ECOGRAPHY

Review and synthesis

Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing?

Christine N. Meynard, Boris Leroy and David M. Kaplan

C. N. Meynard (https://orcid.org/0000-0002-5983-6289) ✉ (christine.meynard@inra.fr), CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montpellier, France. – B. Leroy (https://orcid.org/0000-0002-7686-4302), UMR 7208 BOREA, MNHN, UNICAEN, UA, CNRS, IRD, SU, Paris, France. – D. M. Kaplan (https://orcid.org/0000-0001-6087-359X), IRD, MARBEC, (Univ. Montpellier, CNRS, Ifremer, IRD), Sète, France.

Ecography
42: 2021–2036, 2019
doi: 10.1111/ecog.04385

Subject Editor and
Editor-in-Chief:
David Nogués-Bravo
Accepted 5 June 2019

Species distribution models (SDMs) have become one of the major predictive tools in ecology. However, multiple methodological choices are required during the modelling process, some of which may have a large impact on forecasting results. In this context, virtual species, i.e. the use of simulations involving a fictitious species for which we have perfect knowledge of its occurrence–environment relationships and other relevant characteristics, have become increasingly popular to test SDMs. This approach provides for a simple virtual ecologist framework under which to test model properties, as well as the effects of the different methodological choices, and allows teasing out the effects of targeted factors with great certainty. This simplification is therefore very useful in setting up modelling standards and best practice principles. As a result, numerous virtual species studies have been published over the last decade. The topics covered include differences in performance between statistical models, effects of sample size, choice of threshold values, methods to generate pseudo-absences for presence-only data, among many others. These simulations have therefore already made a great contribution to setting best modelling practices in SDMs. Recent software developments have greatly facilitated the simulation of virtual species, with at least three different packages published to that effect. However, the simulation procedure has not been homogeneous, which introduces some subtleties in the interpretation of results, as well as differences across simulation packages. Here we 1) review the main contributions of the virtual species approach in the SDM literature; 2) compare the major virtual species simulation approaches and software packages; and 3) propose a set of recommendations for best simulation practices in future virtual species studies in the context of SDMs.

Keywords: artificial species, environmental niche models, niche, simulations, species distribution modelling, virtual ecologist

Introduction

By statistically relating occurrence data to environmental gradients, species distribution modelling (SDM) allows one to map the potential distribution of a species of interest over a specific area of interest (Soberón 2010). SDMs have therefore become

NORDIC SOCIETY OIKOS

www.ecography.org

a very powerful tool to project current, future and past distributions of diversity (Calabrese et al. 2014, Albouy et al. 2015), including scenarios of climate change and biological invasions (Guisan and Thuiller 2005, Dawson et al. 2011, Bellard et al. 2012). Despite their popularity, a key issue that remains problematic is that the methodology associated with SDMs is difficult to test and validate with real data. For example, we often lack detailed information on what really determines a species range, or on how to scale relevant physiological or population information at local scales into large-scale range maps (but see Schurr et al. 2012). Furthermore, SDMs are usually applied at regional, continental or sometimes even global scales, making it impossible to design a replicable experiment to test their properties in the real world. Considering all these difficulties, it has become clear that simulations of virtual species (VS) are potentially useful to propose and test SDM methodologies (Zurell et al. 2010, Meynard and Kaplan 2013, Thibaud et al. 2014). In an early review of knowledge gaps and advances in the SDM literature, Austin et al. (2006) and Austin (2007) went as far as to suggest that the simulation of VS should be used systematically to test any new method in this field before applying it to real data, and that those simulations should be made in accordance with ecological theory. Subsequently, VS studies have increased, covering a variety of topics (Table 1). Because of this boom in VS use, especially as it relates to testing methods to model species distributions, we found it useful to review here what topics have been explored with this approach, and to provide some guidelines regarding the methodologies used to this effect.

We started by reviewing the abstracts of 123 papers found in an initial web of science search (performed on 10 September 2018) using different combinations of 'virtual species', 'artificial species' and 'simulations' or 'simulated' + 'species distribution*' in 'ecology', to which we added references that we found relevant by doing some more specific and isolated searches, and by inspecting the references cited in relevant papers. This allowed us to filter out papers that were not relevant to our review, focusing on occurrence distribution models (rather than abundance or population models), and on papers that tested methods related to SDMs. In the rest of this review, we focus on a subset of ~60 papers that we review in more detail (Table 1, Supplementary material Appendix 1 for the full list). An earlier review of VS literature (Miller 2014) emphasized specific methodological choices during the SDM modelling part. Here we update that literature review, but we also emphasize conceptual issues related to VS simulations that can affect results and generalizability. We will therefore focus on key findings and simulation choices, and the implications that they may have for generalizing to more realistic situations. We will start by defining the VS approach with a step-by-step description of the simulation process, and pointing to key stages that can lead to important differences on the VS outcome. We then go on to summarize the major findings and related shortcomings of published VS studies, and end up with a review of the software that is currently available to simulate VS.

The virtual species approach: simulation stages and key choices

Here we will define a VS as a simulated entity that has a known occurrence–environment relationship and for which we can control the sampling strategy, as well as any other desirable property that might be useful in understanding model behavior or in proposing sampling, calibration or validation strategies. Some authors have developed virtual distributions based on simulations of abundances or population dynamics (Austin et al. 2006, Soberón 2010). However, the vast majority of the SDM literature today is based on occurrence data (i.e. presence–absence or presence-only datasets). Therefore, for the purposes of this review, we will limit ourselves to VS studies based on presence–absence simulations. Notice that the SDM literature is vast, and a review of SDMs per se is outside the scope of this article. Interested readers can find detailed accounts of SDM methods and their applications in other recent publications (Franklin 2009, Guisan et al. 2017). Here we will focus solely on a review of the VS approach in the SDM literature.

The general principle of the VS approach is to design a VS, project it into a landscape, and then use the usual SDM procedures to test one or several stages of the SDM modelling process on the VS as if it was a real species (Fig. 1). Therefore, while the VS simulation is unique to VS studies (Fig. 1, stages 1–4), the rest of the work flow (Fig. 1, stage 5) is meant to mimic real-case SDM fitting and testing, and should be familiar to most SDM users. We will therefore largely focus on stages 1–4.

The first stage in the simulation process (Fig. 1, stage 1) is to define a relationship between the VS occurrence and the environment. We will call this the definition of the initial suitability function. VS studies that focus on this stage will usually address the effects of the shape of the suitability function, the effects of the number and complexity of interactions between environmental drivers, or the effects of niche properties such as specialization or niche breadth on model performance (Table 1, Fig. 1, stage 1). The ecological niche concept is often based on the principle that there is an optimal environment for a species to survive, and therefore a Gaussian or a skewed Gaussian distribution (as a function of environment) is often assumed (Austin et al. 2006, Soberón 2010). However, this relationship can take any shape and can depend on different environmental gradients. At this stage, if multiple predictors with different range values are combined into a suitability function, it is often useful to scale them (for example to have all predictors with a mean = 0 and standard deviation = 1 over the landscape) so that their relative impacts on the species response to the environment can be directly compared.

Table 1. Examples of virtual species (VS) studies as classified by the simulation stages and topics defined in Fig. 1. The number under 'stage' refers to the numbers given in Fig. 1 for each simulation stage: 1 = generation of the virtual species; 2 = applying to a landscape; 3 = conversion to presence–absence pattern; 4 = sampling presence–absence data; 5 = SDM fitting and testing. The references mentioned in this table are separated according to their simulation strategy (threshold versus probabilistic simulation approach) in the Supplementary material Appendix 1.

Stage	Potential effects tested	References	Major conclusions
1	Effects of the shape and complexity of the suitability function	Meynard and Quinn 2007, Elith and Graham 2009, Santika 2011, Meynard and Kaplan 2012, García-Callejas and Araújo 2016	• Threshold responses are easier to predict in terms of presence–absences than other types of responses (linear, Gaussian, combination of shapes). • More complex suitability functions (e.g. composed of responses to different variables with a combination of linear and non-linear components) are more difficult to model, recover and predict. • Specific statistical models were designed for certain types of responses (e.g. threshold versus linear versus Gaussian), and they tend to perform better under those circumstances. • Model performance indices based on presence–absence classification success do not detect calibration issues (i.e. the SDM may be fitting response curves that do not match the real species response curves but still have good classification success).
1	Effects of specialization/ niche breadth	Saupe et al. 2012, Valladares et al. 2014, Soultan and Safi 2017, Connor et al. 2018	• Specialist species are more predictable than generalists are. • Simulations based on the threshold VS approach may include confounding effects of prevalence and other factors in their results.
2	Effects of resolution and extent, upscaling and downscaling	Bombi and D'Amen 2012, Lauzeral et al. 2013, Nakazawa and Peterson 2015, Fernandez et al. 2017, Connor et al. 2018, Mertes and Jetz 2018, Moudry et al. 2018	• Predictions are best when the layers used for model calibration are at the same resolution than the species response, and when the whole extent of the species distribution is included. • Upscaling strategies can be successful; downscaling strategies show more mixed results. • Prediction of species occurrences is better when high-resolution datasets are used in model calibration (but see caveats in section 'What have we learnt and what are we missing from virtual species studies?' of this review). • Scaling VS studies using a probabilistic approach are under-represented and may lead to different conclusions than threshold VS studies.
3	Species prevalence or rarity	Real et al. 2006, Albert and Thuiller 2008, Jimenez-Valverde et al. 2009, Meynard and Kaplan 2012, Fukuda and De Baets 2016	• There is a strong effect of sample bias when sample prevalence is different from species prevalence. • Species prevalence seems influential mainly when it is extreme (> 90% or < 10%), especially when sample size is small. • Some have proposed an 'environmental favorability function' that makes model output independent from sample prevalence and could be useful when different species need to be compared in the same scale (e.g. prioritization of conservation sites). • VS studies based on a threshold simulation approach do not separate appropriately rarity (prevalence) from specialization (niche breadth) and other confounding factors.
3	Effects of dispersal patterns/constraints	De Marco et al. 2008, Saupe et al. 2012, Thibaud et al. 2014, Hattab et al. 2017, De Marco and Nobrega 2018	• Predicting distributions from early stages of invasion is difficult because the species has not occupied all of its potential environmental space. At late stages of dispersal, considering dispersal constraints may become less important. • Strategies to take into account the invasion process have been proposed.

(*Continued*)

Table 1. Continued

Stage	Potential effects tested	References	Major conclusions
4	Data quality, sampling strategy, sampling size, sampling bias, imperfect detection issues	Hirzel and Guisan 2002, Jimenez-Valverde et al. 2009, Lauzeral et al. 2012, Sheth et al. 2012, Kramer-Schadt et al. 2013, Owens et al. 2013, Dorazio 2014, Guillera-Arroita et al. 2014a, Lahoz-Monfort et al. 2014, Thibaud et al. 2014, Varela et al. 2014, Stolar and Nielsen 2015, Fei and Yu 2016, Ranc et al. 2017, Liu et al. 2018	• Unbiased presence–absence data is always better than biased presence-only data, but several strategies to correct bias and use presence-only datasets have been proposed. • Representing the full environmental gradient is key to a good model calibration. • Several studies have pointed to a minimum of 50–100 occurrences needed to characterize a species environmental niche; however, there is great variability depending on species properties, whether or not absences are considered, and how is the sample biased or not. • In studies considering multiple factors at a time, sample size usually comes up as the most influential factor determining model performance, especially at the lower end of the spectrum (< 50 occurrences). • Bias in presence-only datasets can be partially corrected by using pseudo-absence bias-correction strategies. • Notice that VS studies using a threshold simulation approach will inadvertently confound several of these factors with species and sample prevalence.
4	Pseudo-absence strategy	Lobo and Tognelli 2011, Barbet-Massin et al. 2012	• The use of a large number of random pseudo-absences or a relatively smaller number of pseudo-absences coupled with a large number of iterations is recommended. • Some exceptions are found at small sample sizes and for some statistical models under special conditions. • Notice that these studies are dominated by the use of indices that are influenced by sample prevalence (e.g. AUC and TSS, Leroy et al. 2018), so we cannot discard the possibility that the increased performance when using a large number of pseudo-absences is an artifact of the indices used.
5	Statistical models	Reineking and Schroder 2006, Meynard and Quinn 2007, Elith and Graham 2009, Guillera-Arroita et al. 2014b, Thibaud et al. 2014, Qiao et al. 2015	• Statistical models may differ greatly on their results. • Some statistical models are better adapted to different types of datasets (e.g. presence-only versus presence–absence), different types of response curves (e.g. threshold, linear, symmetric bell-shaped, asymmetric bell-shaped), or other underlying factors. • Most statistical techniques perform well under the ideal conditions they were designed to solve, but in real situations, we may not have all the relevant information to choose the most pertinent one. • User knowledge of the specific statistical tool under use is fundamental to make appropriate choices. Regularization (i.e. methods that aim to balance model fit and model complexity) may be key to good model performance.
5	Thresholds for presence–absence predictions	Jimenez-Valverde and Lobo 2007, Liu et al. 2013, 2016, Meynard and Kaplan 2013	• Threshold methods based on the maximization of the sum between sensitivity and specificity, or on the minimization of the difference between sensitivity and specificity tend to outperform the others. • There is a tradeoff between sensitivity and specificity, and simulations based on a probabilistic VS approach show higher uncertainty and more caveats to the previous results.
5	Testing indices of performance (based on classification or on predicted probability/ suitability)	Li and Guo 2013, Rapacciuolo et al. 2014, Fieberg et al. 2018	• Several propositions have been made to base model performance metrics on the predicted probability of occurrence and on model calibration, rather than on the success to predict presences and absences. • Probabilistic VS studies have shown that metrics based on presence–absence classification rates are limited by the probabilistic nature of the distribution (e.g. expected AUC values < 1 even for a model that recovers the true probability of occurrence).
Others	Testing biogeographic patterns	Kent and Carmel 2011, Nakazawa and Peterson 2015, Hawkins et al. 2017	• Simulations of individual species distributions stacked to form diversity or community composition patterns have shown promise to test biogeographic/macroecological hypotheses, or test related methodological procedures.

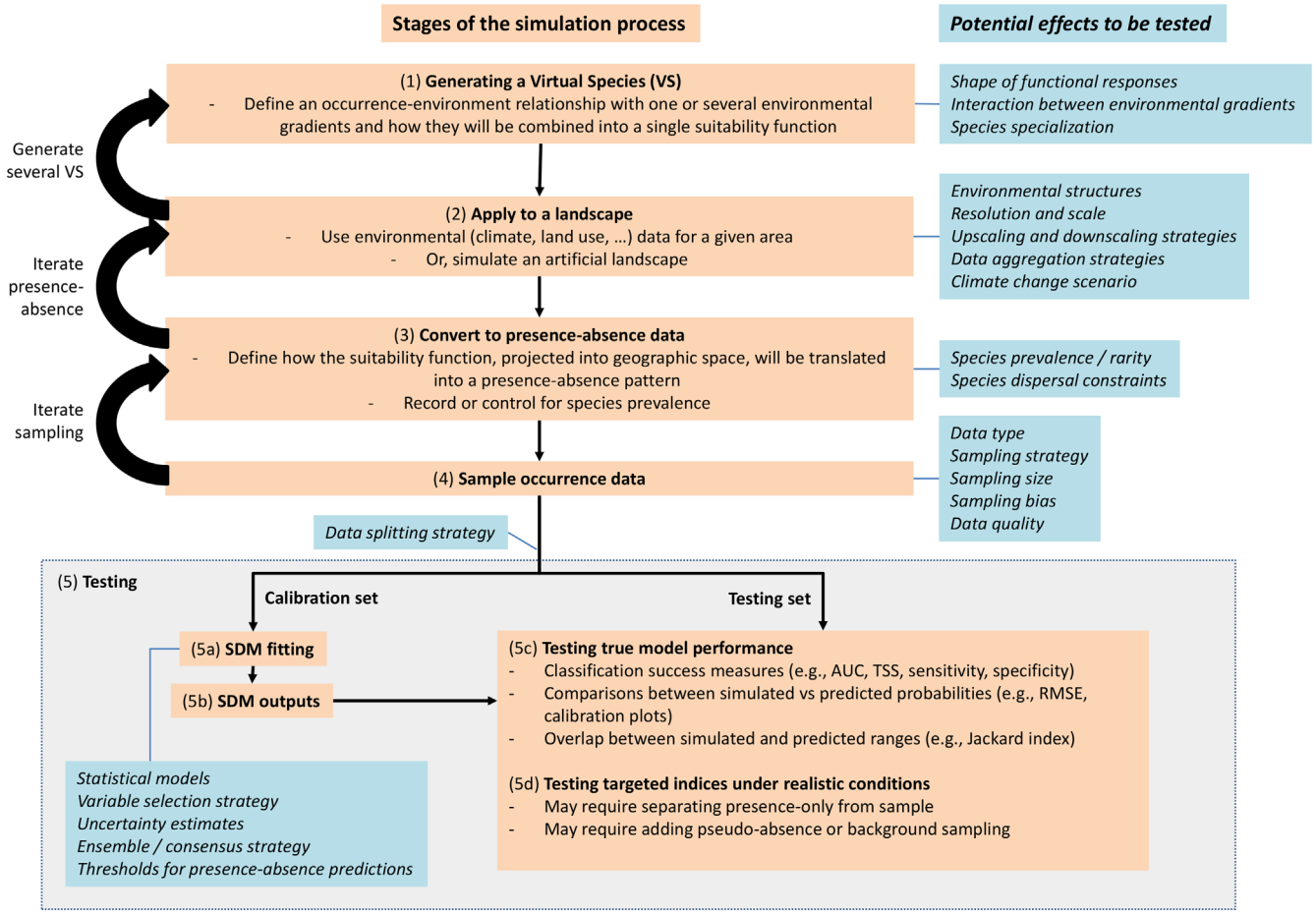

Figure 1. Stages involved in the simulation of a virtual species. The different stages are marked in the orange boxes, whereas the potential effects that can be tested at each step are marked in the blue boxes. Stages 1–4 are unique to the virtual species simulation, whereas stage 5 is common to any species distribution study. Our focus is therefore on the first stages.

A common alternative to this artificial simulation process has been to model a real species using a common SDM, and then use the resulting occurrence–environment relationship as a VS suitability function. For example, Thibaud et al. (2014) used a GLM to model the distribution of real plant species in the Swiss Alps. This resulted in occurrence–environment functional responses estimated from GLMs, which were then used as if they were true suitability functions. In the end, both types of approaches come down to the generation of a suitability function that summarizes habitat quality with respect to one or several environmental gradients (Fig. 1, stage 1).

The second stage of the VS simulation is to project the simulated relationship into a landscape (Fig. 1, stage 2). This landscape can be real or simulated itself, and this step can be accomplished in one or several steps (Fig. 1, stage 2). Using a real landscape (e.g. Worldclim data for a particular continent) has the advantage of being simple and permits a realistic set of explanatory environmental variables with collinearity and interactions that we can then relate to real case studies. However, simulating the landscape itself can have some advantages when we are trying to separate the effects of a particular method from the effects of environmental structure (Thibaud et al. 2014, Yackulic and Ginsberg 2016, Mertes and Jetz 2018). Therefore, simulating the landscape can allow greater flexibility in trying to tease apart the effects of the environmental structure per se with respect to the other components. VS studies focusing on this stage can ask questions related to the influence of environmental structures on SDM performance, resolution and scale, and data aggregation strategies, among other things (Fig. 1, Table 1, stage 2).

The third stage of the VS simulation is to decide how to transform the initial suitability values, which can theoretically have any range of values, into a presence–absence distribution, which is usually the target for modelling in SDM studies. This is a key step in the simulation process because deciding how presences and absences will be determined will affect the shape of the suitability–environment relationship originally simulated, and will also determine what is the true probability of occurrence (which, unlike the suitability function, needs to be between 0 and 1). We will therefore distinguish here the initial suitability function, which describes the species–environment relationship but can have any range of values, from the true probability of occurrence, which is

the final function after any transformation have been applied to the initial suitability function. Therefore, unlike the initial suitability function, the true probability of occurrence is bounded between 0 and 1 and gives the probability that a species is present for a given set of environmental conditions.

At this stage, there have been at least two very different approaches to transform the initial suitability function into a true probability of occurrence, each having very important consequences for the subsequent steps (Fig. 2). The first approach has been called a 'threshold simulation approach' (Meynard and Kaplan 2013) and is well exemplified by Hirzel's original studies, which greatly promoted the use of VS in this field (Hirzel et al. 2001, Hirzel and Guisan 2002). Here, presences are defined with respect to a threshold in the suitability function: any grid cell that has a suitability value greater than a certain threshold will become a presence, whereas any grid cell with a lower suitability value is translated into an absence. Although this approach is attractive because of its simplicity, it inadvertently transforms a complex suitability function (Fig. 2a, blue line) into a probability function that is either 0 (below the threshold) or 1 (above the threshold) (Fig. 2a, black line). Notice that in our example, even though the suitability function was a Gaussian (Fig. 2a, blue line), the true probability of occurrence is composed of two thresholds with a plateau of pure presences between them (Fig. 2a, black line). A common mistake in threshold VS studies has been to compare the initial suitability functions with the predicted probabilities to see if the SDM could recover the simulated environment–occurrence relationship. However, the correlation between those curves will often be low (Hirzel and Guisan 2002) because the true probability of occurrence is quite different from the initial suitability function (Fig. 2, blue versus black line). Therefore, the recovery of the true probability of occurrence (i.e. the black curve in Fig. 2a, when using a threshold) is key when the comparison of simulated and predicted probabilities is part of the model evaluation process.

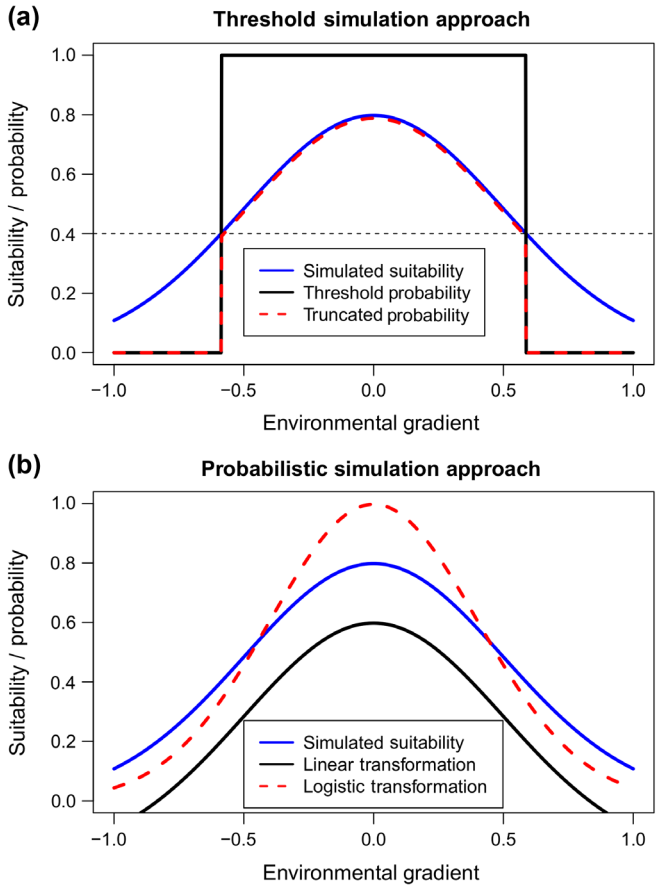

Figure 2. Schematic example of a simulated suitability function and its corresponding simulated probability under (a) a threshold simulation approach and (b) a probabilistic simulation approach. The blue line corresponds to the simulated functional response, whereas black and red lines indicate the corresponding probability of occurrence after the suitability has been converted in different ways.

The second approach to simulating presence–absences from the suitability function has been called a 'probabilistic simulation approach' (Meynard and Kaplan 2013). This approach involves interpreting the initial suitability function as a probability of occurrence, which involves scaling the suitability values to be between 0 and 1. The way the suitability function is scaled to the probability range will affect the original shape of the suitability function to individual environmental gradients. Two alternatives have been commonly used at this stage: a linear transformation (Meynard and Quinn 2007) or a logistic function (Meynard and Kaplan 2012). A linear transformation will preserve the shape of the originally simulated occurrence–environment relationships, uniformly increasing or decreasing the probabilities of occurrence across the landscape (Fig. 2b, black line). However, controlling for species prevalence and making sure that the probability of occurrence is within the 0–1 range may be a little tricky with this method (but see an example in Supplementary material Appendix 2). A logistic function on the other hand, will ensure that the simulated probability is within the 0–1 range and allow easy control of species prevalence (Meynard and Kaplan 2012). However, the logistic function will also flatten out the relationship at extreme suitability values, and narrow or broaden the intermediate probability values depending on the slope of the logistic curve (Fig. 2b, red line). Notice that a logistic function with a steep slope will approximate a threshold response (Meynard and Kaplan 2012), so the parameters of the logistic function may have a great influence on how different the true probability of occurrence is from the initial suitability function. Here again, if the VS study aims at comparing true and predicted probability values, it is important to recover the true probability of occurrence after applying the linear or logistic transformation, instead of directly using the initial suitability function.

In practice, once the suitability function has been translated into a probability of occurrence within a probabilistic VS approach, each grid cell in the landscape can be subjected to a Bernoulli trial by simply generating a random number out of a uniform distribution between 0 and 1 with which to compare its probability. Therefore, under a probabilistic

VS approach, and unlike the threshold VS approach, every iteration of the Bernouilli trial will generate a different distribution pattern. In R, the function rbinom can be used to such effect, and, if the same distribution pattern needs to be reproduced (for example, so that the readers of a paper can replicate a study under exactly the same distribution pattern), the function set.seed can be set to a specific value at the start of the simulation process. Notice that other (non-uniform) statistical distributions can be used to generate random numbers for determining presence (e.g. Thibaud et al. 2014 used random numbers drawn from a Gaussian distribution), but these will further distort the shape of the true probability with respect to the initial suitability function, as well as the intended species prevalence, if it was initially controlled for. It is therefore a good practice to check the match between the initial suitability function, the true probability of occurrence, and the simulated distribution before going further in the simulation process. In more general terms, one should always check at this stage that the properties that we intend to attribute to the VS are in fact translated into the distribution patterns generated in the simulation process, before going into the SDM testing stages.

A number of variants, some of which were intended as combinations of threshold and probabilistic approaches, have been used to simulate presence–absence or presence-only data from the simulated suitability functions. Given the importance of this step, it is interesting to think about the implicit assumptions that some of these variants reflect. For example, Barbet-Massin et al. (2012) used a threshold on the lower end of the suitability function under which the simulated species were always absent. Above that threshold, a Bernoulli trial was used to draw presences according to the simulated probability of occurrence. This combination of threshold and probabilistic processes comes down to a true probability of occurrence that looks like a truncated distribution (Fig. 2a, red line). Soultan and Safi (2017) used yet another variant, where they applied a threshold over the initial suitability function to determine the species range (the species is always present inside and always absent outside); however, they used the initial suitability inside the distribution range to place the sampling locations (i.e. the species was more likely to be sampled in the most suitable sites) and the inverse of that suitability outside the range (sites that were less suitable were more likely to be visited and marked as absent). In the end, this strategy comes down to a threshold simulation approach where sampling effort is diminished near the threshold value that defined the original range of the VS, therefore inadvertently introducing a sampling bias in the environmental gradient that was not accounted for by the authors, and transforming all suitability functions into a threshold. Here again, one should not compare true and predicted probabilities of occurrence if the modelling steps were not well recovered, especially since the initial suitability function does not have a linear relationship to the true probability of occurrence.

A more interesting variant in this step is to simulate the dispersal of the VS in the geographical space (Table 1, stage 3), for example using a spatially explicit cellular automata (De Marco and Nobrega 2018), which will start by seeding the landscape with a few occurrences and will proceed by colonizing further habitat as a function of habitat suitability and a dispersal function. This has the advantage of simulating VS distributions that potentially mimic the real species distributions with stochastic and limited dispersal, allowing one to assess the importance of these processes. However, dispersal limitation is a further confounding factor that needs to be accounted for in model analyses. Finally, another variant in the simulation of presence–absence data is to skip the geographic projection all together, and directly simulate a sampling process of environmental space. Whether that reflects a threshold or a probabilistic simulation approach will depend on how the sampling is simulated. If a simple threshold on suitability determines the presence or absence of the species, this approach is equivalent to a threshold simulation approach and it will have the same properties. If the simulated suitability is compared to a random number of uniform distribution, then the approach is equivalent to a Bernoulli trial, and therefore it falls within the probabilistic simulation approach. In other words, the geographic projection of the simulated range is not strictly necessary in some situations, as long as we understand the underlying assumptions of the sampling generation method used.

In summary, the conversion of the initial suitability function or the true probability of occurrence into a distribution pattern is a step in the simulation process that can have many variants and that can greatly affect the implicit assumptions of the VS study. By applying a threshold over the suitability function, all simulations will be based on a threshold environment–occurrence response, whereas by applying a Bernouilli trial the probability–environment relationship will be preserved. The question then becomes how much distortion, if any, was introduced in the process of converting the suitability function into a probability of occurrence (Fig. 2) or by simulating specific biases in the sampling procedure or occupancy patterns.

General recommendations and guidelines

The previous section shows that there are a number of methodological choices when carrying out a VS study and that diligence is required to make sure that the VS has the properties that were originally intended (Fig. 3). We provide a simple example of some of the most common approaches to VS simulations in a script in Supplementary material Appendix 2. In this section we will summarize our main recommendations for future VS studies (Fig. 3).

Our first recommendation is to introduce multiple checks during the simulation strategy, to make sure that the simulated environment–occurrence relationship, as well as any other key simulation step, was simulated as originally

intended (Fig. 3, blue arrows and text). This may seem as a trivial step. However, as shown in the previous section, if the simulation process is not well understood or explored, a simulation that was intended to match ecological theory or a specific species–environment relationship may well end-up leading to a distribution or sampling pattern that has little in common with such theory or with reality.

Our second recommendation is to clearly define the purpose of the VS study at hand, to define accordingly at which stage of the simulation process most effort should be allocated. When trying to understand the effects of one particular factor on SDM outcomes, it is a good idea to put more emphasis on that stage of the simulation process, leaving everything else constant and using the simplest and most transparent modeling approach. Under ideal and simple conditions, the fit between true and predicted values should be good. If that is not the case, then going back to the previous steps and understanding why the outcome is not as expected is key to providing further insights into the following modeling stages. This will guarantee that the simulation process runs as intended before testing the modelling steps of interest. For example, Jimenez-Valverde et al. (2009) wanted to understand the effects of sample size on SDM results. Therefore, they decided to use the same predictors used for simulating the VS during model calibration. That allowed them to eliminate variable selection and variable omission as a confounding effect in their results and focus on sample size and prevalence instead. Once the target effects had been isolated, then they could add other factors and understand how and when they interacted. Notice, however, that others have argued the opposite, i.e. that the only way to understand the importance of a factor is to consider it within a large set of other potential factors (Thibaud et al. 2014, Soultan and Safi 2017). Regardless of the number of factors that one aims at studying in the end, in terms of checking and understanding the simulation process, starting with the simplest case is, in our view, the best way to make sure the results are consistent with the initial intent, even if in the final analysis we end up looking at multiple factors at a time.

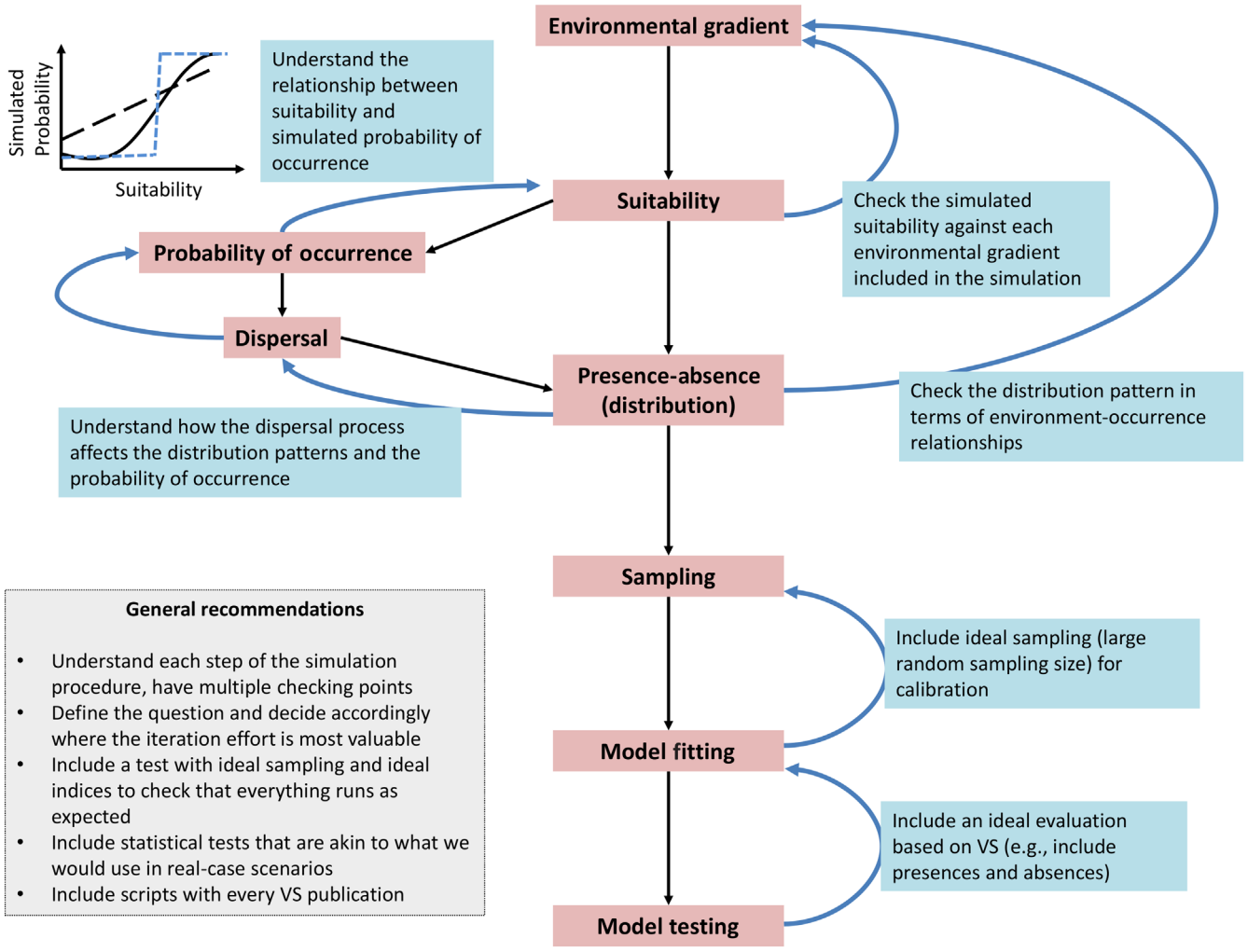

Figure 3. Graphical representation of recommendations and guidelines proposed in section 'General recommendations and guidelines' for future virtual species studies.

Clearly defining the purpose of the VS study is also key to determining the complexity of the simulation itself. In a few cases, the simulation of a distribution per se or a suitability function may not be needed at all. For example, to test different classification success performance indices, Allouche et al. (2006) and Leroy et al. (2018) simply simulated the confusion matrix (a tabulation of the number of predicted versus observed presences and absences used to calculate classification success in SDM predictions) to show that different classification success indices are subject to prevalence biases. In both cases, simulating the confusion matrix was sufficient to point to a problem with the different metrics, so a VS approach was not strictly required. However, the VS approach can add further information regarding the suitability function and its interactions with different landscape properties and sampling schemes (Thibaud et al. 2014, Fernandez et al. 2017).

Another question that often arises in VS studies is at what stage and how many iterations of the modeling processes are necessary to achieve study goals. In a threshold VS study, iterations of the distribution pattern will not produce any variability, so the number and type of iterations is only dependent on the stage of the SDM modeling process that is of interest (sampling, calibration or testing, Fig. 1, stages 4–5). However, when using a probabilistic VS approach, every distribution realization will be different. If the landscape is large enough, as is often the case in simulation studies, iterating this step will have little influence on final results, and it is more interesting to place effort on iterations at the sampling stage (stage 4 in Fig. 1). VS studies based on the probabilistic approach usually present a tradeoff between sample size and variability between iterations, so that the use of a small sample size will require more iterations to capture the full spectrum of variability than one based on larger sample sizes. We would argue that, since most real SDM studies include small sample sizes ($50 < n < 300$), it will often be more interesting to include in a VS study more iterations (i.e. 1000 iterations rather than 10) with small sample sizes, instead of few iterations with large sample sizes.

Our third recommendation has to do with how to test the results in a VS study. There seems to be a belief that testing the performance of models, even under simulated conditions, requires using exactly the same protocols that are used in real case studies (e.g. use presence-only data for VS evaluation if the main question has to do with presence-only data). We contend that one of the main advantages of the VS approach is that it is relatively easy to compare model performance under idealized conditions (i.e. using the true presence–absence and true probability of occurrence) with performance indices obtained when using more realistic data, such as presence-only data. This can, and should, be used to our advantage to separate the target question in two pieces: 1) which method performs the best under ideal conditions (question that should be addressed using true presence–absence or true probability of occurrence) and 2) are these conclusions regarding performance robust to the use of more-realistic imperfect data (e.g. presence-only data) in the evaluation process. In other words, even if the interest is in imperfect data and measures, evaluations based on the VS approach should always include more idealized performance evaluations with which to compare results.

A fourth question that we see occasionally addressed is what kind of statistical analyses should be applied to VS studies. Liu et al. (2019) argued that, as VS represent simulations, standard statistical analyses cannot be used to analyze the results. This was based on an argument advanced by White et al. (2014) and dealing with simulated data at large. In that forum paper, White et al. (2014) contended that, when using simulations, one could use unrealistically large sample sizes and find very significant results when comparing two entities that we know for a fact are different (because they were simulated as different). For example, if we generated two VS that have different suitability functions with the same environmental gradient, and then used a large sample size to test the idea that these two species have indeed different suitability functions, that would be wrong: we already know that the two species are different and differences will always be detected given sufficiently large sample size. However, the interest of the VS approach does not lie in those types of comparisons, which are indeed circular. Instead, the VS approach is useful to test which methods work under realistic conditions: at what sample size can we hope to start seeing those differences and detecting them with our conventional statistical tools? What types of sampling biases are more likely to affect our results? Given a set of problematic datasets, can we correct the results? Those are questions that we cannot answer without a statistical test.

On the opposite extreme, Thibaud et al. (2014) proposed a very specific statistical framework to follow under a VS framework involving the use of linear mixed-effects models. However, given the flexibility of the VS approach and the variety of topics that can be covered, it is unlikely that a single statistical framework will fit all purposes. We therefore recommend being careful about the conceptual pertinence of the question being asked, especially avoiding circularity between simulation and testing, and otherwise using classic statistical analyses to test the results, adapting them to the situation at hand in the same way we would do under a real-case scenario.

Our final recommendation is to provide scripts for every published simulation study. This will ensure that the study is reproducible, as well as enhance understanding of the simulation process and its implications. It will also guarantee that the results of any given study can be improved by subsequent ones. An interesting example of how the availability of scripts can help understand a problem is given by Guillera-Arroita et al. (2014b). In this study the authors were replying to Thibaud et al. (2014), who had found that MAXENT often outperformed the other models tested, including GLM, under situations where GLM should have outperformed MAXENT. By re-working on Thibaud et al. (2014)'s scripts, Guillera-Arroita et al. (2014b) were able to

show that this result was an artifact of the fact that the VS created in Thibaud et al. (2014) had, by chance, a prevalence of about 50%, which makes it a particular case under which MAXENT's regularization algorithm produces a suitability estimate that matches the species true probability of occurrence. Furthermore, whereas MAXENT prioritizes variable importance, Thibaud et al. (2014) had not implemented any variable selection or regularization strategy for GLM, making the methods not fully comparable. Guillera-Arroita et al. (2014b)'s complementary simulations then showed that GLM performed better than MAXENT when the VS had a different species prevalence and when a regularization algorithm was applied to GLM. By including the details of how the simulations were carried out, the authors of both studies provided new opportunities for advancement of value to the entire community. These details are often difficult to present in a methods section that has space constraints and that should be readable by a more generalist audience, but they are all made explicit in a reproducible script. Therefore, the availability of well-documented scripts (and in particular scripts written using tools that enhance reproducibility, such as Rmarkdown) makes it really practical and easy to share, replicate and develop new tests to understand different facets of SDM approaches.

In summary, we propose five general recommendations for future VS studies (Fig. 3): 1) implement multiple checking points at different simulations stages to make sure that the simulated species and sampling schemes have the desired properties; 2) clearly define the goals of the VS study to define accordingly at which stage iterations are the most useful and where more or less complexity should be used in the simulation process to achieve those goals; 3) include true presence–absence and/or true probability of occurrence in the evaluation process to take full advantage of the VS framework; 4) apply adapted statistical tests in the VS analyses; and 5) provide scripts for all simulations wherever possible.

What have we learnt and what are we missing from virtual species studies?

The review of results for each stage of the simulation process is summarized in Table 1. In this section we will only highlight aspects of these results that seem particularly original or different from what has been found in previous reviews (Miller 2014) and in the empirical literature.

Most virtual species studies have focused on issues related to data quality and sampling (size, bias, strategy, effects of prevalence, pseudo-absences), classification success measures (such as Kappa, AUC, TSS), and issues directly related to the modelling process. Not surprisingly, they have generally agreed that there is no replacement for good quality presence–absence data, but, more interestingly, some statistical algorithms and targeted corrective methods can be applied successfully to non-ideal situations (Table 1).

Studies that have tried to tease apart the relative importance of different factors have, in general, agreed that sampling size, especially for $n < 100$, is always one of the main determining factors for a good modelling result, often followed by the statistical model used (Thibaud et al. 2014, Soultan and Safi 2017). Also, the fewer the occurrence points, the more important it becomes that those occurrences represent the environmental gradients that are relevant in explaining the species distribution (Fei and Yu 2016).

Despite the many advances in understanding different factors affecting SDM performance (Table 1), the VS literature has been dominated by implicit assumptions related to the threshold simulation approach (reviewed by Meynard and Kaplan 2013, Meynard et al. 2019), which are the most common in published VS studies (Supplementary material Appendix 1). This shows a disconnect between niche and metapopulation theories on the one hand, which usually assume bell-shaped environment–occurrence relationships and probabilistic occurrence patterns (Hanski 1994, Austin et al. 2006, Austin 2007, Soberón 2010), and the VS simulation strategy on the other. This begs the question of whether or not there is any theoretical basis to think that threshold responses to environmental gradients should be prevalent in any SDM context.

VS studies dealing with scaling issues provide for a good example of how this simulation choice (threshold versus probabilistic simulation approach) can feed into scaling theories, resulting in opposite expectations and becoming ingrained in the current scientific paradigm. If we look at threshold VS studies, they usually conclude that using fine-grain datasets always results in increased classification rates and better model performance as compared to datasets that have been up-scaled to coarser resolutions (Fernandez et al. 2017, Mertes and Jetz 2018). This result is expected when using a threshold simulation approach because the threshold will not introduce any variability in the presence–absence distribution pattern. Under these circumstances, given an ideal sampling and modelling strategy, we can always recover the exact presence–absence pattern at the scale at which the VS was originally designed, and measures of classification rates such as sensitivity, specificity, AUC and others will be perfect. Up-scaling environmental and occurrence data can only blur this perfect predictive pattern. However, the expectation is not the same under a probabilistic simulation approach. Several VS studies have shown that performance measures based on presence–absence classification rates are not good performance measures when the species response is probabilistic rather than threshold (Reineking and Schroder 2006, Elith and Graham 2009, Meynard and Kaplan 2012). Intrinsically linked to stochasticity in the distribution pattern and on the probabilistic properties of statistical models is the fact that SDMs are designed to recover the probability of occurrence, not a specific pattern of presence and absence. Therefore, model success measures based on classification rates (e.g. sensitivity, specificity, AUC, TSS and others) will show poor values if the distribution pattern is stochastic

because the goal is misplaced, not because the model is poor (Meynard and Kaplan 2012). For example, a model that correctly predicts the probability of occurrence of a site at 0.5 will be wrong half of the time if we say that that probability corresponds to a presence or to an absence. In this case, the appropriate measure of model performance should be based on comparisons of true versus predicted probabilities of occurrence or goodness-of-fit measures rather than on predictions of presences and absences per se (Meynard and Kaplan 2012, Thibaud et al. 2014). Meynard and Kaplan (2013) suggested that a distribution that shows a probabilistic pattern at a fine temporal and spatial scale may show a threshold response at coarser scales due to the way data is aggregated in space and time. Considering as a presence any site that was occupied at any point over a long time period and over a large area, regardless of whether the site was continuously and entirely occupied or only partially occupied, has the impact of eliminating the probabilistic signal and transforming any site that has some non-zero probability of occupancy into a presence, potentially mimicking a threshold response. Large-scale threshold responses would therefore be an artifact of data aggregation.

Putting these two pieces together – i.e. the fact that stochastic distribution patterns may become threshold-like when data are aggregated, and that AUC values (or any other metric based on presence–absence classification success) will increase as the response becomes threshold-, one may hypothesize that AUC values will increase as data resolution becomes coarser, which is the opposite of what threshold VS studies have found.

To demonstrate these points, we designed a very simple simulation exercise (scripts available in Supplementary material Appendix 3). We generated a VS with a logistic response to mean annual temperature in Belgium (real gradient). This species is more likely to be found in colder areas of Belgium; real environmental data was downloaded from Worldclim (Hijmans et al. 2005) at 0.5° resolution. To look at the effects of landscape structure, we also simulated two other landscapes: one with the same range and mean values for mean annual temperature but with a perfect east-to-west trend (perfect gradient), and another one where we added random noise around the perfect gradient (noisy gradient). Then we up-scaled distributions and temperature values in 2×2, 4×4 and 8×8 grid cells by marking the larger grid cells as present if any of its constituent smaller grid cells had a presence, and by taking average values of temperature at coarser spatial resolutions. The presence–absence pattern was generated using both a threshold and a probabilistic simulation approach. Then we randomly sampled 1000 sites at the finest resolution for calibration purposes, and a different set of 1000 sites for validation purposes; we used the same sites across scales (which means that the number of sites decreases as we coarsen the resolution).

First, when projected into a perfect gradient, results fit perfectly well with our expectations (Fig. 4a): when using a threshold simulation approach, presence–absence predictions are almost perfect at all scales, especially at the finest resolution, but when a probabilistic approach is used, AUC is actually poorer at the finest resolution and increases as we up-scale the data, although variability between runs also increases. This confirms that expectations regarding model performance are opposite when using a threshold versus a probabilistic approach in this context. When we apply the same methods to a noisy gradient (Fig. 4b), the threshold approach produces a pattern of degradation and then improvement of AUC at larger scales. Overall, however, when using a threshold simulation approach on this noisy gradient, it is always better to use the finer resolution dataset, as concluded by Mertes and Jetz (2018). This contrasts with the probabilistic simulation approach results, where changes in AUC are more erratic across scales, and variance increases greatly as we aggregate over larger scales. Finally, the real temperature gradient (Fig. 4c) is a mixed bag of results where AUC values for the threshold simulation approach decrease with the up-scaling, and AUC values for the probabilistic simulation approach increase and then decrease at coarser resolutions revealing an interaction between occupancy and landscape structure in the up-scaling process.

This simulation exercise demonstrates that expectations regarding scaling issues can change significantly when a probabilistic simulation approach is used as opposed to a threshold approach. Further exploration of these patterns is needed to draw robust conclusions from these results. This exercise also shows that simulating the landscape itself can be very useful to understand the impacts of different processes on VS results. As shown in this example, the general framework proposed by Mertes and Jetz (2018) that separates response grain, environmental grain and modelling grain can be extremely useful in this sense. Finally, the theory and recommendations issued from threshold simulation studies should be interpreted carefully and revisited under the probabilistic simulation approach when the threshold response is not in agreement with ecological theory.

Software and virtual species simulation tools

We hope that the recent publication of software dedicated to simulate VS will help standardize the implementation of VS studies, as well as make the underlying assumptions more explicit.

Most VS studies initially used custom code to carry out their simulation work, and some of them published the scripts as a supplementary material (Meynard and Kaplan 2012, Guillera-Arroita et al. 2014b, Thibaud et al. 2014). However, there have been at least three software packages published for the generation of VS: sdmvspecies (Duan et al. 2015), virtualspecies (Leroy et al. 2016) and Niche Analyst or NicheA (Qiao et al. 2016). The first package was published as an R library, but it contains a limited number of functions, which can be easily recovered from other packages, it is not

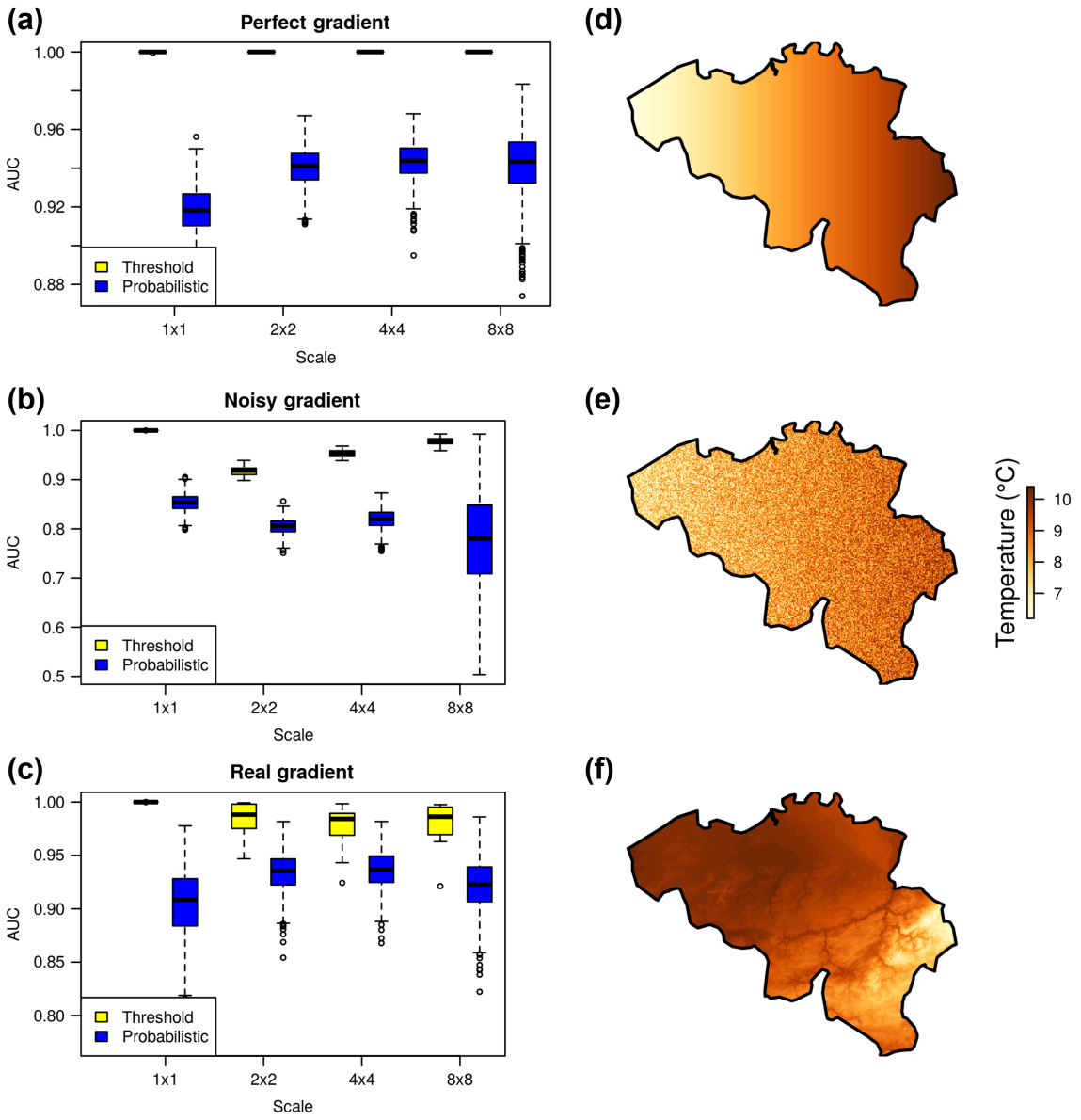

Figure 4. AUC as a function of resolution in an up-scaling VS exercise in (a) a simulated perfect east-to-west gradient of temperature, (b) a gradient with random noise built on top of the east-to-west gradient and (c) a real gradient corresponding to mean annual temperature in Belgium. The corresponding maps of each gradients are represented in (d–f). Boxplots represent results of a threshold simulation approach in yellow, and a probabilistic simulation approach in blue.

associated with a comprehensive manual, and it has not been updated since its publication. Therefore, we focus the rest of this section on virtualspecies and NicheA.

In terms of usage, both packages are fairly distinct. virtualspecies is an R package that was specifically designed to generate VS with features that mimic real-world properties and datasets. Since it is an R package, it integrates easily with the rest of the functionality that R offers and follows the same programming language, making scripting and reproducibility easy. NicheA, on the other hand, is a stand-alone Java software. It was designed to explore the interplay between environmental and geographic spaces in the context of the biological-abiotic-mobility framework proposed by Soberón and Peterson (2005), with an emphasis on multidimensional graphical displays and a threshold hypervolume niche concept. NicheA also includes tools to analyze ecological niches outside of the geographical space (Qiao et al. 2016), which we will not discuss further here. The graphical user interface of NicheA can be easier for non-programmers to understand, but may be limiting in terms of parametrization and options available for manipulation. The command-line interface of virtualspecies is more complex for users who are not used to the R language, but its open-source code offers more transparency, flexibility and possibilities in terms of batch execution

(see e.g. the different scenarios of virtual North America with 2000 species each in Hawkins et al. (2017).

Both virtualspecies and NicheA provide comprehensive tools to generate virtual species, including niche simulation, projection to geographical space, dispersal limitation and simulation of occurrence sampling (see Table 2 for a detailed comparison). Both software now implement a probabilistic simulation approach, although this feature was added later in NicheA and is based on the principle that habitat suitability is equal to one at the center of the multidimensional habitat conditions, and decreases outwards to reach zero at the edge (Qiao unpubl.), which will reduce the size of the realized niche compared to the initial hypervolume.

Regarding functionality, here we highlight three main differences (Table 2). First, both software implement distinct steps to generate virtual niches. Despite those differences, they allow generating similar species. For example, generating niches with a PCA approach in virtualspecies will resemble the 'multidimensional ellipsoid continuous niches' in NicheA. These continuous niches can in turn be transformed into presence–absence, which will be very similar in shape to the 'minimum-volume ellipsoid niches' generated in NicheA. However, notice that virtualspecies uses in this step the probabilistic approach as its default option, whereas NicheA uses the threshold approach as its default option. A second difference lies in the simulation of limited dispersal for the VS. Although both software allow defining an area of dispersal, they differ on the options provided: NicheA proposes a graphical drawing tool to define polygons, whereas virtualspecies can incorporate geographic objects (i.e. shapefiles and raster data) to limit species dispersal. Finally, a third important difference between these two software relates to the data that is recovered to later test the performance of species distribution models: while NicheA focuses on random sampling of presence-only datasets, virtualspecies can sample either presence-only or presence–absence datasets. At this

Table 2. Comparison between the two main software available to generate virtual species: virtualspecies (Leroy et al. 2016) and NicheA (Qiao et al. 2016). Names of functions or tools are indicated in italics.

	Virtual species	Niche analyst
Format	R package	Java software
Interface	R scripts	Graphical User Interface
Current version	ver. 1.4-4, September 2018	ver. 3.0.12, March 2018
Manual	<borisleroy.com/files/virtualspecies-tutorial.html>	<nichea.sourceforge.net/overview.html>
Source code	<github.com/Farewe/virtualspecies>	
Environmental data	Continuous or categorical raster data in any format readable by the R package raster (ESRI ascii, netCDF, GeoTiff, etc.). Environmental variables can be used directly to generate niches or can be transformed with a principal component analysis (PCA).	Continuous raster data in GeoTiff or ESRI ascii format. A function is available to standardise or normalise variables. Variables can be used directly to generate niches (*Draw background cloud*) or can be transformed with a PCA (principal component analysis).
Niche simulation	Virtual niches can be generated either by defining a response to each environmental variable, or by defining a response to axes of a PCA. A function is also available to generate random virtual niches (*generateRandomSp*). Options include: • From multiple variables (*generateSpFromFun*) Response functions (linear, Gaussian or any other shape available through eternal functions in R) are defined for each environmental variable and then combined to obtain an environmental suitability value. No a priori assumption is made on the nature of the niche, thus both scenopoetic and biotic variables can be used. Any formula can be used to combine partial responses into environmental suitability. • From PCA axes (*generateSpFromPCA*) Gaussian responses are defined for any subset of axes of the PCA of environmental conditions. This method generates spherical or ellipsoidal continuous niches in the environmental space on the basis of scenopoetic variables of the Grinnellian niche concept. • From non-analoguous climate (*generateSpFromBCA*): This variant of the PCA method above allows generating a virtual niche in non-analogous conditions and therefore test for extrapolation and transferability of models under conditions that do not exist today but will become available under climate change scenarios. This option uses between-group component analysis (BCA), analogous to a PCA but highlighting non-analogous conditions between two sets of environmental data for the same variables (current and future).	Virtual niches are based on scenopoetic variables in the Grinnellian niche concept. The core assumption is that niches are convex in shape and thus virtual niches are generated as convex polyhedrons or minimum-volume ellipsoids. Options include: • Minimum volume ellipsoids (*Generate virtual N(s) from ellipsoid*) Minimum-volume ellipsoids are generated by drawing the ellipsoids inside environmental space with the interface. • Convex polyhedrons (*Generate virtual N(s) from occurrences*) Convex polyhedrons are generated on the basis of a number of occurrence points in geographic space. From these occurrences, environmental data is extracted to calculate the convex polyhedron in environmental space. When a convex polyhedron is drawn, NicheA automatically calculates the corresponding minimum volume ellipsoid. • Response to one variable (*Virtual species – parameter*) Occurrence points of a species can be generated with the response to a single environmental variable according to one of four available functions: uniform, normal, binomial or Poisson distribution.

stage, virtualspecies also permits spatial biases in sampling effort or detection probability to mimic real-world occurrence datasets.

Conclusions

Here we provide a general overview of the steps involved in VS studies and the different methodological choices available. One key element is the approach used to convert a suitability function into a presence–absence distribution. Though the threshold VS approach dominates the published literature (Supplementary material Appendix 1), its use implicitly depends on a number of underlying assumptions that may not be desirable and that are not required or are more explicit in the probabilistic simulation approach. In general, we recommend using a probabilistic approach, which includes a threshold response as a special case-scenario and allows for the study of a broader spectrum of possible species distributions (Meynard and Kaplan 2013).

Many recent reviews of the SDM literature have pointed out that there is a lack of conceptual understanding of what SDM models can or cannot do, or whether the niche concept is valid at the scale of the range or the scale of individual populations (Soberón 2010, Araújo and Peterson 2012, Yackulic and Ginsberg 2016). We would argue that the VS simulation process is intrinsically linked to these same conceptual questions, and that these questions should therefore be considered at the planning stages of the VS study (Austin et al. 2006, Austin 2007). In particular, the question of whether or not a threshold response arises from aggregating data over large scales has not been fully tested in the literature. Moreover, conceptual questions such as how and when would probabilistic versus threshold strategies apply are important for determining which simulation framework is most appropriate in a particular case. The virtual species approach is powerful in that it allows isolating confounding factors and understanding their effects in different stages of the modelling process. Because of its many advantages, it has a great potential to feed macroecological theories and be applied in real case studies. However, if we are to advance the field, we must establish a rigorous simulation approach, make our assumptions more explicit, and link simulation frameworks to a theoretical macroecological background. We hope the guidelines provided above will help provide some of these links, but there are certainly many areas in which theoretical development is still needed.

Acknowledgements – We greatly appreciate the E4 award committee.
Funding – This research was supported by the INRA – SPE grant NicoTools to CNM.
Author contributions – The three authors contributed to the conceptual ideas developed in the paper and the review. BL performed the review of the VS packages; DMK performed the simulations; CNM organized the ideas and wrote the paper. Everyone contributed to final manuscript editing.

References

Albert, C. H. and Thuiller, W. 2008. Favourability functions versus probability of presence: advantages and misuses. – Ecography 31: 417–422.
Albouy, C. et al. 2015. Projected impacts of climate warming on the functional and phylogenetic components of coastal Mediterranean fish biodiversity. – Ecography 38: 681–689.
Allouche, O. et al. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.
Araújo, M. B. and Peterson, A. T. 2012. Uses and misuses of bioclimatic envelope modeling. – Ecology 93: 1527–1539.
Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. – Ecol. Model. 200: 1–19.
Austin, M. P. et al. 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. – Ecol. Model. 199: 197–216.
Barbet-Massin, M. et al. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods Ecol. Evol. 3: 327–338.
Bellard, C. et al. 2012. Impacts of climate change on the future of biodiversity. – Ecol. Lett. 15: 365–377.
Bombi, P. and D'Amen, M. 2012. Scaling down distribution maps from atlas data: a test of different approaches with virtual species. – J. Biogeogr. 39: 640–651.
Calabrese, J. M. et al. 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models. – Global Ecol. Biogeogr. 23: 99–112.
Connor, T. et al. 2018. Effects of grain size and niche breadth on species distribution modeling. – Ecography 41: 1270–1282.
Dawson, T. P. et al. 2011. Beyond predictions: biodiversity conservation in a changing climate. – Science 332: 53–58.
De Marco, P. and Nobrega, C. C. 2018. Evaluating collinearity effects on species distribution models: an approach based on virtual species simulation. – PLoS One 13: e0202403.
De Marco, P. et al. 2008. Spatial analysis improves species distribution modelling during range expansion. – Biol. Lett. 4: 577–580.
Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. – Global Ecol. Biogeogr. 23: 1472–1484.
Duan, R.-Y. et al. 2015. SDMvspecies: a software for creating virtual species for species distribution modelling. – Ecography 38: 108–110.
Elith, J. and 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.
Fei, S. and Yu, F. 2016. Quality of presence data determines species distribution model performance: a novel index to evaluate data quality. – Landscape Ecol. 31: 31–42.
Fernandez, M. et al. 2017. The importance of temporal resolution for niche modelling in dynamic marine environments. – J. Biogeogr. 44: 2816–2827.
Fieberg, J. R. et al. 2018. Used-habitat calibration plots: a new procedure for validating species distribution, resource selection and step-selection models. – Ecography 41: 737–752.
Franklin, J. 2009. Mapping species distributions: spatial inference and prediction. – Cambridge Univ. Press.
Fukuda, S. and De Baets, B. 2016. Data prevalence matters when assessing species' responses using data-driven species distribution models. – Ecol. Inform. 32: 69–78.

García-Callejas, D. and Araújo, M. B. 2016. The effects of model and data complexity on predictions from species distributions models. – Ecol. Model. 326: 4–12.
Guillera-Arroita, G. et al. 2014a. Ignoring imperfect detection in biological surveys is dangerous: a response to 'fitting and interpreting occupancy models'. – PLoS One 9: e99571.
Guillera-Arroita, G. et al. 2014b. Maxent is not a presence–absence method: a comment on Thibaud et al. – Methods Ecol. Evol. 5: 1192–1197.
Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.
Guisan, A. et al. 2017. Habitat suitability and distribution models: with applications in R. – Cambridge Univ. Press.
Hanski, I. 1994. A practical model of metapopulation dynamics. – J. Anim. Ecol. 63: 151–162.
Hattab, T. et al. 2017. A unified framework to model the potential and realized distributions of invasive species within the invaded range. – Divers. Distrib. 23: 806–819.
Hawkins, B. A. et al. 2017. Structural bias in aggregated species-level variables driven by repeated species co-occurrences: a pervasive problem in community and assemblage data. – J. Biogeogr. 44: 1199–1211.
Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.
Hirzel, A. and Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. – Ecol. Model. 157: 331–341.
Hirzel, A. H. et al. 2001. Assessing habitat-suitability models with a virtual species. – Ecol. Model. 145: 111–121.
Jimenez-Valverde, A. and Lobo, J. M. 2007. Threshold criteria for conversion of probability of species presence to either-or presence–absence. – Acta Oecol. 31: 361–369.
Jimenez-Valverde, A. et al. 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Commun. Ecol. 10: 196–205.
Kent, R. and Carmel, Y. 2011. Presence-only versus presence–absence data in species composition determinant analyses. – Divers. Distrib. 17: 474–479.
Kramer-Schadt, S. et al. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. – Divers. Distrib. 19: 1366–1379.
Lahoz-Monfort, J. J. et al. 2014. Imperfect detection impacts the performance of species distribution models. – Global Ecol. Biogeogr. 23: 504–515.
Lauzeral, C. et al. 2012. Dealing with noisy absences to optimize species distribution models: an iterative ensemble modelling approach. – PLoS One 7: e49508.
Lauzeral, C. et al. 2013. Spatial range shape drives the grain size effects in species distribution models. – Ecography 36: 778–787.
Leroy, B. et al. 2016. virtualspecies, an R package to generate virtual species distributions. – Ecography 39: 599–607.
Leroy, B. et al. 2018. Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. – J. Biogeogr. 45: 1994–2002.
Li, W. and Guo, Q. 2013. How to assess the prediction accuracy of species presence–absence models without absence data? – Ecography 36: 788–799.
Liu, C. et al. 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. – J. Biogeogr. 40: 778–789.
Liu, C. et al. 2016. On the selection of thresholds for predicting species occurrence with presence-only data. – Ecol. Evol. 6: 337–348.
Liu, C. et al. 2018. Detecting outliers in species distribution data. – J. Biogeogr. 45: 164–176.
Liu, C. et al. 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. – Ecography 42: 535–548.
Lobo, J. M. and Tognelli, M. F. 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. – J. Nat. Conserv. 19: 1–7.
Mertes, K. and Jetz, W. 2018. Disentangling scale dependencies in species environmental niches and distributions. – Ecography 41: 1604–1615.
Meynard, C. N. and Quinn, J. F. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. – J. Biogeogr. 34: 1455–1469.
Meynard, C. N. and Kaplan, D. M. 2012. The effect of a gradual response to the environment on species distribution modeling performance. – Ecography 35: 499–509.
Meynard, C. N. and Kaplan, D. M. 2013. Using virtual species to study species distributions and model performance. – J. Biogeogr. 40: 1–8.
Meynard, C. N. et al. 2019. Detecting outliers in species distribution data: Some caveats and clarifications on a virtual species study. – J. Biogeogr. <https://doi.org/10.1111/jbi.13626>.
Miller, J. A. 2014. Virtual species distribution models: using simulated data to evaluate aspects of model performance. – Prog. Phys. Geogr. 38: 117–128.
Moudry, V. et al. 2018. On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs. – Ecol. Model. 383: 3–9.
Nakazawa, Y. and Peterson, A. T. 2015. Effects of climate history and environmental grain on species' distributions in Africa and South America. – Biotropica 47: 292–299.
Owens, H. L. et al. 2013. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. – Ecol. Model. 263: 10–18.
Qiao, H. et al. 2015. No silver bullets in correlative ecological niche modeling: insights from testing among many potential algorithms for niche estimation. – Methods Ecol. Evol. 6: 1126–1136.
Qiao, H. et al. 2016. NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. – Ecography 39: 805–813.
Ranc, N. et al. 2017. Performance tradeoffs in target-group bias correction for species distribution models. – Ecography 40: 1076–1087.
Rapacciuolo, G. et al. 2014. Temporal validation plots: quantifying how well correlative species distribution models predict species' range changes over time. – Methods Ecol. Evol. 5: 407–420.
Real, R. et al. 2006. Obtaining environmental favourability functions from logistic regression. – Environ. Ecol. Stat. 13: 237–245.
Reineking, B. and Schroder, B. 2006. Constrain to perform: regularization of habitat models. – Ecol. Model. 193: 675–690.

Santika, T. 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. – Global Ecol. Biogeogr. 20: 181–192.
Saupe, E. E. et al. 2012. Variation in niche and distribution model performance: the need for a priori assessment of key causal factors. – Ecol. Model. 237: 11–22.
Schurr, F. M. et al. 2012. How to understand species' niches and range dynamics: a demographic research agenda for biogeography. – J. Biogeogr. 39: 2146–2162.
Sheth, S. N. et al. 2012. Understanding bias in geographic range size estimates. – Global Ecol. Biogeogr. 21: 732–742.
Soberón, J. M. 2010. Niche and area of distribution modeling: a population ecology perspective. – Ecography 33: 159–167.
Soberón, J. and Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. – Biodivers. Inform. 2: 1–10.
Soultan, A. and Safi, K. 2017. The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. – PLoS One 12: e0187906.

Supplementary material (available online as Appendix ecog-04385 at <www.ecography.org/appendix/ecog-04385>). Appendix 1–3.

Stolar, J. and Nielsen, S. E. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. – Divers. Distrib. 21: 595–608.
Thibaud, E. et al. 2014. Measuring the relative effect of factors affecting species distribution model predictions. – Methods Ecol. Evol. 5: 947–955.
Valladares, F. et al. 2014. The effects of phenotypic plasticity and local adaptation on forecasts of species range shifts under climate change. – Ecol. Lett. 17: 1351–1364.
Varela, S. et al. 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. – Ecography 37: 1084–1091.
White, J. W. et al. 2014. Ecologists should not use statistical significance tests to interpret simulation model results. – Oikos 123: 385–388.
Yackulic, C. B. and Ginsberg, J. R. 2016. The scaling of geographic ranges: implications for species distribution models. – Landscape Ecol. 31: 1195–1208.
Zurell, D. et al. 2010. The virtual ecologist approach: simulating data and observers. – Oikos 119: 622–635.