Rapacciuolo, Giovanni; Roy, David B.; Gillings, Simon; Purvis, Andy. 2014. **Temporal validation plots: quantifying how well correlative species distribution models predict species' range changes over time.** *Methods in Ecology and Evolution*, 5 (5). 407-420. https://doi.org/10.1111/2041-210X.12181

Temporal validation plots: quantifying how well correlative species distribution models predict species' range changes over time

Giovanni Rapacciuolo[1234*], David B. Roy[3], Simon Gillings[5], Andy Purvis[624]

[1]*Berkeley Initiative in Global Change Biology, University of California Berkeley, 3101 Valley Life Sciences Building, Berkeley, CA 94720-3160, USA*

[2]*Department of Life Sciences, Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, Berkshire, SL5 7PY, UK*

[3]*Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10 8BB, UK*

[4]*Grantham Institute for Climate Change, Imperial College London, South Kensington Campus, Exhibition Road, London, SW7 2AZ, UK*

[5]*British Trust for Ornithology, The Nunnery, Thetford, Norfolk, IP24 2PU, UK*

[6]*Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK*

[*]Corresponding Author:

Giovanni Rapacciuolo

Berkeley Initiative in Global Change Biology

University of California Berkeley

3101 Valley Life Sciences Building

Berkeley, CA 94720-3160, USA

Email: giorapac@gmail.com

Running title: Temporal Validation Plots

Word count: 6,674

SUMMARY

1. The use of data documenting how species' distributions have changed over time is crucial for testing how well correlative species distribution models (SDMs) predict species' range changes. So far, however, little attention has been given to developing a reliable methodological framework for using such data.
2. We develop a new tool – the temporal validation (TV) plot – specifically aimed at making use of species' distribution records at two times for a comprehensive assessment of the prediction accuracy of SDMs over time.
3. We extend existing presence-absence calibration plots to make use of distribution records from two time periods. TV plots visualise the agreement between change in modelled probabilities of presence and the probability of observing sites gained or lost between time periods. We then present three measures of prediction accuracy that can be easily calculated from TV plots.
4. We present our methodological framework using a virtual species in a simplified landscape, and then provide a real-world case study using distribution records for two species of breeding birds from two time periods of intensive recording effort across Great Britain.
5. Together with existing approaches, TV plots and their associated measures offer a simple tool for testing of how well SDMs model species' observed range changes – perhaps the best way available to assess their ability to predict likely future changes.

Keywords: species distribution models, temporal validation, prediction accuracy, range change, calibration plots, historic surveys

INTRODUCTION

Correlative species distribution models (SDMs) are increasingly used to project likely future changes in species' distributions under ongoing global environmental change (Elith & Leathwick 2009). As a result, assessing how well these approaches can predict species' geographic range changes over time is of increasing importance.

Repeated surveys that document species' distributions at multiple time periods represent invaluable opportunities for testing SDM predictions over time (Araújo *et al.* 2005a; b; Kharouba *et al.* 2009; Tingley *et al.* 2009; Rubidge *et al.* 2010; Dobrowski *et al.* 2011; Rapacciuolo *et al.* 2012; Smith *et al.* 2013). A growing number of temporal datasets are emerging from efforts to rescue and digitize natural history museum collections and other historical data sources such as field notes and photographs (Tingley & Beissinger 2009; Pyke & Ehrlich 2010; Drew 2011). So far, however, little attention has been given to how these data should best be used for testing the prediction accuracy of SDMs over time. In this paper, we develop a new type of diagnostic plot, the *temporal validation* (TV) plot, and an associated set of measures, which make use of distribution data at two time periods within a given area to evaluate how well SDMs can predict species' range changes over time.

Although tests of SDM predictions through time are still relatively rare, existing studies have primarily tested how well models built using species distribution data from a first time period (i.e., calibration data) discriminate between the species' observed presences and absences in a second time period (i.e., validation data) using common measures based on a single probability threshold (e.g., Cohen's Kappa, sensitivity, specificity; Araújo *et al.* 2005a; Rapacciuolo *et al.*

2012; Smith *et al.* 2013) or a range of possible thresholds (e.g., AUC; Kharouba *et al.* 2009; Rubidge *et al.* 2010; Dobrowski *et al.* 2011; Smith *et al.* 2013). Such tests of SDM predictions through time are generally used to estimate how well models are likely to predict species' range changes in the future (Araújo *et al.* 2005a; b; Kharouba *et al.* 2009; Tingley *et al.* 2009; Rubidge *et al.* 2010; Dobrowski *et al.* 2011; Rapacciuolo *et al.* 2012; Smith *et al.* 2013). In this context, however, this widely-used approach to temporal validation suffers from two main issues.

The first issue is that converting continuous probabilities of presence to binary presence-absence predictions using a single or multiple thresholds may not alone provide an exhaustive estimate of model prediction accuracy over time. The practice ignores a lot of information generated by the models: all predicted probabilities above the chosen threshold are considered equal, as are all those below, however near or far they are from it. As a result, slight but important changes in the environment may not be captured by binary-converted predictions and prediction accuracy measures based on these converted model predictions may wrongly infer range stability despite the probability of presence being predicted to change.

The second issue is that using calibration and validation datasets collected in different time periods across the same region does not enable fully independent model validation. This is because many modelled factors that correlate with a species' distribution across that region will remain unchanged through the entire study period. As a result, models with high explanatory power in one time period are likely to retain that power in another time period across areas where both observations and model predictions indicate no change in the species' range, regardless of whether the models have captured fundamental drivers of range change over time (Araújo *et al.*

95 2005a; Rapacciuolo *et al.* 2012). Importantly, spurious species-environment correlations
96 identified during model calibration may not be revealed by temporal validation across these
97 unchanged areas. Therefore, measuring prediction accuracy over the entire study area in a second
98 time period – including unchanged areas – may be a misleading measure of how well models are
99 likely to predict to a third time period (e.g., future environmental scenario). This approach should
100 be complemented with measures that focus on how well models predict to areas where species'
101 range changes have actually been observed and/or predicted (Rapacciuolo *et al.* 2012). The issue
102 of examining spatial processes of change with global measures that do not incorporate spatial
103 variation in prediction accuracy within the study region (e.g., Kappa) has been the subject of
104 much scrutiny in the remote-sensing and map comparison literatures (Csillag & Boots 2005;
105 Pontius & Millones 2011; Robertson *et al.* 2014), yet it has been rarely considered in the SDM
106 literature.
107
108 TV plots aim to overcome both issues with existing approaches. First, we extend the method of
109 presence-absence calibration plots – originally developed in the context of statistical medicine
110 (Miller *et al.* 1991; Harrell *et al.* 1996; Harrell 2001) but repeatedly used to quantify the
111 calibration of SDMs (Pearce & Ferrier 2000; Boyce *et al.* 2002; Hirzel *et al.* 2006; Phillips &
112 Elith 2010) – for use with empirical distribution and environmental data from two time periods.
113 Presence-absence calibration plots fit observed presence-absence directly as a function of
114 continuous modelled probabilities, without converting to binary predictions based on any
115 threshold (Phillips & Elith 2010). Thus, our method makes full use of the information generated
116 by the modelling process without ignoring the probabilistic nature of SDM predictions. Second,
117 we focus on assessing model performance only on grid cells where either or both observed data

and model predictions indicate range change over time, whilst disregarding model performance on grid cells where both observations and predictions indicate no range change. TV plots model how well changes in modelled probability of presence between time periods reflect species' observed gains and losses separately, thus incorporating spatial variation in prediction accuracy within the study area. Building on the existing literature, we then present three measures of the agreement between modelled and observed changes that can be easily calculated from TV plots – Acc_{TV}, Cor_{TV}, and $Bias_{TV}$. Together with existing approaches to temporal validation, these measures provide a comprehensive assessment of how well a model predicts observed range changes and, thus, the fullest available picture of how likely the model is to predict future changes. We present our methodological framework using a virtual species in a simplified landscape, then provide a real-world case study using distribution records for two breeding bird species from two time periods of intensive recording effort across Great Britain (Sharrock 1976; Gibbons *et al.* 1993).

VIRTUAL CASE STUDY

Simulated environment

We consider an artificial landscape of 30 x 30 grid cells and generate environmental variation within this grid in an initial time period t using three 'climate' variables – *temperature*, *precipitation* and *covar* – each taking values in the range 0–1. Temperature and covar both exhibit a linear latitudinal gradient and are highly intercorrelated (Pearson's $r = 0.88$), whilst precipitation exhibits a linear longitudinal gradient (Fig. 1). We then simulate change in the environment in a second time period $t + 1$ by updating the values of the three environmental variables across the landscape. We specify alternative change scenarios for each variable – mean

141 temperature increase, mean precipitation decrease and no change in mean covar – by sampling
142 change values from three different normal distributions (temperature: mean ± standard deviation
143 = 0.3 ± 0.25; precipitation: -0.15 ± 0.5; covar: 0 ± 0.5) and summing sampled values with initial
144 environmental values (Fig. S1).
145
146 **Environmental functional relationships**
147 We simulate the distribution of a simple virtual species across this landscape by specifying four
148 alternative functional relationships between the species' probability of presence and the
149 environment – a *true* functional relationship and three potential misspecifications of the truth
150 (Fig. 1). This approach, based on simulations by Phillips & Elith (2010) and Pagel & Schurr
151 (2012), enables us to quantify the effects of alternative model misspecifications on how well
152 models predict the species' true distribution over time. First, we specify the true probability of
153 presence for our virtual species conditional on temperature and precipitation only, but not covar,
154 as: 0.5 x temperature + 0.5 x precipitation. Thus, the variable covar does not bear any functional
155 relationship with the species' probability of presence, although it significantly covaries with the
156 species' presence-absence because of its strong correlation with temperature. We then consider
157 three potential models of our virtual species' probability of presence, which we parameterise
158 statistically based on subsets of the three environmental variables (see Fig. 1).
159 1) The *Incomplete* model estimates probability of presence conditional only on temperature,
160 ignoring precipitation, as: 0.26 + 0.51 x temperature. This model may arise if relevant
161 predictors – in this case precipitation – were unavailable, overlooked, or wrongly
162 excluded during model selection.

2) The *Collinear* model estimates the species' probability of presence conditional on precipitation and covar, ignoring temperature, as: 0.03 + 0.5 x precipitation + 0.5 x covar. This model may arise if irrelevant predictors are naively entered into a model selection algorithm and erroneously selected through their apparent correlation with probability of presence.
3) The *Incomplete and Collinear* model estimates the probability of presence conditional only on covar, ignoring the true predictors temperature and precipitation, as: 0.28 + 0.52 x covar. This model combines both types of misspecification included in the previous two models: it is incomplete, as it only considers a single variable instead of two, and collinear, as it includes a variable correlated but not functionally-related to the species' true probability of presence.

We predict the probability of presence of our virtual species across the landscape in period *t* and *t + 1* based on each of the four environmental functional relationships. To define the true presence-absence of the species across the landscape in both time periods, we convert each grid square's probability of presence to either presence or absence by conducting a Bernoulli trial according to the species' true probability of presence in each grid square.

Temporal validation plots

We extend the approach of presence-absence calibration plots (reviewed by Pearce & Ferrier 2000; Boyce *et al.* 2002; Hirzel *et al.* 2006; Phillips & Elith 2010 in the context of SDMs) to make use of data from two time periods and develop a new plot, the *temporal validation* (TV) plot, for assessing the prediction accuracy of SDMs over time. TV plots show the agreement

between changes in observed presence-absence and changes in modelled probability of presence between t and $t + 1$. This is done in three steps: (i) calculating observed and modelled changes, (ii) estimating gain and loss functions, and (iii) combining gain and loss functions to visualise the agreement between observed and modelled changes.

Step 1: Calculating observed and modelled changes

First, the species' presence-absence (y) across the study area is compared between t and $t + 1$ to identify observed gains (instances where $y_t = 0$ and $y_{t+1} = 1$), losses ($y_t = 1$ and $y_{t+1} = 0$), stable presences ($y_t = 1$ and $y_{t+1} = 1$), and stable absences ($y_t = 0$ and $y_{t+1} = 0$). Figure 2a shows observed changes in the presence-absence of our virtual species between t and $t + 1$. Overall, the species' presence across the landscape has increased: the species has experienced most gains in areas that have become warm enough for the species to expand into and have also remained wet enough for it to occur despite overall decrease in precipitation (i.e., northwest of the landscape). Additionally, there have been localised gains and losses across the entire landscape.

Second, values of change in modelled probability of presence (Δm) are calculated by subtracting modelled probability of presence in t (m_t) from modelled probability of presence in $t + 1$ (m_{t+1}). Importantly, Δm values are not linearly related to the probability that gains or losses are actually observed, even if we assume that a model has captured perfectly a species' environmental functional relationship. For example, consider two absence sites with different m_t: for an equal increase in modelled probability of presence in $t + 1$ ($\Delta m > 0$), the site with a higher m_t will exhibit an inherently higher probability of gain because it already presents a higher probability of finding the species. Similarly, for equal decreases in modelled probability of presence ($\Delta m < 0$),

209 a presence site with a higher initial probability of absence (1 - m_t) has an inherently higher
210 probability of loss. Therefore, weighted, instead of absolute, changes in modelled probability of
211 presence ($\Delta m_{weighted}$) are used in TV plots. $\Delta m_{weighted}$ are calculated by weighting Δm values by m_t,
212 using the following function:

$$\Delta m_{weighted} = f(\Delta m,\ m_t) = \begin{cases} \dfrac{\Delta m}{1 - m_t}, & if\ \Delta m > 0 \\ 0, & if\ \Delta m = 0 \\ \dfrac{\Delta m}{m_t}, & if\ \Delta m < 0 \end{cases} \qquad \text{(eqn 1)}$$

213 Figure 2b shows the species' weighted changes in modelled probability of presence between *t*
214 and *t + 1*. Most increases are predicted in the west and most decreases are predicted in the
215 northeast of the simulated landscape.
216
217 *Step 2: Estimating gain and loss functions*
218 Two separate functions – a *gain* and a *loss* function – are fitted to subsets of the values calculated
219 in step 1. Gain and loss functions (blue and red curves of Fig. 2c, respectively) indicate the
220 probability that gains and losses, respectively, are observed for any given value of $\Delta m_{weighted}$ by
221 interpolating from observed instances. Each of these two functions is generated in a manner
222 analogous to the presence-absence calibration plots of Phillips & Elith (2010): binary 1-0
223 observations are statistically modelled as a function of continuous modelled probabilities using
224 natural splines (Ridgeway 2013). For the gain function, the binary response is calculated by
225 contrasting observed gains (1; the blue tick marks in the top rug plot of Fig. 2c) with observed
226 losses and stable absences (0; the grey tick marks in the top rug plot of Fig. 2c). Notably, stable
227 presences are excluded from the estimation of gain functions since they are uninformative of
228 how well a model predicts *change*: although $\Delta m_{weighted}$ may well increase at these sites, a species

cannot gain sites it already occupies. Similarly, for the loss function, the binary response is calculated by contrasting observed losses (1; the red tick marks in the bottom rug plot of Fig. 2c) with gains and stable presences (0; the grey tick marks in the bottom rug plot of Fig. 2c). Stable absences are not used in the estimation of loss functions since a species cannot lose sites from which it is already absent. For both functions, responses are modelled as a function of values of $\Delta m_{weighted}$ at each site corresponding to a response value. In order to aid visualisation, the loss function is multiplied by -1 before being plotted in TV plots, so that it appears in the negative range of the *y*-axis and can be better contrasted to the gain function (Fig. 2c).

Step 3: Combining gain and loss functions to visualise the agreement between observed and modelled changes

A model that perfectly predicts range change through time should predict a probability of gain of 1 and a probability of loss of 0 in areas where there are no losses and all possible gains are made. Similarly, it should predict a probability of gain of 0 and a probability of loss of 1 where no gains are made and every presence is lost. To verify these expectations, gain and loss functions are combined into a temporal validation curve that quantifies how well a model predicts the probability of observing a given overall change in presence-absence between t and $t + 1$. For any given $\Delta m_{weighted}$, the temporal validation curve (thick black curve of Fig. 2c) equals the gain function minus the loss function. Note that, because probabilities of loss are plotted with a negative sign in TV plots, the model temporal validation curve is actually the sum, not the difference, of plotted gain and loss functions. Using this approach, an ideal model results in an ideal straight line going from (-1,-1) – where every presence is lost and there are no gains – to (1, 1) – where every empty cell is filled and no cell is lost (dashed line of Fig. 2c). The ideal line

252 also passes through the origin (0, 0) – where probability of observing gains and probability of
253 observing losses are equal. It should be noted that, even for an ideal model, the probabilities of
254 observing gains and losses at (0, 0) are not necessarily zero: some grid cells may be gained or
255 lost due to stochastic population processes, even after accounting for all deterministic
256 environmental processes.
257
258 We generate TV plots of the true functional response (Fig. 2c) and the three models (Fig. 2d-f);
259 these visualise the ability of each alternative functional response to model change in the observed
260 distribution of our virtual species between t and $t + 1$. The modelled temporal validation curve
261 can be visually compared to the ideal expectation using ± 2 standard error confidence intervals
262 (orange lines of Fig. 2c). Predictions from the true functional response show near-perfect
263 agreement with observed changes in presence-absence: the ideal curve almost entirely falls
264 within the ± 2 standard error confidence intervals of the model curve and the model curve
265 approaches both (-1, -1) and (1, 1) (Fig. 2c). On the other hand, TV plots of all three alternative
266 models of the species' distribution indicate some level of misprediction (Fig. 2d-f). In particular,
267 the *Incomplete and Collinear* model appears to lack any understanding of the species' drivers of
268 range change: gains and losses are observed with comparable frequencies across the entire range
269 of $\Delta m_{weighted}$ (Fig. 2f).
270
271 **Prediction accuracy measures from TV plots**
272 Visual inspection of TV plots is useful and may be all that is needed for a number of
273 applications, but often repeatable and quantitative measures of predictive accuracy through time
274 are required. This is especially true in studies where many models are used for comparative

purposes and visual inspection is impractical (e.g., Araújo *et al.* 2005a; Kharouba *et al.* 2009; Dobrowski *et al.* 2011; Rapacciuolo *et al.* 2012; Smith *et al.* 2013). How can a model's prediction accuracy be calculated from TV plots? In the context of SDMs, a number of measures have been generated from presence-absence calibration plots; however, few of them offer a comprehensive assessment, as they generally either assume linear model curves (e.g. calibration bias and spread; Pearce & Ferrier 2000) or focus on a single aspect of model calibration whilst ignoring others (e.g., point biserial correlation; Phillips & Elith 2010). Here, we build on the work of Harrell (2001), Pearce & Ferrier (2000) and Phillips & Elith (2010), but also the work of Boyce *et al.* (2002) and Hirzel *et al.* (2006), to develop three simple measures of the agreement between the model and the ideal temporal validation curves – Acc_{TV}, Cor_{TV}, and $Bias_{TV}$. Together, these measures offer a comprehensive assessment of how well a model predicts range change through time. Figure 3 provides visual representations of the three measures, exemplified using the TV plot of the Collinear model of our virtual species.

The first measure, temporal validation accuracy (Acc_{TV}; Fig. 3a), is a measure of the weighted mean distance between the ideal and model temporal validation curves at each observation, subtracted from 1. Acc_{TV} can be calculated using the following equation:

$$\mathrm{Acc_{TV}} = 1 - \frac{\sum_{q=1}^{n} \Delta m_{weighted,q} \left| y_{model,q} - y_{ideal,q} \right|}{\sum_{q=1}^{n} \Delta m_{weighted,q}} \qquad \text{(eqn 2)}$$

where y_{model} and y_{ideal} are the y values of the model curve and ideal curve, respectively, at each observed site q, and $\Delta m_{weighted}$ are the weighted changes in modelled probability of presence at each site q. We use a weighted mean to give more importance to large changes in modelled probability of presence and less importance to minor changes, so as to provide a more rigorous

296 measure of agreement when substantial changes are predicted. Acc_{TV} ranges from a minimum
297 value of 0 – indicating a model whose predictions are on average as distant as possible from
298 probabilities of observing change – to a maximum value of 1 – indicating a perfectly-predictive
299 model whose weighted changes in modelled probability of presence can be taken at face value.
300
301 The second measure, temporal validation correlation (Cor_{TV}; Fig. 3b), is the weighted Pearson's
302 *r* correlation coefficient between y_{model} and y_{ideal} at each observed site q, whereby the weights
303 equal $\Delta m_{weighted,\,q}$. Cor_{TV} can be calculated using the following equation:

$$\mathrm{Cor_{TV}} = \frac{cov(y_{model}, y_{ideal};\ \Delta m_{weighted,q})}{\sqrt{cov(y_{model}, y_{model};\ \Delta m_{weighted,q})\,cov(y_{ideal}, y_{ideal};\ \Delta m_{weighted,q})}} \qquad \text{(eqn 3)}$$

304 where *cov* is the covariance. Our Cor_{TV} measure is similar to the point biserial correlation (COR;
305 Elith *et al.* 2006; Phillips & Elith 2010), except that it correlates predicted probabilities with
306 continuous probability values fitted using natural splines, instead of observed binary values; for
307 this reason, Cor_{TV} values are expected to be considerably higher than corresponding COR values.
308
309 The third measure, temporal validation bias ($Bias_{TV}$; Fig. 3c), quantifies the systematic deviation
310 between the ideal and the model curves. Unlike Acc_{TV} and Cor_{TV}, $Bias_{TV}$ is not simply calculated
311 at each observed site. Instead, it is estimated over the entire interval between minimum and
312 maximum $\Delta m_{weighted}$ values – respectively $min(\Delta m_{weighted})$ and $max(\Delta m_{weighted})$ – using definite
313 integrals evaluating the area between the *ideal* and *model* functions and the *x*-axis. $Bias_{TV}$ can be
314 calculated as:

$$Bias_{TV} = \int_{\min(\Delta m_{weighted})}^{\max(\Delta m_{weighted})} ideal(x)dx \; - \int_{\min(\Delta m_{weighted})}^{\max(\Delta m_{weighted})} model(x)dx \qquad \text{(eqn 4)}$$

A model has a Bias$_{TV}$ of 0 if it perfectly predicts overall change in the probability of observing a species across the entire range of $\Delta m_{weighted}$. A negative Bias$_{TV}$ indicates the model tends to underestimate species' overall presence across the landscape in *t + 1* by underestimating observed gains and/or overestimating observed losses. A positive Bias$_{TV}$ indicates the model tends to overestimate the species' overall presence in *t + 1* by overestimating observed gains and/or underestimating observed losses. Importantly, a model may have a Bias$_{TV}$ of 0 despite substantial deviations from the ideal curve at given $\Delta m_{weighted}$ values. This may occur if overestimates and underestimates of gains are balanced by equal overestimates and underestimates of losses, respectively, and overall change in modelled probability averages out to overall probability of observing change in the species' presence.

Table 1 shows how the three measures derived from TV plots vary across the four environmental functional responses of our virtual species. Unsurprisingly, the true environmental functional response has the highest Acc$_{TV}$ and Cor$_{TV}$ – both close to 1 – and the lowest Bias$_{TV}$ – nearly 0. Amongst the three models, the *Incomplete* model appears to be the best, with a similar Cor$_{TV}$ to the Truth but a lower Acc$_{TV}$ and a large negative Bias$_{TV}$, whilst the *Incomplete and Collinear* model is clearly the least able to predict observed change, with a very low Acc$_{TV}$ and negative Cor$_{TV}$ and Bias$_{TV}$ values. The *Collinear* model has intermediate prediction accuracy, with a Cor$_{TV}$ comparable to the *Truth* but a lower Acc$_{TV}$ than the *Incomplete* model.

What aspects of species and their environment affect measures from TV plots?

336 The calculation of many commonly-used measures of SDM prediction accuracy is affected by
337 the prevalence (i.e., proportion of observed presences) of the modelled species within the study
338 area (McPherson *et al.* 2004; Santika 2011; Lawson *et al.* 2014). In addition, there are
339 indications that the magnitude and extent of environmental change may also affect the
340 assessment of SDM prediction accuracy over time (Fitzpatrick & Hargrove 2009; Elith *et al.*
341 2010). For these reasons, we carried out a sensitivity analysis to test whether temporal prediction
342 accuracy measures from TV plots are sensitive to various aspects of our virtual species and
343 simplified landscape. We investigated the effect of varying three main factors: species' initial
344 prevalence (i.e., number of presences over total number of grid cells), magnitude of
345 environmental change and spatial extent over which environmental change takes place. For the
346 purposes of this sensitivity analysis, we used the same four functional responses and initial
347 environmental values we used in our main virtual case study (see Fig. 1). However, we
348 simplified our environmental change scenario by sampling values of change from a normal
349 distribution with a mean of 0 and a standard deviation of 0.4 for all three variables, unless
350 otherwise specified. First, given the linear relationship between our species' probability of
351 presence and both temperature and precipitation, we varied the species' initial prevalence across
352 the landscape by progressively increasing initial values of temperature and precipitation, with
353 initial *covar* values varying accordingly (25 alternative scenarios). Second, we varied the
354 magnitude of environmental change between time periods by progressively increasing the
355 standard deviation – from 0.01 to 1 – of the normal distribution from which we sampled values
356 of environmental change, concurrently for all three variables (25 alternative scenarios). Finally,
357 we varied the spatial extent over which environmental change occurred by varying the extent of
358 the grid over which we sampled environmental change – from a 1 x 1 grid to the entire 30 x 30

grid (30 alternative scenarios). We ran 100 repeats of each alternative scenario for each factor and present mean values of prediction accuracy measures across those 100 repeats.

Figure 4 shows the effect of varying species' initial prevalence, magnitude and spatial extent of environmental change on temporal validation for the four alternative functional responses of our virtual species. Overall, the three prediction accuracy measures derived from TV plots were not particularly sensitive to any of the three factors: the four alternative functional responses generally maintained their relative rank and values of each measure remained relatively stable across most alternative environmental scenarios of each factor. However, there were two main noteworthy results. First, all models had higher Acc_{TV} than expected compared to the truth at particularly low magnitudes and extents of environmental change (Fig. 4a, second and third columns), suggesting that the reliability of certain measures from TV plots may increase with the amount of environmental change experienced across the study area. Considering alternative measures such as Cor_{TV} and $Bias_{TV}$, which were less sensitive to the magnitude and extent of environmental change, appears to be particularly important for a more consistent picture of temporal validation at low magnitudes and extents of change. Second, all three measures were somewhat sensitive to our virtual species' initial prevalence: at low and high extremes of initial prevalence, $Bias_{TV}$ values were positive and negative, respectively, and Acc_{TV} and Cor_{TV} values were slightly lower than expected (Fig. 4a-c, first column). We suspect these results may be partially explained by the lack of ecological realism in our simulations. In fact, identifying cells as observed gains or losses from given increases or decreases in probability of presence within a Bernouilli trial is less likely when initial probabilities of presence are either extremely low (i.e. low prevalence) or extremely high (i.e. high prevalence), respectively. As a result, mismatches

between observed and modelled changes in our virtual case study are more likely at extremes of prevalence. Nevertheless, it should be noted that the species' initial prevalence, through its effects on the relative probability of observing gains or losses, may have an effect on measures of prediction accuracy from TV plots when using real data.

REAL DATA CASE STUDY

We tested the method of TV plots using observed distribution records for two species of breeding birds – the Pied Wagtail and the Turtle Dove – across Great Britain in two time periods between the 1960s and the 1990s. For those two species, we asked: (1) Does model fit in one time period indicate prediction accuracy over time? (2) Can measures from TV plots – which focus on instances of range change – identify aspects of prediction accuracy over time not apparent from commonly-used range-wide measures?

Species distribution data

We used distribution records for the Pied Wagtail (*Motacilla alba*) and the Turtle Dove (*Streptopelia turtur*) in 2603 British 10-km grid squares at two time periods (*t*: 1968–1972; *t + 1*: 1988–1991), corresponding to the periods of intensive recording effort leading to the publication of two national atlases of breeding birds (Sharrock 1976; Gibbons *et al.* 1993). Although the absence of these species from each 10-km grid square could not be definitively recorded during sampling, most grid squares in Great Britain were meticulously sampled, with high levels of duplicate recording and under-recorded areas being targeted by extra recording schemes (Sharrock 1976; Gibbons *et al.* 1993). Thus, we assumed that each surveyed grid square in which a species was not recorded (i.e., non-detection) represented a true absence.

406 **Climate predictors**

407 We used six climate variables: mean temperature of the coldest month (°C), mean temperature of
408 the warmest month (°C), ratio of actual to potential evapotranspiration (standard moisture index),
409 potential sunshine (hours), total annual precipitation (mm), and the difference between total
410 winter precipitation and total summer precipitation (mm). These were calculated from monthly
411 values of temperature, precipitation and cloud cover for periods *t* and *t + 1* from the Climate
412 Research Unit ts2.1 (Mitchell & Jones 2005) and the Climate Research Unit 61-90 (New *et al.*
413 1999) and did not show strong multicollinearity (i.e., all pairwise Spearman's $\rho < 0.85$).

414

415 **Species distribution models**

416 We modelled the presence-absence of the two bird species in period *t* as a function of climate for
417 the corresponding period using generalised boosted models (GBMs; Ridgeway 1999); we built
418 these using the gbm package (Ridgeway 2013) in R version 2.15.2 (R Core Team 2012), and
419 code provided by Elith *et al.* (2008). We used the species-climate associations identified in
420 period *t* to generate modelled estimates of probability of presence in *t* and *t + 1*, based on
421 observed climate for the corresponding periods.

422

423 **Measures of model performance**

424 We measured how well SDMs fitted species' distributions in the calibration period *t* using the
425 area under the receiver operating characteristic (ROC) curve (AUC; Hanley & McNeil 1982) and
426 the point biserial correlation (COR; Elith et al. 2006) – defined as the Pearson correlation
427 between model values and binary values of observed presence-absence. We measured how well
428 models predicted change between *t* and *t + 1* using Acc_{TV}, Cor_{TV}, and Bias_{TV} derived from TV

plots. In addition to these, we also quantified how well models discriminated between presences and absences across the entire study area in *t + 1* using AUC and COR.

Results

Climate-based SDMs provided an excellent fit to observed distribution records for both bird species in the calibration period *t* (Pied Wagtail: AUC = 0.992, COR = 0.809; Turtle dove: AUC = 0.976, COR = 0.875). However, these two models showed different patterns of prediction accuracy over time. Discrimination across the species' entire range in period *t + 1* indicated a much higher prediction accuracy for the Turtle Dove model (AUC = 0.924; COR = 0.670) than the Pied Wagtail model (AUC = 0.691; COR = 0.335), suggesting that climate models may accurately explain the distribution over time of the Turtle Dove but not the Pied Wagtail. Furthermore, these results also indicate that model fit within one time period may not necessarily indicate a model's ability to predict change over time. Nonetheless, generating TV plots revealed additional aspects of these models and their predictions that could not be identified through focusing on the species' entire ranges.

The Pied Wagtail has expanded in areas of the Northern coast and Islands of Scotland, as well as a few localised areas of Eastern England in period *t + 1* (Fig. 5a), with gains in many of these areas being modelled accurately by our climate-based SDM (Fig. 5b). As a result, the TV plot for this model indicates a near perfect prediction of the species' gains (i.e., the positive range of the *x*-axis), leading to a very high overall precision and correlation (Fig. 5c). This suggests that expansion of the Pied Wagtail's breeding range in these areas may be linked to climate – particularly to an increase in minimum temperature of the coldest month (data not presented).

These findings are consistent with previous studies indicating that higher spring temperatures advance first egg dates in this species (Mason & Lyczynski 1980; Crick & Sparks 1999), potentially leading to higher clutch size and juvenile survival rates (Mason & Lyczynski 1980). However, the Pied Wagtail has also experienced localised losses in areas of Northern Scotland and Central and Western England (Fig. 5a). These losses do not appear to be linked to climate – or at least the climatic variables we considered – since they were not predicted by our climate-based model, which instead predicted stable or even increasing probability of presence in these areas (Fig. 5b). Losses in the Pied Wagtail may be due to loss of suitable breeding habitat (e.g. reed beds) – a driver which our climate-based model could not have captured.

Contrary to the Pied Wagtail, the Turtle Dove model appears to completely lack any understanding of the factors driving both gains and losses in the species (Fig. 6). Despite an overall increase in climatic suitability (Fig. 6b), the Turtle Dove has experienced many losses along the Northern and Western edges of its range (Fig. 6a). This inconsistency between predictions and observations is reflected in the model's TV plot and measures, which indicate a substantial lack of agreement between the ideal and the model curve (Fig. 6c). Previous studies have indicated that range contraction of the Turtle Dove in Great Britain may be a consequence of agricultural intensification (Fuller *et al.* 1995) and changes in farming practice (Browne *et al.* 2004) – drivers that are missing from our climate-based model.

In summary, our real-data case study shows that model fit in one time period does not necessarily indicate a model's ability to predict change over time. The use of empirical data on observed range changes can be used for a more reliable estimate of a model's prediction

475 accuracy over time. TV plots, which focus on instances of change over time, revealed aspects of
476 the relationship between species' range changes and climate that could not be identified through
477 range-wide measures. Therefore, a comprehensive assessment of prediction accuracy over time
478 should include both measures of model fit across the species' entire range and measures that
479 focus on instances where range changes have been observed and/or predicted. Such an integrated
480 approach should provide a better assessment of how useful models are likely to be in predicting
481 to a third time period (e.g., future scenario).
482
483 **DISCUSSION**
484 We have developed a new tool that makes full use of species' distribution records at two time
485 periods over the same geographical area to quantify how well SDMs predict range changes over
486 time. Our TV plots and their associated measures overcome the limitations of current approaches
487 by using all the information generated by SDMs and focusing on predictive accuracy across
488 areas where range changes have actually been observed and/or predicted over time. The
489 approach we developed directly relates the redistribution of a species' suitable environment to
490 the probability of observing it expanding or retracting from a given area. As a result, high
491 predictive accuracy from TV plots can only be achieved by models that accurately capture
492 drivers of *change* in species distributions.
493
494 Here, we have assumed that temporally-replicated survey data include perfect knowledge of both
495 species' presence and absence across a study area; in reality, this assumption never entirely holds
496 and may potentially affect the results of temporal validation tests. In principle, TV plots could be
497 extended to alternative, more common types of temporal distribution data. Often, temporal

498 distribution datasets only hold information on species' presence. Incorporating these data in TV
499 plots could be done through an approach similar to that used by Phillips & Elith (2010) for
500 presence-only calibration plots: background data (i.e., a random sample of sites in the study area)
501 could be used in place of species' absences and a transformation employed to correct for the
502 distortion in the model's gain and loss curves obtained this way. In some cases, including our
503 real data case study, survey data hold more information than just species' presence: they include
504 a list of surveyed sites in which the species of interest was not detected (i.e., non-detections).
505 This additional information can be used to calculate a probability of false absence (PFA) for each
506 recorded non-detection (Tingley & Beissinger 2009). Examples of statistical approaches for
507 doing so are occupancy modelling (MacKenzie *et al.* 2002, 2011; Altwegg *et al.* 2008), if repeat
508 samples at each site within each longer time period are available, or list-based methods (Roberts
509 *et al.* 2007; Szabo *et al.* 2010), if repeat samples are unavailable. Estimates of PFA could be
510 integrated in TV plots in a number of ways. First, absences could be weighted by their certainty
511 (1 – PFA) within the estimation of gain and loss functions in TV plots. Second, hypothesised true
512 absences could be identified from a Bernouilli trial according to absence certainty. Third, PFA
513 estimates could be integrated directly within the response of TV plots so that the new response is
514 no longer binary (i.e., gain vs no-gain or loss vs no-loss) but continuous, incorporating the
515 probability of observing true gains/losses over time given absence certainty. Extending TV plots
516 for use with presence-only and presence-non-detection data would enable taking full advantage
517 of unsystematic historical data sources – such as natural history museum collections, field notes
518 and photographs – for a more exhaustive and taxonomically-broader temporal validation of
519 SDMs aimed at predicting likely future changes.
520

Although the three measures we developed in this paper represent an exhaustive summary of the principal information contained in TV plots, many other measures could be derived from these plots. The choice of predictive accuracy measure should depend on the particular application for which SDMs are being built. Additional measures that we can foresee being useful are measures that contrast how well models predict gains (i.e., the positive range of the *x*-axis) versus losses (i.e., the negative range of the *x*-axis). Indeed, species' gains and losses may not necessarily be driven by the same predictors and models may capture drivers of gain but not loss, or vice versa, as shown by our Pied Wagtail example. The variety of prediction accuracy measures that can be derived from TV plots should enable users to assess model performance in a manner that is better suited to their particular question. Nevertheless, different measures derived from the same TV plot are likely to be correlated to some degree; assessing the level of dependence amongst these will be a necessary step to prevent duplication of information.

We suggest that TV plots are a useful tool for assessing how well SDMs predict species' range changes over time, and thus provide R source code and a simple tutorial for their use (see Supporting Information). Our method complements current range-wide approaches to quantify the prediction accuracy of SDMs over time by focusing on instances where range changes have been observed and/or predicted. Taken together, these approaches should enable a much fuller evaluation of how well SDMs predict species' observed range changes, perhaps the best way available to assess their ability to predict the future.

DATA ACCESSIBILITY

543 The bird distribution data used in these analyses can be accessed via the National Biodiversity
544 Network Gateway (1968–1972 records: https://data.nbn.org.uk/Datasets/GA000600; 1988–1991
545 records: https://data.nbn.org.uk/Datasets/GA000147), whilst the climate data can be accessed via
546 the Climate Research Unit (http://www.cru.uea.ac.uk/cru/data/hrg/).
547

548 **ACKNOWLEDGEMENTS**

549 GR received funding from the Biotechnology and Biological Sciences Research Council
550 (BBSRC) and Old Mutual plc. DR received support under the Biological Records Centre
551 partnership between NERC (through the Centre for Ecology & Hydrology) and the Joint Nature
552 Conservation Committee. This paper is a contribution from the Imperial College grand
553 Challenges in Ecosystems and the Environment initiative. We thank Morgan Tingley and four
554 anonymous reviewers for some insightful comments on previous versions of this paper.
555

556 **REFERENCES**

557 Altwegg, R., Wheeler, M. & Erni, B. (2008). Climate and the range dynamics of species with
558 imperfect detection. *Biology Letters*, **4**, 581–4.

559 Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (2005a). Validation of species-climate
560 impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

561 Araújo, M.B., Whittaker, R.J., Ladle, R.J. & Erhard, M. (2005b). Reducing uncertainty in
562 projections of extinction risk from climate change. *Global Ecology and Biogeography*, **14**,
563 529–538.

564 Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K.. (2002). Evaluating resource
565 selection functions. *Ecological Modelling*, **157**, 281–300.

566 Browne, S.J., Aebischer, N.J., Yfantis, G. & Marchant, J.H. (2004). Habitat availability and use
567 by Turtle Doves Streptopelia turtur between 1965 and 1995: an analysis of Common Birds
568 Census data. *Bird Study*, **51**, 1–11.

569 Crick, H.Q.P. & Sparks, T.H. (1999). Climate change related to egg-laying trends. *Nature*, **399**,
570 423–424.

571 Csillag, F. & Boots, B. (2005). Toward comparing maps as spatial processes. *Developments in*
572 *spatial data handling* (ed P. Fisher), pp. 641–652. Springer, Berlin, Germany.

573 Dobrowski, S.Z., Thorne, J.H., Greenberg, J., Safford, H.D., Mynsberge, A.R., Crimmins, S.M.
574 & Swanson, A.K. (2011). Modeling plant ranges over 75 years of climate change in
575 California, USA: temporal transferability and species traits. *Ecological Monographs*, **81**,
576 241–257.

577 Drew, J. (2011). The role of natural history institutions and bioinformatics in conservation
578 biology. *Conservation Biology*, **25**, 1250–1252.

579 Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J.,
580 Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A.,
581 Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T.,
582 Phillips, S.J., Richardson, K., Scachetti-pereira, R., Schapire, R.E., Soberón, J., Williams,
583 S., Wisz, M.S. & Zimmermann, N.E. (2006). Novel methods improve prediction of species'
584 distributions from occurrence data. *Ecography*, **29**, 129–151.

585 Elith, J., Kearney, M. & Phillips, S. (2010). The art of modelling range-shifting species. *Methods*
586 *in Ecology and Evolution*, **1**, 330–342.

587 Elith, J. & Leathwick, J.R. (2009). Species distribution models: ecological explanation and
588 prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*,
589 **40**, 677–697.

590 Elith, J., Leathwick, J.R. & Hastie, T. (2008). A working guide to boosted regression trees. *The*
591 *Journal of Animal Ecology*, **77**, 802–13.

592 Fitzpatrick, M.C. & Hargrove, W.W. (2009). The projection of species distribution models and
593 the problem of non-analog climate. *Biodiversity and Conservation*, **18**, 2255–2261.

594 Fuller, R.J., Gregory, R.D., Gibbons, D.W., Marchant, J.H., Wilson, J.D., Baillie, S.R. & Carter,
595 N. (1995). Population declines and range contractions among lowland farmland birds in
596 Britain. *Conservation Biology*, **9**, 1425–1441.

597 Gibbons, D., Reid, J. & Chapman, R. (1993). *The New Atlas of Breeding Birds in Britain and*
598 *Ireland: 1988–1991*. Poyser, London, UK.

599 Hanley, J.A. & McNeil, B.J. (1982). The meaning and use of the area under a receiver operating
600 characteristic (ROC) curve. *Radiology*, **143**, 29–36.

601 Harrell, F.E. (2001). Binary logistic regression. *Regression Modeling Strategies: With*
602 *Applications to Linear Models, Logistic Regression, and Survival Analysis* pp. 215–266.
603 Springer-Verlag, New York.

604 Harrell, F.E.J., Lee, K.L. & Mark, D.B. (1996). Multivariable prognostic models: issues in
605 developing models, evaluating assumptions and adequacy, and measuring and reducing
606 errors. *Statistics in Medicine*, **15**, 361–387.

607 Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. (2006). Evaluating the ability of
608 habitat suitability models to predict species presences. *Ecological Modelling*, **199**, 142–152.

609 Kharouba, H.M., Algar, A.C. & Kerr, J.T. (2009). Historically calibrated predictions of butterfly
610 species' range shift using global change as a pseudo-experiment. *Ecology*, **90**, 2213–2222.

611 Lawson, C.R., Hodgson, J. a., Wilson, R.J. & Richards, S. a. (2014). Prevalence, thresholds and
612 the performance of presence-absence models (R. Freckleton, Ed.). *Methods in Ecology and*
613 *Evolution*, **5**, 54–64.

614 MacKenzie, D.I., Bailey, L.L., Hines, J.E. & Nichols, J.D. (2011). An integrated model of
615 habitat and species occurrence dynamics. *Methods in Ecology and Evolution*, **2**, 612–622.

616 MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A.
617 (2002). Estimating site occupancy rates when detection probabilities are less than one.
618 *Ecology*, **83**, 2248–2255.

619 Mason, C.F. & Lyczynski, F. (1980). Breeding biology of the Pied and Yellow Wagtails. *Bird*
620 *Study*, **27**, 1–10.

621 McPherson, J., Jetz, W. & Rogers, D.J. (2004). The effects of species' range sizes on the
622 accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of*
623 *Applied Ecology*, **41**, 811–823.

624 Miller, M.E., Hui, S.L. & Tierney, W.M. (1991). Validation techniques for logistic regression
625 models. *Statistics in Medicine*, **10**, 1213–26.

626 Mitchell, T.D. & Jones, P.D. (2005). An improved method of constructing a database of monthly
627 climate observations and associated high-resolution grids. *International Journal of*
628 *Climatology*, **25**, 693–712.

629 New, M., Hulme, M. & Jones, P. (1999). Representing Twentieth-Century Space – Time Climate
630 Variability. Part I: Development of a 1961 – 90 Mean Monthly Terrestrial Climatology.
631 *Journal of Climate*, **12**, 829–856.

632 Pagel, J. & Schurr, F.M. (2012). Forecasting species ranges by statistical estimation of ecological
633 niches and spatial population dynamics. *Global Ecology and Biogeography*, **21**, 293–304.

634 Pearce, J. & Ferrier, S. (2000). Evaluating the predictive performance of habitat models
635 developed using logistic regression. *Ecological Modelling*, **133**, 225–245.

636 Phillips, S.J. & Elith, J. (2010). POC plots: calibrating species distribution models with presence-
637 only data. *Ecology*, **91**, 2476–84.

638 Pontius, R.G. & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and
639 allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*,
640 **32**, 4407–4429.

641 Pyke, G.H. & Ehrlich, P.R. (2010). Biological collections and ecological/environmental
642 research: a review, some observations and a look to the future. *Biological reviews of the*
643 *Cambridge Philosophical Society*, **85**, 247–66.

644 R Core Team. (2012). R: A language and environment for statistical computing. Retrieved from
645 http://www.r-project.org/

646 Rapacciuolo, G., Roy, D.B., Gillings, S., Fox, R., Walker, K. & Purvis, A. (2012). Climatic
647 associations of British species distributions show good transferability in time but low
648 predictive accuracy for range change. *PLoS ONE*, **7**, e40212.

649 Ridgeway, G. (2013). gbm: generalized boosted regression models. R package version 2.1.
650 Retrieved from http://cran.r-project.org/package=gbm

651 Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, **31**, 172–181.

652 Roberts, R.L., Donald, P.F. & Green, R.E. (2007). Using simple species lists to monitor trends in
653 animal populations: new methods and a comparison with independent data. *Animal*
654 *Conservation*, **10**, 332–339.

655 Robertson, C., Long, J. a., Nathoo, F.S., Nelson, T. a. & Plouffe, C.C.F. (2014). Assessing
656 Quality of Spatial Models Using the Structural Similarity Index and Posterior Predictive
657 Checks. *Geographical Analysis*, **46**, 53–74.

658 Rubidge, E.M., Monahan, W.B., Parra, J.L., Cameron, S.E. & Brashares, J.S. (2010). The role of
659 climate, habitat, and species co-occurrence as drivers of change in small mammal
660 distributions over the past century. *Global Change Biology*, **17**, 696–708.

661 Santika, T. (2011). Assessing the effect of prevalence on the predictive performance of species
662 distribution models using simulated data. *Global Ecology and Biogeography*, **20**, 181–192.

663 Sharrock, J. (1976). *The atlas of breeding birds of Britain and Ireland*. Poyser, Berkhamsted,
664 UK.

665 Smith, A.B., Santos, M.J., Koo, M.S., Rowe, K.M.C., Rowe, K.C., Patton, J.L., Perrine, J.D.,
666 Beissinger, S.R. & Moritz, C. (2013). Evaluation of species distribution models by
667 resampling of sites surveyed a century ago by Joseph Grinnell. *Ecography*, **36**, 1–15.

668 Szabo, J.K., Vesk, P. a, Baxter, P.W.J. & Possingham, H.P. (2010). Regional avian species
669 declines estimated from volunteer-collected long-term data using List Length Analysis.
670 *Ecological applications : a publication of the Ecological Society of America*, **20**, 2157–69.

671 Tingley, M.W. & Beissinger, S.R. (2009). Detecting range shifts from historical species
672 occurrences: new perspectives on old data. *Trends in Ecology & Evolution*, **24**, 625–633.

673 Tingley, M.W., Monahan, W.B., Beissinger, S.R. & Moritz, C. (2009). Birds track their
674 Grinnellian niche through a century of climate change. *Proceedings of the National*
675 *Academy of Sciences of the United States of America*, **106**, 19637–19643.

676

677 **Tables**

678 **Table 1:** Prediction accuracy measures derived from temporal validation plots of the four

679 environmental functional responses of our virtual species

Prediction accuracy measures			
	Acc_{TV}	**Cor_{TV}**	**$Bias_{TV}$**
Truth	0.930	0.996	-0.004
Incomplete	0.789	0.976	0.213
Collinear	0.603	0.993	-0.424
Incomplete and Collinear	0.424	-0.187	-0.271

680

681

682

683

684

685

686

687

688

689

690

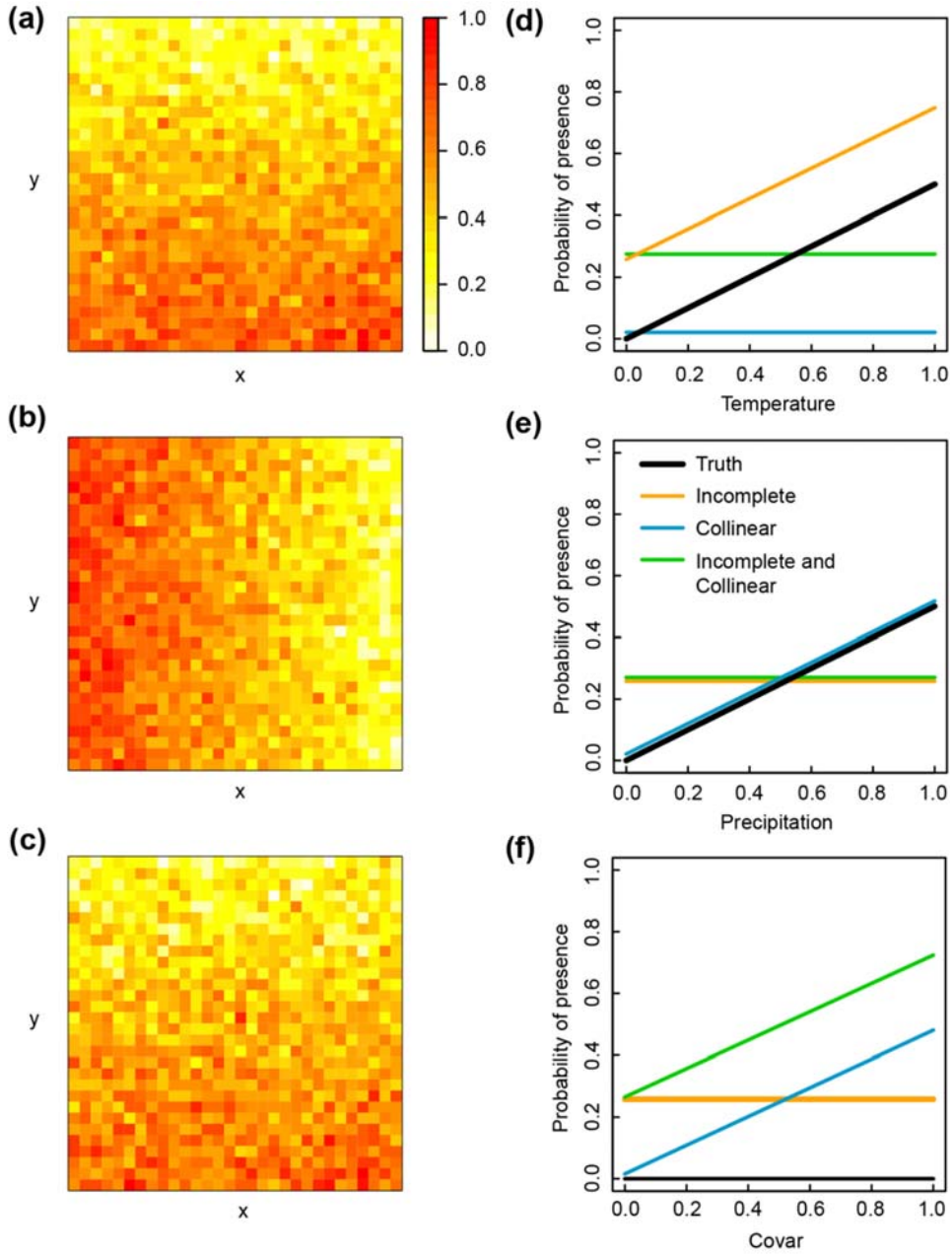
691

692

693

694

695 **Figures**



696

697 **Fig. 1**

698

699

700 **Figure 1:** Four alternative environmental functional responses of a virtual species to three

701 simulated variables over a simplified landscape of 30 x 30 grid cells. Right panels show

simulated values for (a) temperature, (b) precipitation, (c) covar across the simplified landscape; hotter colours indicate higher values (see figure legend). Right panels show how probability of presence varies with (d) temperature, (e) precipitation, (f) covar (whilst keeping all other variables constant at 0) according to each functional response – the Truth (thick black), the Incomplete model (orange), the Collinear model (blue), and the Incomplete and Collinear model (green).

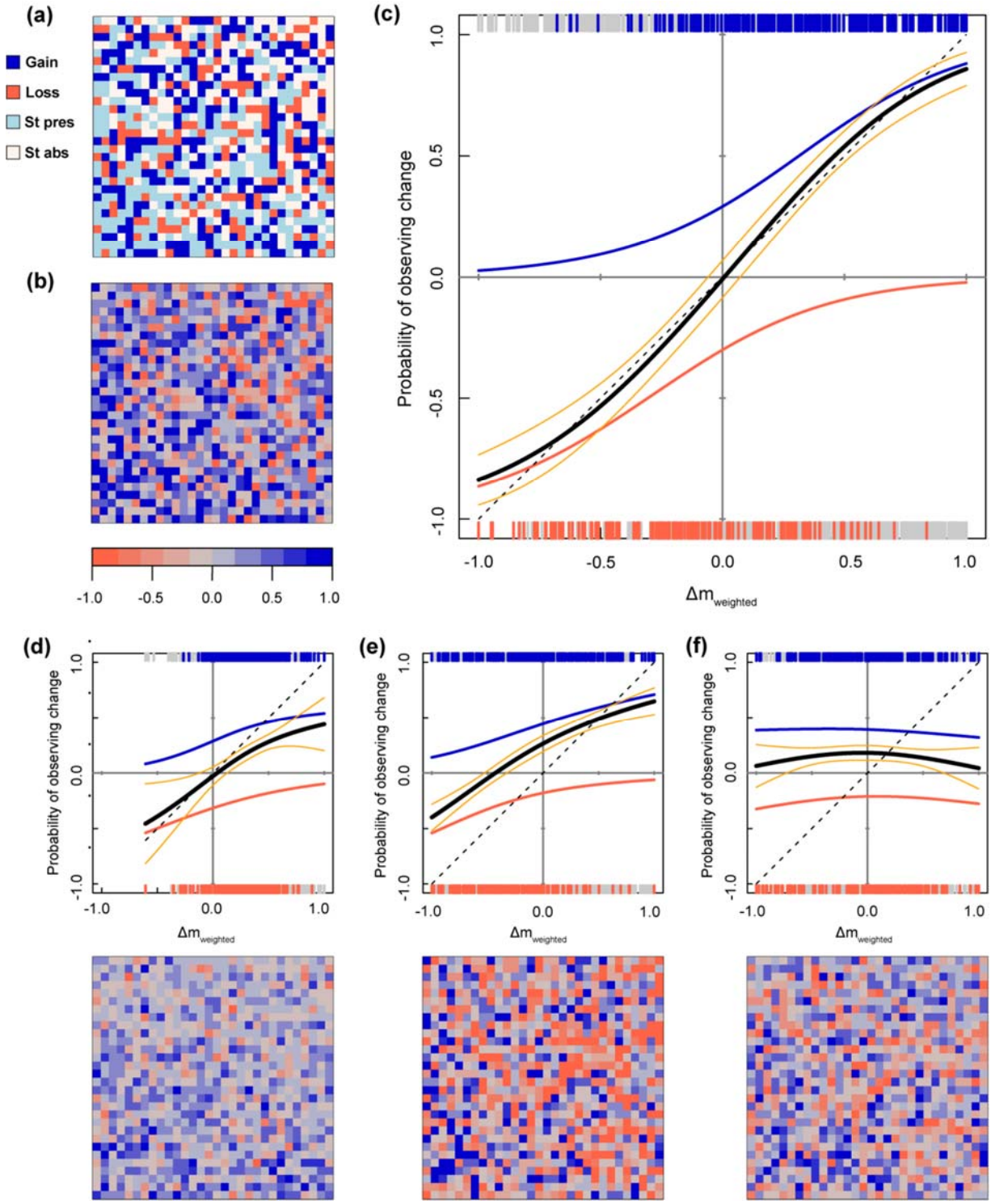
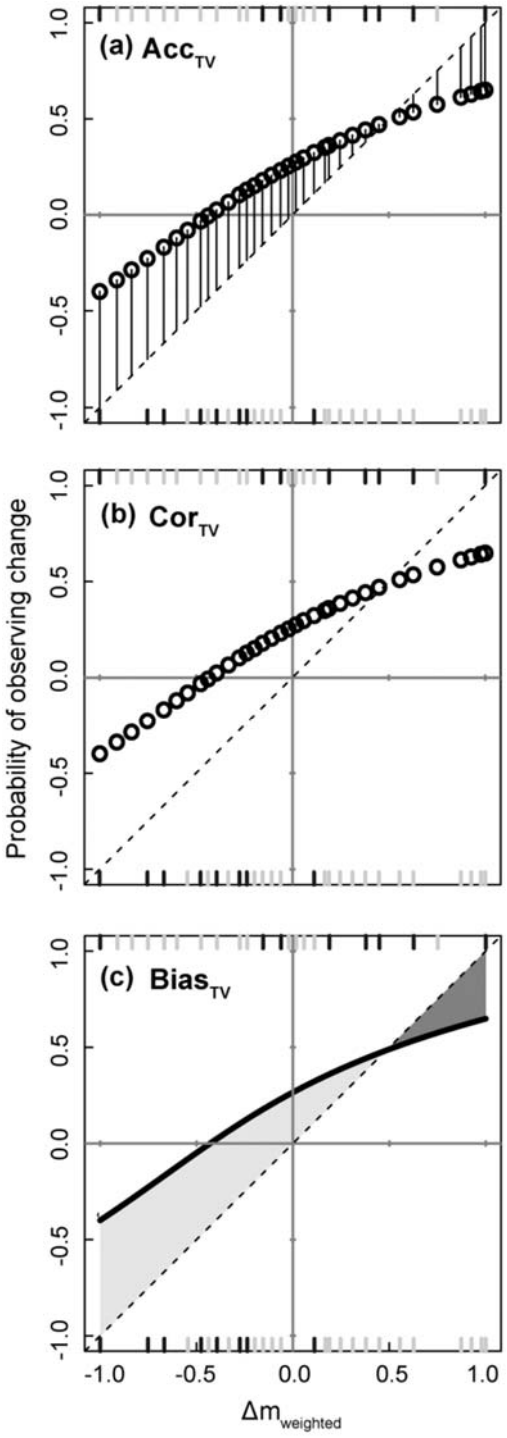

(a)
Gain
Loss
St pres
St abs
(b)
-1.0
-0.5
0.0
0.5
1.0
(c)
Probability of observing change
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
Δm weighted
(d)
Probability of observing change
1.0
0.0
-1.0
-1.0
0.0
1.0
Δm weighted
(e)
Probability of observing change
1.0
0.0
-1.0
-1.0
0.0
1.0
Δm weighted
(f)
Probability of observing change
1.0
0.0
-1.0
-1.0
0.0
1.0
Δm weighted

Fig. 2

728 **Figure 2:** Quantifying the agreement between observed distribution changes and weighted
729 changes in modelled probabilities of presence ($\Delta m_{weighted}$) between time periods t and $t + 1$ for
730 the four functional responses of our virtual species using TV plots. (a) Observed distributional
731 changes in simulated space of our virtual species (gains, losses, stable presences and stable
732 absences) between time periods. (b) $\Delta m_{weighted}$ values across the landscape according to the true
733 functional response of our virtual species. Bluer and redder colours indicate increases and
734 decreases in probability of presence, respectively. (c) TV plot for the true functional response of
735 our virtual species. Shown are the model temporal validation curve (thick black) – the sum of the
736 plotted gain function (blue curve) and loss function (red curve) – and confidence intervals of ± 2
737 standard errors of the mean (orange). The dashed black line represents the expectation for an
738 ideal temporal validation curve. The rug plots show model values at observed gain sites (blue,
739 top of the plot), loss sites (red, bottom of the plot) and stable absences/losses (grey, top of the
740 plot) and stable presences/gains (grey, bottom of the plot). (d-f) TV plots (top panels) and
741 $\Delta m_{weighted}$ (bottom panels) for (d) the Incomplete model, (e) the Collinear model, and (f) the
742 Incomplete and Collinear model.

743
744
745
746
747
748
749
750



751

752 **Fig. 3**

753

754

Figure 3: Visualisations of the three measures of prediction accuracy from TV plots (Acc_{TV}, Cor_{TV} and $Bias_{TV}$), exemplified using the TV plot for the Collinear model. (a) Acc_{TV} equals 1 minus the mean absolute distance between the model's and the ideal y values (black lines), weighted by the corresponding x values, at each observed site (tick marks). (b) Cor_{TV} is the Pearson's r coefficient between the model's and the ideal y values, weighted by the corresponding x values, at each observed site (tick marks). (c) $Bias_{TV}$ is the difference between the area under the model curve (thick black) and the area under the ideal curve (dashed black); it is equivalent to the dark grey minus the light grey area. Note that observed sites shown in scatter and rug plots have been subsampled to aid visualisation.

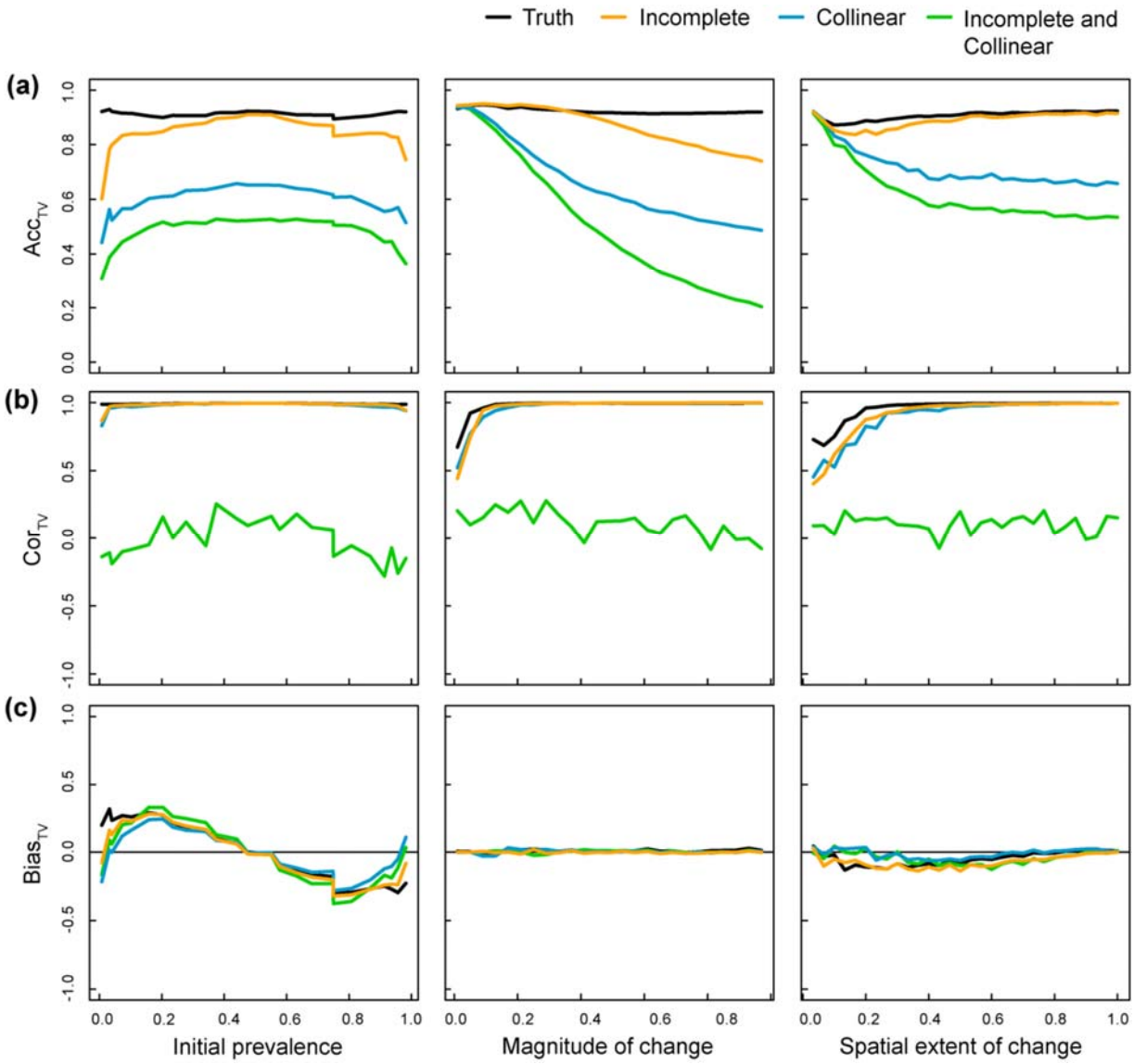
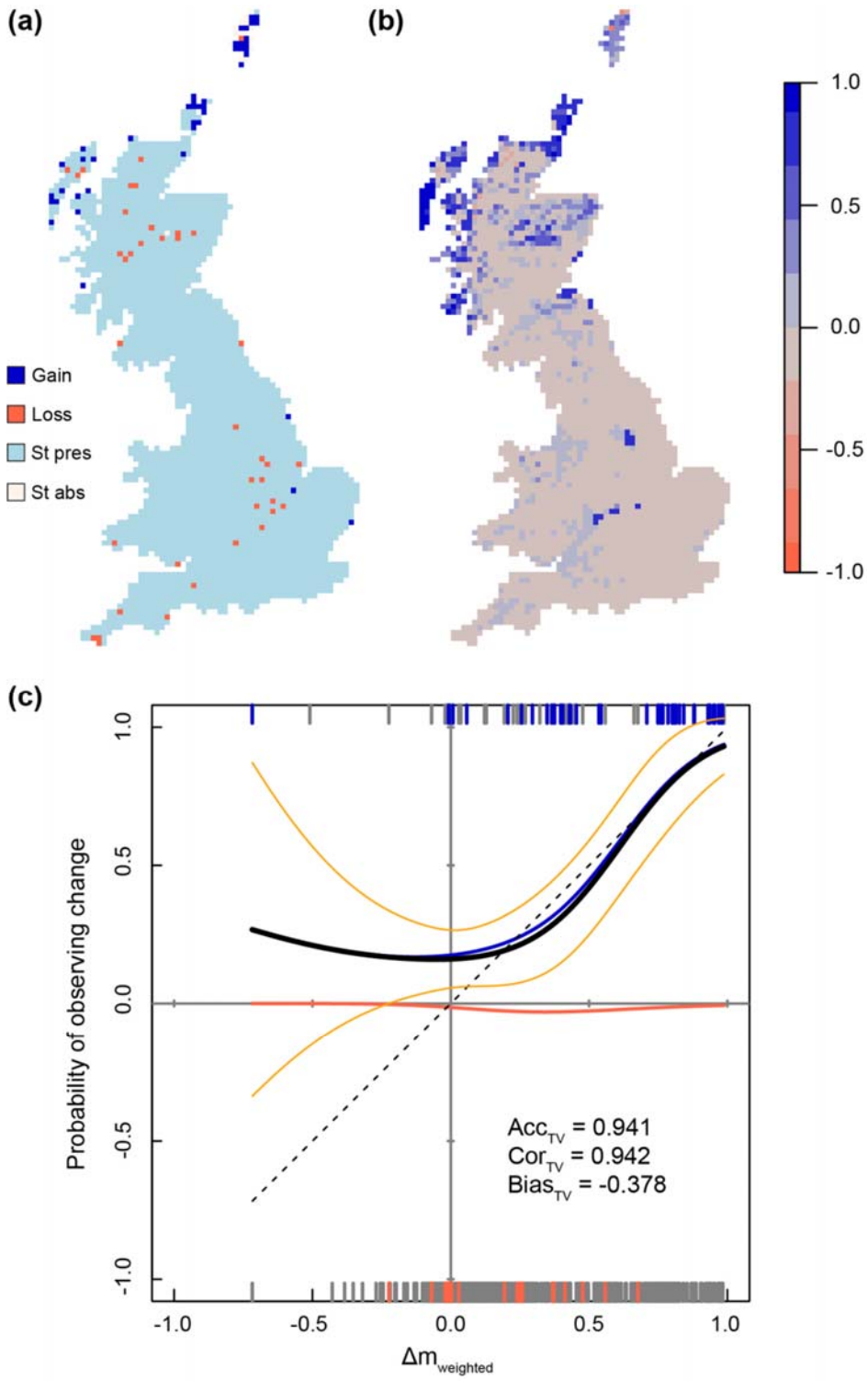

Truth
Incomplete
Collinear
Incomplete and Collinear
(a)
(b)
(c)
Acc_TV
Cor_TV
Bias_TV
Initial prevalence
Magnitude of change
Spatial extent of change

Fig.4

Figure 4: Sensitivity analysis of the effect of species' initial prevalence, magnitude and spatial extent of environmental change on (a) Acc_{TV}, (b) Cor_{TV}, and (c) $Bias_{TV}$ measured from TV plots of the four functional responses of our virtual species. Initial prevalence is the number of species' presences in *t* divided by the total number of grid cells (n = 25). Magnitude of environmental change corresponds to the standard deviation of the normal distribution from which we sampled environmental change values (n = 25). Spatial extent of change is the number of grid cells over which we sampled environmental change divided by the total number of grid cells (n =30). For each measure, values shown represent the mean values of 100 randomisations of each alternative environmental scenario.

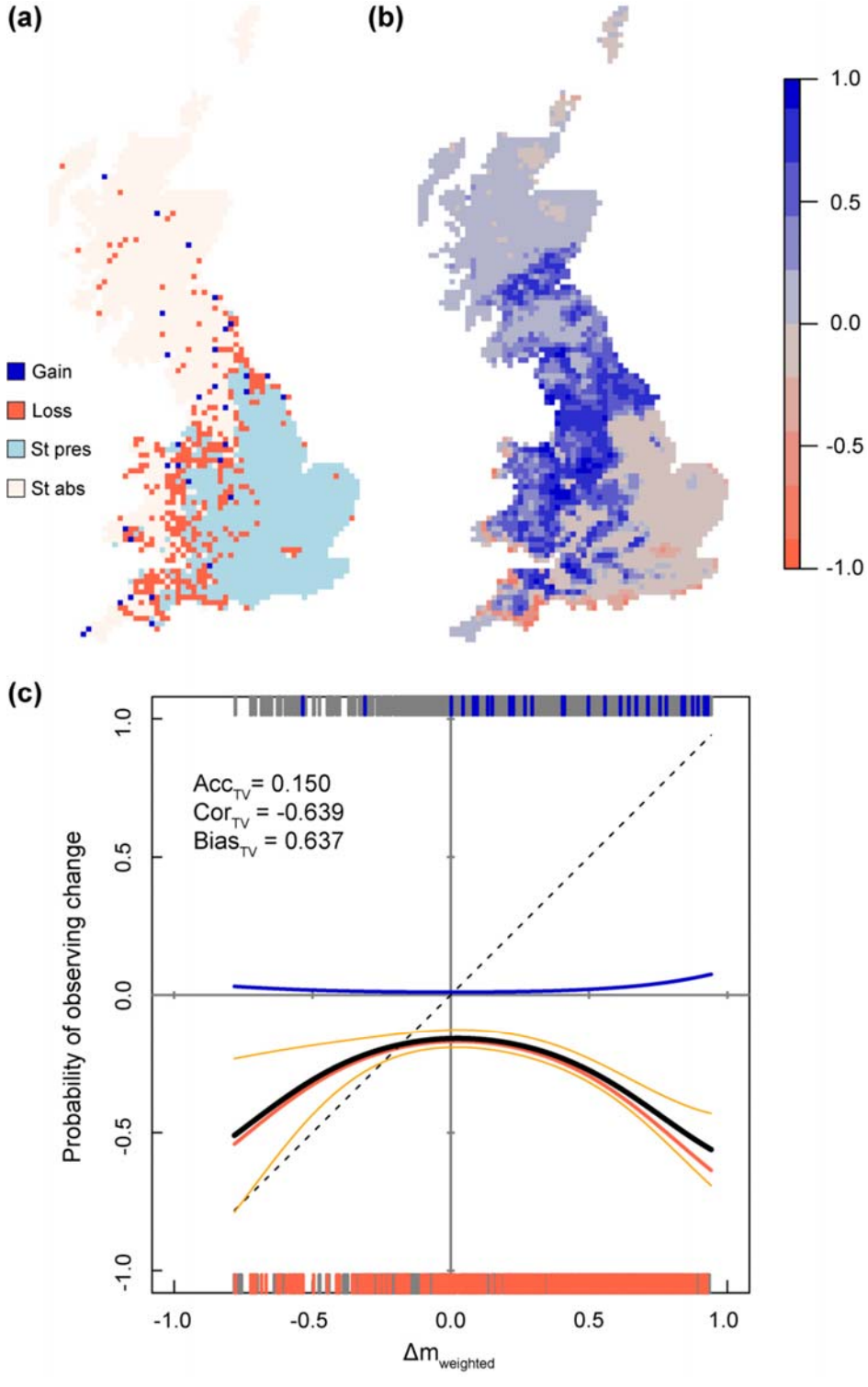


809

810 **Fig. 5**

811

Figure 5: Temporal validation of a climate-based species distribution model of the Pied Wagtail across Great Britain between *t* and *t +1*. (a) Observed changes in the distribution of the Pied Wagtail between time periods. (b) Weighted changes in modelled probability of presence ($\Delta m_{weighted}$) from a climate-based SDM. Bluer and redder colours indicate increases and decreases in probability of presence, respectively. (c) TV plot of the climate-based SDM. Shown are the model temporal validation curve (thick black) – the sum of the plotted gain function (blue curve) and loss function (red curve) – and confidence intervals of ± 2 standard errors of the mean (orange). The dashed black line represents the expectation for an ideal temporal validation curve. The rug plots show model values at observed gain sites (blue, top of the plot), loss sites (red, bottom of the plot) and no-gain and no-loss sites (grey, top and bottom of the plot).


(a)
(b)
Gain
Loss
St pres
St abs
1.0
0.5
0.0
-0.5
-1.0
(c)
Acc$_{TV}$ = 0.150
Cor$_{TV}$ = -0.639
Bias$_{TV}$ = 0.637
Probability of observing change
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
Δm$_{weighted}$

822

823 **Fig. 6**

824 **Figure 6:** Temporal validation of a climate-based species distribution model of the Turtle Dove
825 across Great Britain between *t* and *t +1*. (a) Observed changes in the distribution of the Turtle
826 Dove between time periods. (b) $\Delta m_{weighted}$ from a climate-based SDM. Bluer and redder colours
827 indicate increases and decreases in probability of presence, respectively. (c) TV plot of the
828 climate-based SDM. Shown are the model temporal validation curve (thick black) – the sum of
829 the plotted gain function (blue curve) and loss function (red curve) – and confidence intervals of
830 ± 2 standard errors of the mean (orange). The dashed black line represents the expectation for an
831 ideal temporal validation curve. The rug plots show model values at observed gain sites (blue,
832 top of the plot), loss sites (red, bottom of the plot) and no-gain and no-loss sites (grey, top and
833 bottom of the plot).