Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 412–423 doi: 10.1111/2041-210X.12217

SPECIAL FEATURE
NEW OPPORTUNITIES AT THE INTERFACE BETWEEN ECOLOGY AND STATISTICS

A climate of uncertainty: accounting for error in climate variables for species distribution models

Jakub Stoklosa[1]*, Christopher Daly[2], Scott D. Foster[3], Michael B. Ashcroft[4] and David I. Warton[1]

[1]*School of Mathematics and Statistics, Evolution & Ecology Research Centre, The University of New South Wales, Sydney, NSW 2052, Australia;* [2]*PRISM Climate Group, College of Engineering, Oregon State University, Corvallis, OR 97331, USA;* [3]*CSIRO's Division of Computational Informatics, CSIRO's Wealth from Oceans Flagship, Hobart, TAS 7001, Australia; and* [4]*Australian Museum, Sydney, NSW 2010, Australia*

Summary

1. Spatial climate variables are routinely used in species distribution models (SDMs) without accounting for the fact that they have been predicted with uncertainty, which can lead to biased estimates, erroneous inference and poor performances when predicting to new settings – for example under climate change scenarios.

2. We show how information on uncertainty associated with spatial climate variables can be obtained from climate data models. We then explain different types of uncertainty (i.e. classical and Berkson error) and use two statistical methods that incorporate uncertainty in climate variables into SDMs by means of (i) hierarchical modelling and (ii) simulation–extrapolation.

3. We used simulation to study the consequences of failure to account for measurement error. When uncertainty in explanatory variables was not accounted for, we found that coefficient estimates were biased and the SDM had a loss of statistical power. Further, this bias led to biased predictions when projecting change in distribution under climate change scenarios. The proposed errors-in-variables methods were less sensitive to these issues.

4. We also fit the proposed models to real data (presence/absence data on the Carolina wren, *Thryothorus ludovicianus*), as a function of temperature variables.

5. The proposed framework allows for many possible extensions and improvements to SDMs. If information on the uncertainty of spatial climate variables is available to researchers, we recommend the following: (i) first identify the type of uncertainty; (ii) consider whether any spatial autocorrelation or independence assumptions are required; and (iii) attempt to incorporate the uncertainty into the SDM through established statistical methods and their extensions.

Key-words: climate maps, errors-in-variables, hierarchical statistical models, measurement error, prediction error, PRISM, SIMEX

Introduction

Species distribution models (SDMs, Elith & Leathwick 2009) are of fundamental importance to many aspects of biological and ecological sciences as well as to environmental management. SDMs quantify the relationship between the environment and a species' distribution. The environment is quantified using spatial climate variables, such as maximum/minimum temperature, temperature in warmest month, amongst many others (Soria-Auza *et al.* 2010). These variables are often obtained by querying GIS data bases. Example uses of a SDM are to predict a species' distribution of a study region (Pearson & Dawson 2003), or to project potential change in distribution under climate change scenarios (Forester, DeChaine & Bunn 2013; Wenger *et al.* 2013).

Most spatial climate data sets in use today have been developed using one of several interpolation techniques, which represent a mixture of general numerical methods and specific models. These include the following: inverse-distance weighting (Matheron 1971; Isaaks & Srivastava 1989); various forms of kriging (Phillips, Dolph & Marks 1992; Dodson & Marks 1997); tri-variate splines (Wahba & Wendelberger 1980; Cressie 2003; Hijmans *et al.* 2005; Xu & Hutchinson 2012); local regression (Daly 2006); and regional regression models (Goodale, Aber & Ollinger 1998; Johansson & Chen 2005; Ashcroft & Gollan 2012). These spatial climate data sets are estimates (or predictions) of the true spatial climate and are therefore subject to uncertainty, which itself can also have spatial structure with some regions consistently overestimated and others consistently underestimated (Fernández, Hamilton & Kueppers 2013). In this article, we use PRISM (Parameter–ele-

*Correspondence author. E-mail: j.stoklosa@unsw.edu.au

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

vation Relationships on Independent Slopes Model) as an illustrative example. PRISM is a weighted, local regression technique that accounts for physiographic factors affecting spatial climate variations, and has been used extensively in the United States, Europe and Asia (Daly, Neilson & Phillips 1994; Daly *et al*. 2002; Daly, Helmer & Quinones 2003; Daly *et al*. 2008; Bishop & Beier 2013).

Even if the uncertainty arising from spatial climate variables can be estimated, there remain questions about how this information can be used in SDMs. Can uncertainty in climate variables be incorporated? If so, how? What happens if the uncertainty is ignored? What is the type of change in predictions and/or inference expected if uncertainty is incorporated? How might extrapolation (for example a changed climate) behave under an uncertain model? This paper sets out to answer these questions.

Accounting for uncertainty in explanatory variables (through what is commonly referred to as *measurement error* models or *errors-in-variables* models) is a well-known and important topic in many applied fields, such as engineering and medical studies (Fuller 1987; Carroll *et al*. 2006). Uncertainty in explanatory variables has two main implications: bias in estimates of regression coefficients, and a loss of power (to determine whether explanatory variables are important), which combined, Carroll *et al*. (2006) refer to as the 'double whammy'. Generally, more uncertainty in the explanatory variables induces more bias in the estimates of the model's parameters, which can have adverse consequences for model predictions too. Errors-in-variables models aim to avoid the 'double whammy' using one of a variety of statistical methods (Carroll *et al*. 2006). In order for these methods to be applicable, some *known* information on the uncertainty in the explanatory variables is required (e.g. the variance) which is usually obtained from the measuring device/procedure/model, or some validation data set, or from repeated measures. However, it is critical that we specify the type of underlying error in the explanatory variables. In section 'Classical vs. Berkson Errors', we discuss two common types (classical and Berkson errors) in greater detail and highlight their implications for SDMs.

In the SDM context, several attempts have been made to either examine or account for uncertainty in spatial climate variables – for example: Elston *et al*. (1997) proposed an adjustment in regression coefficients; Foster, Shimadzu & Darnell (2012) used errors-in-variables models to account for explanatory variables that are overly smooth; Denham, Falk & Mengersen (2011) considered a conditional independence model in a hierarchical Bayesian framework using a Gibbs sampler where uncertainty in the explanatory variables was accounted for using a validation data set; McInerny & Purves (2011) investigated uncertainty in explanatory variables attributed to fine-scale environmental variation, and proposed a general correction for regression dilution (or attenuation) also based on Bayesian methods; Fernández, Hamilton & Kueppers (2013) examined the influence of interannual variability, topographic heterogeneity and the distance to nearest weather station; and Hefley *et al*. (2014) investigated the presence of location uncertainty in presence-only data.

We use two statistical errors-in-variables methods: (i) hierarchical modelling and (ii) simulation–extrapolation (SIMEX) – both of which are well developed. In contrast to the existing approaches (those referenced above), our presented methodology differs from (and complement) in the assumptions made about the underlying prediction process. We present a case study where estimates of uncertainty in temperature variables are available, via the PRISM software (Daly *et al*. 2008), and we relate them to the species distribution of the Carolina wren *Thryothorus ludovicianus* in the United States. Additionally, we present simulation studies to investigate bias, efficiency and statistical power, and look at how well SDMs predict and project to new scenarios when prediction error is both ignored and accounted for.

Species distribution modelling and data

In this article, we focus on SDMs fitted using generalized linear models (GLMs; McCullagh & Nelder 1989) using logistic regression of presence/absence data. SDMs are currently implemented using a variety of different methods: for example MaxEnt (Phillips & Dudík 2008); hierarchical Bayes (Clark 2005); generalized additive models; boosted regression trees; or multivariate adaptive regression splines (Hastie, Tibshirani & Friedman 2001). However, most of these are generalisations of GLMs (in fact MaxEnt is exactly a penalized Poisson GLM; Fithian & Hastie 2013; Renner & Warton 2013), and there is an opportunity to extend errors-in-variables models to these other modelling frameworks.

SPATIAL CLIMATE VARIABLES DATA

PRISM was used to develop grids that reflected, as closely as possible, the current state of knowledge of spatial climate patterns in the USA. PRISM calculated a local climate-elevation regression function for each grid cell on a digital elevation model, and stations entering the regression were assigned weights based primarily on the physiographic similarity of the station to the grid cell. Factors considered were distance, elevation, coastal proximity, topographic facet orientation, vertical atmospheric layer, topographic position and orographic effectiveness of the terrain. Information on these physiographic factors was provided to PRISM by means of grids generated by models of marine intrusion into adjacent inland areas (Daly, Helmer & Quinones 2003), topographic orientation (Daly *et al*. 2002), relative position on the topography (Daly *et al*. 2007) and others.

We used PRISM to obtain the predicted spatial climate variables and the uncertainty estimates (see section 'Obtaining uncertainty information from PRISM'). These estimates were generated as part of a USA Department of Agriculture project to interpolate 1971–2000 monthly averages of minimum and maximum temperature and precipitation to a regular grid covering the conterminous United States (Daly *et al*. 2008). Grid cell resolution was 30 arc-seconds, which averages to about 800 m on a side. Specifically, we obtained model-generated 1971–2000 mean minimum tem-

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

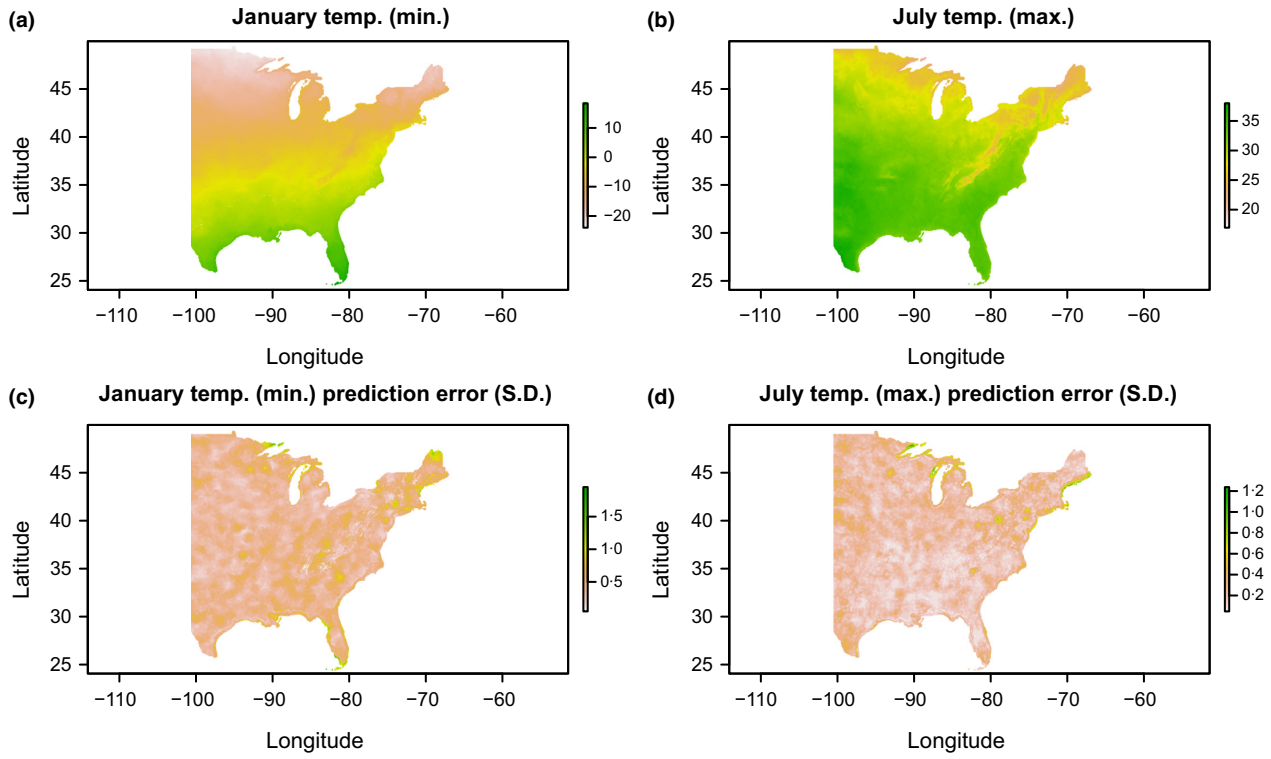


Fig. 1. USA temperature map data for: (a) January minimum; (b) July maximum; (c) predicted standard deviations for January minimum; and (d) predicted standard deviations for July maximum. Temperature is measured in degrees Celsius. Note that these data were standardized in our analysis.

peratures in January, and 1971–2000 mean maximum temperatures in July for conterminous USA. These data are plotted in Fig. 1(a,b).

PRESENCE/ABSENCE DATA FOR THE CAROLINA WREN

Similar to Royle *et al.* (2012), we obtained presence/absence data collected on the Carolina wren (*Thryothorus ludovicianus*) from the North American Breeding Bird Survey (BBS). The presence/absence points were obtained from observers counting all bird species seen or heard from surveyed BBS routes at several points along transects across North America. As the spatial location data were only available for the first points along transects, we used these in our analysis. We considered data from 2010, where $n = 1048$ presence/absence points were recorded. In Fig. 2, we plot the observed presence/absence points. Our analyses differ from those of Royle *et al.* (2012) in a number of ways – in the year of sampling and explanatory variables considered, and in the methodology used to analyse the data. Temperature variables were used as explanatory variables because they were available at a suitably fine resolution and because uncertainty information (which we will also refer to as prediction error) was available for both explanatory variables, see section 'Obtaining uncertainty information from PRISM'.

GENERALIZED LINEAR MODELS

Throughout the article, we will denote observable quantities by lower case and unobservable quantities by upper case. Let $y = (y_1, \ldots, y_n)^T$ be the observable response variable (such as count or presence/absence data) collected from site $i = 1, \ldots, n$ which is related to some set of true and unobservable climate variables $X = (X_1, \ldots, X_n)^T$. Our objective was to understand the nature of the relationship between y and X. The problem is that X is not measured directly; instead, we have the predicted climate variables $w = (w_1, \ldots, w_n)^T$ which have been predicted with uncertainty denoted by $U = (U_1, \ldots, U_n)^T$, and so w only approximates the actual climate experienced by species.

The first model we will consider is one which does not take into account this uncertainty. That is, a model which naïvely treats w (the error contaminated climate variable) as if it were the true climate. Let w_i be a q-length vector of explanatory variables with associated regression parameters $\beta = (\beta_1, \ldots, \beta_q)^T$. For the GLM, we incorrectly assume $f(y_i|w_i;\beta)$ where $f(\cdot\,|\,\cdot)$ belongs to the exponential family, and write $\mu_i = E(y_i|w_i) = h(w_i^T\beta)$, where h is the inverse logit function.

In our case study, we assume Carolina wrens respond to climate and the problem we have is that we are predicting the climate imperfectly (or subject to some prediction error). Initially, we naïvely fitted the above GLM using both the min./max. temperatures explanatory variables (which were standardized prior to fitting) as quadratic effects to the Carolina wren data presented in section 'Presence/absence data for the Carolina wren'. The predicted presence probabilities obtained from this GLM fit are plotted in Fig. 2. In section 'Incorporating uncertainty from spatial climate variables into SDMs', we will develop models which take into account the uncertainty from the explanatory variables.

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

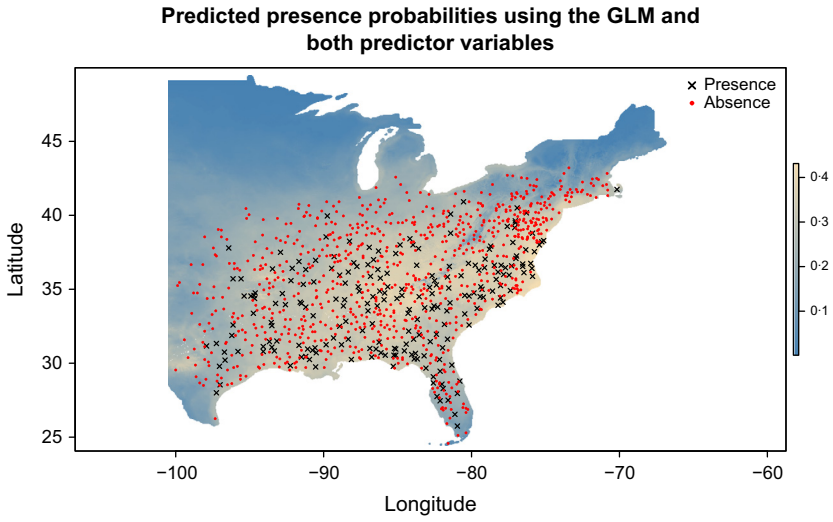


Fig. 2. Presence (black cross) and absence (red circle) points (n = 1048) for the Carolina wren plotted on the predicted presence arising from modelling the GLM using both temperature variables as quadratic terms.

Obtaining uncertainty information from PRISM

PRISM interpolation uncertainties were estimated by Daly *et al.* (2008) using two methods: single-deletion jack-knife cross-validation with replacement, and the prediction interval of the PRISM climate-elevation regression function. The jack-knife method involved removing, in turn, each station value from the data set, estimating it in its absence, and returning the station to the data set. While jack-knife error estimation is a useful independent measure of interpolation uncertainty, the disadvantage is that information is provided at point station locations only and not as a continuous grid.

In contrast, model-based uncertainty estimates have the advantage of being available as continuous grids. However, these estimates rely at least partly on the very same assumptions used in the interpolation process itself and therefore typically underestimate the true interpolation error. As PRISM uses weighted linear regression to estimate precipitation or temperature as a function of elevation, standard methods for calculating prediction intervals (PI) for the response variable could be used.

Unlike a confidence interval (CI), the PI takes into account both the variation in the possible location of the expected value of the response variable for a given explanatory variable, and variation of individual values of the response variable around the expected value. We used a 70% prediction interval (PI70) – further details on the calculation of PI70 are available in section 5 of Daly *et al.* (2008).

The premise behind interpreting the PI70 spatially is that it is relatively large when there is a high degree of scatter about the local regression line, indicating a poor relationship between climate and elevation and suggesting a poor prediction. This tends to occur at locations far from stations, in areas within transition zones between two or more climatic regimes (such as coastal temperature boundaries), or at elevations in the vertical transition between the boundary layer and free atmosphere during temperature inversions. PI70 also increases the farther the prediction is extrapolated away from the mean regression elevation. This is seen in high-mountain areas that are well above the highest stations in the vicinity and thus have relatively large intervals.

To account for uncertainty in the temperature data, we make use of the available predicted standard deviations (obtained from the PI70s, see Daly *et al.* (2008) by incorporating them into the proposed errors-in-variables SDMs presented in section 'Incorporating uncertainty from spatial climate variables into SDMs'. These predicted standard deviations are of the same resolution as the spatial climate variables discussed in section 'Spatial climate variables data' and are plotted in Fig. 1(c,d).

We note that PRISM can calculate regression prediction intervals for any variable that is being interpolated by the model, so other environmental variables, such as precipitation, could also estimate prediction error similar to the above temperature variables. Other interpolation methods and their software may also estimate some form of uncertainty from the predicted environmental variables, for example: kriging provides estimation variances with each grid cell prediction; and WORLDCLIM (Hijmans *et al.* 2005) produces single-value uncertainty estimates (e.g. R^2 or RMSE values) across the entire study area, although realistically one would expect the uncertainty to vary spatially.

Incorporating uncertainty from spatial climate variables into SDMs

In this section, we discuss two different types of uncertainty associated with errors-in-variables models. We then present two statistical approaches: both of which take into account uncertainty from spatial climate variables in SDMs.

CLASSICAL VS. BERKSON ERRORS

The two most common types of underlying uncertainty (sometimes referred to as 'error') in the explanatory variables are as follows: (i) *classical* error and (ii) *Berkson* error (Fuller 1987; Carroll *et al.* 2006). In this article, we focus on classical error and refer to Carroll *et al.* (2006), McInerny & Purves (2011)

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

and Foster, Shimadzu & Darnell (2012) for Berkson errors; however, as the analyst can choose which error to consider in their analysis, we will discuss and distinguish both error types.

A classical error model considers the predicted (or observed) explanatory variables as noisy realisations of the true explanatory variables – that is $w = X + U$ where the errors are centred around zero, $E(U|X) = 0$. For SDMs, this model is usually appropriate when the true climate variables are thought to be an 'average'. Ecologically, the model is appropriate if the species is assumed to respond to the expected value but not the realisation. For example, a species may tolerate individual years that are colder than the mean January minimum, but prolonged exposure may be intolerable (i.e. a colder expectation).

A Berkson error model considers that the predicted explanatory variables are an overly smooth realisation of the true explanatory variables – that is $X = w + U$ where errors for a given prediction of the explanatory variable are centred around zero, $E(U|w) = 0$. For SDMs, it may be appropriate to assume Berkson errors when the true explanatory variables are thought to be noisier than the predicted explanatory variable, see McInerny & Purves (2011) and Foster, Shimadzu & Darnell (2012). For example, a species that is intolerant of cold weather may be absent from relatively warm sites (as measured by average temperature) because the temperature sometimes falls below the species' cold tolerance.

We also make the standard assumptions that U: (i) has some known distribution and (ii) is additive. Note that if no distributional assumption is made on the prediction errors U, then nonparametric alternatives could also be considered, see Aitkin & Rocci (2002) and Carroll *et al.* (2006).

Which of these two types of error models to consider will depend on what the analyst believes to be the 'true underlying explanatory variable', and how the data were collected/measured. The analyst must take into account: how and whether the species responds to a particular climate observation (Berkson); or that it might respond to an average, such that relatively minor deviations from this are immaterial (classical).

If the analyst believes that the species responds to average explanatory variable (e.g. average min. winter temperature), then the relevant uncertainty measure describes the average – the standard error. Alternatively, if the analyst believes that the species responds to the actual explanatory variable (which is predicted but not observed), then the relevant uncertainty measure describes the spread of the covariate around its prediction – the standard deviation. Note that the standard deviation will always be larger than the standard error.

As we are assuming that Carolina wrens respond to climate (which is an expectation), we use classical error. This also implies that predicted standard errors of the predicted climate should be used. However, as predicted standard errors were not available through PRISM, we used the available predicted standard deviations as an approximate alternative. These predicted standard deviations serve as upper-bounds to the required standard errors. It should be noted however that additional bias in model estimates can arise if the predicted standard deviations are too large.

HIERARCHICAL MODELLING

Hierarchical models, which are constructed as joint conditional probabilities of the underlying process, are commonly used when accounting for different sources of uncertainty in an ecological setting (Cressie *et al.* 2009). This ideology falls quite naturally in our framework, such that the uncertainty in the explanatory variables can be modelled and carried over to SDMs.

Suppose now that $f(y_i|X_i;\beta)$ arises from some hierarchical structure generated by X_i. Following Schafer (1987) and Aitkin & Rocci (2002), we have some $f(w_i|X_i)$ and $f(X_i)$. Recall that a classical error model assumes the following additive error structure:

$$w_i = X_i + U_i,$$

where $U_i|X_i \sim N(0, \sigma_u^2)$ is the prediction error with variance σ_u^2. In our case study, σ_u^2 is treated as a heteroskedastic variance, with a different variance estimate available in each grid cell of PRISM output. The joint probability density function is given by:

$$\begin{aligned} f(y, w; \beta) &= \prod_{i=1}^{n} f(y_i, w_i; \beta) \\ &= \int \left\{ \prod_{i=1}^{n} f(y_i, w_i, X_i; \beta) \right\} dX \\ &= \int \left\{ \prod_{i=1}^{n} f(y_i|X_i; \beta) f(w_i|X_i) f(X_i) \right\} dX \end{aligned} \qquad \text{eqn 1}$$

We aim to estimate the parameters of interest β using maximum likelihood estimation and therefore must integrate out the latent X in the estimation procedure. For non-normal response data, a closed form expression for the marginal likelihood of (eqn 1) – that is the joint likelihood after integrating out the latent X – is not obtainable. However, there are a number of different estimation methods which can be used, such as Markov chain Monte Carlo (Cressie & Wikle 2011; Gelman *et al.* 2013), or the expectation–maximization (EM) algorithm, following the works of Schafer (1987) and Li, Tang & Lin (2009).

We used a variation of the EM-algorithm known as Monte Carlo EM (MCEM, Wei & Tanner 1990). In our MCEM approach, we simulated replicate Monte Carlo values for measurement error (from the prior distribution, $N(0, \sigma_u^2)$), then weighted these observations proportional to $f(y_i|X_i;\beta)f(X_i)$, and fitted a GLM on the subsequent estimated explanatory variables. This method has the advantages that it was quite computationally efficient and it is quite general. It can be readily modified to handle a range of variations on the standard GLM – such as including interaction or quadratic terms, smoothers, GAMs, mixed effects – and could in principle handle MARS, LASSO, etc. (Hastie, Tibshirani & Friedman 2001) with little technical difficulty. Further details on the computation are given in first section of Appendix S1.

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

SIMULATION–EXTRAPOLATION

Simulation–extrapolation (SIMEX, Cook & Stefanski 1994; Carroll *et al*. 2006) is a popular tool when dealing with error in the explanatory variables, particularly if the response is non-normal. It has the advantage that software is currently available to fit errors-in-variables GLMs, and it shares with the MCEM algorithm the advantage that (in principle) it can be applied to any parametric model without the need for modification of the underlying model-fitting algorithm. It also avoids having to integrate out X in (eqn 1) using a straightforward simulation method which we briefly describe in second section of Appendix S1. It is not however a maximum likelihood approach, and its estimation algorithm can incur some loss in efficiency, as investigated in our simulations.

ADDITIONAL REMARKS ABOUT UTILITY OF ERRORS-IN-VARIABLES MODELS

As stated in section 2.6 of Carroll *et al*. (2006), 'Generally, there is no need for the modelling of measurement error to play a role in the prediction problem' – that is if the contaminated explanatory variables are only available as the prediction (or test) data w_{test}, then the error-free model (e.g. a GLM) will generally result in better predictions. This is expected as $\hat{\beta}_{GLM}$ is estimated conditional on w_{train}; then, it follows that the best predictions will arise from the GLM when using w_{test} (assuming that the training and test data come from the same population).

This may seem quite reassuring in the SDM context as prediction is usually the aim. However, it turns out that there are some important cases when naïve models predicting from w_{test} will not work well. First, if test data were measured in a different way with a different amount of prediction error, the errors-in-variables models could be expected to be better. Secondly, and more importantly, when making projections from the fitted model, for example when making climate change projections, we would expect projections from naïve models to be biased, and for the bias to increase as the extent of projection increased. The reason for this is that parameters are biased and hence projections of changes as X changes will be biased. We explore this further in the simulations.

Finally, likelihood-based model selection criteria such as AIC or BIC can be used for both the MCEM approach and SIMEX, but require using Monte Carlo to approximate the marginal likelihood. Alternatively, other measures such as generalized cross-validation (Hastie, Tibshirani & Friedman 2001) could be also employed and used for model selection.

Simulations

To investigate the effects on SDMs with uncertainty in explanatory variables, we conducted several simulation studies. We considered logistic regression with two explanatory variables both generated from the normal distribution with mean 0 and variance 1. The error (U) in the explanatory variables was also assumed to be normally distributed with mean 0 and variance σ_u^2 (over a range from 0·01 to 1).

BIAS, EFFICIENCY AND PREDICTIVE PERFORMANCE SIMULATIONS

First, we considered two scenarios to examine the bias, mean square error (MSE) and coverage probabilities (CP) for the regression coefficients, and predicted performance when predicting to new data. For the first simulation scenario, we set the true intercept and the two (linear) regression coefficients to $\beta = (\beta_0, \beta_1, \beta_2)^T = (0.5, 1, 1)^T$, and in the second simulation scenario, we set $\beta = (0{\cdot}5,0,1)^T$. We investigated the predictive performance by simulating additional (test) data and calculating the MSE of the linear predictor on to the predicted test data. In both of the above scenarios, we set $n_{\text{train}} = 200$ and $n_{\text{test}} = 800$, and considered two types of test data: (i) w_{test} which was generated exactly the same way as the training data and (ii) $w^c_{\text{test}} = w_{\text{test}} + 3$ (e.g. an increased climate change scenario of 3 °C).

We fitted the GLM, SIMEX and MCEM (discussed in section 'Incorporating uncertainty from spatial climate variables into SDMs') and performed 1000 simulations. In Fig. 3, we plotted: (a) the bias, (b) the MSE, and (c) the 95% CP for β_1 against increasing values of error variance for both scenarios. When a slope coefficient was required in the model (as in the left panel of Fig. 3a), the estimates for the GLM were biased, and in general, the 95% CI did not include the true value of the parameter a majority of the time (e.g. 95% CP covered only 20% for β_1 when $\sigma_u^2 = 0{\cdot}5$, Fig. 3c) – the poor coverage for the GLM is a result of the large bias and short CIs. This suggests that estimates of, and inferences about, parameters in a model, and about predicted species distributions (see below), are quite sensitive to classical errors. The MSE and 95% CP were similar for all models until $\sigma_u^2 > 0{\cdot}20$ where the differences between the GLM and errors-in-variables models were more apparent.

When a slope coefficient was not needed in the model (as in the right panel of Fig. 3a), it was estimated with little bias, and accurate CPs were obtained irrespective of whether or not the error in the explanatory variables was accounted for. This implies that a naïve model, which does not account for the error in the variables, will still handle *unimportant* explanatory variables adequately, although see Hefley *et al*. (2014). For the MCEM approach, both the MSE and 95% CP worsened as the error variance had increased (Fig. 3b,c) – as the MSE is a sum of the squared bias and the variance, this suggested that the MCEM yielded larger variances for the coefficient estimates, and may be due to the additional uncertainty involved in accounting for error in explanatory variables (which can be understood as a type of bias-variance trade-off).

The evaluation of the predictive performance is given in Fig. 4 where we plotted the MSE on the linear predictor against increasing values of error variance for both simulation scenarios and both types of test data sets: (i) w_{test} and (ii) w^c_{test}, see above. As expected, in both simulation scenarios, the

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

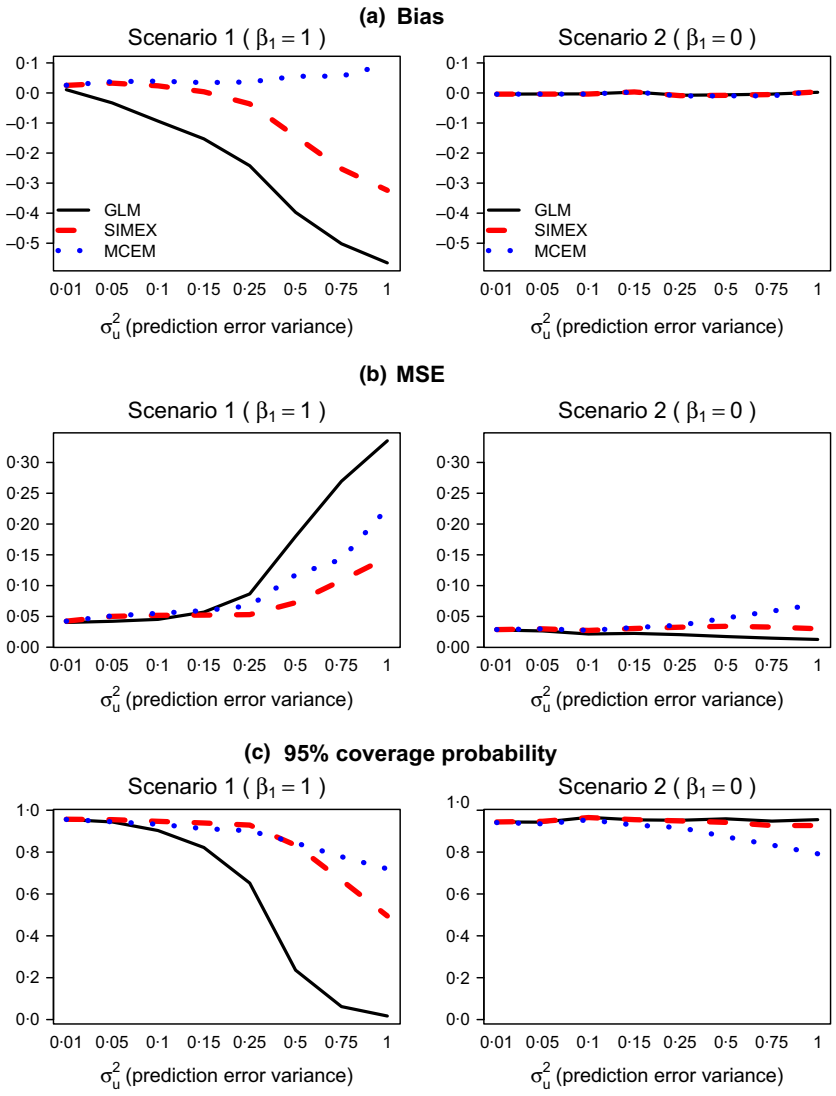


Fig. 3. Plots of the: (a) bias; (b) MSE; and (c) 95% CP for β_1 against increasing values of the prediction error variance for simulation scenarios 1 and 2 (both after 1000 simulations), see text for further details. Notice that when an explanatory variable is in the model (scenario 1), the GLM gives the largest bias and MSE, and poor 95% CP as the prediction error variance increases.

predictive performance was worse for the errors-in-variables models when using the test data w_{test} (see left panel of Fig. 4). However, when making climate change projections using w^c_{test} (see right panel of Fig. 4), the MSE for the GLM had substantially increased; it was reported largest in comparison to the errors-in-variable models for simulation scenario 1, and comparable with MCEM for simulation scenario 2.

STATISTICAL POWER SIMULATIONS

We also examined the statistical power with varying effect sizes and errors-in-variables. A separate simulation study was conducted here because our interest is in investigating the statistical power for different sample sizes. We used the same coefficient values as scenario 1, and looked at two cases where $\sigma^2_u = 0{\cdot}25$ and $\sigma^2_u = 0{\cdot}5$. The null hypothesis assumes the regression coefficients are zero. In Fig. 5, we plotted the statistical power against increasing sample sizes (using 1000 simulations for each sample size) for β_1. In both cases, the MCEM had substantial statistical power compared with the GLM and SIMEX, with SIMEX giving greater statistical power over the GLM when the error in explanatory variables was increased.

PROJECTED CLIMATE CHANGE SIMULATIONS

We further investigated the predictive performance for an increasing climate scenario but now constructed simulated data using the Carolina wren case study (see section 'Presence/absence data for the Carolina wren'). We only used the min. temperature explanatory variable (denoted here as w) and generated new response data by treating w as the true climate explanatory variable and $\hat{\beta}_{\text{MCEM}}$ (see Table 1) as the true coefficient values. We then generated prediction error (using the estimated σ^2_u from PRISM) and added these to both w and $w + 3^{\circ}$C, to create the new observed training and test data, respectively. Each model was fit using the simulated training data, and the MSE of the linear predictor was calculated on the simulated test data. The largest MSE (when using the $w + 3^{\circ}$C test data) was reported for the GLM (97·50), which was clearly outperformed by SIMEX (46·78) and MCEM (40·90).

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

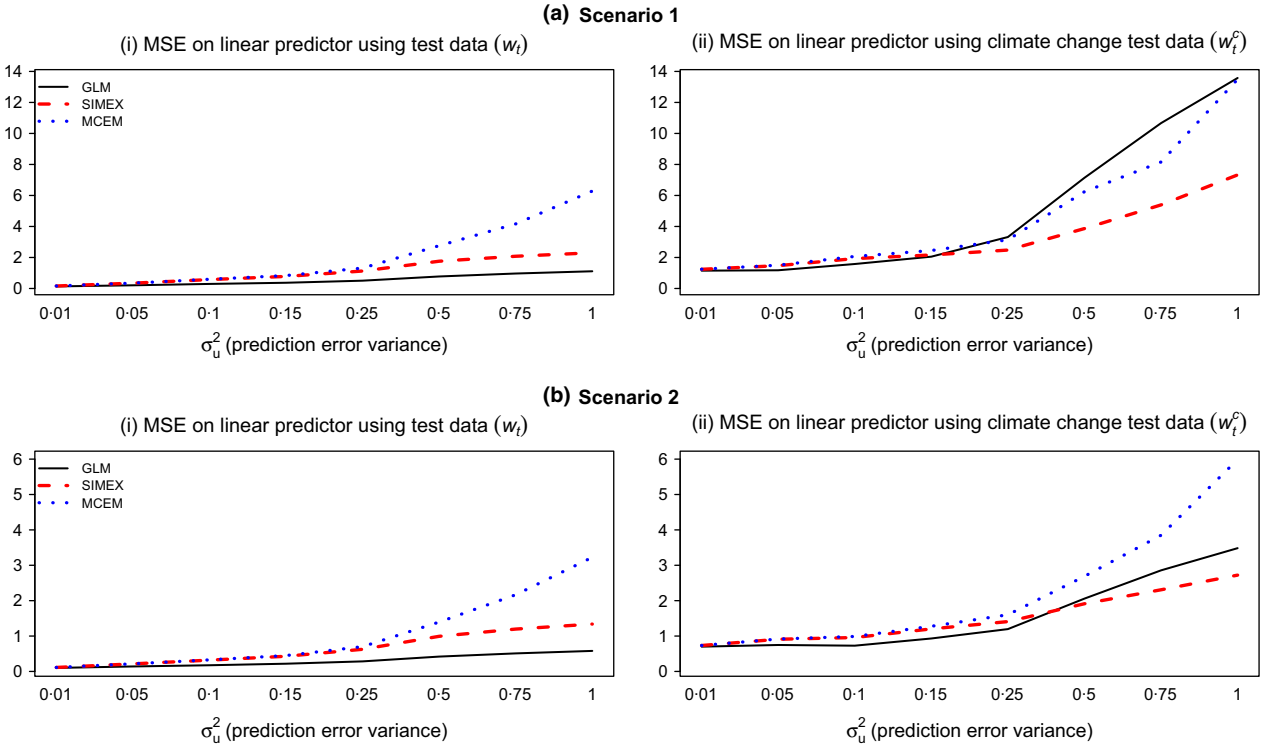


Fig. 4. Plots for the MSE on the linear predictor against increasing values for the error variance for: (a) simulation scenario 1 and (b) simulation scenario 2, using (i) test data w_{test} and (ii) test data under a climate change scenario w_{test}^c after 1000 simulations, see text for further details. Notice the difference in MSE when using w_{test}^c.

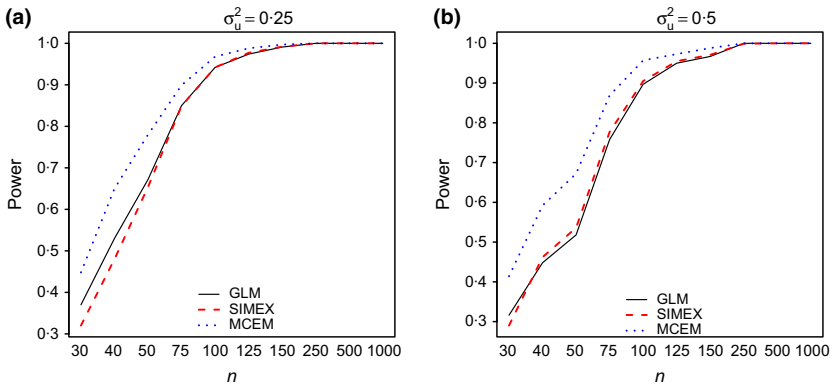


Fig. 5. Statistical power against increasing sample sizes for each model where (a) $\sigma_u^2 = 0{\cdot}25$ and (b) $\sigma_u^2 = 0{\cdot}5$ after 1000 simulations, see text for further details. Notice that the MCEM had substantial statistical power compared with the GLM and SIMEX.

Case study: incorporating uncertainty to the Carolina wren data

To account for uncertainty in the climate variables, we fitted the MCEM and SIMEX methods using both max. and min. temperature covariates. First, we compared BIC values (see section 'Additional remarks about utility of errors-in-variables models') for all models (including the GLM), which contained either both or one temperature climate variable only and modelled these as quadratic terms. We found that the BIC was smallest for quadratic models with the min. temperature climate variable only. Thus, we excluded the max. temperature climate variable, and only fitted quadratic models using the min. temperature climate variable. In Table 1, we reported parameter estimates with 95% CI (in parentheses) and calculated the log-likelihood ($\log f(y_{\text{test}}|w_{\text{test}}; \hat{\beta}_{\text{train}})$) using a block-type cross-validation – that is we divided the data into 16 grids and selected four random grids as the test data. Note that this log-likelihood measure (denoted by CV-LL) was employed as the true linear predictor is unknown for the test data.

First, there was a difference in the min. temperature slope for the errors-in-variables models compared to GLM. This reflects the simulation study results. Also, the standard error estimates for the errors-in-variables models were larger compared with the GLM, which resulted in larger 95% CI, reflecting the additional uncertainty in the model when accounting for error in the climate variable. Not surprisingly, the blocking CV-LL was marginally better for the GLM; however, as

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

Table 1. Parameter estimates with 95% CI (in parentheses) for quadratic models (using the min. temperature covariate only) after fitting GLM, SIMEX and MCEM using the Carolina wren data. The blocking cross-validation log-likelihood (CV-LL) is also reported to evaluate the predictive performance

	GLM	SIMEX	MCEM
$\hat{\beta}_{intercept}$	−1·09 (−1·28, −0·90)	−0·99 (−1·21, −0·77)	−1·07 (−1·27, −0·878)
$\hat{\beta}_{linear}$	0·90 (0·67, 1·14)	1·25 (0·92, 1·58)	1·33 (1·01, 1·66)
$\hat{\beta}_{quadratic}$	−0·46 (−0·63, −0·29)	−0·77 (−1·03, −0·50)	−0·84 (−1·11, −0·57)
CV-LL	−0·603	−0·629	−0·643

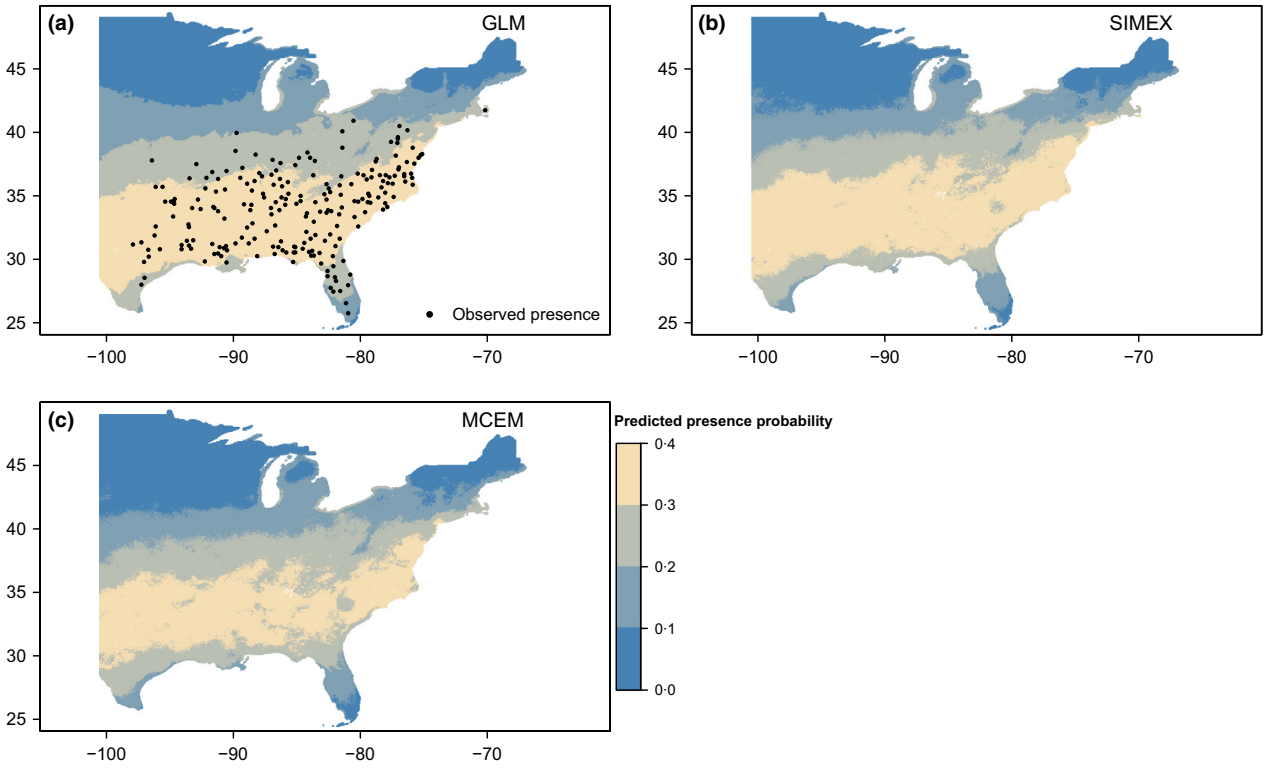


Fig. 6. Predicted presence probabilities for: (a) GLM; (b) SIMEX; and (c) MCEM, using only the min. temperature explanatory variables as quadratic terms for the entire temperature climate map. Plots are presented on the same scale. The observed presences have also been included in (a). Notice there are some differences in the general shape of the maps. The main difference is in the magnitude, especially for the more southern dense areas.

demonstrated in section 'Projected climate change simulations', the predictive performance becomes worse under a future climate change scenario.

In Fig. 6, we plotted the predicted presence probabilities using the entire temperature climate map for each model. The predicted presence probabilities are presented on the same scale. We observed some slight difference in all three species distribution maps, particularly in the magnitude for the more southern less dense areas, for example comparing Fig. 6(a,b). To examine the uncertainty in the predictions for each model/map, we plotted the standard errors of the linear predictor for the entire temperature climate map in Fig. 7, for further details see Appendix S1. The largest standard errors were observed on the boundaries of the north-western and some southern areas, where very few or no presence/absence records were observed and where temperatures were at the extremes of the observed range. Otherwise, uncertainty in the predictions was fairly constant across each map. We also inspected how the estimated models responded to min. temperature in Fig. 8, where there is a clear distinction between GLM and the errors-in-variables models. This is also in keeping with simulation results, where we found downward-biased estimates of slope parameters when prediction error was ignored.

Discussion

Explanatory variables considered in SDMs, in particular climate variables, are predicted with uncertainty. We have investigated the impact of such prediction error, and ways to account for it in the context of SDMs. The main impact of failing to account for error in variables is bias (Fig. 3a), but there is also a loss of power (Fig. 5) as the error increases. Different

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

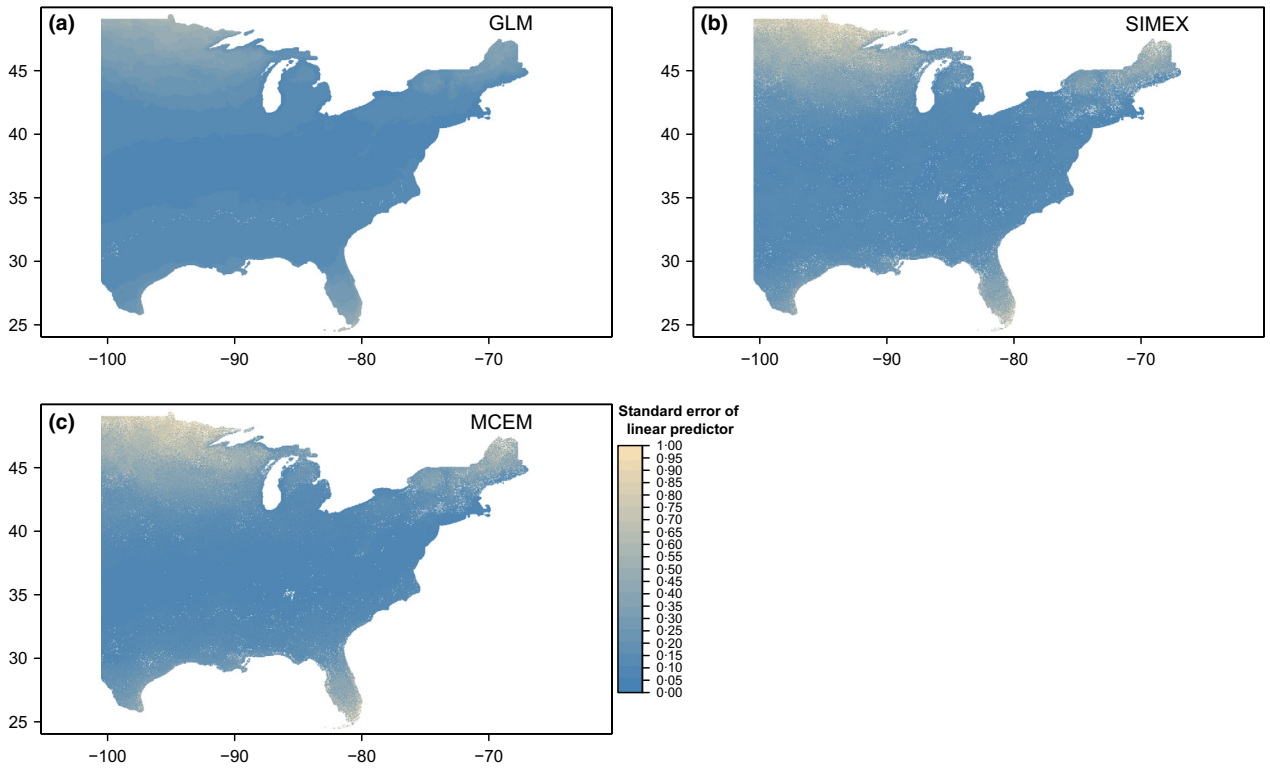


Fig. 7. Standard error of the linear predictors for: (a) GLM; (b) SIMEX; and (c) MCEM for the entire temperature climate map. Note that the white blank dots are NA values.

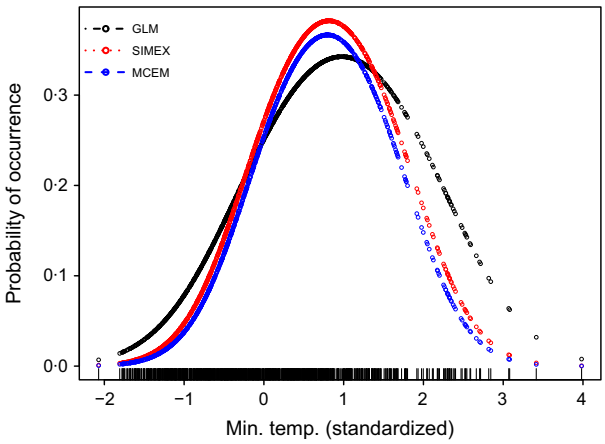


Fig. 8. Fitted occurrence probabilities plotted against min. temperature for the GLM, SIMEX and MCEM.

conclusions could be drawn depending on the model used and whether or not we ignore the errors-in-variables assumption. But while explanatory variables that are informative for species response experienced the 'double whammy' of bias and low power when errors-in-variables were ignored, uninformative variables appeared to be unaffected.

An important consequence of biased parameter estimates is biased projections under changes of environmental variables – for example under different climate change scenarios, as in section 'Additional remarks about utility of errors-in-variables models'. This result has wide ramifications as it is common to use SDMs fitted without uncertainty in explanatory variables for climate change projections, and a reasonably likely consequence of failing to account for such uncertainty is underestimation of climate change effects. The reason being that when prediction error is ignored, climate responses more often than not are estimated to be attenuated (as in Fig. 8); thus, projected climate change effects could also be expected to often be attenuated.

Producing reliable uncertainty estimates from climate models is a challenging task but most climate modelling software does provide uncertainty estimates. We obtained maps of uncertainty in climate variables from PRISM software. For most errors-in-variable approaches, some components of the uncertainty must be assumed known or estimated to a reasonable degree of accuracy. In our case, we obtained an upper-bound of the prediction error variance from the PRISM climate model, which was the best estimate available to us. Recall that these uncertainty estimates are generated from some climate model, hence predictions are based on what the climate model knows and assumes. This is analogous to asking a student to grade their own final exam (Daly 2006). In our case, PI from linear regression will only be accurate if the assumed model is 100% correct. If this assumption fails, it would be difficult to get reliable PIs from simple linear regression (i.e. under the normality assumption). In addition, the form of the uncertainty statistic varies from climate model to climate model, so they may not be comparable, for example

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

PRISM PI70 vs. kriging estimation variance. Obtaining more accurate and efficient uncertainty estimates remains an issue for ongoing research, and we hope there will be improvements in the near future.

On the question of which method to use for fitting errors-in-variables models, SIMEX is a well-known and flexible approach: it can be computationally fast and has an easy to use R-package (Lederer & Kuchenhoff 2006). On the other hand, MCEM can be more naturally extended to spatial models (or more general hierarchical structures), and the design matrix can be easily modified to handle more general regression structures (e.g. interaction terms) and shares similar properties to classical maximum likelihood theory. When considering more sophisticated models, we recommend SIMEX as a natural first step, unless the model is inherently hierarchical, in which case MCEM might be preferable.

POSSIBLE EXTENSIONS

In some cases, the assumption of no spatial autocorrelation in the response variable may be reasonable; however, recent studies have discussed the importance of including spatial correlation in SDMs (Record *et al*. 2013). Further, the spatial autocorrelation may just be a manifestation of an errors-in-variables process where the errors are spatially dependent (Foster, Shimadzu & Darnell 2012). Adding a spatial component to prediction error is also important when an estimated climate map is likely to have a patchy error distribution, with climate variables being consistently over- or under-estimated in particular regions. Therefore, the first and perhaps most important extension to the methods presented in section 'Incorporating uncertainty from spatial climate variables into SDMs' is to include spatiality into the SDM analysis. In Appendix S2, we show how the hierarchical models given in section 'Hierarchical modelling' can be modified to: (i) account for spatial autocorrelation in the response variable and (ii) include spatially autocorrelated prediction error (see Foster, Shimadzu & Darnell 2012), or an environmentally structured error, but further work is needed to implement and evaluate these methods.

We also ignored possible temporal uncertainty in both the climate mapping variables and presence/absence data, which of course could vary if the sampling is conducted at different times. For example, we could follow Xia & Carlin (1998) who included both spatiotemporal effects with uncertainty in covariates, although in this case the response data were normally distributed.

Finally, while we considered presence–absence data here, the issue of prediction error in explanatory variables also arises in presence-only data, and the methods implemented here can be readily extended to handle presence-only data. In fact, presence-only analysis via a point process model can be implemented using GLM software (Baddeley & Turner 2005) and MaxEnt. Hence, methods developed here can be applied to presence-only data relatively easily. We hope to explore these extensions elsewhere.

Acknowledgements

We would like to thank Robert O'Hara, Trevor Hefley and an anonymous referee for providing some helpful suggestions and feedback. We would also like to thank Noel Cressie for his helpful comments and Aritra Sengupta for providing some useful R-code. This project is a product of QUEST (Quantifying Uncertainty in Ecosystem Studies) Research Coordination Network (http://www.quantifyinguncertainty.org/), which is funded by the US National Science Foundation. JS and DIW work was supported by the Australian Research Council Discovery Project (project no. DP0985886, DP130102131), and DIW was supported by Future Fellow (project no. FT120100501). SDF was supported by Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Research Programme (NERP). NERP Marine Biodiversity Hub partners include the Institute for Marine and Antarctic Studies, University of Tasmania; CSIRO Wealth from Oceans National Flagship, Geoscience Australia, Australian Institute of Marine Science, Museum Victoria, Charles Darwin University and the University of Western Australia.

Data accessibility

The North American BBS data were obtained from https://www.pwrc.usgs.gov/bbs, and PRISM climate data sets are available from http://prism.oregonstate.edu.

References

Aitkin, M. & Rocci, R. (2002) A general maximum likelihood analysis of measurement error in generalized linear models. *Statistics and Computing*, **12**, 163–174.

Ashcroft, M.B. & Gollan, J.R. (2012) Fine-resolution (25 m) topoclimatic grids of near-surface (5 cm) extreme temperatures and humidities across various habitats in a large (200×300 km) and diverse region. *International Journal of Climatology*, **32**, 2134–2148.

Baddeley, A. & Turner, R. (2005) Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **6**, 1–42.

Bishop, D.A. & Beier, C.M. (2013) Assessing uncertainty in high-resolution spatial climate data across the US northeast. *PLoS One*, **8**, e70260.

Carroll, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C.M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edn. Chapman & Hall, London.

Clark, J.S. (2005) Why environmental scientists are becoming Bayesians. *Ecology Letters*, **8**, 2–14.

Cook, J.R. & Stefanski, L.A. (1994) Simulation–extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, **89**, 1314–1328.

Cressie, N. (2003) *Statistics for Spatial Data*, revised edn. John Wiley & Sons, Inc, New York.

Cressie, N. & Wikle, C.K. (2011) *Statistics for Spatio-temporal Data*. John Wiley & Sons Inc., Hoboken, NJ.

Cressie, N., Calder, C.A., Clark, J.S., Ver Hoef, J.M. & Wikle, C.K. (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, **19**, 553–570.

Daly, C. (2006) Guidelines for assessing the suitability of spatial climate data sets. *International Journal of Climatology*, **26**, 707–721.

Daly, C., Neilson, R.P. & Phillips, D.L. (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology*, **33**, 140–158.

Daly, C., Gibson, W.P., Taylor, G.H., Johnson, G.L., & Pasteris, P. (2002) A knowledge-based approach to the statistical mapping of climate. *Climate Research*, **22**, 99–113.

Daly, C., Helmer, E.H. & Quinones, M. (2003) Mapping the climate of Puerto Rico, Vieques, and Culebra. *International Journal of Climatology*, **23**, 1359–1381.

Daly, C., Smith, J.W., Smith, J.I. & McKane, R.B. (2007) High-resolution spatial modeling of daily weather elements for a catchment in the Oregon Cascade Mountains, United States. *Journal of Applied Meteorology and Climatology*, **46**, 1565–1586.

Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J. & Pasteris, P.P. (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, **28**, 2031–2064.

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,

Denham, R.J., Falk, M.G. & Mengersen, K.L. (2011) The Bayesian conditional independence model for measurement error: applications in ecology. *Environmental and Ecological Statistics*, **18**, 239–255.
Dodson, R. & Marks, D. (1997) Daily air temperature interpolated at high spatial resolution over a large mountainous region. *Climate Research*, **8**, 1–20.
Elith, J. & Leathwick, J. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.
Elston, D.A., Jayasinghe, G.J., Buckland, S.T., Macmillan, D.C. & Aspinall, R.J. (1997) Adapting regression equations to minimise the mean squared error of predictions made using covariate data from a GIS. *International Journal of Geographic Information Science*, **11**, 265–280.
Fernández, M., Hamilton, H. & Kueppers, L.M. (2013) Characterizing uncertainty in species distribution models derived from interpolated weather station data. *Ecosphere*, **61**. doi: 10.1890/ES13-00049.1.
Fithian, W. & Hastie, T. (2013) Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, **7**, 1837–1939.
Forester, B.R., DeChaine, E.G. & Bunn, A.G. (2013) Integrating ensemble species distribution modelling and statistical phylogeography to inform projections of climate change impacts on species distributions. *Diversity and Distributions*, **19**, 1480–1495.
Foster, S.D., Shimadzu, H. & Darnell, R. (2012) Uncertainty in spatially predicted covariates: is it ignorable? *Journal of the Royal Statistical Society, Series C,* **61,** 637–652.
Fuller, W.A. (1987) *Measurement Error Models.* John Wiley & Sons, Inc, New York.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2013) *Bayesian Data Analysis*, 3rd edn. Chapman & Hall/CRC, London.
Goodale, C., Aber, J. & Ollinger, S. (1998) Mapping monthly precipitation, temperature, and solar radiation for Ireland with polynomial regression and a digital elevation model. *Climate Research*, **10**, 35–49.
Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning*. Springer, New York.
Hefley, T.J., Baasch, D.M., Tyre, A.J. & Blankenship, E.E. (2014) Correction of location errors for species distribution models. *Methods in Ecology and Evolution*, **5**, 207–214.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.
Isaaks, E.H. & Srivastava, R.M. (1989) *An introduction to Applied Geostatistics.* Oxford University Press, New York.
Johansson, B. & Chen, D. (2005) Estimation of areal precipitation for runoff modelling using wind data: a case study in Sweden. *Climate Research*, **29**, 53–61.
Lederer, W. and Kuchenhoff, H. (2006) A short introduction to the SIMEX and MCSIMEX. *R News*, **6.4**, 26–31.
Li, Y., Tang, H. & Lin, X. (2009) Spatial linear mixed models with covariate measurement errors. *Statistica Sinica*, **19**, 1077–1093.
Matheron, G. (1971) *The Theory of Regionalized Variables and Its Applications.* Cahiers du Centre de Morphologie Mathematique, Ecole des Mines, Fountainebleau, France.
McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall, London.
McInerny, G.J. & Purves, D.W. (2011) Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.
Pearson, R.G. & Dawson, T.P. (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361–371.
Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.
Phillips, D.L., Dolph, J. & Marks, D. (1992) A comparison of geostatistical procedures for spatial analysis of precipitation in mountainous terrain. *Agricultural and Forest Meteorology*, **58**, 119–141.
Record, S., Fitzpatrick, M.C., Finley, A.O., Veloz, S. & Ellison, A.M. (2013) Should species distribution models account for spatial autocorrelation? A test of model projections across eight millennia of climate change. *Global Ecology and Biogeography*, **22**, 760–771.
Renner, I.W. & Warton, D.I. (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**, 274–281.
Royle, A.R., Chandler, R.B., Yackulic, C. & Nicholas, J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.
Schafer, D.W. (1987) Covariate measurement error in generalized linear models. *Biometrika*, **72**, 385–391.
Soria-Auza, R.W., Kessler, M., Bach, K., Barajas-Barbosa, P.M., Lehnert, M., Herzog, S.K. & Böhner, J. (2010) Impact of quality of climate models for modelling species occurrences in countries with poor climatic documentation: a case study from Bolivia. *Ecological Modelling*, **221**, 1221–1229.
Wahba, G. & Wendelberger, J. (1980) Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review*, **108**, 1122–1143.
Wei, G.C.G. & Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**, 699–704.
Wenger, S.J., Som, N.A., Dauwalter, D.C., Isaak, D.J., Neville, H.M., Luce, C.H. *et al.* (2013) Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. *Global Climate Change*, **19**, 3343–3354.
Xia, H. & Carlin, B.P. (1998) Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Statistics in Medicine*, **17**, 2025–2043.
Xu, T. & Hutchinson, M. (2012) *ANUCLIM 5.1 User's Guide*. Australian National University, Canberra, ACT.

Received 15 May 2014; accepted 10 June 2014
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Algorithms and estimation details.

Appendix S2. Extension to spatial models.

Appendix S3. R-code for the models discussed in section 'Incorporating uncertainty from spatial climate variables into SDMs'.

© 2014 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of the British Ecological Society.,