Received: 17 April 2018 | Revised: 24 July 2018 | Accepted: 28 September 2018

DOI: 10.1111/ddi.12868

BIODIVERSITY RESEARCH

WILEY Diversity and Distributions

Effects of simulated observation errors on the performance of species distribution models

Rui F. Fernandes[1] | Daniel Scherrer[1] | Antoine Guisan[1,2]

[1]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[2]Institute of Earth Surface Dynamics, Geopolis, University of Lausanne, Lausanne, Switzerland

Correspondence
Rui F. Fernandes, University of Lausanne, Lausanne, Switzerland.
Email: rui.fff24@gmail.com

Editor: Boris Schröder

Abstract

Aim: Species distribution information is essential under increasing global changes, and models can be used to acquire such information but they can be affected by different errors/bias. Here, we evaluated the degree to which errors in species data (false presences–absences) affect model predictions and how this is reflected in commonly used evaluation metrics.

Location: Western Swiss Alps.

Methods: Using 100 virtual species and different sampling methods, we created observation datasets of different sizes (100–400–1,600) and added increasing levels of errors (creating false positives or negatives; from 0% to 50%). These degraded datasets were used to fit models using generalized linear model, random forest and boosted regression trees. Model fit (ability to reproduce calibration data) and predictive success (ability to predict the true distribution) were measured on probabilistic/binary outcomes using Kappa, TSS, MaxKappa, MaxTSS and Somers'D (rescaled AUC).

Results: The interpretation of models' performance depended on the data and metrics used to evaluate them, with conclusions differing whether model fit, or predictive success were measured. Added errors reduced model performance, with effects expectedly decreasing as sample size increased. Model performance was more affected by false positives than by false negatives. Models with different techniques were differently affected by errors: models with high fit presenting lower predictive success (RFs), and vice versa (GLMs). High evaluation metrics could still be obtained with 30% error added, indicating that some metrics (Somers'D) might not be sensitive enough to detect data degradation.

Main conclusions: Our findings highlight the need to reconsider the interpretation scale of some commonly used evaluation metrics: Kappa seems more realistic than Somers'D/AUC or TSS. High fits were obtained with high levels of error added, showing that RF overfits the data. When collecting occurrence databases, it is advisory to reduce the rate of false positives (or increase sample sizes) rather than false negatives.

KEYWORDS
artificial data, AUC, ecological niche models, evaluation metric, habitat suitability models, Kappa, model fit, predictive accuracy, TSS, uncertainty

1 | INTRODUCTION

As biodiversity and ecosystems are under growing pressure by global changes, we need to urgently increase our understanding of, and associated capacity to model, the main factors driving changes in the distributions of species, assemblages and ecosystems (Dawson, Jackson, House, Prentice, & Mace, 2011). Species distribution models (SDMs; Guisan, Thuiller, & Zimmermann, 2017) allow modelling the distribution of species and their assemblages at different spatial and temporal scales (D'Amen, Rahbek, Zimmermann, & Guisan, 2017; Ferrier & Guisan, 2006). SDMs statistically correlate species observations (presence–absence or presence-only) with environmental data (Guisan & Thuiller, 2005) and are commonly evaluated by assessing their predictive performance and accuracy (Peterson et al., 2011). The most used metric is, by far, the area under the receiver operating characteristic curve (AUC-ROC; Fourcade, Besnard, & Secondi, 2018). It is calculated by plotting a model's sensitivity against its false-positive rate at all possible thresholds (Hanley & McNeil, 1982), measuring the model's performance in discriminating between species presences and absences (Lobo, Jimenez-Valverde, & Real, 2008). Alternative metrics have also been proposed, mainly due to the known limitations of the AUC (e.g., dependence on the calibration area, ignores spatial distribution of errors, relies on the ranking of sensitivity/specificity across thresholds and ignores the probability values given by a model or equally weights omission/commission errors; Lobo et al., 2008; Peterson, Papes, & Soberon, 2008; Jiménez-Valverde, 2012; Jiménez-Valverde, Acevedo, Barbosa, Lobo, & Real, 2013). The most common alternatives are Cohen's Kappa (Kappa; Cohen, 1960) and the true skill statistic (TSS; Allouche, Tsoar, & Kadmon, 2006). Kappa corrects the overall accuracy of model predictions by the accuracy expected to occur by chance while TSS corrects Kappa's dependency on prevalence (see Table 1 for more information). Moreover, SDMs can contain uncertainty from various sources (reviewed by e.g., Barry & Elith, 2006; Beale & Lennon, 2012), including errors associated with species data (e.g., unavailable absence data, small or insufficient sample sizes, unexplored geographical bias or spatial errors; e.g., Fielding & Bell, 1997; Pearce & Ferrier, 2000; Jenkins, Powell, Bass, & Pimm, 2003), environmental variables (e.g., missing important ones; Mod, Scherrer, Luoto, & Guisan, 2016) or modelling techniques (e.g., Guisan, Zimmermann, et al., 2007; Thibaud, Petitpierre, Broennimann, Davison, & Guisan, 2014). One problem commonly affecting SDMs concerns the inability to separate potentially false and true species' absences obtained through field surveys (Lahoz-Monfort, Guillera-Arroita, & Wintle, 2014) leading to underestimation of species occupancy (i.e., when occupied sites are misclassified as unoccupied; Guillera-Arroita, Ridout, & Morgan, 2010), incorrect inference about species distributions or inaccurate predictions (Lahoz-Monfort et al., 2014). The wrongly recorded absences (false absences) in presence–absence datasets or the omission of presences in presence-only models can then lead to predictions that will reflect where the species is more or less likely to be detected instead of the locations where if should occur or not (Kéry, 2011; Lahoz-Monfort et al., 2014). This means that one would eventually model what is called the "apparent distribution" and not the true distribution (Kéry, 2011). Additionally, some environmental relationships that are important to explain species occurrence and distribution might be wrongly identified or completely missed when false absences/presences are recorded (Kéry, 2011). The effect of detection errors on model performance is likely to depend on the modelling techniques used as those differ in their ability to fit complex response curves (i.e., species–environment relationships; Guisan, Zimmermann, et al., 2007; Merow et al., 2014).

TABLE 1 Detailed information about the evaluation metrics used to assess the predictive performance of SDMs (adapted from Liu et al., 2005 and Allouche et al., 2006), *a* is true positives (or presences), *b* is false positives (or presences), *c* is false negatives (or absences), *d* is true negatives (or absences), *n* (=*a* + *b* + *c* + d) is the total number of sites. Sensitivity is the probability that the model will correctly classify a presence (*a*/*a* + *c*). Specificity is the probability that the model will correctly classify an absence (*d*/*b* + *d*)

Metric	Acronym	Definition/Formula	Scale	References
Area under the receiver operating curve	AUC	Calculated by plotting a model's sensitivity against its false-positive rate at all possible thresholds	0/+1	Hanley and McNeil (1982)
Somers' rank correlation	Somers'D	2(AUC − 0.5)	−1/+1	Harrell (2015)
Cohen's Kappa	Kappa	$\frac{\left(\frac{a+d}{n}\right)-\frac{(a+b)(a+c)+(c+d)(d+b)}{n2}}{1-\frac{(a+b)(a+c)+(c+d)(d+b)}{n2}}$	−1/+1	Cohen (1960)
True skill statistic	TSS	Sensitivity +Specificity − 1	−1/+1	Allouche et al. (2006)
Kappa maximization	MaxKappa	Kappa statistic is maximized	−1/+1	Guisan et al. (1998); Huntley et al. (1998)
TSS maximization	MaxTSS	TSS statistic is maximized	−1/+1	Liu et al. (2005)

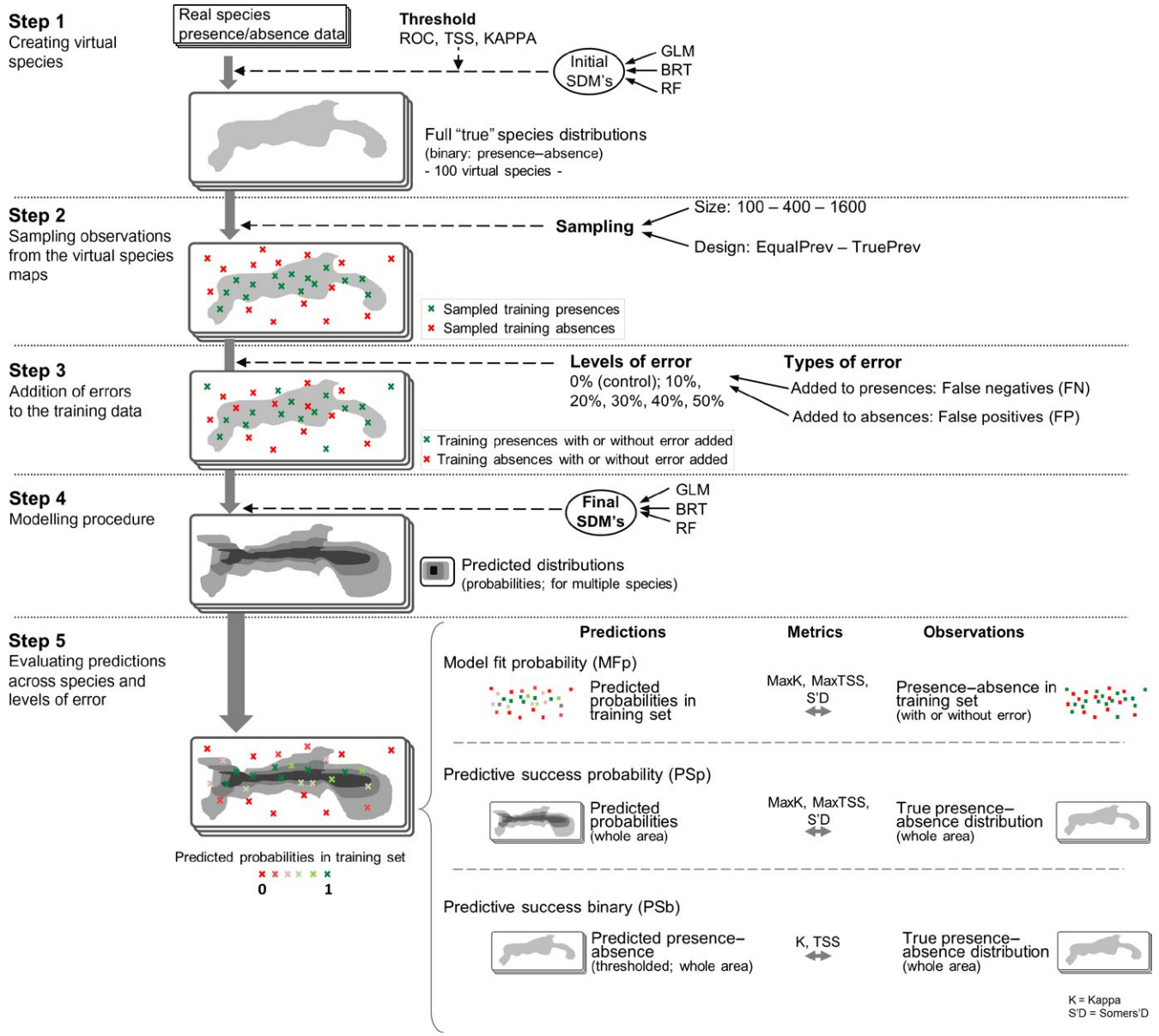

FIGURE 1 Workflow of the analytical steps followed in the study. Step 1—We started by creating binary distribution maps for 100 virtual species from models based on real species' data (using either generalized linear models (GLM), boosted regression trees (BRT) or random forests (RF) as modelling techniques and the receiver operating characteristic (ROC), true skill statistic (TSS) or KAPPA as thresholding techniques). Step 2—For each species, we sampled presence–absence data using three different sample sizes (100–400–1,600) and two sampling designs (EqualPrev and TruePrev). Step 3—To each of the sampled datasets, errors were added according to six different levels (0%—training data without error added, the control; 10%, 20%, 30%, 40% and 50%—training data with error added) and two different types of error (errors added to presences, creating false negatives or added to absences, creating false positives). Step 4—Each occurrence dataset was used to create single species distribution models (probability and binary maps), using three different modelling techniques (GLM, BRT and RF). Step 5—The predictions for each species were then evaluated with three evaluation approaches: *model fit probability (MFp)*, *predictive success probability (PSp)* and *predictive success binary (PSb)*, using different metrics: maximized Kappa (MaxKappa), maximized TSS (MaxTSS) and Somers'D (rescaled measure of AUC) for MFp and PSp; Kappa and TSS for PSb

Several of these issues have received considerable attention in recent years, providing information to improve survey designs, proposing approaches to account for imperfect detection and evaluating the impacts of non-detection of species in models of individual species (Gu & Swihart, 2004; Guillera-Arroita et al., 2010; MacKenzie et al., 2002). However, the majority of the studies focusing on uncertainties in SDMs used real species observations, putting a limit to proper assessment of model accuracy because the complete distribution and the species–environment relationships cannot be entirely known and may result from factors that cannot be controlled. A way to avoid these limitations is to use artificial data (Austin, Belbin, Meyers, Doherty, & Luoto, 2006) in a virtual

ecologist approach (see Zurell et al., 2010 for a review), where all the information necessary for a study can always be obtained in a fully artificial or semi-artificial world, allowing complete or at least partial control on the data and models being tested (Austin et al., 2006). In one of the first application to SDMs, Hirzel, Helfer, and Metral (2001) created virtual species to test different habitat suitability methods and their predictive power under different scenarios. Virtual species have been used to test different ecological models and assumptions, to test different approaches to sample species data (Hirzel & Guisan, 2002), to downscale coarse-grain data into high-resolution predictions (Bombi & D'Amen, 2012) or to measure the relative effect of different factors affecting predictions (Thibaud et al., 2014).

In this study, we take a virtual ecologist approach, using 100 virtual species defined from real observations in a real mountain landscape with large environmental gradients, to investigate: (a) the effect of sample size when error is added to the data; (b) the model performance behaviour when different levels of errors are added to the training data (to presences or absences) and how different evaluation approaches influence the conclusions of that performance, (c) how different metrics traditionally used to evaluate SDM predictions perform with those errors (d) what are the implications for interpreting the performance/reliability of models when using those metrics, (e) how different modelling techniques deal with degraded training data and (f) how different types of errors affect models and metrics. Taking into account the frequent use of SDMs in ecology, evolution and conservation, this paper provides an essential analysis of the potential effects of errors in species data on SDM reliability and on the interpretation of common evaluation metrics.

2 | METHODS

2.1 | Analytical framework

We implemented a virtual ecologist approach (see Figure 1), based initially on real data in a real landscape (i.e., which can also be considered as a semivirtual study; Albert et al., 2010) in the western Swiss Alps (a priority research area; https://rechalpvd.unil.ch), covering approximately 700 km^2. We defined the distributions of virtual species based on predictions of models fitted on real data in this study area to keep ecological realism (see Step 1 below). The approach consisted of five steps:

2.1.1 | Step 1. Creating virtual species

From a set of real species data (previously sampled in the study area), we generated 100 virtual species, by fitting SDMs (initial SDMs in Figure 1) using presence–absence data against five environmental predictors: summer mean monthly temperatures (2–19°C), sum of winter precipitation (65–282 mm), annual sum of potential solar radiation (KJ), slope (°) and topographic position (unitless; indicating ridges and valleys; see Supporting Information Appendix S1).

The models were fitted using generalized linear models (GLMs; McCullagh & Nelder, 1989), random forests (RFs; Breiman, 2001) or boosted regression trees (BRTs; Friedman, Hastie, & Tibshirani, 2000) as modelling techniques. These modelling techniques were chosen because GLMs allow hump-shaped and linear response curves that can be easily justified by ecological niche theory while RFs and BRTs have been increasingly used in recent years as they allow for more complex combinations and interactions of environmental factors, which can result in more complex species–environment relationships. This study set-up allowed us to check if the complexity of those relationships could influence the outcome of our study.

The resulting probability distributions were transformed into presence–absence data (considered as our "true" virtual species distribution) using three thresholding approaches: (a) threshold that corresponded to the point on the receiver operating characteristic plot (ROC; sensitivity against 1 specificity across successive thresholds; Hanley & McNeil, 1982; Swets, 1988) with the shortest distance to the top-left corner (0,1) of that plot (Cantor, Sun, Tortolero-Luna, Richards-Kortum, & Follen, 1999); (b) threshold maximizing Kappa (MaxKappa; Huntley, Berry, Cramer, & McDonald, 1995; Guisan, Theurillat, & Kienast, 1998); and (c) threshold maximizing TSS (MaxTSS; which is equivalent to the sensitivity–specificity sum maximization described in Liu, Berry, Dawson, & Pearson, 2005). By using a number of different thresholding techniques, we minimize the bias of thresholding techniques on the interpretation of the results.

In this study, all initial environmental and species data were available at a 25 m resolution. In real-world studies, the spatial resolution can have an important influence on model predictions, with diverging results being observed between small- and large-scale studies (e.g., Meyer & Thuiller, 2006; Mertes & Jetz, 2018; Record et al., 2018), or when changing resolution or extent (e.g., Thuiller, Brotons, Araujo, & Lavorel, 2004; Guisan, Graham, Elith, & Huettmann, 2007). This can, for instance, result from the scale dependency of the environmental predictors (Vicente et al., 2014) and spatial stochastic effects at smaller spatial scales (Scherrer et al., 2018; Steinmann, Eggenberg, Wohlgemuth, Linder, & Zimmermann, 2011). As a result, the distribution of real species cannot usually be fully explained by the abiotic predictors, as dispersal and biotic factors also play a role and interact with scale (Soberon & Nakamura, 2009). Here, we avoid this problem by using a virtual species approach, with the same predictors being used to create the species and fit their distribution models, and therefore, the initial species distributions are fully explained by the chosen predictors at the study scale (extent and resolution). This approach guaranteed that the virtual species showed realistic response curves for our landscape resulting in realistic species assemblages. In theory, the resolution should thus not matter in our study and should not affect our findings. All models were run in R software version 3.3.3 (R Core Team, 2017), using biomod2 default settings (Thuiller, Lafourcade, Engler, & Araujo, 2009), as in most published studies.

2.1.2 | Step 2. Sampling observations from virtual species maps

Virtual species presence–absence data were randomly sampled to create training datasets of different sizes (100–400–1,600) using two sampling designs: (a) selection of species data with equal number of presences–absences (equal prevalence; "EqualPrev") and (b) selection of species data taking into account a presence–absence ratio reflecting the true prevalence of the species (species true prevalence; "TruePrev"). These datasets served as control (0% error) to establish a baseline of the potential (re-)sampling bias for the different sampling schemes, modelling techniques and species (Step 2, Figure 1).

2.1.3 | Step 3. Addition of errors to the training data

For each sampling design and size, errors were randomly added (using software R) to the training data according to six different levels (i.e., "levels of error"; 0% (no error added), 10%, 20%, 30%, 40% and 50%). These errors were added either to presences only (creating false negatives [FN] by changing presences to absences), or to absences only (creating false positives [FP] by changing absences to presences; Step 3, Figure 1).

2.1.4 | Step 4. Modelling procedure

The control dataset (without error) and all the datasets with errors added were used to create SDMs (final SDMs in Step 4, Figure 1) using the same environmental predictors and modelling techniques employed to initially create the virtual species. This ensures that—without error added (i.e., controls)—the models can potentially replicate perfectly the distributions of our species, since all information that initially defined these distributions is available (i.e., same predictors), and the response curves could be fitted perfectly (i.e., if using the same technique). The only factors that can affect the performance are therefore the sample size, the change of modelling technique, the threshold method and the errors added, which we can untangle through the control and the known full distribution. In other words, having SDM predictions for our control and degraded datasets allowed us to distinguish decreases in model performance only caused by resampling (using the control dataset), the thresholding effect and from the effects caused by the errors added to the presences (FN) and/or the absences (FP).

2.1.5 | Step 5. Evaluating predictions across species and levels of errors

Finally, we evaluated all predictions built for each sample size, sampling design, modelling technique and threshold approach by measuring model fit on probability (MFp) at sampled sites and predictive success for probabilistic (PSp) and binary predictions (PSb) across the whole area (i.e., evaluation approaches; see description below), using five widely used agreement/evaluation metrics (for more information see Table 1 and Liu et al., 2005): Cohen's Kappa (Kappa), true skill statistic (TSS), maximized Kappa (MaxKappa), maximized TSS (MaxTSS) and a rescaled measure of AUC, Somers' rank correlation (Somers'D; Harrell, 2015). Somers'D was used instead of AUC, because its rescaled between −1 and +1, making it directly comparable to the other used evaluation metrics (and is therefore also intuitively interpretable along the same scale as a correlation coefficient).

Depending on the evaluation data used (i.e., evaluation approach hereafter), different evaluation metrics were used. For MFp/PSp, we calculated MaxKappa, MaxTSS and Somers'D, while for PSb only observed Kappa and TSS under a chosen threshold could be calculated (Step 5, Figure 1):

1. *Model fit probability* (MFp) corresponds to the ability of the model to reproduce the training data. It was measured by comparing predicted probabilities of the different models (control and the various levels of errors) to the data used to fit those models and thus was conducted on the same set of points used to build the models (presence–absence in training dataset with errors added; and without errors for the control).
2. *Predictive success probability* (PSp) is the potential of the model to recreate the complete true distribution of a species when the model is trained with degraded (or not) training data. It was calculated by comparing predicted probabilities of the different models (control and various levels of errors) to the original true species distribution map (presence–absence), giving Somers'D, MaxTSS and MaxKappa across the whole study area.
3. *Predictive success binary* (PSb) is the ability of the model to predict the complete true distribution of the species based on the degraded (or not) training data, using only information available to the model (no information about the truth available for threshold selection). It was calculated by comparing binary predictions of the different models (control and various levels of errors) to the complete true distribution dataset. To create binary predictions, MaxTSS (for the calculation of TSS) and MaxKappa (for the calculation of Kappa) thresholds were selected based on the predicted probabilities and the training data used in each model (calibration data with error).

Evaluating model predictions with the control data (no error added) allows to measure the effect of sampling and, more particularly, since the sampling design was random, to assess the effect of sample size. Also, to assess if evaluation values decrease with increasing errors in the training data, we standardized all our degraded models with the corresponding control (0% error) to eliminate resampling effects (see Results; difference = [evaluation value of degraded model—evaluation value of control model]). Therefore, negative values indicate that model performance decreased compared to the control (i.e., the higher the decrease, the higher the effects of errors added).

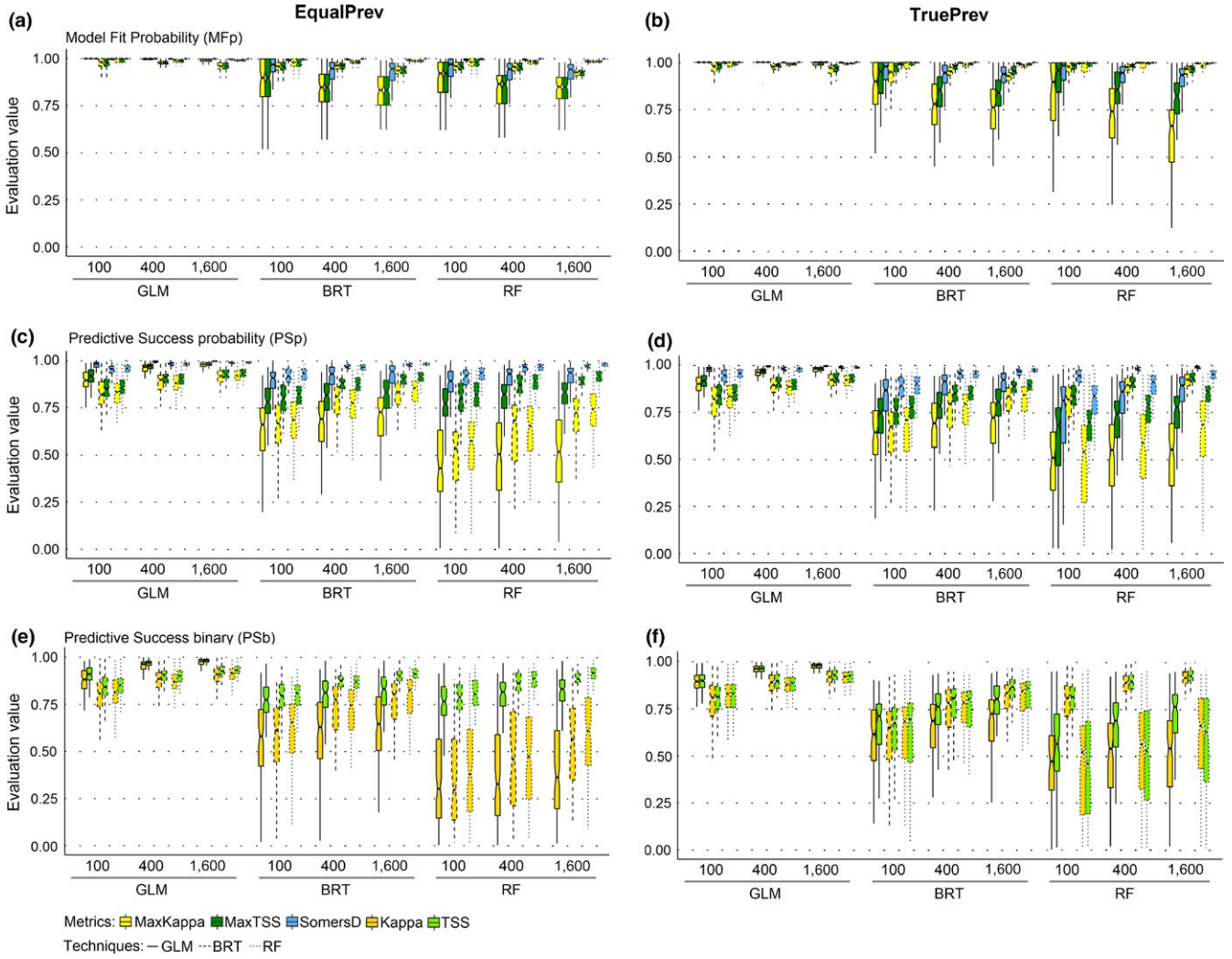

FIGURE 2 Evaluation values of control model (using training data without errors added; 0%) for "EqualPrev" (left column; a, c and e) and "TruePrev" (right column; b, d and f) sampling designs, with measured MFp, PSp and PSb for virtual species (n = 100), created using generalized linear models (GLM), boosted regression trees (BRT) or random forests (RF; initial SDMs) and with different sample sizes (100–400–1,600). Model fit probability (MFp) and predictive success probability (PSp) were measured using maximized Kappa (MaxKappa; yellow), maximized TSS (MaxTSS; green) and Somers'D (blue), while predictive success binary (PSb) was measured using Kappa (gold) and true skill statistic (TSS; light green). For each sample size, three sets of three box plots are displayed, corresponding to models fitted (final SDMs) using either GLMs (solid plots), BRTs (dashed plots) or RFs (dotted plots) and evaluated with corresponding metrics. The same applies to PSb, but only two box plots are displayed in each of the three model sets, corresponding to the two metrics used

3 | RESULTS

3.1 | Model evaluation using training data without errors added: effects of sampling

Evaluation values increased with increasing sample size, regardless of the sampling design ("EqualPrev" and "TruePrev"; Figure 2) with the exception of model fit probability (MFp) which decreased when models were fitted by GLMs/BRTs.

The MFp for the initial models (i.e., 0% with no errors added) was always above 0.75 for all modelling techniques and metrics (except MaxKappa in "TruePrev") and mostly close to 1 (which can be considered an excellent model) when species were created by GLMs or fitted using BRT/RF. In contrast, the predictive success probability (PSp) and predictive success binary(PSb) showed much higher variation, ranging from 0.75 to 1 for all metrics when species were created by GLMs, but from 0.25 to 1 when created by BRT/RF (Figure 2).

Somers'D presented always the highest evaluation values, usually followed by MaxTSS and MaxKappa (Figure 2; MFp/PSp). MaxKappa was the metric that presented the greatest range of variation, while models evaluated by Somers'D presented very similar values. When PSb was measured, TSS had the highest values and Kappa the lowest, independently of the modelling technique used (Figure 2).

Models fitted using species created by GLMs showed the highest evaluation values (usually above 0.75 for all metrics; Figure 2). However, models fitted using virtual species created by BRTs/RFs

presented a wider range of values, with model performance being worse than when species were created by GLMs. Independently of the modelling technique used to create the species, we observed that models fitted by RFs had higher evaluation values of MFp/PSp while models fitted by GLMs presented the highest values of PSb.

3.2 | Effects on model evaluation of adding errors to the training data

As the patterns observed across sample sizes were similar, we only report results on the intermediate sample size (i.e., 400; but see

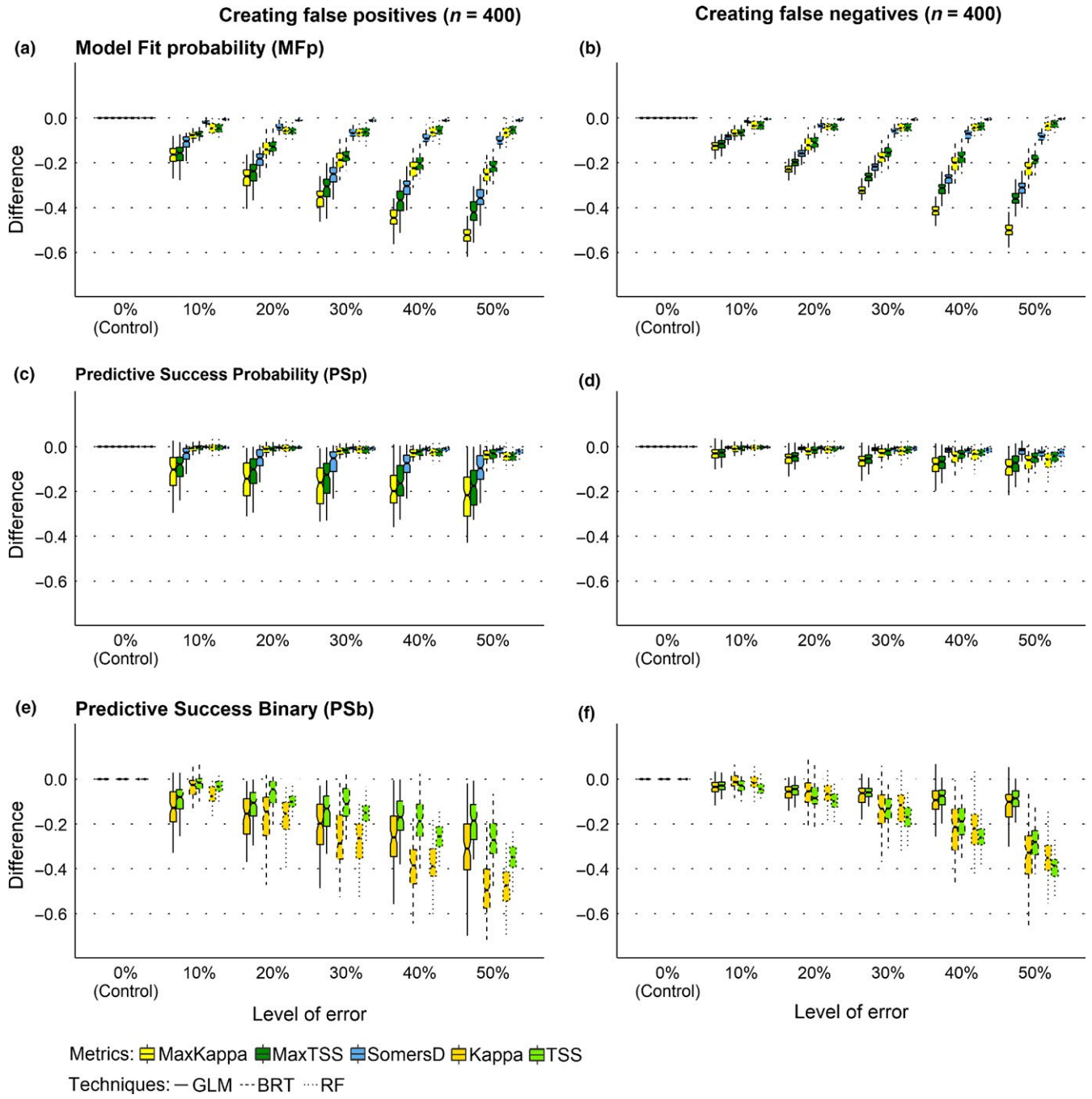

FIGURE 3 Observed difference of measured model fit probability (MFp), predictive success probability (PSp) and predictive success binary (PSb) between control (training data without errors added; 0%—sampled data) and degraded data (training data with errors added) models, under the sampling design EqualPrev and sample size 400, for virtual species created using GLM (generalized linear models). Errors were added to the occurrence dataset, creating either false positives (errors added only to absences; left column; a, c and e) or false negatives (errors added only to presences; right column; b, d and f). MFp and PSp were measured using maximized Kappa (MaxKappa; yellow), maximized TSS (MaxTSS; green) and Somers'D (blue), while PSb was measured using Kappa (gold) and true skill statistic (TSS; light green). For each level of error, three sets with three plots are observed, corresponding to models fitted using either GLMs (solid plots), BRTs (dashed plots) or RFs (dotted plots). For PSb, only two plots are present in each of the three sets

Supporting Information Appendix S2–S3 for complete results on "EqualPrev" and "TruePrev" sampling designs, respectively). The effect of error added decreased with sample size, with more accurate models being observed at higher sample sizes (i.e., difference between control and degraded models was smaller).

Regardless of the evaluation approach, as errors were increasingly added to training data, evaluation values decreased when compared with the control models (Figure 3). This decrease in model performance was more pronounced in model fit probability and predictive success binary (MFp/PSb). Still, models whose performance

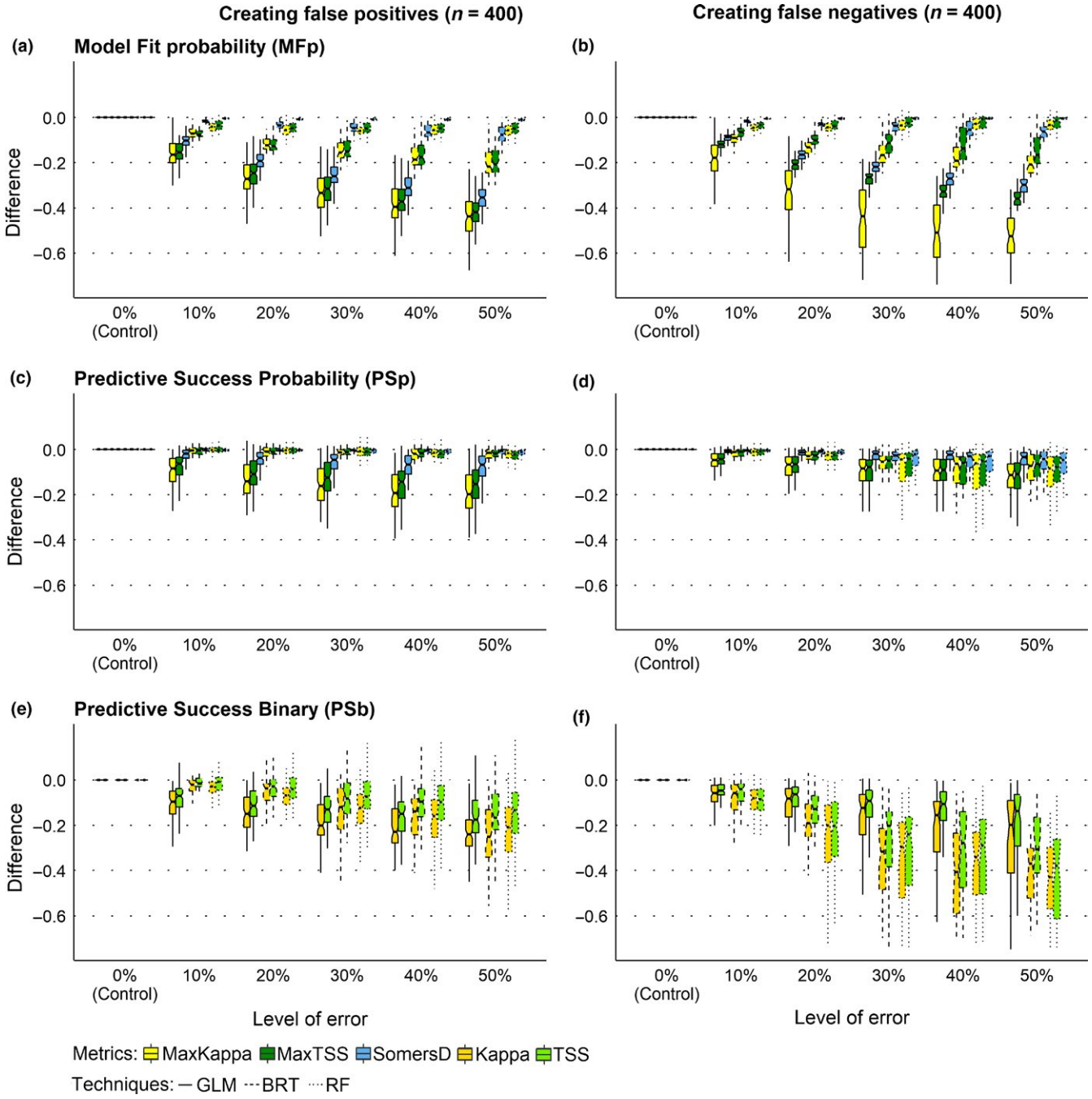

FIGURE 4 Observed difference of measured model fit probability (MFp), predictive success probability (PSp) and predictive success binary (PSb) between control (training data without errors added; 0%—sampled data) and degraded data (training data with errors added) models, under the sampling design TruePrev and sample size 400, for virtual species created using GLM (generalized linear models). Errors were added to the occurrence dataset, creating either false positives (errors added only to absences; left column; a, c and e) or false negatives (errors added only to presences; right column; b, d and f). MFp and PSp were measured using maximized Kappa (MaxKappa; yellow), maximized TSS (MaxTSS; green) and Somers'D (blue), while PSb was measured using Kappa (gold) and true skill statistic (TSS; light green). For each level of error, three sets with three plots are observed, corresponding to models fitted using either GLMs (solid plots), BRTs (dashed plots) or RFs (dotted plots). For PSb, only two plots are present in each of the three sets

decreased the most in each approach depended on the modelling technique used (Figures 3 and 4). As a result, random forests (RF) presented higher model performance when MFp/PSp were measured and generalized linear models (GLM) when PSb was measured.

In general, the creation of false positives (FP; Figure 3, left) had a stronger negative effect on model performance than false negatives (FN; Figure 3, right), but in some rare cases, the creation of FN could have a stronger effect on model performance (e.g., Figures 3e–f, PSb evaluated by TSS in models fitted by RFs).

Somers'D displayed the smallest decrease in model performance when errors were added to the calibration data (for MFp/PSp), regardless the technique used to fit the models (Figure 3) or to create the virtual species (Supporting Information Appendix S2). The strongest decrease in model performance (for MFp/PSp) was usually presented by MaxKappa, while MaxTSS presented intermediate values (Figure 3). When measuring PSb, true skill statistic (TSS) usually showed a smaller decrease in model performance (Figures 3e–f), except when creating false negatives for species generated either by BRTs (e.g., Supporting Information Figure S5: Appendix S2) or RFs (e.g., Supporting Information Figure S8: Appendix S2).

The results obtained with sampling design "TruePrev" (Figure 4 and Supporting Information Appendix S3) did not differ from those previously described in "EqualPrev" (Figure 3 and Supporting Information Appendix S2), except when models were evaluated with MaxKappa (Figure 4a–b; FN have a stronger effect on model performance when fitted by GLMs) or Kappa (Figure 4f). However, this is most likely an artefact of the difference in the number of species with successful models, which was lower when FN were created. This difference in number was due to the presence reduction in some species when adding FN, making it impossible to correctly fit a model.

Additionally, the use of different threshold techniques (to create initial models) did not bias the results and their interpretation, with the same patterns being observed across techniques (see Supporting Information Appendix S4).

After increasingly degrading the data and when MFp/PSp were measured, models fitted by GLMs (Figures 3 and 4) presented the highest decrease when compared with control models. On the opposite side, models fitted by RFs (Figures 3 and 4) were the least affected by the addition of degraded data. Still, when measuring PSb, the decrease in model performance was higher for models fitted by RFs and more stable for models fitted by GLMs (especially as the errors added increased).

We performed an additional evaluation approach, *predictive success on calibration data* (PSc), not providing the results here since it is a subset of PSb and accordingly yielded similar patterns (but see Supporting Information Appendix S5).

4 | DISCUSSION

We used a virtual ecologist approach with artificial species data to evaluate the degree to which errors in presences/absence data (see Graham, Ferrier, Huettman, Moritz, & Peterson, 2004; Guillera-Arroita et al., 2010; Tyre et al., 2003 for examples of causes like false-negative errors or imperfect detection, taxonomic inaccuracies or biases in the spatial coverage of data) can affect SDM predictions and assess the reliability of currently used evaluations metrics. By using artificial data, we prevented limitations of real-world data (most previous studies used real species data from surveys, herbaria or museums; e.g., Hernandez, Graham, Master, & Albert, 2006; Osborne & Leitao, 2009; Mitchell, Monk, & Laurenson, 2017), allowing us to have complete knowledge of the full species distribution and to simulate errors in presence/absence data with complete control of the factors affecting their distribution. The models must then find a signal in the degraded (or not) training data and be able to predict to the known remaining distribution which is largely unaffected by errors. Our work revealed four main findings. First, as expected, the effect of degraded data decreased as sample size increased. Second, the classification of a model along a range of performance (e.g., poor, fair, good, excellent) strongly depended on the metric used to evaluate it. Models evaluated by Somers'D (a rescaled measure of the AUC) still corresponded to high values of predictive performance (according to the interpretation scales as in Araujo, Pearson, Thuiller, & Erhard, 2005; fail: AUC < 0.7, fair: >0.7, good >0.8, excellent >0.9; refined from the initial scale by Swets, 1988, note nr 11). This suggests that whatever the modelling technique used, AUC, Somers D and related metrics produce over-optimistic evaluations, potentially affecting the conclusion of studies that rely solely on it (e.g., conservation prioritization studies, assessment of future climate change impacts on plants or animals, current and future threats and spread of invasive species). However, other metrics, such as Kappa (or MaxKappa), can provide more realistic evaluation. Third, we confirmed that predictions with too good model fit (MFp) usually presented low predictive success (PSb), with data-driven techniques such as RF usually tending towards higher overfitting and lower prediction success, while model-driven technique like GLMs showing the opposite (Petitpierre, Broennimann, Kueffer, Daehler, & Guisan, 2017; Randin et al., 2006). Fourth, the creation of false positives had a stronger effect in decreasing model performance than the creation of false negatives. We discuss these findings below.

4.1 | Confirming the effect of sample size and controlling for it

We found that the effects of degraded data consistently decreased with sample size, showing sample size as an important factor affecting model performance. This relationship between model performance and sample size is well known (e.g., Stockwell & Peterson, 2002; Wisz et al., 2008; Thibaud et al., 2014; Mitchell et al., 2017). It can be partially explained by the fact that with greater number of presence/absence data, a more complete (broader) information about the occupied environmental space will likely be available. This improves parameter definition, leading to more accurate predictions (Carroll & Pearson, 1998). Our results could be useful since we showed that accurate models (i.e., when all the metrics show high

evaluation values) could be generated even when substantial levels of errors (>30%) are present in the training data (if a large enough sample is collected and the adequate modelling techniques are used).

4.2 | Importance of contrasting model fit and predictive success

Different conclusions about model performance can be inferred depending on how model performance is measured (i.e., our different evaluation approaches). To our knowledge, this is the first study to formally test and compare the outputs of these evaluation approaches to assess model predictions. This was possible through the use of virtual species allowing us to simultaneously assess how well models reproduced the partially degraded training data (MFp), how well they predicted the true distribution of species despite the added errors, taking into account restrictions of real-world data (PSb) and without those restrictions when the evaluation with the complete distribution knowledge was available (PSp). Studies with real data have contrasted model fit, internal validation and external validations (e.g., Petitpierre et al., 2017; Randin et al., 2006; Wenger & Olden, 2012), which is distinct from what was done here using and only possible with artificial data.

As expected, all evaluation approaches showed a decrease in model performance with increasing degraded data (and in both sampling designs, EqualPrev and TruePrev). However, we showed that the different evaluation approaches are complementary, since predictions with good (i.e., high values) model fit (MFp) usually presented a bad (i.e., or low values) predictive success (PSb). Additionally, the same pattern is reflected in the different modelling techniques (i.e., techniques with good MFp had poor PSb and vice versa). This reflects the classical trade-off between model (over-) fitting and model predictive performance, and is supported by previous works showing a decrease in evaluation values between model fit and independent evaluation (e.g., Randin et al., 2006; transferability test, where General Additive Models (GAMs) fit better than GLMs but predict worse to independent data).

4.3 | How do evaluation metrics reflect model performance?

A consistent pattern was identified, with models evaluated by Somers'D (rescaled AUC) always yielding the highest evaluation values, usually followed by MaxTSS and MaxKappa, or TSS and Kappa (for probabilistic and binary predictions, respectively). Within the same model, Somers'D values had very small differences when compared with the control model (even with errors >30%). Somers'D (rescaled AUC; from −1 to 1) was used instead of the widely used AUC to allow direct comparisons to the other evaluation metrics, as they all range between −1 and +1, being interpreted roughly in a same way as correlation coefficients. This means that when considering Somers'D (or AUC, with even higher evaluation values, concentrated between 0.5 and 1), all models evaluated in this study would be considered at least fair (based on thresholds proposed by Swets (1988, note nr 11); i.e., models with AUC values above 0.7 are considered "useful for some purpose," while models with AUC > 0.9 are considered as being "of rather high accuracy"). However, when evaluated by the other metrics, a large amount of these models would be considered poor or not different than random. Therefore, concluding whether a model is good, fair or poor partly depends on the evaluation metric used and not only on model performance. In particular, our results suggest a strong tendency of Somers'D (i.e., AUC) to yield over-optimistic evaluations. We also observed, although in a lesser measure, a tendency of TSS (resp. MaxTSS) to yield over-optimistic values, whereas Kappa (resp. MaxKappa) proved to better reflect the level of errors added to the training data. These results are supported by recent findings showing that AUC/TSS are not the most efficient metrics to assess model performance (being over-optimistic or unrealistic) and that these could be classified as having good performance even when "dummy" data (e.g., pseudopredictors derived from paintings; Fourcade et al., 2018) or wrong information (e.g., locational uncertainty; Graham et al., 2008; Mitchell et al., 2017) was used. As a result, many models could be considered as satisfactory despite generating partially wrong spatial predictions. Our results confirm these previous criticisms and show how important it is to take into account these drawbacks in future uses of AUC (or Somers'D)—and to a lesser extent of TSS—to assess model performance. Some suggested approaches might be to assess the spatial predictions when comparing models (Mitchell et al., 2017; Randin et al., 2006) or accounting for the most relevant section of the ROC curve (Peterson et al., 2008; assuming that true absences and independent data exist). However, as noted by Fourcade et al. (2018), this "perfect" data are usually unavailable and detailed screening of ROC plots can be difficult when modelling multiple species. Therefore, the use of AUC needs to be considered with great care in future studies and the interpretation scales (Araujo et al., 2005; Swets, 1988) used to assign a level of model performance to its values need to be revisited. We believe it is probably more effective and productive to investigate new ways/methods to correctly evaluate model performance and predictions, with the use of artificial data being a useful tool to completely assess the value of these new methods.

4.4 | How do different modelling techniques deal with the degraded training data?

The contrasted results of predictions with high model fit (random forests) presenting low values of predictive success (i.e., higher with generalized linear models) and vice versa clearly show that some techniques (like RF/BRT) are good at finding a signal in the degraded training data (i.e., can fit complex responses; Merow et al., 2014) and still deliver a good MFp (as seen in Figures 3 and 4). However, these techniques are not as good at predicting to independent data (in our case to the rest of the distribution, largely unaffected by errors). On the other hand, techniques like GLMs reflect better the errors in training data (though showing a drop in MFp), but are still fairly good (within a reasonable range of error added) at predicting the true distribution of the species across the whole

study area (PSb). Considering predictive success binary (PSb) as the expected aim for any predictive model, it turns out that some modelling techniques (here GLM) are able to "compensate" for errors added to the training data (i.e., still fit a similar response curve, e.g., unimodal, with increase error bonds) while others are not (random forests; i.e., might fit totally different response curves, adapted to the modified training data). So, models with simpler response curves (like GLM) tend to better manage errors when present in the species data, resulting in better predictions to independent data (see e.g., Randin et al., 2006 when compared to GAMs), and better fit to ecological theory (Austin, 2002, 2007). More complex methods (here RF/BRT) seem to overfit the degraded data, maintaining a good/high model fit (MFp) but at the cost of a poorer/low predictive success (PSb; see also Merow et al., 2014).

4.5 | How do different types of errors affect models and metrics?

The creation of false positives (FP) had a stronger negative effect on model performance than when false negatives (FN) were created. This is especially true when species had the same number of presences–absences ("EqualPrev"), not being obvious when sampling true prevalence ("TruePrev"), possibly due to the characteristic low prevalence of some species. False positives had a stronger negative effect because presences are expected to be on average more informative. They generally occur in a unimodal and limited way along environmental gradients, contributing to a fairly clear signal that can be captured in a model. On the other hand, absences are usually less informative since they can span entire environmental gradients and thus be found for example, on both sides of the mode of a species' occurrences (i.e., would need a bimodal response to be captured). Depending on the species, absences can still hold a signal in some cases (e.g., low elevations for alpine plants), but it is likely to be on average much weaker than that of presences. We can think of the creation of false negatives (in "EqualPrev") the same way as one uses pseudo-absences (i.e., when real absences are not available), setting the weights of those pseudo-absences to 0.5 (therefore ensuring equal prevalence). As the addition of errors to presences/absences decreased model performance in both cases, it is important to account for imperfect detection in models (see Guélat & Kéry, 2018; Lahoz-Monfort et al., 2014 for recommendations).

4.6 | Conclusions, recommendations and perspectives

Our study showed that much can be learned by using artificial data where truth is known, especially by contrasting model fit and predictive success, and modellers would gain much by using these virtual approaches more systematically in the future in complement to real data. Several important findings emerged specifically from this study:

1. The effect of errors added to species data decreased with increasing sample size.
2. Different conclusions about model performance can be inferred depending on how it is measured (i.e., which metrics and which data):
3. Models with high evaluation values can be obtained even with high levels of error artificially added to the training data;
4. The classical classification or interpretation of a model as excellent, good, fair or poor strongly depends on the metric used to evaluate it and thus can be misleading;
5. Evaluation metrics matter: We identified AUC as a particularly over-optimistic metric and in a lesser measure MaxTSS (TSS for binary predictions), with often high evaluation values produced even with high levels of errors in the training data, thus not necessarily translating a good predictive success; therefore, we recommend the use of MaxKappa.
6. Modelling techniques were differently affected by added error, with some delivering better measures of predictive success (GLMs here) and others delivering better model fit (RFs).
7. The creation of false positives had a stronger effect on the measured evaluation approaches than the creation of false negatives.

A particularly important finding in our study is thus the need to seriously reconsider the current use of AUC (here rescaled, Somers'D) and its scale of interpretation. We advise caution when models are solely evaluated with this metric (and to a lesser extent by TSS and MaxTSS) as the interpretation of their quality, reliability and transferability might be too optimistic and lead to biased conclusions. The incorrect interpretation of how good/accurate a model is might have serious consequences if not considered. For example, the prioritization of specific areas for conservation can be wrong if the models used for that prioritization are over-optimistic or biased. The same can be said if invasive species prevention/eradication efforts are occurring, with an over-optimistic prediction possibly leading to management being directed to areas where those efforts are unnecessary. Taking into account previous and current studies, the most appropriate measure might be to completely cease to use AUC and instead focus on more effective evaluation metrics. Based on our results, we recommend using MaxKappa (resp. Kappa) if one wants a metric that better reflects the actual level of errors in the predictions. As it is usually preferable to evaluate models using spatially independent data (Guisan et al., 2017; James, Witten, Hastie, & Tibshirani, 2013), our results suggest that techniques that are better at reproducing ecological theory (Austin et al., 2006), like GLMs here, tend to show a better overall behaviour for modelling species distributions. However, additional modelling techniques (e.g., as found in Elith et al., 2006) should also be tested to determine the most suitable ones. Additionally, effort should be put in minimizing false-positive rates when collecting training data (e.g., improving species identification or detectability). Finally, research using a virtual ecologist approach could also be employed to further develop more reliable evaluation metrics that could be properly tested in a "controlled environment." In a general manner, a more systematic use of artificial data bears the potential to improve methodological developments considerably in future ecological and evolutionary research.

ACKNOWLEDGEMENTS

This work was supported by SNF project "SESAM'ALP—Challenges in simulating alpine species assemblages under global change" (nr 31003A-1528661). Computations were performed at Vital-IT Centre for high-performance computing of the Swiss Institute of Bioinformatics. We thank Damaris Zurell and Catherine Graham for their valuable comments on an early stage of this work.

ORCID

Rui F. Fernandes https://orcid.org/0000-0002-8578-5665
Daniel Scherrer https://orcid.org/0000-0001-9983-7510
Antoine Guisan https://orcid.org/0000-0002-3998-4815

REFERENCES

Albert, C. H., Yoccoz, N. G., Edwards, T. C., Graham, C. H., Zimmermann, N. E., & Thuiller, W. (2010). Sampling in ecology and evolution – bridging the gap between theory and practice. *Ecography*, *33*, 1028–1037. https://doi.org/10.1111/j.1600-0587.2010.06421.x

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Araujo, M. B., Pearson, R. G., Thuiller, W., & Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology*, *11*, 1504–1513. https://doi.org/10.1111/j.1365-2486.2005.01000.x

Austin, M. P. (2002). Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling*, *157*, 101–118. https://doi.org/10.1016/S0304-3800(02)00205-3

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, *200*, 1–19. https://doi.org/10.1016/j.ecolmodel.2006.07.005

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. https://doi.org/10.1016/j.ecolmodel.2006.05.023

Barry, S., & Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, *43*, 413–423. https://doi.org/10.1111/j.1365-2664.2006.01136.x

Beale, C. M., & Lennon, J. J. (2012). Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 247–258. https://doi.org/10.1098/rstb.2011.0178

Bombi, P., & D'Amen, M. (2012). Scaling down distribution maps from atlas data: A test of different approaches with virtual species. *Journal of Biogeography*, *39*, 640–651. https://doi.org/10.1111/j.1365-2699.2011.02627.x

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

Cantor, S. B., Sun, C. C., Tortolero-Luna, G., Richards-Kortum, R., & Follen, M. (1999). A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. *Journal of Clinical Epidemiology*, *52*, 885–892. https://doi.org/10.1016/S0895-4356(99)00075-X

Carroll, S. S., & Pearson, D. L. (1998). The effects of scale and sample size on the accuracy of spatial predictions of tiger beetle (Cicindelidae) species richness. *Ecography*, *21*, 401–414. https://doi.org/10.1111/j.1600-0587.1998.tb00405.x

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. https://doi.org/10.1177/001316446002000104

D'Amen, M., Rahbek, C., Zimmermann, N. E., & Guisan, A. (2017). Spatial predictions at the community level: From current approaches to future frameworks. *Biological Reviews of the Cambridge Philosophical Society*, *92*, 169–187. https://doi.org/10.1111/brv.12222

Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C., & Mace, G. M. (2011). Beyond predictions: Biodiversity conservation in a changing climate. *Science*, *332*, 53–58. https://doi.org/10.1126/science.1200303

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, *43*, 393–404. https://doi.org/10.1111/j.1365-2664.2006.01149.x

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. https://doi.org/10.1017/S0376892997000088

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*, 245–256. https://doi.org/10.1111/geb.12684

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, *28*, 337–407.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Peterson, A. T., Loiselle, B. A., & Gro, N. P. S. W. (2008). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, *45*, 239–247. https://doi.org/10.1111/j.1365-2664.2007.01408.x

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. https://doi.org/10.1016/j.tree.2004.07.006

Gu, W. D., & Swihart, R. K. (2004). Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biological Conservation*, *116*, 195–203. https://doi.org/10.1016/S0006-3207(03)00190-3

Guélat, J., & Kéry, M. (2018). Effects of spatial autocorrelation and imperfect detection on species distribution models. *Methods in Ecology and Evolution*, *9*, 1614–1625. https://doi.org/10.1111/2041-210X.12983

Guillera-Arroita, G., Ridout, M. S., & Morgan, B. J. T. (2010). Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution*, *1*, 131–139. https://doi.org/10.1111/j.2041-210X.2010.00017.x

Guisan, A., Graham, C. H., Elith, J., & Huettmann, F. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, *13*, 332–340.

Guisan, A., Theurillat, J. P., & Kienast, F. (1998). Predicting the potential distribution of plant species in an Alpine environment. *Journal of Vegetation Science*, *9*, 65–74. https://doi.org/10.2307/3237224

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in R*. Cambridge, UK: Cambridge University Press.

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S., & Peterson, A. T. (2007). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, *77*, 615–630.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Heidelberg, Germany: Springer International Publishing.

Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, *29*, 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

Hirzel, A., & Guisan, A. (2002). Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, *157*, 331–341. https://doi.org/10.1016/S0304-3800(02)00203-X

Hirzel, A. H., Helfer, V., & Metral, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, *145*, 111–121. https://doi.org/10.1016/S0304-3800(01)00396-9

Huntley, B., Berry, P. M., Cramer, W., & McDonald, A. P. (1995). Modelling present and potential future ranges of some European higher plants using climate response surfaces. *Journal of Biogeography*, *22*, 967–1001.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Berlin, Germany: Springer.

Jenkins, C. N., Powell, R. D., Bass, O. L., & Pimm, S. L. (2003). Why sparrow distributions do not match model predictions. *Animal Conservation*, *6*, 39–46. https://doi.org/10.1017/S1367943003003068

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, *21*, 498–507. https://doi.org/10.1111/j.1466-8238.2011.00683.x

Jiménez-Valverde, A., Acevedo, P., Barbosa, A. M., Lobo, J. M., & Real, R. (2013). Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. *Global Ecology and Biogeography*, *22*, 508–516. https://doi.org/10.1111/geb.12007

Kéry, M. (2011). Towards the modelling of true species distributions. *Journal of Biogeography*, *38*, 617–618. https://doi.org/10.1111/j.1365-2699.2011.02487.x

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Wintle, B. A. (2014). Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, *23*, 504–515. https://doi.org/10.1111/geb.12138

Liu, C. R., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *28*, 385–393. https://doi.org/10.1111/j.0906-7590.2005.03957.x

Lobo, J. M., Jimenez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*, 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*, 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

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

Merow, C., Smith, M. J., Edwards, T. C., Guisan, A., McMahon, S. M., Normand, S., … Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, *37*, 1267–1281. https://doi.org/10.1111/ecog.00845

Mertes, K., & Jetz, W. (2018). Disentangling scale dependencies in species environmental niches and distributions. *Ecography*, *41*, 1604–1615. https://doi.org/10.1111/ecog.02871

Meyer, C. B., & Thuiller, W. (2006). Accuracy of resource selection functions across spatial scales. *Diversity and Distributions*, *12*, 288–297. https://doi.org/10.1111/j.1366-9516.2006.00241.x

Mitchell, P. J., Monk, J., & Laurenson, L. (2017). Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. *Methods in Ecology and Evolution*, *8*, 12–21. https://doi.org/10.1111/2041-210X.12645

Mod, H. K., Scherrer, D., Luoto, M., & Guisan, A. (2016). What we use is not what we know: Environmental predictors in plant distribution models. *Journal of Vegetation Science*, *27*, 1308–1322. https://doi.org/10.1111/jvs.12444

Osborne, P. E., & Leitao, 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. https://doi.org/10.1111/j.1472-4642.2009.00572.x

Pearce, J., & Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, *133*, 225–245. https://doi.org/10.1016/S0304-3800(00)00322-7

Peterson, A. T., Papes, M., & Soberon, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, *213*, 63–72. https://doi.org/10.1016/j.ecolmodel.2007.11.008

Peterson, A. T., Soberon, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araujo, M. B. (2011). *Evaluating model performance and significance. Ecological niches and geographic distributions* (pp. 150–181). Princeton, NJ: Princeton University Press.

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., & Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions. *Global Ecology and Biogeography*, *26*, 275–287. https://doi.org/10.1111/geb.12530

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Core Team.

Randin, C. F., Dirnböck, T., Dullinger, S., Zimmermann, N. E., Zappa, M., & Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography*, *33*, 1689–1704. https://doi.org/10.1111/j.1365-2699.2006.01466.x

Record, S., Strecker, A., Tuanmu, M. N., Beaudrot, L., Zarnetske, P., Belmaker, J., & Gerstner, B. (2018). Does scale matter? A systematic review of incorporating biological realism when predicting changes in species distributions. *PLoS ONE*, *13*, e0194650. https://doi.org/10.1371/journal.pone.0194650

Scherrer, D., Mod, H. K., Pottier, J., Litsios-Dubuis, A., Pellissier, L., Vittoz, P., … Guisan, A. (2018). Disentangling the processes driving plant assemblages in mountain grasslands across spatial scales and environmental gradients. *Journal of Ecology*. 1–14. https://doi.org/10.1111/1365-2745.13037

Soberon, J., & Nakamura, M. (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, *106*(Suppl 2), 19644–19650. https://doi.org/10.1073/pnas.0901637106

Steinmann, K., Eggenberg, S., Wohlgemuth, T., Linder, H. P., & Zimmermann, N. E. (2011). Niches and noise—Disentangling habitat diversity and area effect on species diversity. *Ecological Complexity*, *8*, 313–319. https://doi.org/10.1016/j.ecocom.2011.06.004

Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, *148*, 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, *240*, 1285–1293. https://doi.org/10.1126/science.3287615

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C., & Guisan, A. (2014). Measuring the relative effect of factors affecting species distribution model predictions. *Methods in Ecology and Evolution*, *5*, 947–955. https://doi.org/10.1111/2041-210X.12203

Thuiller, W., Brotons, L., Araujo, M. B., & Lavorel, S. (2004). Effects of restricting environmental range of data to project current and

future species distributions. *Ecography*, *27*, 165–172. https://doi.org/10.1111/j.0906-7590.2004.03673.x

Thuiller, W., Lafourcade, B., Engler, R., & Araujo, M. B. (2009). BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, *32*, 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., & Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications*, *13*, 1790–1801. https://doi.org/10.1890/02-5078

Vicente, J. R., Goncalves, J., Honrado, J. P., Randin, C. F., Pottier, J., Broennimann, O., … Guisan, A. (2014). A framework for assessing the scale of influence of environmental factors on ecological patterns. *Ecological Complexity*, *20*, 151–156. https://doi.org/10.1016/j.ecocom.2014.10.005

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, *3*, 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & Distribut, N. P. S. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*, 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Munkemuller, T., … Grimm, V. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos*, *119*, 622–635. https://doi.org/10.1111/j.1600-0706.2009.18284.x

BIOSKETCH

Rui Fernandes is a PhD student and Daniel Scherrer is a postdoc in the spatial ecology group at the University of Lausanne (https://www.unil.ch/ecospat). Antoine Guisan leads this group, which specializes in spatial modelling of species, diversity and community distributions, using empirical data, statistical models and more dynamic approaches. A strong focus is given on the use of models and their predictions to support conservation management.

Authors' contributions: R.F., D.S. and A.G. conceived the ideas and developed the methodological framework; R.F. and D.S. developed the tools to create and run the data; R.F. analysed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Fernandes RF, Scherrer D, Guisan A. Effects of simulated observation errors on the performance of species distribution models. *Divers Distrib*. 2019;25:400–413. https://doi.org/10.1111/ddi.12868