Title	Dynamic selection of environmental variables to improve the prediction of aphid phenology: A machine learning approach
Authors	Holloway, Paul;Kudenko, Daniel;Bell, James R.
Publication date	2018-02-19
Original Citation	Holloway, P., Kudenko, D. and Bell, J. R. (2018) 'Dynamic selection of environmental variables to improve the prediction of aphid phenology: A machine learning approach', Ecological Indicators, 88, pp. 512-521. doi: 10.1016/j.ecolind.2017.10.032
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://www.sciencedirect.com/science/article/pii/S1470160X17306696 - 10.1016/j.ecolind.2017.10.032
Rights	© 2017 Elsevier Ltd. All rights reserved. This manuscript version is made available under the CC-BY-NC-ND 4.0 license. - http://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2025-09-08 15:21:59
Item downloaded from	https://hdl.handle.net/10468/5630

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

41 **Keywords:** Entropy, Scale; Weather; Decision Trees; Generalised Additive Models, GAM; First Flight

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 extent (i.e.

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

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

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

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

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

130 **Methodology**

131 *Data Collection*

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

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

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

Data Analysis – Machine Learning

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:

$$Entropy(S) = -p(a) * \log(p(a)) - p(b) * \log(p(b))$$
<div align="right">Equation 1.</div>

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:

$$g\big(E(Y)\big) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n) + \varepsilon$$
<div align="right">Equation 2.</div>

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.

Model Evaluation

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

231 **Results**

232 ***Modelling Approach***

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

243 ***Data Representation (Variable Selection)***

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

258 When the frequency distribution of environmental variables with the highest information gain for
259 each species were analysed, we saw differences in both the thresholds (e.g. ADD value) and temporal
260 extents (e.g. number of days) reported, with certain patterns emerging when species were analysed
261 based on phenology (Figure 2). In general, the lower ADD thresholds resulted in higher information
262 gain (Figure 2a), with 20 of the 51 species best predicted by a threshold of 11°C. A dynamic temporal
263 window of 60 days immediately prior to FF was the most informative temporal extent for both host
264 alternating and non-host alternating species (Figure 2b). The average maximum temperature reported

265 higher information gains across the winter metrics, with neither NAO nor pressure resulting in the
266 highest gain for any species (Figure 2c). Shorter dynamic temporal winter metrics were most
267 informative, with longer periods of time prior to an observation indicative of a larger gain (Figure 2d).

268 *Machine Learning and Entropy*

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

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

Discussion

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

308 We identified from expert-opinion and well-established aphid literature (Harrington *et al.* 2007; Bell
309 *et al.* 2015) that a measure of spring development (ADD16 in the previous 60 days) and a measure of

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

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

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

354 An advantage of decision tree models is that they have the ability to handle more variables than
355 commonly applied parametric approaches, in part due to the fact that they are not as sensitive to
356 issues relating to multi-collinearity. Our results suggest that suitable information pertaining to the

prediction of aphid first flight can be ascertained from both voluminous daily data and smoothed monthly data (Figure 3). This has important practical consequences for implementations of these predictive models; projecting species-environment relationships into future space and time at a monthly resolution would reduce the resources and cost of such an endeavour compared to using daily data. Moreover, as ecological phenomena are often influenced by drivers operating across multiple temporal scales, the use of variables generated at different temporal resolutions allowed for both longer- term conditions that enforce an overarching influence and shorter-term variations that describe finer-scale patterns to be included in the model (Figure 4). It should be noted that the use of a large number of meteorological variables within such models could result in the identification of an incorrect spuriously correlated variable that has nothing to do with aphid lifecycles. However, the high accuracy obtained from these models in predicting aphid first flight and the use of 10-fold cross-validation to control for such concerns suggests that our models do well in identifying the meteorological drivers of this phenomena. One caveat to the study is that ML methods require a rich data set for the training of the algorithm and thus this technique may be limited to longer-term studies.

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

Acknowledgements

References

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J., Aide, T.M., 2009. Automated classification of bird and amphibian calls using machine learning: A comparison of 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.

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

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

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

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

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

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

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

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

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

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

424   Fancourt, B.A., Bateman, B.L., VanDerWal, J., Nicol, S.C., Hawkins, C.E., Jones, M.E., Johnson, C.N.
425   2015. Testing the role of climate change in species decline: is the eastern quoll a victim of a change in
426   the weather?. *PloS one 10*, p.e0129420.

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

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

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

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

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

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

441 Harrington, R., Fleming, R.A., Woiwod, I.P. 2001. Climate change impacts on insect management and
442 conservation in temperate regions: can they be predicted?. *Agricultural and Forest Entomology* 3, 233-
443 240.

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

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

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

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

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

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

458 Lima, M., Harrington, R., Saldaña, S., Estay, S., 2008. Non-linear feedback processes and a latitudinal
459 gradient in the climatic effects determine green spruce aphid outbreaks in the UK. *Oikos 117*, 951-
460 959.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

501

502 **Tables**

503 Table 1. Information gain for different accumulated degree day (ADD) temperature thresholds
504 calculated at a range of temporal extents for first flight of *Myzus persicae* for both maximum
505 temperature and mean temperature, and for winter harshness measures. Maximum information gain
506 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

507

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

509

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

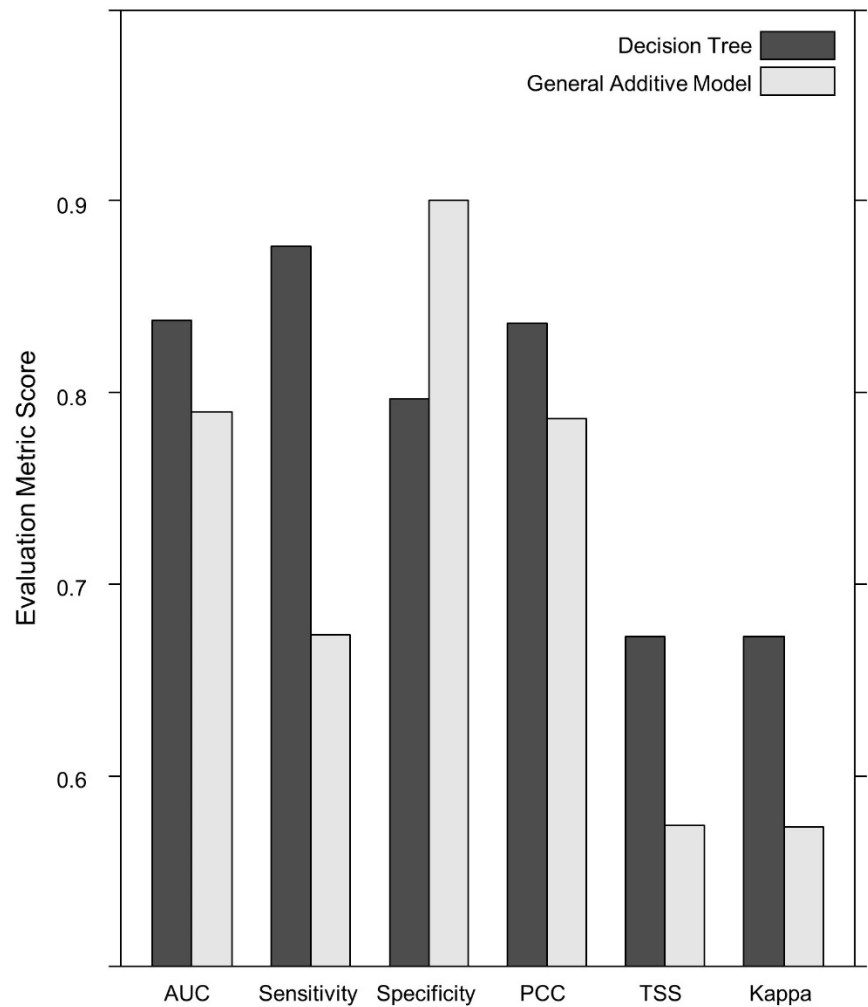|		
combined	0.836	0.863
heteroecious	0.824	0.851
monoecious	0.852	0.881

518

519

520

521
522 **Figures**

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

533

534
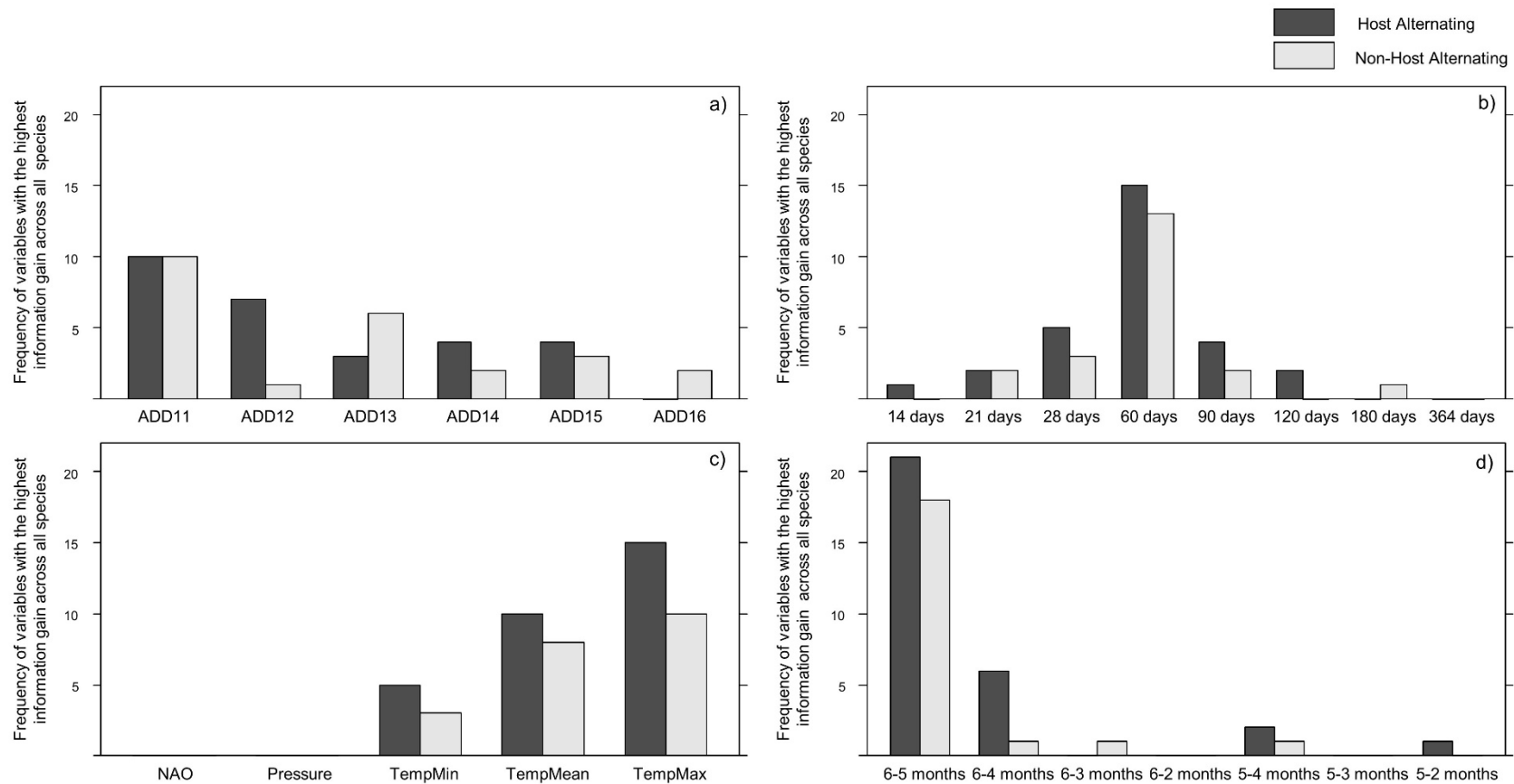
Figure 2: Frequency distribution of the variables attaining the highest information gain measured using entropy for each of the 51 species. The variables
representative of spring development were a) accumulated degree days (ADD) above certain temperature thresholds, and these were calculated at a number
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.

539

540

541
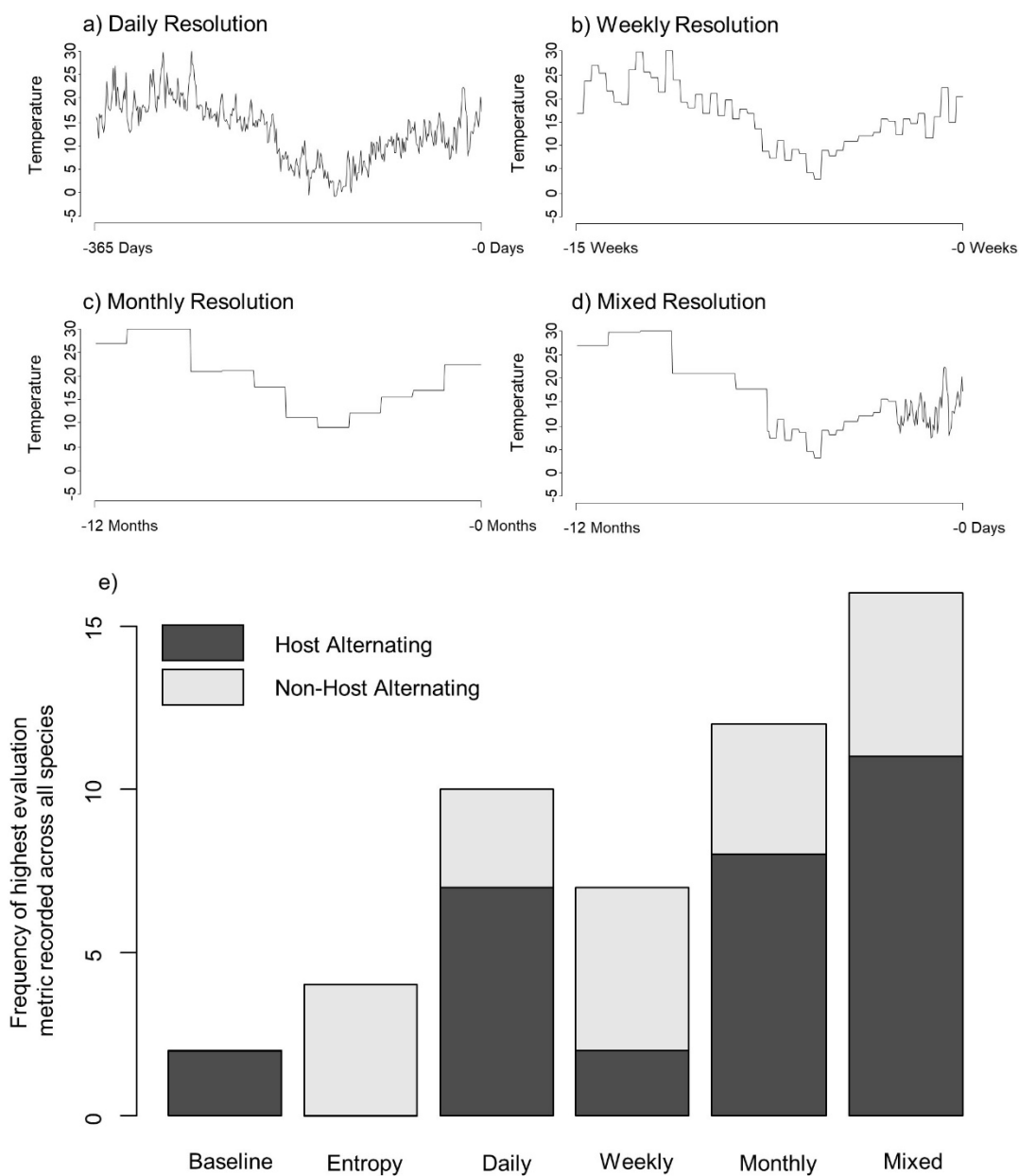
Figure 3: Illustration of the different resolutions that were used within the decision tree models to predict aphid first flight. Resolutions include a) fine-scale daily data, b) intermediate-scale weekly data, c) coarse-scale monthly data, and d) a mix of resolutions, including daily data going back 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.

550

551
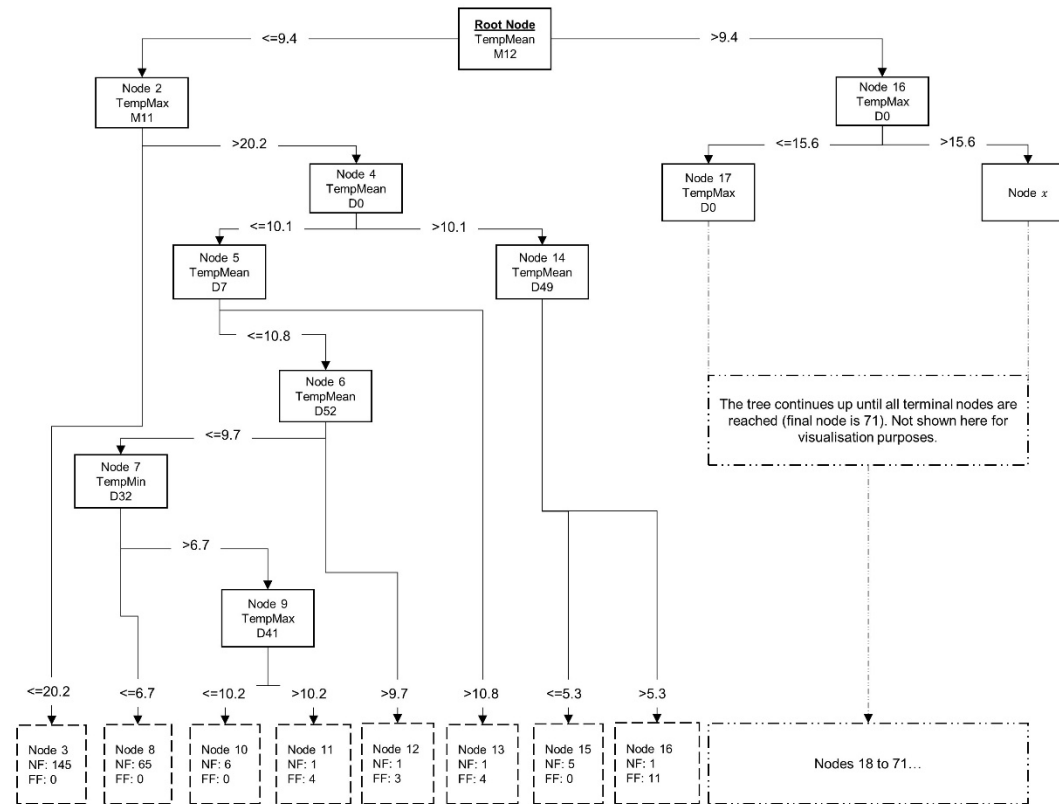
Figure 4: A decision tree for *Myzus persicae* fit using temperature data of mixed resolutions. The aim of a decision tree is to classify the response data correctly as either first flight (FF) or no flight (NF). The tree begins with the root node, that separates training data based on whether the mean 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 17 in order to highlight the main decisions and the ability of decision tree modelling to predict hierarchical ecological systems.
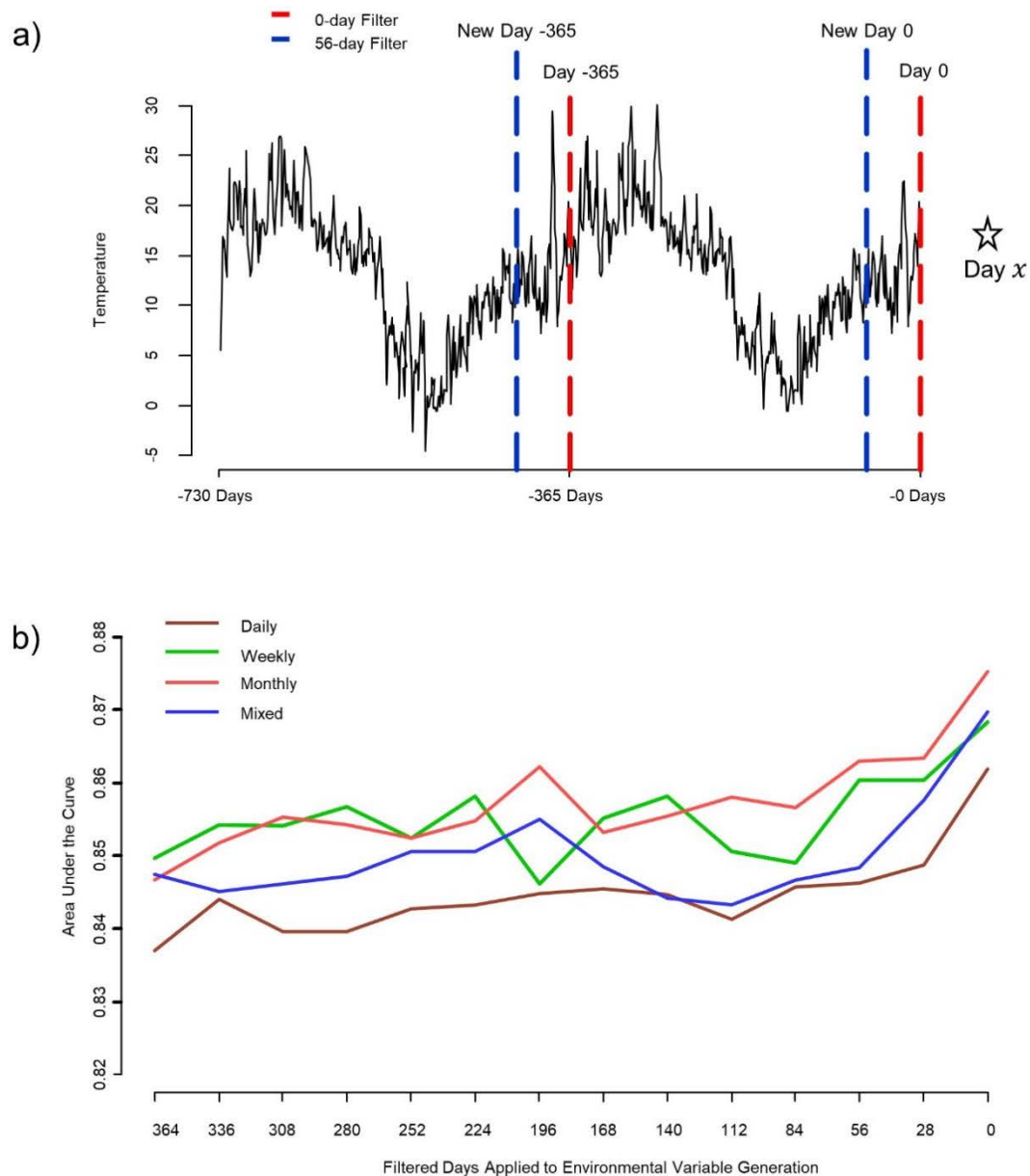
Figure 5: Illustration and results of the application of filters to the environmental variables in predicting aphid first flight. a) A two-year period of daily temperature data preceding a first flight or no flight recording at day 0, with a 0-day (red) and 56-day (blue) filter applied to the temperature 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.