Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 1246–1257 doi: 10.1111/2041-210X.12590

Identifying the best climatic predictors in ecology and evolution

Martijn van de Pol[1,2,3]*, Liam D. Bailey[1], Nina McLean[1], Laurie Rijsdijk[1,2,4], Callum R. Lawson[2] and Lyanne Brouwer[1,2]

[1]*Department of Evolution, Ecology & Genetics, Research School of Biology, The Australian National University, Canberra, ACT 0200, Australia;* [2]*Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Droevendaalsesteeg 10, 6708PB, Wageningen, The Netherlands;* [3]*Centre for Avian Population Studies, Nijmegen, the Netherlands; and* [4]*Department of Animal Ecology and Physiology, Radboud University, Heyendaalseweg 135 6525 AJ, Nijmegen, The Netherlands*

Summary

1. Ecologists and many evolutionary biologists relate the variation in physiological, behavioural, life-history, demographic, population and community traits to the variation in weather, a key environmental driver. However, identifying which weather variables (e.g. rain, temperature, El Niño index), over which time period (e.g. recent weather, spring or year-round weather) and in what ways (e.g. mean, threshold of temperature) they affect biological responses is by no means trivial, particularly when traits are expressed at different times among individuals.

2. A literature review shows that a systematic approach for identifying weather signals is lacking and that the majority of studies select weather variables from a small number of competing hypotheses that are founded on unverified *a priori* assumptions. This is worrying because studies that investigate the nature of weather signals in detail suggest that signals can be complex. Using suboptimal or wrongly identified weather signals may lead to unreliable projections and management decisions.

3. We propose a four-step approach that allows for more rigorous identification and quantification of weather signals (or any other predictor variable for which data are available at high temporal resolution), easily implementable with our new R package '*climwin*'. We compare our approach with conventional approaches and provide worked examples.

4. Although our more exploratory approach also has some drawbacks, such as the risk of overfitting and bias that our simulations show can occur at low sample and effect sizes, these issues can be addressed with the right knowledge and tools.

5. By developing both the methods to fit critical weather windows to a wide range of biological responses and the tools to validate them and determine sample size requirements, our approach facilitates the exploration and quantification of the biological effects of weather in a rigorous, replicable and comparable way, while also providing a benchmark performance to compare other approaches to.

Key-words: bias, climate change, climate sensitivity, cross-validation, false positive, precision, R package *climwin*, sample size, sliding window, weather

Introduction

Ecology and parts of evolutionary biology concern the study of how organisms interact with their environment. Consequently, a core task is to relate the variation in physiological, behavioural, life-history, demographic, population, species and community responses (henceforth referred to as traits) to the variation in environmental variables, such as food abundance, competitor density and weather conditions. Particularly for studies on climate change and variability, the best choice of environmental predictor is not always obvious, even in well-studied systems. Which weather variables (e.g. rain, temperature) affect the expression of traits, and over which time period (e.g. recent weather, spring or year-round weather) and in what ways (e.g. mean or maximum of temperature)? In some cases, these factors can be experimentally manipulated, but in many cases experiments are impossible or misrepresent responses to climate change in the wild (Wolkovich *et al.* 2012), and weather drivers will need to be identified using observational data.

Using observational data to capture how organisms are responding to a history of multidimensional weather variation is by no means trivial (Stenseth & Mysterud 2005) and requires a systematic approach; yet no such approach is currently available. Studies that investigate the nature of weather signals in detail suggest that signals can be complex (e.g. Gienapp, Hemerik & Visser 2005; Biro, Beckmann & Stamps 2010;

*Correspondence author. E-mail: martijn.vandepol@anu.edu.au

Kruuk, Osmond & Cockburn 2015). However, in most studies, the choice of which weather variables to consider, over which period of the year and which metric to use seems to have no strong justification and is largely based on *a priori* assumptions that are rarely validated. Furthermore, they tend to focus on a narrow range of competing hypotheses. This is concerning because one generally has limited *a priori* knowledge, while there are potentially large numbers of plausible competing weather signal hypotheses.

Using overly simplistic, suboptimal or wrongly identified weather signals that ignore these biological realities can lead to unreliable projections and consequently to inappropriate conservation decisions. For example, if a trait displays no response to weather, it is difficult to determine whether this is evidence of climatic insensitivity or a flawed choice of time period. Even when we find a relationship between weather and the biological response, we cannot be sure that we have selected the period where the response is *most* sensitive. These problems not only hamper projections for single species, but also cloud whether reported interspecific variation in weather sensitivities reflects biological or methodological differences (van de Pol *et al.* 2013). Therefore, a systematic and rigorous method to identify and quantify the weather signals affecting biological processes is urgently needed.

In this paper, we first perform a literature review to describe conventional weather signal selection approaches – and their associated limitations – focusing on the three defining characteristics of weather signals: (i) the identity of the weather variables, (ii) the critical time windows affecting the trait expression and (iii) the aggregate statistics (e.g. mean, max) that best describe the influence of the weather variables over the critical period. Subsequently, we propose a stepwise approach using our new and easy-to-use R package *climwin* (Bailey & van de Pol 2015) to investigate these characteristics in a systematic way and provide worked examples using empirical data set. Finally, we perform simulations to show how our approach can be used to quantify unbiased and precise weather signals, and the sample size required to do so, while avoiding spurious results.

Conventional approaches and their limitations

IDENTITY OF WEATHER SIGNALS

Weather typically affects ecological processes through a mixture of variables (Remmert 1980); consequently when considering candidate weather variables for a signal, the number of possibilities is substantial. In ecology, this problem is tackled using either a confirmatory or exploratory approach. A confirmatory approach uses pre-existing biological knowledge to limit the number of potential variables to a few testable hypotheses (e.g. Frederiksen *et al.* 2014). Although sufficient biological knowledge may be likely for environmental drivers such as food or predator abundance (one can observe what an organism eats or is eaten by and use this to decide what prey or predator species' abundance to include as environmental driver), this is more difficult for weather variables. In some model systems, the ecophysiology or behaviour of an organism may provide clues to identify candidate variables; for example, snow cover is known to affect the feeding behaviour of herbivores (Stenseth & Mysterud 2005). However, we often have limited *a priori* knowledge about weather influences because direct weather effects are typically hard to observe, may exhibit time-lags and weather may affect organisms indirectly (e.g. via food). In such situations, a more exploratory approach, in which a wider range of weather signal hypotheses is being tested, may be preferable.

A systematic review of the literature (see Appendix S1 for methods; $N = 50$ studies, unless stated otherwise) showed that often the choice of weather variables is confirmatory (66%), typically based on a previous study on a different population or species. However, making choices from previous studies can be fraught as weather sensitivity might vary between environments (Phillimore *et al.* 2010), differ between closely related species (van de Pol *et al.* 2013), and the choice of weather variable in the reference study may also lack justification. Only 6% of studies specifically stated that they used an exploratory approach, while 28% of studies gave no justification for the choice of weather variables.

Furthermore, most studies only considered a single weather variable (Fig. 1a; variables considered: 59% temperature, 20% precipitation, 8% large-scale oceanic climatic indices, 13% other). The studies that did consider multiple variables generally lacked methods to deal with collinearity (91% of studies; 20 out of 22), despite the fact that strong correlations are often expected (e.g. sunny warm weather generally means low rainfall).

CRITICAL TIME WINDOWS

More often than not the period over which weather is deemed to be important for a trait (critical time window) appeared to be chosen *a priori* and little justification is provided, with most studies (62%) not refining the time window beyond an annual or seasonal mean. Moreover, few studies considered competing time windows (Fig. 1b). For example, in birds, the variation in the timing of egg laying is typically related to spring temperatures (e.g. Crick & Sparks 1999) and annual survival rate to winter temperatures (e.g. Grosbois *et al.* 2008), yet temperatures during other periods and shorter resolution time windows are rarely considered (e.g. McLean *et al.* 2016). Considering a variety of time windows is not only important to identify the 'best' possible window, but it also helps to distinguish the potentially co-occurring effects of short-lag (more recent) and long-lag (more distant) weather signals that could be acting at different stages of an organism's life cycle (Fig. 2; van de Pol & Cockburn 2011). For example, high temperatures during winter may have positive effects on summer reproductive performance, while a recent sequence of hot summer days can have negative effects when tolerance thresholds are exceeded (Kruuk, Osmond & Cockburn 2015).

In addition to considering competing time windows that vary in duration and lag time, the choice of the type of time window – absolute or relative (Fig. 2; Box 1) – becomes

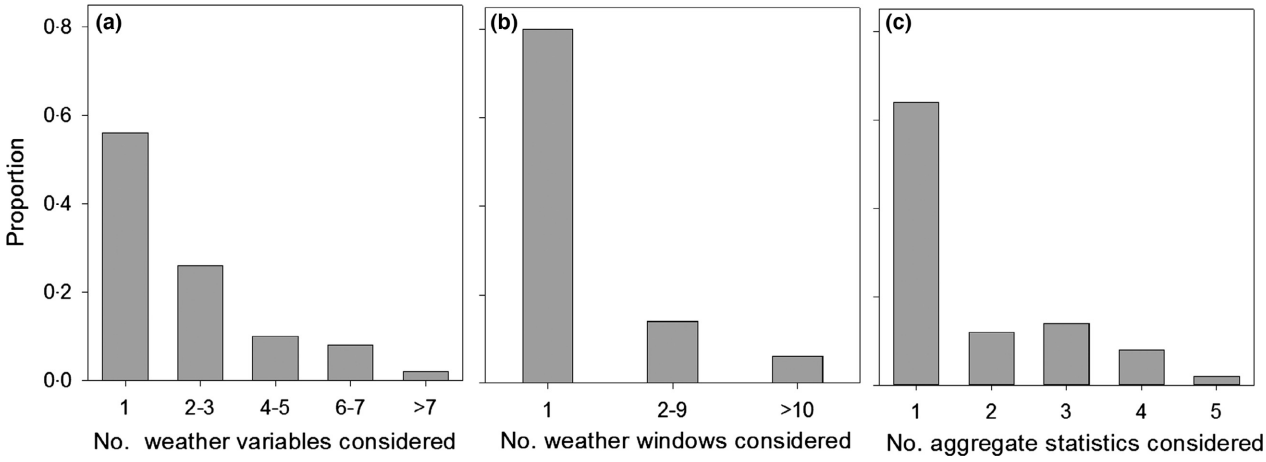


Fig. 1. Histograms of the number of different (a) weather variables, (b) weather windows and (c) aggregate statistics considered per study ($N = 50$ studies, see Appendix S1 for details).

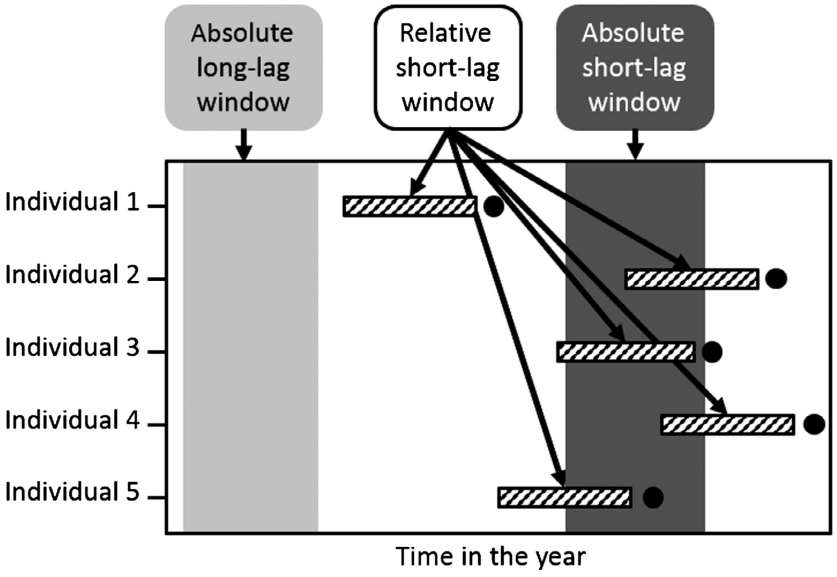


Fig. 2. Graphical explanation of the difference between short- and long-lag time windows and between absolute and relative windows. Short- and long-lag time windows are, respectively, more recent or distant since the timing of trait expression. Individual variation in the timing of expression of a trait (black circles) affects the choice of whether to use absolute time windows (i.e. the same window for all individuals; grey areas) or relative time windows that depend on the individual and the time when the trait was expressed (striped areas). Relative time windows assume that, instead of using an absolute period (e.g. March–April temperatures), windows are relative to the timing of expression of a trait for each individual (e.g. the temperature in the month preceding the time of trait expression by each individual).

particularly crucial when traits are expressed at different times among individuals (Gienapp, Hemerik & Visser 2005; van de Pol & Cockburn 2011). However, relative windows were rarely considered (6% of studies).

CHOICE OF AGGREGATE STATISTICS

In our review, the most common choice of aggregate statistic was the mean of a weather variable over a given period (55% of studies). However, developmental studies have focused on cumulative measures such as (growing) degree or chill days (19% of studies) and studies where physiological tolerance limits of an organism can be exceeded have used maximum or minimum weather values (15% of studies; see extreme events literature; Bailey & van de Pol 2016). Sometimes it is not the absolute value of a weather variable that affects trait expression, but the seasonal rate of change or daily range (Biro, Beckmann & Stamps 2010; Schaper *et al.* 2012). Finally, not all days within a period are necessarily equally important, with more recent weather potentially having a stronger influence than weather in the more distant past (Gienapp, Hemerik & Visser 2005; van de Pol, Osmond & Cockburn 2012). Such patterns – reflective of a fading memory – can be described using weighted means (van de Pol & Cockburn 2011). Evidently, several biologically plausible choices of aggregate statistic exist, but few studies actually compared competing hypotheses (Fig. 1c). This is concerning, because studies that have made this comparison illustrate that results can strongly depend on the choice of aggregate statistic used (Charmantier *et al.* 2008; Husby *et al.* 2010).

A stepwise systematic approach towards more rigorous weather signals using R package *climwin*

Our literature review showed that there is currently no systematic approach to identify the weather signals affecting biological processes. Furthermore, the typical practice of considering only a limited range of hypotheses, often founded on unverified *a priori* assumptions, seems at odds with how little we still

Box 1. Absolute vs. relative time windows.

The choice of the type of time window – absolute or relative (Fig. 2) – becomes particularly crucial when traits are expressed at different times among individuals (Gienapp, Hemerik & Visser 2005; van de Pol & Cockburn 2011). The reason for this is that individuals that express traits at different times of the year are likely to have been affected by weather over different time windows. Even quantitative traits such as offspring size and reproductive success are often expressed at variable times among individuals, because the moment traits can be quantified typically depends on phenological events (e.g. individuals from the same cohort vary in their natal weather conditions if some offspring are born earlier than others and weather varies during the season). Assuming the same absolute time window (e.g. June temperature) for all individuals is unlikely to be appropriate if the timing of trait expression varies substantially among individuals and if the time-lag is short (i.e. if some individuals reproduce in May instead of July, then they cannot be affected by June temperatures). In such cases, the use of relative time windows (e.g. temperature during the month preceding reproduction) is needed that cover different periods for early- and late-reproducing individuals (Fig. 2; van de Pol & Cockburn 2011).

An advantage of using absolute windows is that the weather windows are easier to interpret and use in future projections, as there is only one weather window for the entire population. However, instead of viewing this as a drawback of using relative windows, it should be seen as a biological reality that time windows can be heterogeneous. Future projections on the biological consequences of climate change in situations of relative windows can still be made using numerical simulations (van de Pol, Osmond & Cockburn 2012). Finally, the choice of either using an absolute or relative time window can amount to asking slightly different biological questions (see Appendix S2 for more details).

know about how weather affects organismal functioning in most species. Therefore, we propose a four-step approach that investigates a broader set of competing hypotheses concerning the choice of weather variable, time window and aggregates statistic used (Fig. 3). This more exploratory approach is not meant to be exhaustive, but primarily to widen the number of competing hypotheses beyond the small number of confirmatory hypotheses typically considered.

Our stepwise approach is easily implementable with the new R package *climwin* (Bailey & van de Pol 2015; Fig. 3), and we illustrate our approach with an example data set. The R code and the detailed description of analyses are available in Appendix S2; here, we will focus on the key outcomes and their interpretation. Although our approach is developed for weather variables, it can also be used for any predictor variable for which data are available at high temporal resolution. For example, studies that repeatedly measure dominance scores, food abundance or body size during the year could use the tools developed here to investigate over which period such variables best explain the trait variation.

STEP 1: DETERMINE A BASELINE MODEL STRUCTURE WITHOUT WEATHER EFFECTS AS A NULL HYPOTHESIS

Our overall approach will be to compare the support by the data for competing hypotheses formalized into regression models. To assess the performance of competing models, one needs to have a yardstick, which will be a baseline regression (null) model containing no weather effects, but that can include other confounding predictor variables (e.g. the sex of an individual, a random effect of study site; see Fig. 3 for R code example). The *climwin* package allows for baseline models using most types of regression models that can be fitted in R (models that return a likelihood or AIC; we have tested '*lm*', '*glm*', '*lmer*', '*glmer*', *'coxph';* the main constraint in adding more type of models is differences in syntax used among packages).

STEP 2: CREATE A CANDIDATE MODEL SET BY IDENTIFYING ALL COMPETING HYPOTHESES THAT REQUIRE TESTING

In the second step of our approach (Fig. 3), we create a candidate model set by identifying all competing hypotheses. The first substep (step 2a) in this process involves identifying all weather variables (temperature, precipitation, etc.) that could be of potential interest. Ultimately, our aim will be to limit the candidate variables to a reasonable number (Grosbois *et al.* 2008). Furthermore, weather variables often exhibit strong collinearity, which needs to be dealt with. However, before reducing the number of weather variables, for each weather variable of interest we first have to make various choices about – and therefore a more detailed investigation into – the critical time window (step 2b and c), the best aggregate statistic (step 2d) and function of the relationship (step 2e) for describing the biological response to each weather variable. Since the choice of best weather variable may depend on the choice of time window, aggregate statistic or response function of the relationship, it is not clear that one aspect can be investigated sequentially or independently of the other. Therefore, we propose to try all combinations of choices made in step 2b-e for each weather variable to identify the time window(s), aggregate statistic(s) and response function(s) that are best supported by the data. Once such candidate signals are identified for each variable (using single variable models; step 3), their number can be subsequently reduced using conventional methods for multiple variable model selection and dealing with collinearity (step 4).

To compare the different time windows, one needs to decide in step 2b whether it is biologically more appropriate to use absolute or relative time windows (Fig. 2; Box 1). If there is large individual variation in the timing of expression or measurement of a trait, or short lag times of the weather signal are expected, a relative window (e.g. temperature during the month preceding reproduction) may be more biologically relevant than an absolute time window (e.g. June temperature). If there is no clear *a priori* expectation, one can try both window types and compare their model support in step 3.

In step 2c, we must decide the period over which we should look at each weather variable. Computationally, it is now possible to try many time windows and decide from such an analysis what the most appropriate window is (sometimes called a sliding or moving window approach; e.g. Husby *et al.* 2010). Consequently, we suggest testing a wide diversity of time

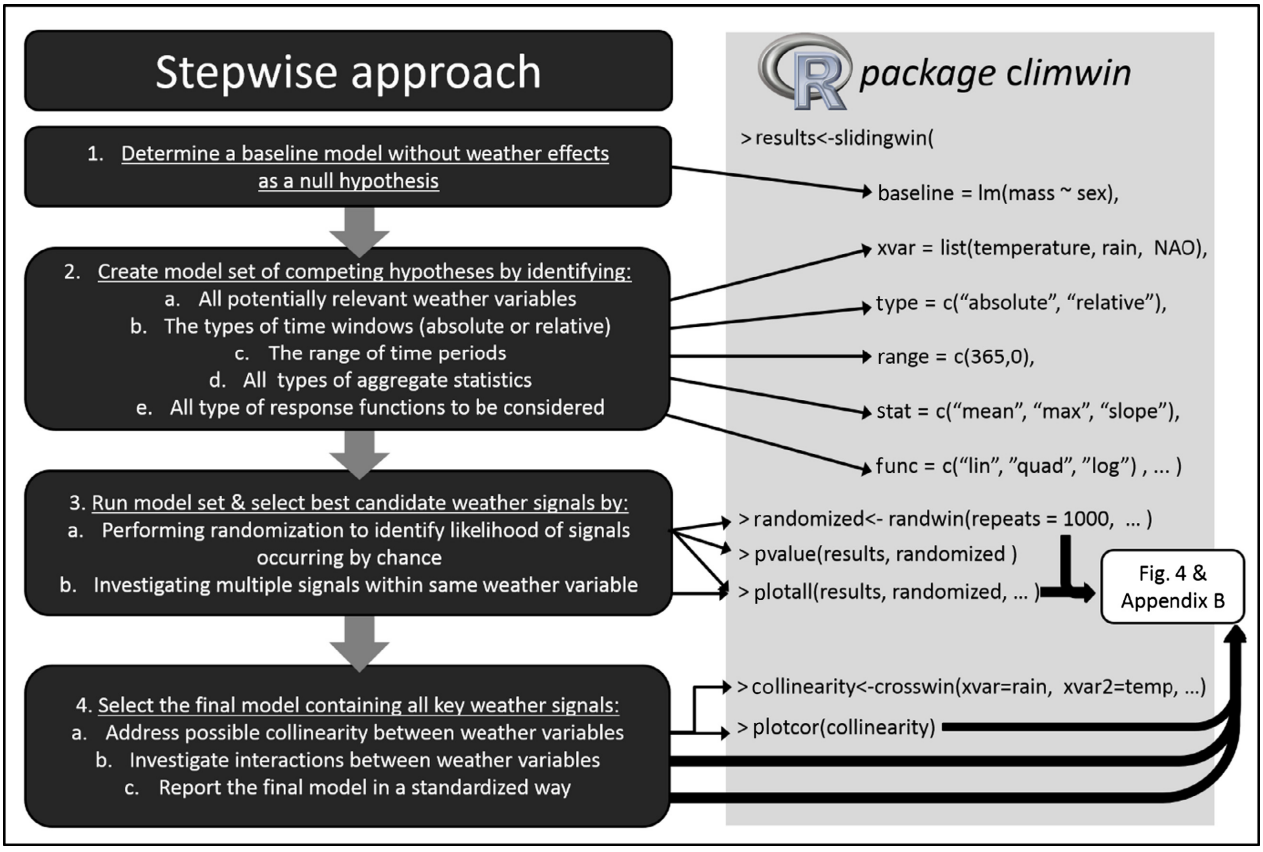


Fig. 3. The four steps used in our systematic approach to determine the best weather signals for a specific response, and an illustration of how the R code from package *climwin* implements those steps.

windows, also to investigate the possible effects of weather signals with both short and long lag times (Fig. 2). For example, when looking at the effects of temperature on summer body mass, all possible combinations of time windows within the previous year can be investigated (see R code in Fig. 3), as summer body mass may depend on spring temperatures, but carry-over effects of winter temperature may also be plausible (Harrison *et al.* 2011). Ideally, time windows are varied at a daily resolution, as this avoids the rather arbitrary use of monthly or seasonal data typically used in existing studies and allows for the detection of short signals of a few days (Kruuk, Osmond & Cockburn 2015).

Next, in step 2d, a decision is made on the aggregate statistic used to summarize the weather variable over each time window. The choice of aggregate statistic(s) to be considered can be driven by the possible biological mechanism involved, while in systems with limited mechanistic knowledge one could explore several statistics.

In the final step of model set identification (step 2e), we choose the response functions to be considered. Many fields strongly focus on linear relationships (e.g. reaction norms), but this is probably mostly driven by the need for simplification. In reality, trait values often peak at a certain optimum weather value (e.g. thermal performance curves; Angilletta 2009) and the fact that threshold values are regularly used as an aggregate statistic emphasizes that the responses of traits to weather signals can be nonlinear. Sometimes the shape of the response curve may even be of interest in itself: the effects of environmental variability on population dynamics may depend on the curvature of the response of demographic or population growth rates to weather (Lawson *et al.* 2015).

STEP 3: RUN MODEL SET AND SELECT BEST CANDIDATE WEATHER SIGNALS

In the right panel of Fig. 3, we illustrate how the function '*slidingwin*' from the *climwin* package can be used to automatically translate all hypotheses considered in step 2 into a set of many thousands of single variable regression models (see Appendix S2 for details). In step 3, we fit each of these models to the biological data and compare and interpret their output to (a) distinguish real weather signals from false-positive signals inevitably occurring by chance due to the testing of a large model set and (b) identify multiple (short- and long-lag) weather signals within the same weather variable. The time it takes to run all the models can vary from minutes to days, depending on the sample size, model complexity and computer speed. We can then use the results from these steps to select typically a few candidate weather signals for each weather variable for further analysis. To compare the empirical support for each of the different regression models, *climwin* uses the information-theoretic model selection criteria AICc, with the option of using *K*-fold cross-validation to address issues of overfitting (Box 2).

Box 2. Criteria to determine which model is 'best'.

The R package *climwin* uses the Akaike Information Criterion (AIC; Akaike 1973) to compare support for the different models. Information-theoretic criteria trade off goodness-of-fit with model complexity and allow for direct comparison of non-nested models that have variable numbers of parameters (e.g. models with different response functions):

$$\mathrm{AICc}_{\mathrm{model}} = -2\log(L) + 2P + \left(\frac{2P(P+1)}{N-P-1}\right), \quad \text{eqn 1}$$

where L is the likelihood of the data given the model, P is the number of estimated model parameters, and the term $(\frac{2P(P+1)}{N-P-1})$ is a small-sample size correction that is negligible if the sample size N is large (Burnham & Anderson 2002). To facilitate the comparisons among models, *climwin* compares the AICc for each model i relative to the support for the baseline model without a weather effect:

$$\Delta\mathrm{AICc}_{\mathrm{model}\,i} = \mathrm{AICc}_{\mathrm{model}\,i} - \mathrm{AICc}_{\mathrm{baseline\ model}} \quad \text{eqn 2}$$

This metric is used to decide which model in the model set has the strongest model support (the model with the lowest $\Delta\mathrm{AICc}_{\mathrm{model}\,i}$).
As a metric for model diagnostics, *climwin* also determines the Akaike weight of each model:

$$\text{Akaike weight}_{\mathrm{model}\,i} = e^{-\frac{1}{2}\Delta\mathrm{AICc_i}} / \sum_{j=1}^{J} e^{-\frac{1}{2}\Delta\mathrm{AICc_j}}, \quad \text{eqn 3}$$

where the sum of all weights across all models J considered add up to one ($\sum_{j=1}^{J}$ Akaike weight$_j$ = 1). Simulations showed that the proportion of all models from the candidate set that is in the 95% model confidence set (i.e. together account for the top 95% of the total Akaike weight across all models) is a very useful measure (henceforth metric C) to distinguish false from true positives. If C is close to zero, this means that a small subset of all tested models receives 95% of all model support (in terms of Akaike weights) and this is what we typically found to be the case for true signals, while if C is close to one, then almost all models are roughly equally well (or poorly) supported and this is what one would expect if there is no true climate signal. By comparing C of the observed data to the distribution of C in randomized data in combination with the sample size N, one can quantify the probability P_C whether the candidate signal in the observed data is likely to be due to chance or not (see worked example and Appendix S3 for details).

Alternative Model Selection Based on Cross-Validation

The model set typically involves a huge number of models, which consequently increases the risk of overfitting (i.e. the amount of variation that the best weather signal will explain in the response variable (R^2) may be systematically overestimated). Using K-fold cross-validation for model selection can reduce such biases (Arlot & Celisse 2010) and can be implemented in *climwin* by setting the argument 'K' in the *slidingwin* function to, for example, K = 10. K-fold cross-validation divides the data set into K training data sets (of length N-N/K) and K test data sets (of length N/K, with $K \leq N$). Subsequently, each model is first fitted to one of the training data sets and its predictive accuracy tested on the corresponding test data set. To measure the predictive accuracy, the mean square error (MSE) of the training fit to the test data is used to calculate the AICc:

$$\mathrm{AICc}_{\mathrm{model}} = N\log(\mathrm{MSE}) + 2P + \left(\frac{2P(P+1)}{N-P-1}\right) \quad \text{eqn 4}$$

and subsequently compared to the fit of the baseline model (sensu eqn 2). This procedure is repeated K times (once for each test data set), after which the $\Delta\mathrm{AICc}_{\mathrm{model}}$ is averaged across all folds to obtain the cross-validated $\Delta\mathrm{AICc}_{\mathrm{model}\,i}$. Similar to above, this cross-validated $\Delta\mathrm{AICc}_{\mathrm{model}\,i}$ can be subsequently used for identifying the best supported model in the model set and to calculate metric P_C used for model diagnostics.

The '*plotall*' function provides several tools for visual interpretation of results (see code Fig. 3). In Fig. 4a, we illustrate how ΔAICc (the AICc difference between candidate and null models) can be used to compare the effects of mean temperature on the egg-laying date of British Chaffinches over different absolute time windows (data 1966–2012; Baillie *et al.* 2014). The best supported time window during which mean temperature affects laying date is effectively the 2 months before egg laying (Fig. 4a). Mean annual laying dates have advanced with 4·3 days/°C over this critical period (Fig. 4b). Many neighbouring windows are almost equally well supported (broad red peak in Fig. 4a), as could be expected due to their overlapping periods and due to temporal autocorrelation in weather, and their biological effect size is very similar (Fig. 4c).

The model support for the best time window (Fig. 4a) can be directly compared to other models using different response functions, aggregate statistics or types of windows (absolute or relative). For example, there was equal support for both a quadratic and linear response of mean temperature on Chaffinch laying date, while models using mean temperature were much better supported than models using temperature degree days or rate of temperature increase (Table 1). In cases where the mean of a weather variable is the best supported aggregate statistic, it can be worthwhile to explore the use of a weighted mean, as this may allow for further refinement of the weather's temporal signal (Box 3).

In step 3a, *climwin*'s randomization function can be used to quantify the likelihood of obtaining such a strong model support by chance (in this case for a linear effect of mean temperature on Chaffinches laying dates) due to the high number of models tested (step 3a, see Bailey & van de Pol 2015). Ideally, one performs thousands of randomizations and compares the ΔAICc of the best model fitted to the observed data to the distribution of ΔAICc values from the best model in each randomized data set. It should be noted that by chance even some of the many models fitted using the randomization method can achieve ΔAICc scores that would be considered evidence for strong model support by conventional standards

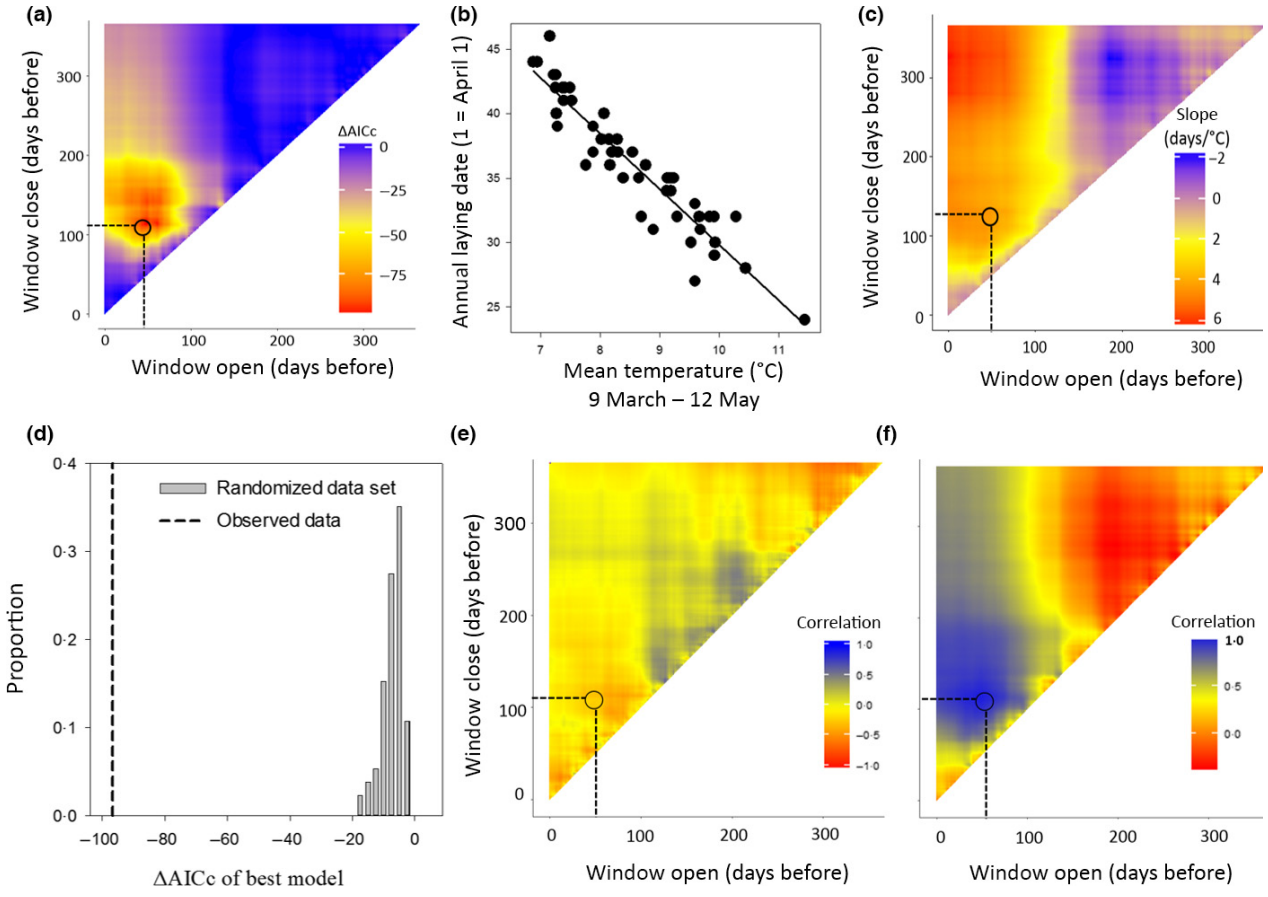


Fig. 4. Illustration of the functions and visual output available in *climwin* to select the best weather window models from a model set. Output shown is produced by functions 'plotall' (a–d), 'crosswin' (e) and 'autowin' (f). (a) The difference in model support (ΔAICc) for the different time windows of an effect of mean temperature on Chaffinch' laying date compared to a null model with no weather effect included. The circle and dotted lines point towards the time window that was best supported by the data (from day 49–113 [12 May-9 March]). (b) The relationship between temperature and Chaffinch annual mean egg-laying date for the best supported time window. (c) The slope estimates for the relationship between temperature and egg-laying date for each time window in the model set. (d) The distribution of the ΔAICc values of the best supported model in each of the 1000 randomized data sets (grey bars) can be compared to the ΔAICc value of the best supported model in the observed data set (dashed line) to determine the likelihood an observed signal is real. The models assumed absolute time windows going back 365 days from 30th of June (British Chaffinches generally lay in May) and investigated the linear effects of mean temperature. (e) The correlation between mean temperature and rainfall sum over different time windows. (f) The correlation between the mean temperature during each window with the mean temperature in the best supported time window.

Table 1. Model support (ΔAICc compared to a model with no temperature effects included) for the best time windows using different aggregate statistics and response curves. 'Mean' refers to mean temperature, 'degree days' to a model that sums the temperatures of days above 10 °C and 'rate of increase' to the mean rate of temperature increase per day during the critical window. The values in between brackets indicate the time window belonging to the lowest ΔAICc value, as illustrated in Fig 4a

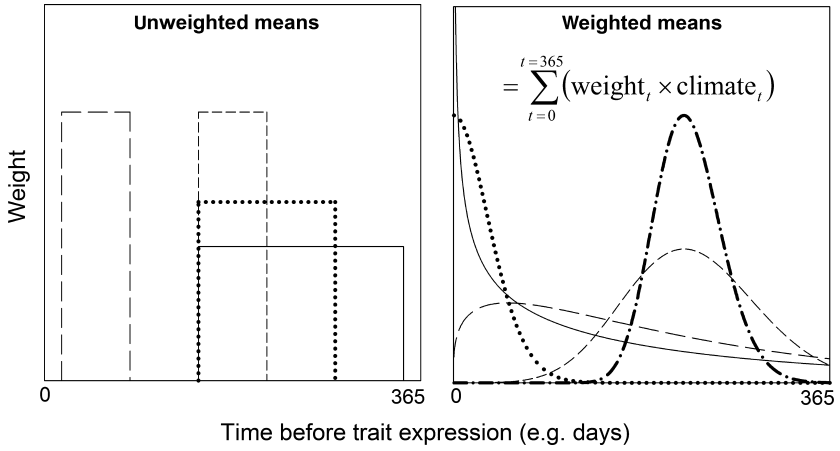
	Aggregate statistic		
	Mean	Degree days	Rate of increase
Response curve			
Linear	−96·8 (49–113)	−58·1 (36–148)	−55·5 (46–322)
Quadratic	−96·9 (49–113)	−59·4 (71–152)	−53·5 (46–322)

(ΔAICc < −5; Burnham & Anderson 2002). This shows that randomization is a necessary step to assess the chance of a candidate signal being a false positive (and minimize type I errors). Comparing the ΔAICc value from our Chaffinch example (ΔAICc = −97) to randomized data sets with no weather signal shows that none of the 1000 randomized data sets displayed such a high level of model support (Fig. 4d), indicating that such a strongly supported temperature signal is very unlikely to have occurred by chance ($P_{\Delta AICc} \leq 0{\cdot}001$; see Appendix S2 for other examples with less strong weather signals). In practice, for some data sets carrying out so many randomizations may take too much computation time. For such situations, we have developed an alternative statistic for $P_{\Delta AICc}$, namely P_C, that requires much less randomizations (5–10) but still gives a reliable indication of whether a signal is spurious or not (here P_C = 1·7E-05; for details, see Appendix S3).

In some situations, the ΔAICc landscape of the different time windows shows multiple peaks instead of a clear single peak as in Fig. 4a. This can indicate the presence of multiple (e.g. both long- and short-lag) weather signals within the same weather variable, but it can also occur due to collinearity or

Box 3. Weighted means in climwin.

Arguably, aggregating weather over a certain time window by weighting each day in that window equally (i.e. by calculating the unweighted mean) may generally be biologically unrealistic, as more recent weather could have a stronger influence than weather in the more distant past (Gienapp, Hemerik & Visser 2005; van de Pol, Osmond & Cockburn 2012). Weighted means can account for such gradual decay effects and also do not have the abrupt change in influence that normal (unweighted) means have at the beginning and end of the window.



climwin allows for testing of weighted means via the function '*weightwin*' based on the methods described in van de Pol & Cockburn (2011). Since the weight function used to calculate the weighted mean needs to be estimated (it has a shape, location and width parameter, reflecting, respectively, the decay in weight/importance, the lag time and duration of the time window), and can take on an infinite number of forms, *weightwin* uses different optimization methods than *slidingwin* to find the best window described by the weighted mean function. Nonetheless, the output from *weightwin* that describes the best supported weather signal can be directly compared to the output from models fitted by the *slidingwin* function to investigate whether a weighted mean model is better supported by the data than, for example, a model with the aggregate statistic unweighted mean (see Appendix S2). For alternative nonparametric methods using smoothing, see Roberts (2008) and Teller *et al.* (2016).

chance. In step 3b, the evidence for multiple signals can be investigated by adding the best supported of the two weather windows to the baseline model, and re-fitting all the different time windows again: this tests whether there is still strong model support for the second best (e.g. short-lag) weather window once the other best supported (e.g. long-lag) weather window has been accounted for in the baseline model (see Appendix S2).

By repeating step 3 a&b for each weather variable, we can select the candidate signals for each weather variable. For some weather variables, there will be no candidate signals if the model support for the best combination of time window, aggregate statistic and response function is no higher than those observed in the randomized data. For other weather variables, there may be either one (Fig. 4a) or possibly several candidate signals (Appendix S2; it should be noted that multiple effects of a single variable such as temperature may be biologically plausible, but nonetheless statistically hard to detect).

STEP 4: PERFORM MODEL SELECTION TO SELECT THE FINAL MODEL CONTAINING ALL WEATHER SIGNALS

Sometimes we may end up with a large number of candidate signals. In step 4, we aim to (a) reduce the number of potentially intercorrelated weather signals and (b) explore the possible interactions between weather signals in order to (c) report in a standardized way the final multiple variable model that contains all important weather signals. Reducing the number of collinear variables is a common problem and other papers describe established methods well (e.g. Freckleton 2011; Grueber *et al.* 2011). Notwithstanding, *climwin* offers two specific functions to explore the degree of correlation among and within weather variables over different time windows: '*crosswin*' and '*autowin*'. Figure 4e illustrates that the correlation between weather variables (here mean temperature and sum of rainfall in the UK) can be weak in some parts of the year but strong in others, highlighting that dealing with collinearity is most sensible once the critical time windows are known. Figure 4f illustrates that mean temperatures are typically strongly correlated among nearby overlapping time windows, which explains why a wide range of adjacent time windows can receive high model support (see red peak in Fig. 4a).

Interactions between weather signals have rarely been explored, and it is thus unclear how common and strong they might be. One way to investigate the interactions by means of proxy is to replace, for example, the temperature and rainfall variables in the model set (step 2a) by a single weather-derived variable that integrates the interactions between temperature and rainfall (e.g. a drought severity index). A more direct way is to include two-way interactions between the temperature and rainfall candidate signals. In step 4b, one could investigate such interactions as part of the model selection procedure to identify the final model containing all important weather signals.

In the final step, 4c, we suggest that the output should be reported in a standardized way such that effect sizes and hypotheses considered can be easily compared among studies. Reported effect sizes could be based either on the estimates

from the single final model containing all important weather signals, or on model averaging of effect size estimates (Grueber *et al*. 2011) across all models considered in the model selection process of step 4 (but see Cade 2015). To improve the interpretation of effect sizes and interactions, we suggest rescaling covariates (Schielzeth 2010). Furthermore, for comparisons among studies, it is important to report the model set considered (steps 2 & 4). Finally, for future meta-analyses, one should report the variability in both the weather signal and biological response variable to facilitate standardized comparison, and present all model parameters that are needed to reconstruct response functions (e.g. including the estimates of intercepts and random effects in the case of logistic or Poisson regression). Online archiving of *climwin* R code and the data used also contributes to these goals.

Performance of our approach and sample size considerations

A crucial step is to assess how well our approach actually performs in identifying weather signals. This requires a statistic to decide whether a weather signal is real or not, and what its associated rate of misclassification is (i.e. type I/II error rate). Furthermore, a question that has received surprisingly little attention in the literature is: 'How many different environments (years, sites; i.e. sample size) should be measured to identify and estimate weather signals precisely and accurately?' Investigating this question requires deciding what one would like to estimate precisely or accurately. Statistical modelling in general, and studies on climate change ecology in particular, have two goals: explanation and prediction (Shmueli 2010). To explain which weather signals are most important, we need unbiased estimates of their explanatory value (R^2). To predict (via extrapolation) what future effects of weather on the response variable will be, we need to estimate the slope of the relationship between the weather signal and biological trait both accurately and precisely, and identify the period of the time window correctly (as climate change can cause different parts of the year to change variably).

To assess the rate of misclassification (false positives and negatives) and determine the accuracy and precision of model characteristics (R^2, slope and time window location), we generated simulated data sets. We created data sets with one biological response measurement for each sampled environment (e.g. the mean laying date in a given year or site) based on the previously introduced Chaffinch data set (i.e. assuming a linear effect of mean temperature between March 9th – May 12th). We generated 1000 data sets each for a wide range of sample sizes (10, 20, 30, 40 or the original 47 data points) and effect sizes (the 'true' R^2 of the underlying model was set to either very high (0·80), high (0·40) and moderate (0·20), while keeping the slope constant). Additionally, we generated and analysed data without any weather signal (for R code and details, see Appendix S4). Note that a sample size of 30 may reflect a single location followed over 30 years, 30 locations with varying climates measured in 1 year, or any combination in between. Subsequently, we randomized (step 3a in Fig. 3) each simulated data set five times to calculate the P_C statistics from *climwin* (see Appendix S3) to classify signals as either true or not. The rate of false negatives was calculated as the proportion of simulated data set that contained a true weather signal, but was misclassified as containing no signal (i.e. $P_C \geq 0{\cdot}5$). The rate of false positives was calculated as the proportion of simulated data set that contained no weather signal, but was misclassified as containing a weather signal (i.e. $P_C < 0{\cdot}5$).

When the effect sizes were high ($R^2 = 0{\cdot}80$ or 0·40), we found the rate of false negatives to be very low, even with a low sample size (Fig. 5a-i, a-ii). However, when effect sizes were moderate ($R^2 = 0{\cdot}20$), low levels of false-positive rates (<10%) were only achieved with a relatively high sample size ($N > 47$; Fig. 5a-iii). The rate of false negatives also strongly depended on sample size, with low false-negative rates (<10%) only achieved with relatively large sample sizes ($N \geq 30$; Fig. 5a-iv). Notably, we had to set an arbitrary cut-off point for our statistic during these simulations ($P_C < 0{\cdot}5$) to decide whether we considered a climate signal to be 'real'. However, in practice, the value of P_C will give additional information on the certainty of a given climate signal (i.e. $P_C = 0{\cdot}1$ and $P_C = 0{\cdot}4$ are both likely a true signal, but the likelihood is much higher for $P_C = 0{\cdot}1$), and users can use different cut-off values depending on whether they think false positives or negatives are most problematic for their study.

In the simulations, R^2 was greatly overestimated at low sample sizes ($N \leq 20$) (Fig. 5b). This bias could be expected due to overfitting and was indeed substantially reduced by using 10-fold cross-validation (see Box 2; some bias remains because the best model has itself been selected based on its cross-validation score from a large set of candidate models; Gelman, Hwang & Vehtari 2014).

Simulations were able to estimate the true slope of the climate signal with high accuracy, as long as only those climate signals that were classified as true signals were considered (i.e. $P_C < 0{\cdot}5$; Fig. 5c). Nonetheless, some underestimation, particularly at low effect size, occurred when cross-validation was used (Fig. 5c). The choice of using cross-validation may therefore depend on whether one is most interested in explanation or prediction (i.e. accurate R^2 or slope, respectively). Generally, our method selected for windows that were too short (Fig. 5d; although this bias largely disappeared with increasing effect and sample size). The reason for this bias seems to be that in situations of small sample or moderate effect size, very short spurious windows were best supported by the data (as suggested by the duration of 'best windows' in the randomized data Fig. 5d-iii).

Precision of the slope was generally low at low sample size ($N = 10$) and substantially improved as sample size increased ($N = 20$–30), although precision did not improve much past this point (Fig. 5c). For example, when $N = 10$, $R^2 = 0{\cdot}2$ and only signals classified as 'true' are considered (i.e. $P_C < 0{\cdot}5$; without using cross-validation), 19% of the simulations over- or underestimated the slope by a factor of two or more, and 5% of the simulations estimated the direction of the slope incorrectly (Fig. 5c-iii).

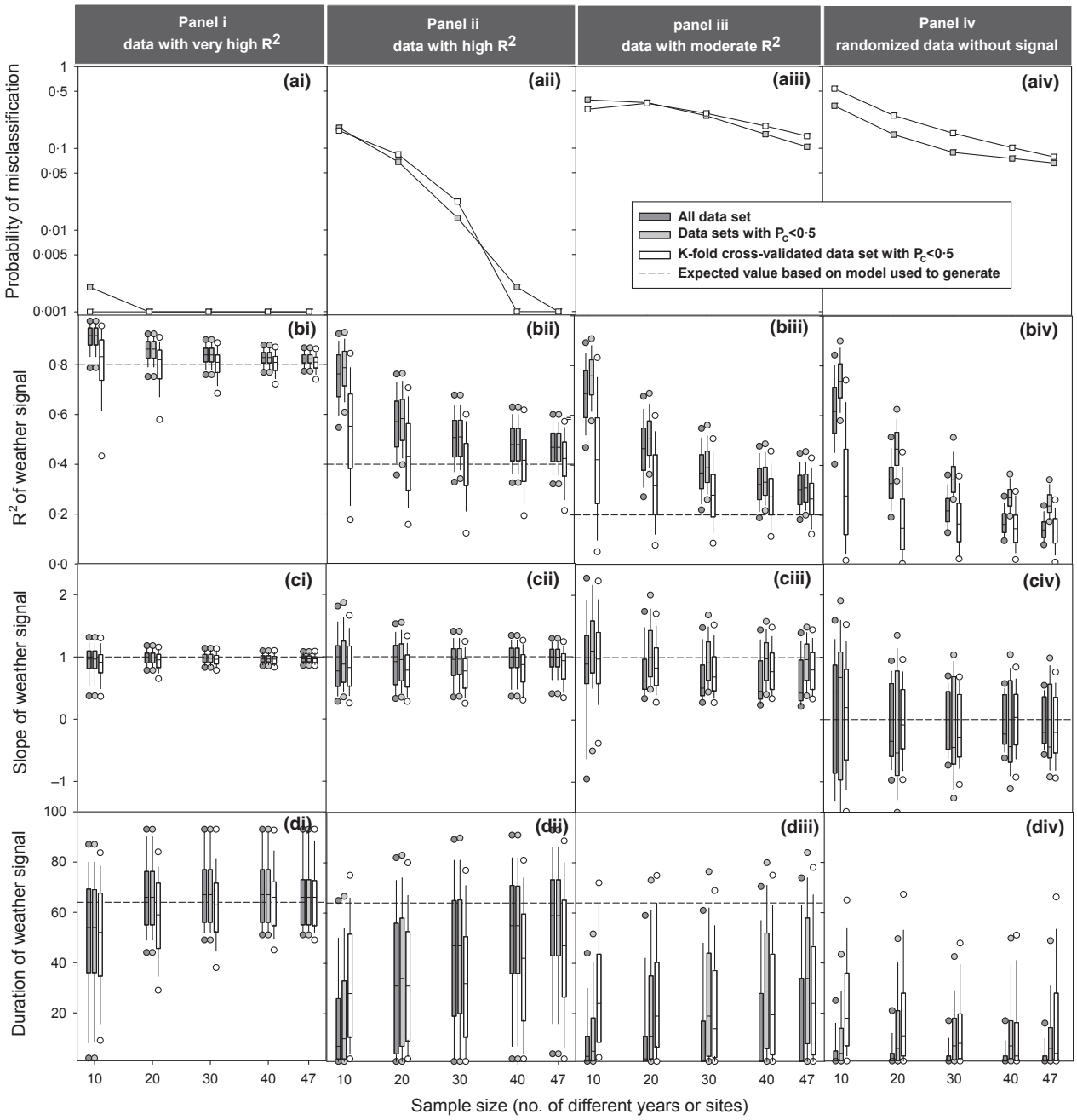


Fig. 5. Performance of *climwin* package in estimating the 'true' weather signal in situations with very high $R^2 = 0{\cdot}80$ (panel i), high $R^2 = 0{\cdot}40$ (panels ii), moderate $R^2 = 0{\cdot}20$ (panels iii) or no weather signal (panel iv) under varying sample sizes (x-axes). Simulations focused on the rate of (a-i to a-iii) false negatives and (a-iv) false positives, (b) the R^2 and (c) the slope of the weather signal, and on (d) the duration of the time window. In panels (b)-(d), each box represents the 50 quantile range (with line for the median), with bars for the 75 quantile and dots for the 95 quantile, based on 1000 analysed test data sets. The different series represent estimates from either all data sets, data sets in which *climwin* classified a signal to be present ($P_C < 0{\cdot}5$), either with or without using cross-validation (see main text and legend). Note that in panels (a) *y*-axis are logarithmic and values are truncated at ≤0·001.

Overall, our simulations show that our approach can detect and estimate weather signals without substantial bias when sample and/or effect size is large. Inaccurate estimation is a problem when sample and/or effect size is low, but most bias can be avoided as long as one uses the right methods. We encourage the use of our *P*-statistic to filter out 'false' signals and avoid biased slopes, and recommend cross-validation to avoid overestimation of R^2 when sample size is low ($N \leq 20$). Furthermore, our simulations show that the precision of the weather slope estimate is low at small sample sizes, reminding us that measurements often need to be collected over long time periods or in many sites before reliable conclusions can be reached.

Our simulations covered a wide range of scenarios, including the challenging case of using *climwin* with a very small data set and moderate effect size. However, within this context, our

results should be seen as a best-case scenario: we assumed that the aggregate statistic and response function were known, that there were no confounding variables, nor multiple or interacting weather signals, and we did not consider binary or count response variables. Furthermore, results may depend on the structure of the data, such as the number of measurements per year or the degree of temporal autocorrelation in the weather variable (e.g. rainfall typically has lower autocorrelation than temperature). This analysis should thus be taken as a first step towards identifying the potential pitfalls of weather window selection and methods to circumvent them; further simulation studies incorporating a wider range of weather data and biological response structures would help to expand and generalize these basic principles.

Alternatives, limitations and future avenues

Roberts (2008) and Teller *et al.* (2016) have suggested alternative explorative methods to identify the critical time window, but their ability to distinguish true from false signals and accuracy and precision of most of the key metrics are unknown. These studies used multiple regression methods in which each daily, weekly or monthly mean temperature is used as a separate predictor variable, and subsequently identified which predictor variables over which time window best explain the variation in the response variable. They employed penalized (ridge) regression and smoothing functions to deal with collinearity and identify contiguous predictor variables (e.g. months) during which the weather signal is strongest. The results from these alternative methods can be used to derive a weighted mean (sensu Box 3), but are not applicable to other aggregate statistics. The advantage of multiple variable methods over our single variable method is that they utilize statistical frameworks (LASSO, machine learning) that are particularly suitable for dealing with correlated variables, meaning they can identify the best time window of multiple weather variables simultaneously within a single model, instead of sequentially as in our method.

Further research is needed to determine the performance of different methods on the same simulated data over a wider part of the parameter space and different data structures, while keeping in mind that different biologists are interested in optimizing the reliability of different metrics (slope, R^2, false-positive or negative rate). Our aim is to extend *climwin* to include a variety of methods and provide the tools and benchmarks to compare them, as the question of what constitutes the best method may depend on the biological question (Teller *et al.* 2016; this study). Another interesting avenue would be to adapt our approach to the question of over which spatial window one should aggregate environmental predictors (Mesquita *et al.* 2015), as for species moving between various locations, the locations at which the weather influence is strongest may in fact need to be determined (note that *climwin* can already incorporate weather data from different locations in a single model, see Appendix S2).

Conclusion

We have developed a stepwise approach and accompanying statistical tools to quantify how biological responses are affected by weather drivers, or any other intrinsic or extrinsic environmental variable for which high temporal resolution data are available. Our approach is predominantly exploratory, avoiding the need to make untested *a priori* assumptions and to consider only a small number of competing hypotheses. Crucially, however, this exploration is both systematic and statistically grounded, such that the detected effects of weather reflect biological patterns rather than potentially arbitrary decisions made by the modeller. Although this more open-ended approach has some drawbacks, such as the risk of overfitting and bias that can occur at low sample and effect sizes, these issues can be addressed with the right knowledge and tools. Our simulation approach, focused on a diversity of performance metrics, provides a much needed benchmark to facilitate future objective comparison across methods. By providing both the tools to fit weather windows to a wide range of biological responses, and the methods to validate them and determine sample size requirements, we hope that the *climwin* package will make it easier for researchers to explore and quantify the biological effects of weather in a rigorous, replicable and comparable way.

Author contribution

All authors contributed to the development of ideas and discussed the manuscript; NM, LR, LDB and MvdP performed the literature review; LDB and MvdP programmed the R code; and MvdP performed the simulations and wrote the paper.

Acknowledgements

We are grateful to David Leech and James Pearce-Higgins of the British Trust for Ornithology for providing access to the Chaffinch data, Stephen Ellner for discussion and the reviewers for their comments. MvdP and LB were supported by an Australian Research Council Future (FT120100204) and DECRA fellowship (DE130100174), respectively. The authors declare they have no conflict of interest.

Data accessibility

The Chaffinch data set and weather data are available as part of R package *climwin*. The R code used to generate the simulation data and results are available as part of online Appendix S4.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd international symposium on information theory*, 267–281.

Angilletta, M.J. (2009) *Thermal Adaptation: A Theoretical and Empirical Synthesis*. Oxford University Press, Oxford.

Arlot, S. & Celisse, A. (2010) A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4**, 40–79.

Bailey, L. & van de Pol, M. (2015) climwin: Climate window Analysis. URL http://cran.r-project.org/web/packages/climwin/index.html

Bailey, L. & van de Pol, M. (2016) Tackling extremes: challenges for ecological and evolutionary research on extreme climatic events. *Journal of Animal Ecology*, **85**, 85–96.

Baillie, S.R., Marchant, J.H., Leech, D.I., Massimino, D., Sullivan, M.J.P., Eglington, S.M. *et al.* (2014) BirdTrends 2014: trends in numbers, breeding

success and survival for UK breeding birds. BTO Research Report 662. British Trust for Ornithology, Thetford.
Biro, P.A., Beckmann, C. & Stamps, J.A. (2010) Small within-day increases in temperature affects boldness and alters personality in coral reef fish. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 71–77.
Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media, New York.
Cade, B.S. (2015) Model averaging and muddled multimodel inferences. *Ecology*, **99**, 2370–2382.
Charmantier, A., McCleery, R.H., Cole, L.R., Perrins, C.M., Kruuk, L.E.B. & Sheldon, B.C. (2008) Adaptive phenotypic plasticity in response to climate change in a Wild Bird population. *Science*, **320**, 800–803.
Crick, H.Q.P. & Sparks, T.H. (1999) Climate change related to egg-laying trends. *Nature*, **399**, 423–424.
Freckleton, R.P. (2011) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology*, **65**, 91–101.
Frederiksen, M., Lebreton, J., Pradel, R., Choquet, R. & Gimenez, O. (2014) Identifying links between vital rates and environment: a toolbox for the applied ecologist. *Journal of Applied Ecology*, **51**, 71–81.
Gelman, A., Hwang, J. & Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, **24**, 997–1016.
Gienapp, P., Hemerik, L. & Visser, M.E. (2005) A new statistical tool to predict phenology under climate change scenarios. *Global Change Biology*, **11**, 600–606.
Grosbois, V., Gimenez, O., Gaillard, J., Pradel, R., Barbraud, C., Clobert, J., Møller, A. & Weimerskirch, H. (2008) Assessing the impact of climate variation on survival in vertebrate populations. *Biological Reviews*, **83**, 357–399.
Grueber, C., Nakagawa, S., Laws, R. & Jamieson, I. (2011) Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology*, **24**, 699–711.
Harrison, X.A., Blount, J.D., Inger, R., Norris, D.R. & Bearhop, S. (2011) Carry-over effects as drivers of fitness differences in animals. *Journal of Animal Ecology*, **80**, 4–18.
Husby, A., Nussey, D.H., Visser, M.E., Wilson, A.J., Sheldon, B.C. & Kruuk, L.E. (2010) Contrasting patterns of phenotypic plasticity in reproductive traits in two great tit (*Parus major*) populations. *Evolution*, **64**, 2221–2237.
Kruuk, L.E., Osmond, H.L. & Cockburn, A. (2015) Contrasting effects of climate on juvenile body size in a Southern Hemisphere passerine bird. *Global Change Biology*, **21**, 2929–2941.
Lawson, C.R., Vindenes, Y., Bailey, L. & van de Pol, M. (2015) Environmental variation and population responses to global change. *Ecology Letters*, **18**, 724–736.
McLean, N., Lawson, C., Leech, D. & van de Pol, M. (2016) Predicting when climate-driven phenotypic change affects population dynamics. *Ecology Letters*, **19**, 595–608.
Mesquita, M.D., Erikstad, K.E., Sandvik, H., Reiertsen, T., Barrett, R., Anker-Nilssen, T., Hodges, K.I. & Bader, J. (2015) There is more to climate than the North Atlantic Oscillation: a new perspective from climate dynamics to explain the variability in population growth rates of a long-lived seabird. *Frontiers in Ecology and Evolution*, **3**, 43.
Phillimore, A.B., Hadfield, J.D., Jones, O.R. & Smithers, R.J. (2010) Differences in spawning date between populations of common frog reveal local adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 8292–8297.
van de Pol, M. & Cockburn, A. (2011) Identifying the critical climatic time-window for trait expression. *American Naturalist*, **177**, 698–707.
van de Pol, M., Osmond, H.L. & Cockburn, A. (2012) Fluctuations in population composition dampen the impact of phenotypic plasticity on trait dynamics in superb fairy-wrens. *Journal of Animal Ecology*, **81**, 411–421.
van de Pol, M., Brouwer, L., Brooker, L.C., Brooker, M.G., Colombelli-Négrel, D., Hall, M.L. *et al.* (2013) Problems with using large-scale oceanic climate indices to compare climatic sensitivities across populations and species. *Ecography*, **36**, 249–255.
Remmert, H. (1980) *Ecology - A Textbook*. Springer Verlag, Berlin-Heidelberg.
Roberts, A. (2008) Exploring relationships between phenological and weather data using smoothing. *International Journal of Biometeorology*, **52**, 463–470.
Schaper, S.V., Dawson, A., Sharp, P.J., Gienapp, P., Caro Samuel, P. & Visser, M.E. (2012) Increasing temperature, not mean temperature, is a cue for avian timing of reproduction. *The American Naturalist*, **179**, E55–E69.
Schielzeth, H. (2010) Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, **1**, 103–113.
Shmueli, G. (2010) To explain or to predict? *Statistical Science*, **25**, 289–310.
Stenseth, N.C. & Mysterud, A. (2005) Weather packages: finding the right scale and composition of climate in ecology. *Journal of Animal Ecology*, **74**, 1195–1198.
Teller, B.J., Adler, P.B., Edwards, C.B., Hooker, G. & Ellner, S.P. (2016) Linking demography with drivers: climate and competition. *Methods in Ecology and Evolution*, **7**, 171–183.
Wolkovich, E.M., Cook, B.I., Allen, J.M., Crimmins, T.M., Betancourt, J.L., Travers, S.E. *et al.* (2012) Warming experiments underpredict plant phenological responses to climate change. *Nature*, **485**, 494–497.

Received 6 April 2016; accepted 27 April 2016
Handling Editor: Olivier Gimenez

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Methods of systematic literature review of papers analysing biological responses to weather.

Appendix S2. Examples of analyses on real data sets with R code from package c*limwin*.

Appendix S3. A slow and fast randomization method to quantify the likelihood that a signal is due to chance or not.

Appendix S4. R code used to generate the simulated data sets.