model to produce the highest accuracy (Proportion Correctly Classified – PCC) among models fit  
552 using these different resolutions, as well as both baseline variables (North Atlantic Oscillation – NAO  
553 at 6-2 months and Accumulated Degree Days – ADD above 16°C) and entropy selected variables for  
554 the 51 aphid species.

555



556

557 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  
 558 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  
 559 for the twelfth month prior to an observation was  $\leq$  or  $>$  9.4°C. For values with a mean temperature  $\leq$  9.4°C, the tree moves to the second node, which  
 560 splits the response data based on whether the maximum temperature for the eleventh month prior was  $\leq$  or  $>$  20.2°C. Node 3 is a terminal node, whereby  
 561 if the maximum temperature recorded was  $\leq$  20.2°C, then the observation is classified as no flight. Based on our data, this was a pure node, with 145  
 562 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  
 563 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  
 564 the ability of decision tree modelling to predict hierarchical ecological systems.



565

566 Figure 5: Illustration and results of the application of filters to the environmental variables in  
 567 predicting aphid first flight. a) A two-year period of daily temperature data preceding a first flight or  
 568 no flight recording at day 0, with a 0-day (red) and 56-day (blue) filter applied to the temperature  
 569 data used to generated predictor variables. All models fit on these new environmental variables are  
 570 predicting first flight or no flight at day  $x$ . b) the average Area Under the Curve (AUC) score of  
 571 decision tree models fit on temperature data (at four different resolutions) with the filters applied  
 572 for the 51 species. As the number of filtered days applied to the environmental variable generation  
 573 increases, the recorded AUC value subsequently decreases; however this decrease is minimal. The  
 574 largest recorded difference in AUC between a 0-day and a 364-day filter is 0.0286 for the monthly  
 575 resolution decision tree.

576