Journal of Biogeography (J. Biogeogr.) (2011) 38, 1497–1509

ORIGINAL ARTICLE

Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling

Babak Naimi[1,2]*, Andrew K. Skidmore[1], Thomas A. Groen[1] and Nicholas A. S. Hamm[1]

[1]*Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE, Enschede, The Netherlands,* [2]*Graduate School of the Environment and Energy, Science and Research Branch, IAU, Tehran, Iran*

*Correspondence: Babak Naimi, Department of Natural Resources, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE, Enschede, The Netherlands.
E-mail: naimi@itc.nl

ABSTRACT

Aim To investigate the impact of positional uncertainty in species occurrences on the predictions of seven commonly used species distribution models (SDMs), and explore its interaction with spatial autocorrelation in predictors.

Methods A series of artificial datasets covering 155 scenarios including different combinations of five positional uncertainty scenarios and 31 spatial autocorrelation scenarios were simulated. The level of positional uncertainty was defined by the standard deviation of a normally distributed zero-mean random variable. Each dataset included two environmental gradients (predictor variables) and one set of species occurrence sample points (response variable). Seven commonly used models were selected to develop SDMs: generalized linear models, generalized additive models, boosted regression trees, multivariate adaptive regression spline, random forests, genetic algorithm for rule-set production and maximum entropy. A probabilistic approach was employed to model and simulate five levels of error in the species locations. To analyse the propagation of positional uncertainty, Monte Carlo simulation was applied to each scenario for each SDM. The models were evaluated for performance using simulated independent test data with Cohen's Kappa and the area under the receiver operating characteristic curve.

Results Positional uncertainty in species location led to a reduction in prediction accuracy for all SDMs, although the magnitude of the reduction varied between SDMs. In all cases the magnitude of this impact varied according to the degree of spatial autocorrelation in predictors and the levels of positional uncertainty. It was shown that when the range of spatial autocorrelation in the predictors was less than or equal to three times the standard deviation of the positional error, the models were less affected by error and, consequently, had smaller decreases in prediction accuracy. When the range of spatial autocorrelation in predictors was larger than three times the standard deviation of positional error, the prediction accuracy was low for all scenarios.

Main conclusions The potential impact of positional uncertainty in species occurrences on the predictions of SDMs can be understood by comparing it with the spatial autocorrelation range in predictor variables.

Keywords

Artificial dataset, error propagation, positional uncertainty, spatial autocorrelation, species distribution models, stochastic simulation.

INTRODUCTION

Species distribution models (SDMs) are used to infer the ecological requirements of species as well as to predict their geographic distribution. These models have become important in a range of applications including regional biodiversity assessment, conservation biology, evolutionary biology, epidemiology, wildlife management and conservation planning (Skidmore *et al.*, 1996; Segurado & Araújo, 2004; Elith *et al.*, 2006; Elith & Leathwick, 2009; Franklin, 2010). Species distribution modelling is usually based on statistical relationships between species occurrences and corresponding environmental variables. A key component in this process is estimation of species distribution in ecological space, which can be useful to predict their potential distribution in geographic space (Peterson, 2006). Many models are now available to explore this relationship, although these techniques may differ in their ability to summarize useful relationships between response and predictor variables (Segurado & Araújo, 2004). These models vary in the kind of species data they use (e.g. presence/absence versus presence only), the form of their output (a continuous or a binary prediction), the type of relationship they assume (from simple linear to complex nonlinear), how they estimate the distribution of the species (using parametric versus nonparametric approaches), how they select relevant predictor variables, whether variable contributions are weighted, and whether they allow for interactions of explanatory variables (Guisan & Zimmermann, 2000; Elith *et al.*, 2006; Austin, 2007). Different models may yield different results, even when calibrated with the same response and predictor variables (Araújo & Guisan, 2006).

Many SDMs have been developed using presence/absence or presence-only species occurrence data. The great majority of these data, especially in the form of presence-only data from museum or herbarium collections (Graham *et al.*, 2004) or from volunteer observation networks, are available increasingly over the Internet. One of the problems with these data is the uncertainty about where the occurrence was located. This uncertainty is caused by a variety of factors, including inaccuracy in the measurement of location, failure to specify the geographical datum, georeferencing error and operator error (Graham *et al.*, 2004, 2008; Rowe, 2005).

Recent studies have addressed the impact of positional error in species occurrences on SDM accuracy. Graham *et al.* (2008) explored whether positional error in species occurrence data affected SDM performance, with a focus on comparison of models. They introduced a random error (up to 5 km) to the location of presence-only species data and evaluated how it influenced the prediction accuracy of 10 different models. They concluded that SDMs are, in general, robust to positional errors. Johnson & Gillingham (2008) assessed the sensitivity of a logistic-regression SDM to 20 levels of errors (from 50 to 1000 m) in presence/absence locations of a species, sampling bias, error in environmental data (misclassification of land classes) and model order. Their results showed that the species positional error made the greatest contribution to the reduction in prediction accuracy (Johnson & Gillingham, 2008). Osborne & Leitão (2009) explored the impact of extreme and typical positional errors in both species and environmental data on the performance and ecological interpretation of three different SDMs. Concurring with Graham *et al.* (2008), their results showed that species positional errors had a small effect on the predictions from many models. Osborne & Leitão (2009) also raised an interesting issue, namely that the impact of positional errors on SDMs may be understood by examining spatial autocorrelation in predictor variables. They examined the relationship between spatial autocorrelation (quantified by Moran's *I*) in predictors and the consistency of the contribution of variables to models for four scenarios of positional error. They proposed that, if high spatial autocorrelation reduces the impact of positional error, it should show a link between autocorrelation of a variable and the consistency of that variable's contribution to a model. They found a weak but significant relationship, $R^2 = 0.32$, and concluded that the degree to which spatial autocorrelation confers resilience to positional error is more complex than they first expected and needs further study.

Spatial autocorrelation is a statistical property of most ecological variables (Legendre, 1993) and represents the relationship between values of the given variable at different geographical separations. It is hypothesized that, in species distribution modelling, errors in species location will matter less if nearby locations have similar environmental characteristics to the true location (Osborne & Leitão, 2009). Therefore, the robustness of a SDM to species positional uncertainty is expected to be affected positively by spatial autocorrelation in environmental variables.

This study aims to assess and test the interaction between spatial autocorrelation in predictors and positional uncertainty in species occurrences. Different scenarios were designed to assess the propagation of positional uncertainty through SDMs. This experiment was conducted by implementing a series of commonly used species distribution models.

MATERIALS AND METHODS

The study was implemented in four main stages, as follows.

1. Designing scenarios and simulating artificial datasets. Each scenario included two environmental variables and the distribution of a virtual species.

2. Introducing error at the locations of species occurrence and generating realizations of uncertainty in the positions.

3. Positional uncertainty propagation using Monte Carlo simulation for seven SDMs.

4. Evaluating the results for prediction performance and assessing their interaction with spatial autocorrelation in predictors.

Simulating artificial datasets

To explore the interaction between spatial autocorrelation and positional uncertainty both were varied in a controlled way

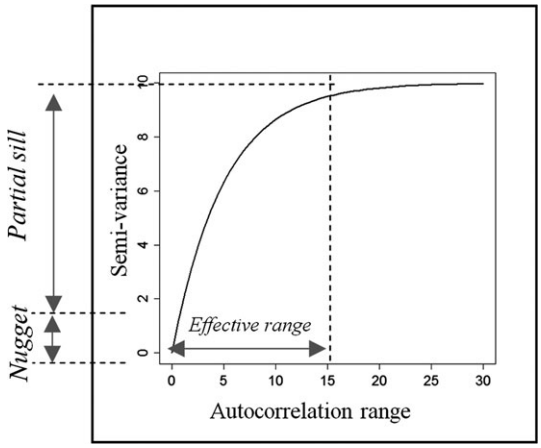

Figure 1 The exponential variogram model for range = 15, nugget = 0 and partial sill = 10.

over a range of values. To achieve this, environmental variables (predictors) and species occurrences (response variable) were simulated. The variogram was used to control the spatial autocorrelation in the predictors. The variogram is commonly used to model the spatial structure in a single variable. Formally, it is defined as half the expected squared difference (half the variance of the difference) in the variable value at a specific geographical separation. The variogram parameters are (Webster & Oliver, 2007): (1) the *sill*, which is the total variance and represents the variability in the absence of spatial correlation; (2) the *range*, which is the distance at which the variogram approaches the *sill*; and (3) the *nugget effect*, which is a combination of spatially unstructured variance (e.g. attribute error) and spatially structured variance at distances shorter than the minimum measurement separation. The *sill* minus *nugget* is known as *partial sill* or structural variance (Fig. 1).

Datasets were generated covering a range of spatial autocorrelation in predictor variables. Each dataset included two artificial environmental gradients and the distribution of one virtual species. Unconditional simulation was used to construct regular grids of 150 × 150 cells for each environmental gradient. Unconditional simulation is a geostatistical technique that generates a realization of a spatially correlated variable, where the spatial correlation is defined by a variogram (Dungan, 1999). Conditional simulation generates a realization using a defined variogram and measurements from the field. Various computational algorithms are available for implementing conditional and unconditional simulation, but all have the objectives of generating a surface with the appropriate correlation structure, as defined by the variogram. For this research the circulant-embedding algorithm (Dietrich & Newsam, 1993) implemented in the RandomFields package v. 1.3.41 in the R programming environment (Schlather, 2009) was used. An exponential variogram model with a *sill* of 10 and a *nugget* of 0 was used for all datasets. The variogram models were assigned different values for the *range* parameter to control the extent of spatial autocorrelation in the predictor variables. In total, 30 levels of range size, from 1 to 30 grid cells, were used, giving a transition from minimum spatial autocorrelation (range = 1) to relatively large-scale spatial autocorrelation (range = 30). Additionally, a white-noise surface (range = 0) was simulated, giving a total of 31 scenarios of spatial autocorrelation.

The distribution of the virtual species was simulated based on assumed species response curves, representing the probability of presence along an environmental gradient. They were constructed with a Gaussian and a decreasing linear function in response to the first and second predictor, respectively (Fig. 2). These functions were used to create an individual measure of habitat suitability for each predictor. The suitability index based on a Gaussian response curve was modelled using equation 1 (ter Braak & Looman, 1986):

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 = p_{\max} - \frac{(x-u)^2}{2t^2} \quad (1)$$

where β_0, β_1 and β_2 are regression parameters, x is the environmental variable, $p(x)$ is the individual suitability index, u is the species' optimum, which is the value of the environmental variable where $p(x)$ is at a maximum ($p_{\max}$), t is the species' tolerance (a measure of ecological amplitude). The parameters $p_{\max}$, u and t are defined as follows (ter Braak & Looman, 1986):

$$\begin{aligned} u &= -\beta_1/(2\beta_2) \\ t &= 1/\sqrt{(-2\beta_2)} \\ p_{\max} &= 1/[1+\exp(-\beta_0 - \beta_1 u - \beta_2 u^2)] \end{aligned} \quad (2)$$

Parameter values of $p_{\max} = 1$, $u = 50$ and $t = 1.1$ were used for the Gaussian response in this study.

Equation 3 was used to model the habitat suitability based on a linear response curve:

$$p(x) = \begin{cases} 1, & x < x_{\text{low}} \\ 1 - \left(\frac{x - x_{\text{low}}}{x_{\text{high}} - x_{\text{low}}}\right), & x_{\text{low}} \le x \le x_{\text{high}} \\ 0, & x > x_{\text{high}} \end{cases} \quad (3)$$

where x represents the environmental variable and x_{low} and x_{high} represent the range between where the species starts to experience negative effects from the environmental variable, causing a decline in its probability of occurrence. Parameter values of $x_{\text{low}} = 45$ and $x_{\text{high}} = 55$ were used for the linear response in this study.

Finally, the habitat suitabilities (Hirzel *et al.*, 2001) for both environmental variables were combined at a grid cell resolution using equation 4:

$$HS = \sum_{i=1}^{2} w_i x_i \quad \text{with} \quad \sum_{i=1}^{2} w_i = 1 \quad (4)$$

where HS represents the habitat suitability score for the virtual species, w_i denotes the weight or importance of variable i and x_i denotes the habitat suitability index based on environmental

variable *i*. The same weight (0.5) was applied for both variables. *HS* can be interpreted as the probability of species occurrence. The habitat suitability values were used to realize presence and absence of the virtual species by applying a threshold of 0.5. One hundred and fifty randomly selected sites were used to train the SDMs and another 150 sites were randomly selected to evaluate the predictive performance of the models. The prevalence (the proportion of presence sample points) for both training and test data was 0.5. To provide sample points for methods that require presence-only data, the presence records in presence–absence samples were used. Figure 2 illustrates schematically the procedure used to simulate the artificial datasets.

Species distribution modelling

To develop SDMs, several commonly implemented models that use presence–absence or presence-only records of species occurrences were selected. The presence–absence models were generalized linear models (GLMs; McCullagh & Nelder, 1989), generalized additive models (GAMs; Hastie & Tibshirani, 1990), boosted regression trees (BRTs; Friedman, 2001), multivariate adaptive regression splines (MARS; Friedman, 1991) and random forests (RFs; Breiman, 2001); the presence-only models were maximum entropy (Maxent; Phillips *et al.*, 2006) and genetic algorithm for rule-set production (GARP; Stockwell & Noble, 1992).

The GLM, GAM, MARS, BRT and RF models were implemented in the R environment v. 2.8.1 (R Development Core Team, 2008) using the bioclimatic niche modelling (BIOMOD) v. 1.1-6.1 package (Thuiller *et al.*, 2009). This enables SDMs to be run simultaneously and incorporates several features for evaluating and examining species–environment relationships. To run the GARP model, the openModeller framework v. 1.0.9 was used. This framework was developed to perform the most common tasks related to species distribution modelling (de Souza Muñoz *et al.*, 2009). All versions of GARP are available in this framework (see below). Maxent was run by using the Maxent software v. 3.3.1 that was developed and introduced by Phillips *et al.* (2006). The specifics of each model are summarized in Table 1.

Spatial autocorrelation in the model residuals can be a problem for regression-based techniques where random sampling and independent residuals are important assumptions (Dormann *et al.*, 2007). Moran's *I* test and correlogram were used to estimate the correlation in the GLM residuals as a function of geographic distance (Schabenberger & Gotway, 2005). The technique for linearly recovered errors (residuals) (LRE), described by Schabenberger & Gotway (2005, p. 315) was followed.

Model evaluation

The predicted distributions of both presence-only and presence–absence SDMs were evaluated for their performance using independent presence–absence data. It is important to use more than one metric to assess model performance because

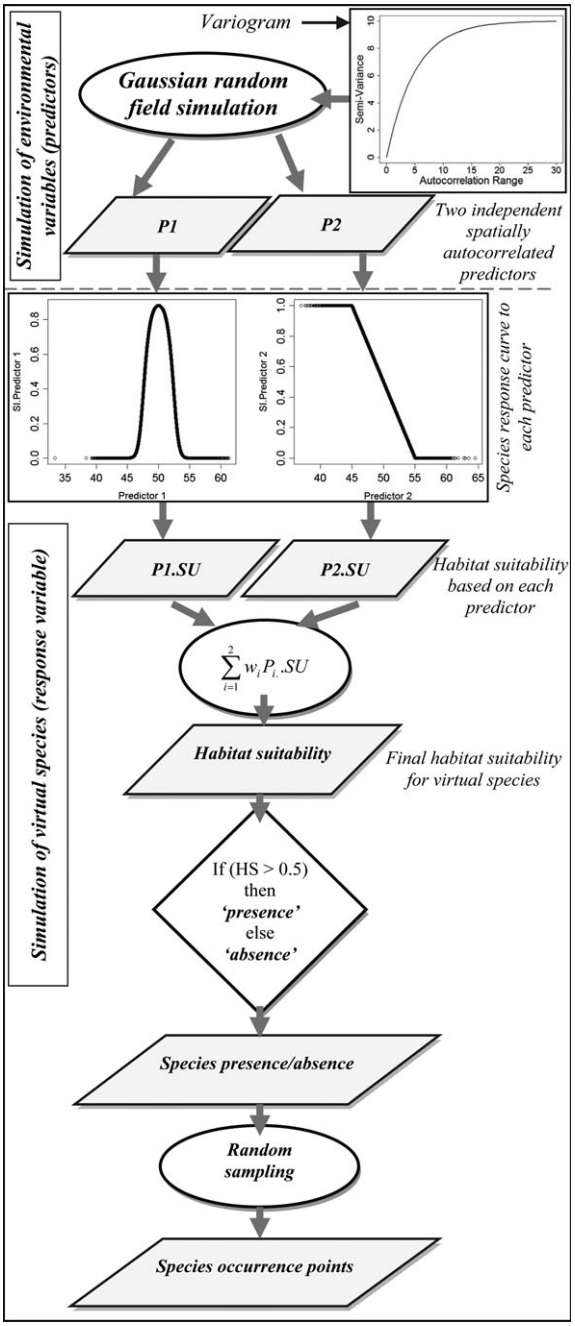

Figure 2 Flow diagram showing the procedure for simulating datasets.

each quantifies different aspects of predictive performance (Elith & Graham, 2009). Two methods were therefore used to measure the predictive performance of models: area under the curve (AUC) of a receiver operating characteristic (ROC) plot, and Cohen's Kappa. A ROC curve plots sensitivity values (true positive fraction) on the *y*-axis against '1 – specificity' values (false positive fraction) for all thresholds on the *x*-axis (Fielding & Bell, 1997). AUC is a threshold-independent metric and provides a single measure of model performance. AUC scores vary from 0 to 1. AUC values of less than 0.5 indicate discrimination worse than chance; a score of 0.5

Table 1 The details and settings of model implementation.

Model	Acronym	Data	Specifics and settings	Reference for more explanation
Generalized linear models	GLM	PA	Uses parametric functions to link the response variable to a linear, quadratic and/or cubic combination of explanatory variables. Here, a GLM ordinary polynomial with an automatic stepwise model selection based on the Akaike information criterion (AIC) was used. For simplicity we refer to this as the 'GLM' approach	McCullagh & Nelder (1989), Austin (2002)
Generalized additive models	GAM	PA	Uses nonparametric and data-defined, smoother to fit nonlinear functions. Here, a GAM with a cubic spline smoother and an automated stepwise process was used	Hastie & Tibshirani (1990), Austin (2002)
Boosted regression trees	BRT	PA	Fits complex nonlinear relationships by combining two algorithms of regression trees (relates a response to their predictors by recursive binary splits) and boosting (an additive method to combine many single models to improve the performance). The recommended default settings (maximum number of trees = 3000, learning rate = 0.001) were used	Friedman (2001), Elith *et al.* (2008)
Multivariate adaptive regression spline	MARS	PA	Similar to GAM but uses a piece-wise linear basis function	Friedman (1991), Leathwick *et al.* (2005)
Random forests	RF	PA	Selects many bootstrap samples from the data and generates and fits a large number of regression trees to each of these subsamples. Each tree is used to predict the out-of-bag observations (i.e. those that were not selected as bootstrap samples). The classification given by considering each tree as a 'vote', and the predicted class of an observation is determined by the majority vote among all trees. Models presented here had 500 trees with one variable randomly selected from the two candidates at each split	Breiman (2001), Cutler *et al.* (2007)
Genetic algorithm for rule-set production	GARP	PO	GARP uses a genetic algorithm with an iterative process to produce a set of conditional rules in the form of 'if-then' statements that describe the ecological niches of the species under study. The openModeller-GARP followed by the 'best subset' procedure was used in this study. The 'best subset' procedure was originally developed to sift through the model-to-model variation generated by the random-walk nature of the GARP algorithm	Stockwell & Noble (1992), Anderson *et al.* (2003), Peterson *et al.* (2007)
Maximum entropy	Maxent	PO	Uses a maximum entropy density estimation algorithm to approximate the true distribution of species as a probability distribution which respects a set of constraints where the mean of each environmental variable is required to be close to the empirical average over the presence sites. The recommended default settings were used to run the models except: the regularization multiplier was set to 2.5 as this produced the predicted response shapes visually closest to the ones used to simulate datasets	Phillips *et al.* (2006), Phillips & Dudík (2008)

PA, presence–absence; PO, presence-only.

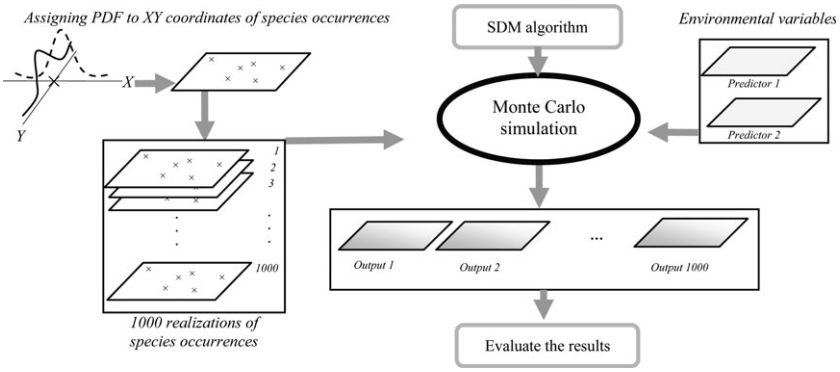

Figure 3 Conceptual framework of species positional error propagation analysis. PDF, probability density function; SDM, species distribution model.

implies random predictive discrimination; and a score of 1 indicates perfect discrimination.

Kappa is a proportional agreement between predictions and observations after removing the agreement expected to occur by chance (Cohen, 1960). Kappa ranges from −1 to +1, where +1 indicates perfect agreement, a value of 0 implies agreement expected by chance, and a value of less than 0 indicates agreement worse than chance. This statistic is calculated from a confusion matrix which is a cross-tabulation of observed and predicted values. The calculation, therefore, is dependent on a threshold to reclassify predicted probabilities into binary values (presence–absence). In this study, a constant threshold of 0.5 was applied to all cases. This threshold is the same as that adopted when simulating the virtual species (see section Simulating artificial datasets). The evaluation statistics were calculated using the PresenceAbsence package v. 1.1.3 (Freeman & Moisen, 2008) implemented in R (R Development Core Team, 2008).

Positional uncertainty propagation

Positional uncertainty in species occurrence leads to a shift in the point's position in the *x*- and *y*-directions (Heuvelink *et al.*, 2007). A probabilistic approach was used to introduce a positional error (ε) in species occurrences that had no directional bias. Taking $\varepsilon \sim N(0,1)$ gives a normally distributed unbiased error with a standard deviation equal to one map unit (here one grid cell). This was added to the easting and northing of each location (Hamm *et al.*, 2004):

$$\begin{aligned} xi &= \text{Easting} + \varepsilon xi \\ yi &= \text{Northing} + \varepsilon yi \end{aligned} \quad (5)$$

where *i* refers to each individual species occurrence. Different realizations of the sample were simulated and used to explore the effect of positional uncertainty (Hamm *et al.*, 2004). These were termed the 'perturbed' datasets. Five scenarios with increasing standard deviations, from one to five grid cells, were applied to explore the performance of the models over a range of positional uncertainties.

For each level of positional error, 1000 realizations of perturbed occurrence points were simulated. These realizations were used to train the models. The idea was to compute the result of the model repeatedly using varied input values (Heuvelink, 1999) and then to assess the accuracy of each. This so-called Monte Carlo simulation allowed the assessment of uncertainty. The conceptual framework for running the Monte Carlo simulation is illustrated in Fig. 3.

Interaction of positional uncertainty and spatial autocorrelation

To assess the interaction of positional uncertainty and spatial autocorrelation, all combinations of the scenarios were considered for each SDM algorithm. Therefore, in total 155 scenarios (5 levels of positional uncertainties × 31 spatial autocorrelation ranges) were applied to seven SDM algorithms. The performance of 1000 model runs was then calculated and compared with the performance of models with unperturbed data at different ranges of spatial autocorrelation at each level of positional error. To assess whether spatial autocorrelation range in predictors reduces the impact of positional uncertainty, a two-way Friedman's test (Friedman, 1937) with spatial autocorrelation range and positional uncertainty scenarios as factors was applied for each model.

RESULTS

The Moran's *I* correlograms for the GLM residuals, showed that the residuals were not spatially autocorrelated (see Fig. S1 in Appendix S1 in the Supporting Information).

Three simulated environmental gradients (as an example of the generated dataset) with different spatial autocorrelation ranges (5, 15 and 25) together with their corresponding variogram are illustrated in Fig. 4. They exemplify continuous environmental predictors (e.g. temperature, digital elevation model). The distribution of Kappa and AUC values as a result of the Monte Carlo simulation, and the response of this distribution to changes in spatial autocorrelation range of the environmental variables (Figs 5, 6 & S2 in Appendix S1) were generally consistent in the trend they showed. A similar interaction effect of positional uncertainty and spatial autocorrelation range in predictors on model performance was detected for both AUC and Kappa. Hence only the figures for Kappa are presented here, although the AUC results are provided in Table S1 and Figs S2–S4 in Appendix S1.

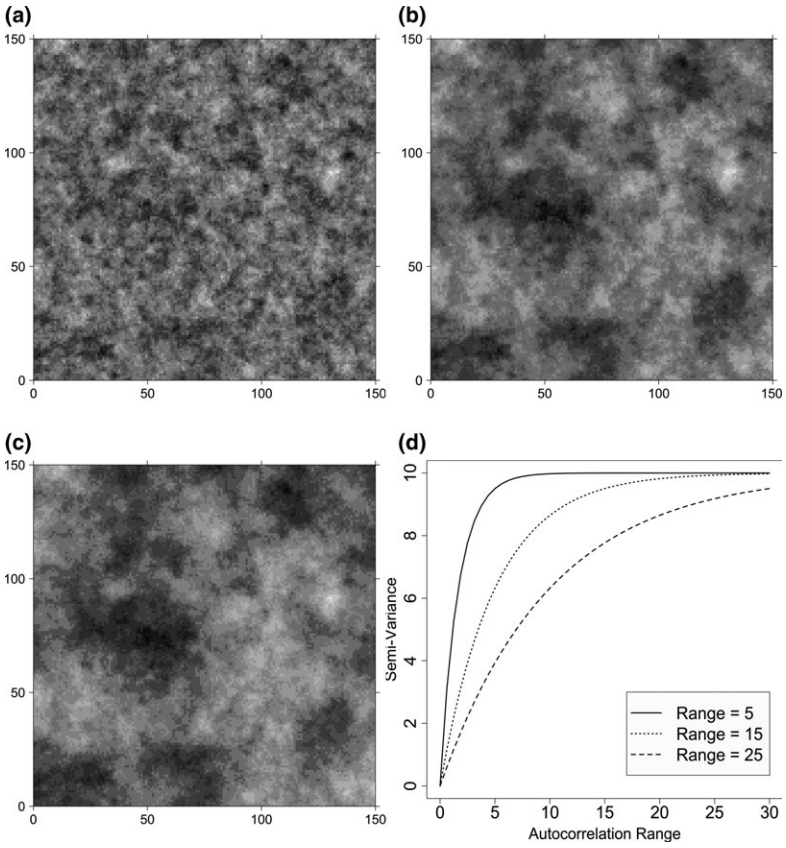

Figure 4 Three examples of simulated environmental gradients with spatial autocorrelation range of (a) 5, (b) 15, (c) 25 and (d) corresponding variogram models.

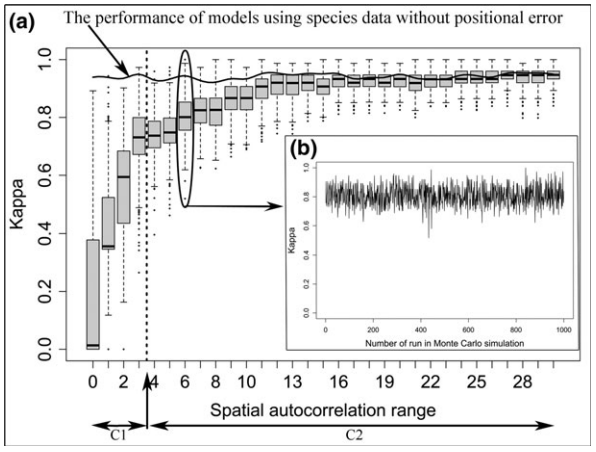

Figure 5 Variation of model accuracy (Kappa) over the Monte Carlo simulation for different ranges of spatial autocorrelation (*x*-axis) under the positional uncertainty. (a) For the generalized linear model (GLM) under the positional uncertainty with error $\varepsilon \sim N(0,1)$, the vertical dashed line represents the limit of spatial autocorrelation range that divided the results into two categories to compare the scenarios (Table 1). (b) Each box represents the results for 1000 Monte Carlo runs.

The mean Kappa and AUC for the presence–absence models fitted with the original data were 0.92 and 0.99, respectively, and for the presence-only models were 0.74 and 0.96, respectively. The effect of positional error on model performance was strongly influenced by the spatial autocorrelation (variogram) range. The level of accuracy depends on the spatial autocorrelation range and the level of error. Comparison of the graphs suggests that the models in general behave consistently over the range of autocorrelation and level of error.

As a normal probability density function (PDF) was used to introduce error in the locations, 99.7% of the perturbed points are expected to be within a distance equal to three times the standard deviation away from the original point. It was therefore expected that the performance of the SDMs would become stable when the range of the spatial autocorrelation of the predictors is larger than three times the standard deviation of the spatial error. This is because the value of the covariate at the perturbed location would then still be similar to that at the true location. This was used to group the results presented in each graph into two categories and to compare the performance of the models between these two categories, and over the different positional uncertainty scenarios. The first category included the simulations using predictors with an effective autocorrelation range equal to or lower than three times the standard deviation of the imposed positional error, and the second category included the simulations using predictors with a larger spatial autocorrelation range. Performance of the models was expected to be lower than model performance with the unperturbed location dataset when in the first category, and equal to performance of the unperturbed location dataset when in the second category. The summary statistics for the performance measures of the SDMs were then

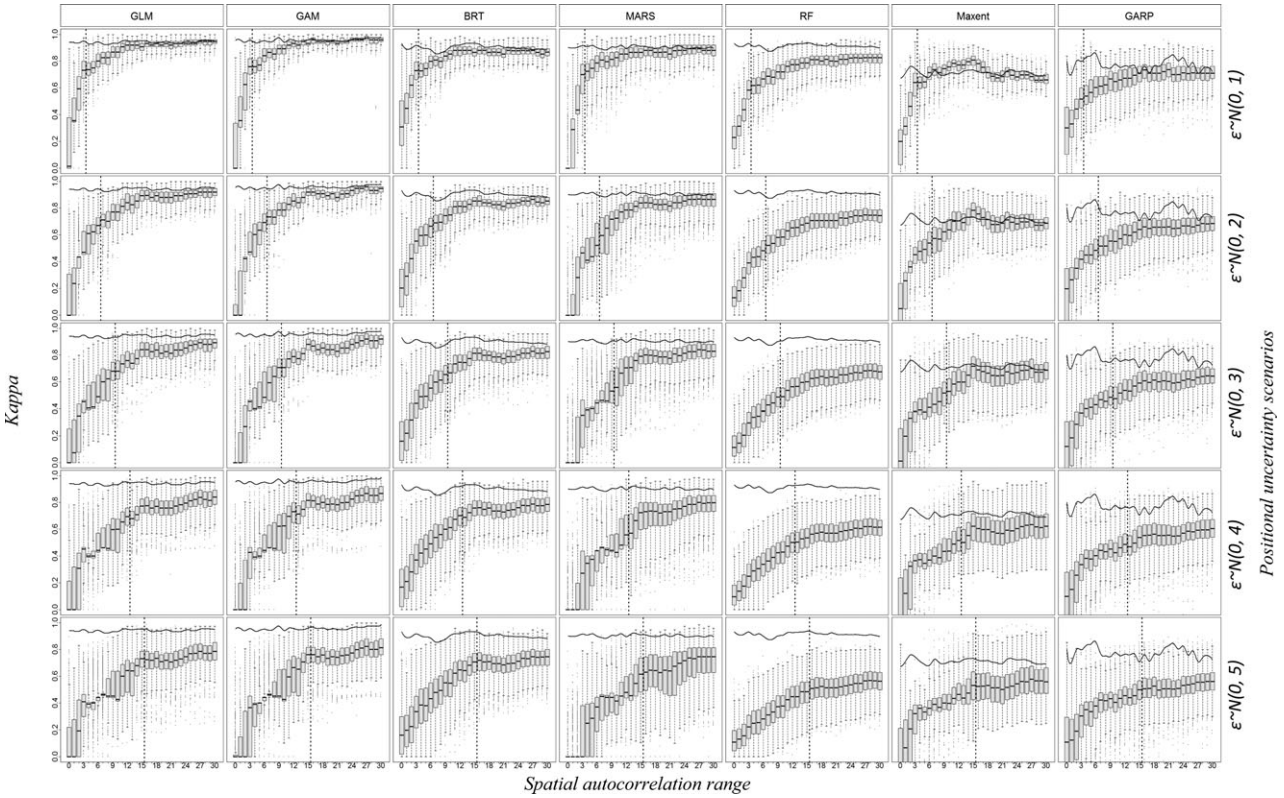

Figure 6 Variation of the model accuracy (Kappa) for five positional uncertainty scenarios with increasing standard deviation of error, from one to five grid cells, and seven species distribution models. Each row corresponds to the level of positional error. The different components of the graph are described in Fig. 5.

calculated for each category (Table 2). The Kappa and AUC values from the first category showed that the models were highly influenced by the positional error. The decline in the Kappa values, in the first category, was in a range between a minimum of 41% and a maximum of 65% in comparison with the Kappa for the model using the unperturbed data. The drop in AUC values was in a range between a minimum of 18% and a maximum of 32% in comparison with the AUC for the model using unperturbed data. Comparing the Kappa and AUC values for the second group showed that the drop was smaller than in the first group, ranging between a minimum of 0.1% and a maximum of 42% for the Kappa, and a minimum of 0% and a maximum of 17% for the AUC, depending on the level of positional error and the model used.

Table 2 Mean and standard deviations (mean | SD) of Kappa for the model outputs resulting from the Monte Carlo simulation for different positional error scenarios (PE1–PE5) and the Kappa for the model using unperturbed data. C1 and C2 specify the categories of Monte Carlo simulations where the spatial autocorrelation range was less than three times the standard deviation in positional error (C1) and those of which the spatial autocorrelation range was more than three times the standard deviation in positional error (C2). See Table 1 for details of each model.

		The positional error scenarios									
	Unperturbed	PE1		PE2		PE3		PE4		PE5	
Model	data	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
GLM	0.941	0.48 | 0.27	0.90 | 0.07	0.44 | 0.24	0.86 | 0.09	0.44 | 0.23	0.82 | 0.09	0.44 | 0.23	0.78 | 0.10	0.45 | 0.23	0.74 | 0.11
GAM	0.956	0.47 | 0.29	0.92 | 0.07	0.43 | 0.27	0.88 | 0.08	0.42 | 0.26	0.84 | 0.09	0.43 | 0.26	0.81 | 0.09	0.44 | 0.26	0.76 | 0.11
BRT	0.901	0.53 | 0.22	0.85 | 0.06	0.47 | 0.21	0.81 | 0.08	0.45 | 0.21	0.78 | 0.08	0.46 | 0.22	0.75 | 0.08	0.47 | 0.22	0.71 | 0.09
MARS	0.903	0.36 | 0.29	0.84 | 0.08	0.32 | 0.25	0.80 | 0.12	0.32 | 0.24	0.76 | 0.13	0.33 | 0.24	0.72 | 0.14	0.33 | 0.24	0.66 | 0.16
RF	0.913	0.39 | 0.17	0.77 | 0.09	0.34 | 0.16	0.68 | 0.10	0.32 | 0.16	0.62 | 0.10	0.32 | 0.16	0.58 | 0.10	0.32 | 0.16	0.53 | 0.10
MAXENT	0.711	0.40 | 0.24	0.71 | 0.07	0.36 | 0.24	0.68 | 0.10	0.35 | 0.24	0.65 | 0.11	0.34 | 0.23	0.59 | 0.13	0.34 | 0.23	0.54 | 0.13
GARP	0.774	0.39 | 0.2	0.68 | 0.10	0.36 | 0.21	0.63 | 0.11	0.35 | 0.21	0.59 | 0.11	0.35 | 0.21	0.56 | 0.11	0.34 | 0.21	0.53 | 0.11

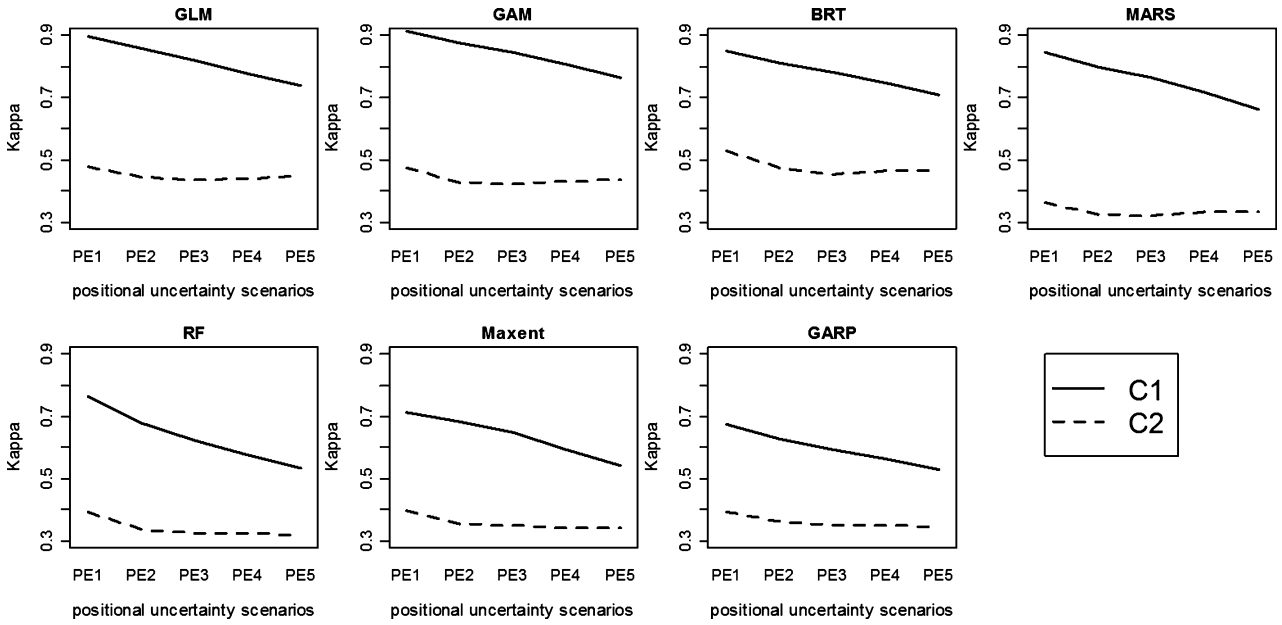

Figure 7 Interaction plots based on the Friedman test – difference of Kappa mean between two categories of spatial autocorrelation range (C1 and C2) through different levels of positional error (PE1–PE5). C1 and C2 specify the categories of Monte Carlo simulations where the spatial autocorrelation range was less than three times the standard deviation in positional error (C1) and those of which the spatial autocorrelation range was more than three times the standard deviation in positional (C2). GLM, generalized linear model; GAM, generalized additive model; BRT, boosted regression trees; MARS, multivariate adaptive regression spline; RF, random forest; GARP, genetic algorithm for rule-set production; Maxent, maximum entropy.

The decline in the performance with an increase of positional error was not equal for all models. For the GLM, GAM and BRT algorithms, the Kappa values for the second category dropped by a similar amount of 21% for the scenarios where the highest level of positional error was introduced (an increase of almost 4% for each level of error). In the same situation, the AUC dropped 5% for both the GLM and GAM, and 8% for the BRT. For the MARS and RF, the decline in Kappa was 27% and 42%, respectively, for the scenarios where the highest level of positional error was introduced (almost 5% for the MARS and 8% for the RF for each level of error), and this decline in the AUC was 11% and 17%, respectively. For presence-only algorithms, Maxent and GARP, the decline in the Kappa was 24% and 31%, respectively, and in the AUC was 13% and 11%, respectively. Of the models employed for this study, RF was the most sensitive algorithm to species positional error. The results, furthermore, indicated that, in many cases, RF, GARP and Maxent performed less well than GAM, GLM and BRT models.

The results from the Friedman test with two factors (two categories of spatial autocorrelation range in predictors as the first factor and five scenarios of positional uncertainty as the second factor) indicated that the difference between the spatial autocorrelation range groups for the measures of model performance (Kappa and AUC) for all modelling techniques were significant ($P < 0.001$). The differences between positional uncertainty groups were also significant ($P < 0.001$). The interaction plots based on the Friedman test (Figs 7 & S3 in Appendix S1) show the difference between two categories of spatial autocorrelation range through different levels of positional error as well as the trend of decline in performance as described above.

In order to provide a framework to link the positional uncertainty to model robustness, the results of all scenarios for each SDM were interpolated into a contour plot (Figs 8 & S4 in Appendix S1). This graph shows the interaction of positional uncertainty and spatial autocorrelation range in environmental variables. By considering the level of positional uncertainty (*y*-axis) and the spatial autocorrelation range in the environmental variables (*x*-axis), an accuracy measure is provided which can be compared with the accuracy measure for the same range of spatial autocorrelation but with no positional error in species occurrences. The difference between these two values can be interpreted as an expected decline in performance, and shows the potential impact of positional uncertainty, given a level of spatial autocorrelation in the explanatory variables.

DISCUSSION

This study linked the robustness of the SDM to positional uncertainty in species occurrences and spatial autocorrelation range in predictors. The results show that the impact of positional uncertainty can be assessed by examining spatial autocorrelation in the predictors. Graham *et al.* (2008) and Osborne & Leitão (2009) found that positional error had a small effect on predictive performance of models as judged by AUC and, consequently, useful predictions can be made even

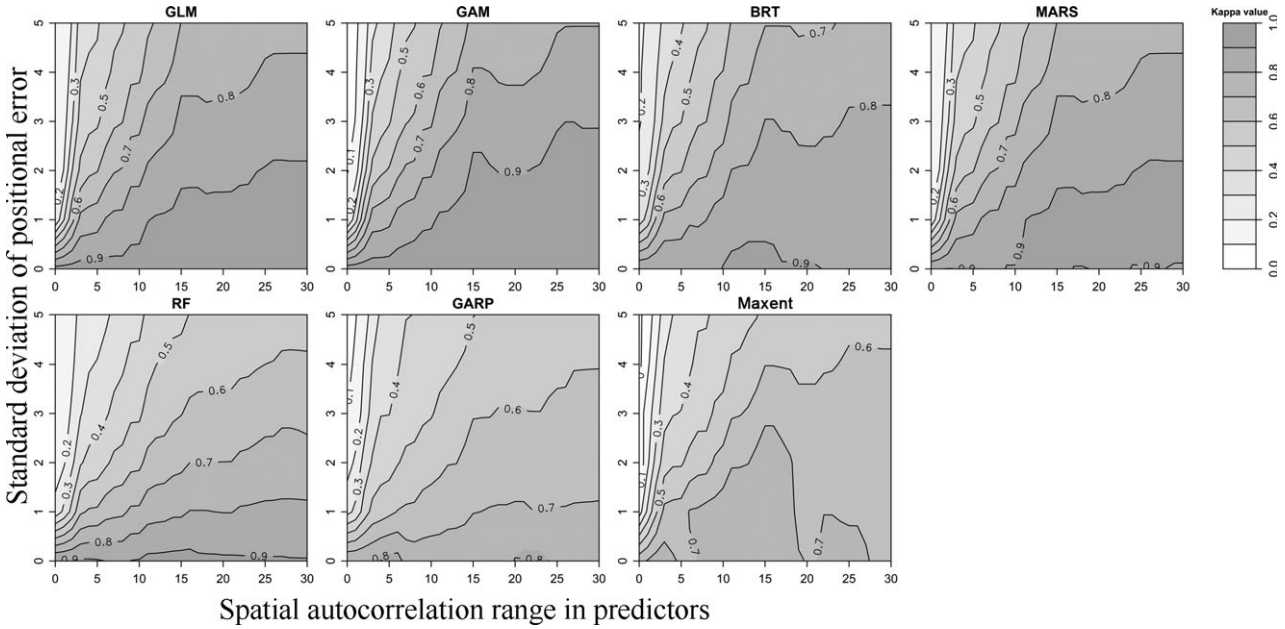

Figure 8 Interaction of species occurrence positional error and spatial autocorrelation range in predictors; the *x*-axis represents the range of spatial autocorrelation in predictors, and the *y*-axis represents the standard deviation of positional error in species occurrences. GLM, generalized linear model; GAM, generalized additive model; BRT, boosted regression tree; MARS, multivariate adaptive regression spline; RF, random forest; GARP, genetic algorithm for rule-set production; Maxent, maximum entropy.

when species data are subjected to some positional error. Based on the results of the current study this is not always true, and this depends on the range of spatial autocorrelation in the predictors relative to the level of positional error.

It was expected that larger spatial autocorrelation ranges in the predictors would diminish the impact of positional uncertainty on prediction accuracy. Comparing the Kappa and AUC values in the second category of spatial autocorrelation range over the positional uncertainty scenarios showed that, in all scenarios, positional error led to a drop in performance of the models, but most of them would still be regarded as acceptable and useful (Manel *et al.*, 2001). The drop, however, was greater for the extreme positional error. Although spatial autocorrelation range can reduce the effect of positional uncertainty, it cannot completely compensate for it.

For all scenarios the interaction between spatial autocorrelation range and positional uncertainty showed a consistent trend. The prediction accuracy of the models using predictors with a spatial autocorrelation range greater than the standard deviation of positional error was high because the perturbed points should generate similar correlations between environmental predictors and presence–absence compared with the case when they would not be perturbed. It was shown that the performance of the SDMs became stable when the range of the spatial autocorrelation of the predictors is greater than three times the standard deviation of the spatial error (second category), because more than 99% of the points remained relatively unaffected by the perturbation in their correlation with the environmental predictors. The graphs demonstrate this trend and the results of the Friedman test support this argument.

The results indicated there is a difference in the performance and behaviour of the models, especially between the two groups of presence–absence and presence-only algorithms. The models using the presence–absence data used more information than the presence-only models, and therefore this information allowed the models to be better calibrated (Elith & Graham, 2009). This may explain why presence–absence models performed better as judged by the Kappa and AUC. The variation in the performance measures decreased when the range of spatial autocorrelation in predictors increased or when the level of positional uncertainty decreased. This is not the same for all models. For instance, comparing the standard deviation of the Kappa measures in both categories for the RF (Table 2) showed that the variation is lower only in the first error scenario and, for the other scenarios, remained approximately the same. This variation in accuracy for the RF model, when judged by AUC (see Table S1 in Appendix S1), changed slightly from 0.04 to 0.06 for the second category through the scenarios. The behaviour of Maxent showed that it might be sensitive to spatial autocorrelation in predictors as the variation is not consistent through different levels of spatial autocorrelation. These results showed how the models differed under the simulated conditions, and encourage further study using other conditions, such as different patterns of spatial autocorrelation in predictors or cross-correlated autocorrelated predictors.

The two most commonly used methods were selected for evaluating the SDMs (i.e. AUC and Kappa) (Elith & Leathwick, 2009). AUC is a threshold-independent measure (Fielding & Bell, 1997) which has been used in most recent papers as an evaluation method, because the major drawback of the Kappa

statistic is related to the selection of a threshold, which is affected by prevalence and bias in data. Biased training, because of either low or high prevalence, affects the optimal value for the cut-off threshold (Manel *et al.*, 2001; Jiménez-Valverde *et al.*, 2009). The selection of the threshold, however, was not an issue in this study since a constant threshold of 0.5 was applied to generate the artificial datasets and the prevalence was controlled at 0.5 to avoid bias. For all the models, except Maxent, the Kappa and AUC statistics were consistent. Based on the AUC (Table S1 in Appendix S1), Maxent provided high accuracy (the AUC measures ranged between 0.86 and 0.97) in the second category of results, whereas it gave only moderate accuracy based on Kappa (the Kappa measures ranged between 0.54 and 0.71).

Different levels of unknown complexity existing in real data make completely accurate inference in species distribution modelling impossible. Furthermore, statistical methods employed in SDMs are also difficult to evaluate because the 'truth' is unknown (Austin *et al.*, 2006). As a result, simulated data have increasingly been used as a tool for developing and evaluating SDMs (Austin *et al.*, 2006; Meynard & Quinn, 2007; Elith & Graham, 2009; Jiménez-Valverde *et al.*, 2009; Santika & Hutchinson, 2009; Beale *et al.*, 2010). In this research, the simulated datasets for both species and predictors provided full control over the training data as well as providing reference data for accuracy assessment. Clearly they do not cover all possible complications that are likely to be found in real data. For example, the degree of spatial autocorrelation might vary locally over a real study area, whereas a homogeneous area was simulated in this study. In a heterogeneous area the global spatial autocorrelation measures may not properly model its effect on positional uncertainty. Under this circumstance, measuring local spatial autocorrelation in the environmental predictors might be a promising solution to find the locations with a high degree of influence. Also, the spatial autocorrelation range might be different for the different predictor variables. In this study, the range was the same for both the simulated predictors. Further study is required to analyse the effect of positional uncertainty when there is variation in spatial autocorrelation range between the contributing environmental variables. It is expected that the variables with higher contribution to the SDM will have a bigger effect on reducing the impact of positional uncertainty. Under such circumstances, it is advisable that the minimum spatial autocorrelation range amongst all predictors, or the predictors with the highest contributions to the model, be considered when using the contour plots (Fig. 8).

Spatial autocorrelation has been widely cited as a problem for regression-based techniques (for a review see Dormann *et al.*, 2007). In particular, this arises when care has not been taken to address the assumption that the residuals should be independent and identically distributed. In this study, a random sampling design was employed. Furthermore, analysis of Moran's *I* for the model residuals, showed that the residuals were not spatially correlated. It should also be noted that there is no conflict between the potential problems of autocorrelation and the arguments of this paper, which regard it as an opportunity.

Finally, the range of spatial autocorrelation in the environmental predictors can be linked to positional uncertainty, since both can be expressed in distance units. Of key importance here is the interplay between the level of positional error in the response, the spatial resolution of the predictor variable (e.g. pixel size in a raster grid) and the range of spatial autocorrelation of the predictor. For a given pixel size, an increase in the positional error will increase the probability that the response will fall within an incorrect grid cell, and hence be associated with incorrect predictor attribute values (Hamm *et al.*, 2003). The results presented in this paper show that the extent to which this reduces prediction accuracy depends primarily on the range of spatial autocorrelation of the predictors and, secondarily, on the SDM employed. Importantly, the autocorrelation was modelled at the same resolution as the predictor variable. It should be noted that autocorrelation range is likely to change with spatial resolution (Atkinson, 1993). Hence a user who wishes to apply the approach employed in this paper should take care that the autocorrelation is assessed at the resolution of their predictor variable and not at a finer resolution (Graham *et al.*, 2008).

CONCLUSIONS

This study has explored how different degrees of positional uncertainty in species occurrence data influence the prediction accuracy of SDMs for varying levels of the spatial autocorrelation range of predictors. Spatial autocorrelation in predictors reduced the impact of positional uncertainty on prediction accuracy of the SDMs when the range of spatial autocorrelation was greater than three times the standard deviation of positional error. In such circumstances, although positional error led to a decline in predictive performance they yielded the maximum achievable performance given the level of uncertainty in the presence.

It is argued that examining the spatial autocorrelation in predictors to find the effective autocorrelation range can give insight into whether predictions are likely to be affected by the uncertainty in the sample locations. This is especially important when the output of these models is to be used in conservation and planning as it provides a basis to use them with a clear level of confidence.

ACKNOWLEDGEMENTS

We would like to thank the editor and the three anonymous referees for valuable comments on an earlier draft of this manuscript. We thank Bert Toxopeus for comments on an earlier draft. We also thank Eva Skidmore for her English editing. This work was supported by the European Union, Erasmus Mundus programme (2007/1139/001-001 MUN ECW) PhD award to Babak Naimi.

REFERENCES

Anderson, R.P., Lew, D. & Peterson, A.T. (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, **162,** 211–232.

Araújo, M.B. & Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33,** 1677–1688.

Atkinson, P.M. (1993) The effect of spatial-resolution on the experimental variogram of airborne MSS imagery. *International Journal of Remote Sensing*, **14,** 1005–1011.

Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200,** 1–19.

Austin, M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157,** 101–118.

Austin, M.P., Belbin, L., Meyers, J.A., Doherty, M.D. & Luoto, M. (2006) Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. *Ecological Modelling*, **199,** 197–216.

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J. & Elston, D.A. (2010) Regression analysis of spatial data. *Ecology Letters*, **13,** 246–264.

ter Braak, C.J.F. & Looman, C.W.N. (1986) Weighted averaging, logistic-regression and the Gaussian response model. *Vegetatio*, **65,** 3–11.

Breiman, L. (2001) Random forests. *Machine Learning*, **45,** 5–32.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20,** 37.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A. & Hess, K.T. (2007) Random forests for classification in ecology. *Ecology*, **88,** 2783–2792.

Dietrich, C.R. & Newsam, G.N. (1993) A fast and exact method for multidimensional Gaussian stochastic simulations. *Water Resources Research*, **29,** 2861–2869.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M. & Wilson, R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30,** 609–628.

Dungan, J. (1999) Conditional simulation: an alternative to estimation for achieving mapping objectives. *Spatial statistics for remote sensing* (ed. by A. Stein, F. Meer and B. Gorte), pp. 135–152. Springer, Dordrecht.

Elith, J. & Graham, C.H. (2009) Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, **32,** 66–77.

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

Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29,** 129–151.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77,** 802–813.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24,** 38–49.

Franklin, J. (2010) *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge, UK.

Freeman, E.A. & Moisen, G.G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, **217,** 48–58.

Friedman, J.H. (1991) Multivariate adaptive regression splines. *Annals of Statistics*, **19,** 1–67.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29,** 1189–1232.

Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, **32,** 675–701.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19,** 497–503.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Peterson, A.T. & Loiselle, B.A. (2008) The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, **45,** 239–247.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135,** 147–186.

Hamm, N., Atkinson, P.M. & Milton, E.J. (2003) The combined effect of spatial resolution and measurement uncertainty on the accuracy of empirical atmospheric correction. *Proceedings of the 2003 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2003), Learning from Earth's Shapes and Sizes*, pp. 2082–2084. IEEE, New York.

Hamm, N., Atkinson, P.M. & Milton, E.J. (2004) On the effect of positional uncertainty in field measurements on the atmospheric correction of remotely sensed imagery. *geoENV IV – geostatistics for environmental applications* (ed. by X. Sánchez-Vila, J. Carrera and J.J. Gómez-Hernández), pp. 91–102. Springer, Dordrecht.

Hastie, T. & Tibshirani, R. (1990) *Generalized additive models*. Chapman and Hall, London.

Heuvelink, G.B.M. (1999) Propagation of error in spatial modelling with GIS. *Geographical information systems* (ed. by P. Longley, M. Goodchild, D. Maguire and D. Rhind), pp. 207–217. John Wiley and Sons, Chichester.

Heuvelink, G.B.M., Brown, J.D. & van Loon, E.E. (2007) A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science*, **21,** 497–513.

Hirzel, A.H., Helfer, V. & Metral, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145,** 111–121.

Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. (2009) The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology*, **10,** 196–205.

Johnson, C.J. & Gillingham, M.P. (2008) Sensitivity of species-distribution models to error, bias, and model design: an application to resource selection functions for woodland caribou. *Ecological Modelling*, **213,** 143–155.

Leathwick, J.R., Rowe, D., Richardson, J., Elith, J. & Hastie, T. (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, **50,** 2034–2052.

Legendre, P. (1993) Spatial autocorrelation – trouble or new paradigm. *Ecology*, **74,** 1659–1673.

Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38,** 921–931.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*, 2nd edn. Chapman and Hall, London.

Meynard, C.N. & Quinn, J.F. (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, **34,** 1455–1469.

Osborne, P.E. & Leitão, P.J. (2009) Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions*, **15,** 671–681.

Peterson, A.T. (2006) Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics*, **3,** 59–72.

Peterson, A.T., Papes, M. & Eaton, M. (2007) Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography*, **30,** 550–560.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31,** 161–175.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190,** 231–259.

R Development Core Team (2008) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Rowe, R.J. (2005) Elevational gradient analyses and the use of historical museum specimens: a cautionary tale. *Journal of Biogeography*, **32,** 1883–1897.

Santika, T. & Hutchinson, M.F. (2009) The effect of species response form on species distribution model prediction and inference. *Ecological Modelling*, **220,** 2365–2379.

Schabenberger, O. & Gotway, C.A. (2005) *Statistical methods for spatial data analysis*. Chapman & Hall/CRC Press, Boca Raton, FL.

Schlather, M. (2009) *RandomFields: simulation and analysis of random fields. R package version 1.3.41*. Available at: http://cran.r-project.org/web/packages/RandomFields/index.html.

Segurado, P. & Araújo, M.B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31,** 1555–1568.

Skidmore, A.K., Gauld, A. & Walker, P. (1996) Classification of kangaroo habitat distribution using three GIS models. *International Journal of Geographical Information Systems*, **10,** 441–454.

de Souza Muñoz, M., De Giovanni, R., de Siqueira, M., Sutton, T., Brewer, P., Pereira, R., Canhos, D. & Canhos, V. (2009) openModeller: a generic approach to species' potential distribution modelling. *GeoInformatica*, **15,** 111–135.

Stockwell, D.R.B. & Noble, I.R. (1992) Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. *Mathematics and Computers in Simulation*, **33,** 385–390.

Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, **32,** 369–373.

Webster, R. & Oliver, M.A. (2007) *Geostatistics for environmental scientists*, 2nd edn. John Wiley and Sons, Chichester, UK.

SUPPORTING INFORMATION

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

Appendix S1 Additional results.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCH

Babak Naimi is a PhD candidate in the Department of Natural Resources, Faculty of Geo-Information Science and Earth Observation, University of Twente, The Netherlands. He is also a research assistant at the University of IAU, Iran. His research focuses on the effects of uncertainty on species distribution modelling. He is interested in the application of spatial statistics and machine learning techniques in biodiversity studies.

Author contributions: B.N., A.K.S., T.A.G. and N.A.S.H. conceived the ideas and reviewed the results. B.N. implemented the study, analysed the results and drafted the manuscript. A.K.S., T.A.G. and N.A.S.H. supervised the research and reviewed the manuscript.

Editor: Miguel Araújo