Journal of Biogeography (J. Biogeogr.) (2012)

SPECIAL ISSUE

Correlation and process in species distribution models: bridging a dichotomy

Carsten F. Dormann[1,2*], Stanislaus J. Schymanski[3,4], Juliano Cabral[5], Isabelle Chuine[6], Catherine Graham[7], Florian Hartig[8], Michael Kearney[9], Xavier Morin[10], Christine Römermann[11,12], Boris Schröder[13,14] and Alexander Singer[7]

[1]*Helmholtz Centre for Environmental Research – UFZ, Department Computational Landscape Ecology, 04318 Leipzig, Germany,* [2]*Biometry and Environmental System Analysis, Faculty of Forest and Environmental Sciences, University of Freiburg, D-79106 Freiburg, Germany,* [3]*Max Planck Institute for Biogeochemistry, 07745 Jena, Germany,* [4]*Swiss Federal Institute of Technology Zurich, CH-8092 Zürich, Switzerland,* [5]*Free Floater Group Biodiversity, Macroecology and Conservation Biogeography, 37077 Göttingen, Germany,* [6]*Equipe BIOFLUX, Centre d'Ecologie Fonctionnelle et Evolutive–CNRS, 34293 Montpellier Cedex 05, France,* [7]*Department of Ecology and Evolution, Stony Brook, NY 11794-5245, USA,* [8]*Helmholtz Centre for Environmental Research – UFZ, Department Ecological Modelling, 04318 Leipzig, Germany,* [9]*Department of Zoology, The University of Melbourne, Melbourne, Vic. 3010, Australia,* [10]*ETH Zürich, Forest Ecology, Institut für Terrestrische Ökosysteme, 8092 Zürich, Switzerland,* [11]*Institute for Physical Geography, Goethe-University Frankfurt am Main, D-60438 Frankfurt am Main, Germany,* [12]*Theoretical Ecology, Faculty of Biology and Preclinical Medicine, University of Regensburg, 93040 Regensburg, Germany,* [13]*Potsdam University, Institute of Geoecology, D-14476 Potsdam, Germany,* [14]*Technical University Munich, Landscape Ecology, 85354 Freising, Germany*

*Correspondence: Carsten Dormann, Biometry and Environmental System Analysis, Faculty of Forest and Environmental Sciences, University of Freiburg, Tennenbacherstrasse 4 D-79106 Freiburg, Germany.
E-mail: carsten.dormann@biom.uni-freiburg.de

ABSTRACT

Within the field of species distribution modelling an apparent dichotomy exists between process-based and correlative approaches, where the processes are explicit in the former and implicit in the latter. However, these intuitive distinctions can become blurred when comparing species distribution modelling approaches in more detail. In this review article, we contrast the extremes of the correlative–process spectrum of species distribution models with respect to core assumptions, model building and selection strategies, validation, uncertainties, common errors and the questions they are most suited to answer. The extremes of such approaches differ clearly in many aspects, such as model building approaches, parameter estimation strategies and transferability. However, they also share strengths and weaknesses. We show that claims of one approach being intrinsically superior to the other are misguided and that they ignore the process–correlation continuum as well as the domains of questions that each approach is addressing. Nonetheless, the application of process-based approaches to species distribution modelling lags far behind more correlative (process-implicit) methods and more research is required to explore their potential benefits. Critical issues for the employment of species distribution modelling approaches are given, together with a guideline for appropriate usage. We close with challenges for future development of process-explicit species distribution models and how they may complement current approaches to study species distributions.

Keywords

Hypothesis generation, mechanistic model, parameterization, process-based model, species distribution model, SDM, uncertainty, validation.

INTRODUCTION

The most commonly used approaches to describe distributions of species and biodiversity are known as correlative (syn. phenomenological) species distribution models (Elith & Leathwick, 2009). These methods aim to describe the patterns, not the mechanisms, in the association between species occurrences and environmental data (mainly climatic data).

Journal of Biogeography

They have provided useful insights for conservation of biodiversity and ecological understanding of large-scale patterns. However, predictions based on such correlative models are usually limited in their biological realism and their transferability to novel environments (Loehle & Leblanc, 1996; Davis *et al.*, 1998; Vaughan & Ormerod, 2003).

Process-based distribution models (here used synonymously with mechanistic models) can address these deficits by explicitly including processes omitted from the correlative approach (Kearney & Porter, 2009). However, process-based models often demand a large number of parameters to be estimated, many requiring data of limited availability at often high spatio-temporal resolution. Thus, to date, such models have been used for far fewer species than have correlative models.

Here we will show that current approaches to modelling species distributions represent a continuum with respect to the explicit inclusion of processes. The aim of this paper is to compare correlation and process in species distribution modelling, thereby exposing strengths and weaknesses, and differences and similarities, between these approaches. From this comparison we identify some key challenges for species distribution modelling and indicate promising avenues of integration.

DEFINITIONS

Correlative species distribution models statistically relate environmental variables *directly* to species occurrence or abundance. In contrast, process-based models formulate the ecology of a species as mathematical functions in a reductionist sense, defining causality; the species' occurrence or abundance is an *indirect*, emergent consequence. These functions are also often empirical correlations, not related to the species' occurrence or abundance, but to the species' functional traits (morphology, behaviour and physiology) and associated life history (development, growth, reproduction).

Some process-based models are developed entirely 'forward', i.e. without any calibration of the model (Kleidon & Mooney, 2000; Morin *et al.*, 2007), while correlative models are necessarily data-driven. However, correlative models employ explanatory variables that are expected to represent causal mechanisms (Austin, 2002). Furthermore, many process-based models also use distributional data to evaluate model structure or to calibrate and fine-tune some unmeasurable parameters. The common perception that process-based models are generally more complex is not true either, as machine learning-based correlative models are usually of high complexity (Elith *et al.*, 2006) while process-based models can be structurally simple (e.g. Kleidon & Mooney, 2000). We hence propose the following criteria and definitions.

In correlative models, parameters have no a priori defined ecological meaning and processes are implicit. In contrast, process-based models are built around explicitly stated mechanisms and parameters have a clear ecological interpretation that is defined a priori. Functional relationships in process-based models are specified as *causal*: *x* affects *y*. This is not the case in correlative models, although their *post hoc* interpretation is usually (and sometimes erroneously) causal.

This definition allows us to differentiate models that are described as process models (e.g. Heisey *et al.*, 2010), but are not always seen as such (Hodges, 2010; Lele, 2010), from models that are explicitly process-based. While some models can clearly be placed at the extreme ends of the correlation–process continuum, most models will fall somewhere in between, depending on the extent to which they represent processes explicitly (Fig. 1). For example, by adding dispersal to the results of a correlative projection, hybrid models can be constructed (see Appendix S1 in Supporting Information for examples). There is, as yet, no consensus what defines *hybrid* models, as opposed to *integrated* models. Here, we use the term 'hybrid model' to refer to the sequential application of different models (e.g. dispersal after correlation, Thuiller *et al.*, 2006; process-derived explanatory variables subsequently used in correlative models, Rickebusch *et al.*, 2008). As a subset of hybrid models, 'integrated models' refer to models where both modelling strategies are fitted simultaneously to data (e.g. demography within suitable habitats, Pagel & Schurr, 2011).

Process-based models also differ in the degree of calibration. We distinguish between 'forward' process-based models, where no parameter is fitted to the data to be explained, and fitted (statistically calibrated or manually tuned) process-based models, where some parameters are adjusted to match at least a subset of the data to be predicted. Examples for the former include PHENOFIT (Chuine, 2000; Chuine & Beaubien, 2001), most individual-based models (Grimm & Railsback, 2005) and the Jena diversity (JeDi) model developed by Kleidon & Mooney (2000). The latter case is more common, as unknown parameters are always easier to fit to data than to estimate independently. Examples in the context of the distribution of species functional types include CLIMEX (Sutherst & Maywald, 1985), LPJ (Sitch *et al.*, 2003), LPJ-GUESS (Smith *et al.*, 2001) and ORCHIDEE (Krinner *et al.*, 2005). In the extreme, a process-based model may be completely parameterized by

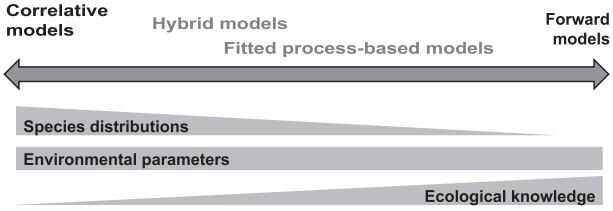

Figure 1 The correlative–process model continuum. In the most extreme case, correlative models can be applied to existing species distribution maps without any ecological knowledge (left). They commonly require a large amount of information on a species' distribution and environmental variables in order to extract useful information about the drivers. Hybrid and so-called fitted process-based models use species distribution data for parameter calibration, but they always include ecological knowledge based on other observations and/or theory. In the most extreme case, some process-based models do not require any information about a species' distribution as input data (so-called forward models).

distribution data (e.g. Van Oijen *et al.*, 2005; Arhonditsis *et al.*, 2007; also Hartig *et al.*, unpublished).

To date, there have been few direct comparisons between correlative and process-based models for the same species (Hijmans & Graham, 2006; Buckley, 2008; Morin & Thuiller, 2009; Elith *et al.*, 2010; Kearney *et al.*, 2010; Keenan *et al.*, 2011; see Appendix S1). We review the two modelling approaches from a theoretical perspective with a focus on how they have been put into practice in the past (Table 1).

PURPOSES OF SPECIES DISTRIBUTION MODELS

An obvious but sometimes forgotten point is that the usefulness of a model must be assessed with respect to its purpose. Species distribution models are used to ask a wide range of questions that can be broadly categorized as seeking understanding or seeking prediction (see also Perry & Millington, 2008). For example, understanding what limits the distribution and abundance of species is a classical question in ecology and evolution as well as in conservation and biosecurity. In very few cases can we claim to understand distributional constraints, and species distribution models are thus valuable for *generating hypotheses* that may be tested experimentally (e.g. Angert & Schemske, 2005; Doak & Morris, 2010). Process-based models can also be used to *falsify hypotheses* by formulating a hypothesis as a model and comparing it formally with data (e.g. Morin *et al.*, 2007).

Identification of the environmental factors influencing distributional limits is also useful in conservation and management. Addressing these questions using correlative models may involve *interpolation within the environmental domain* used to develop the model, and *extrapolation beyond this domain*. Correlative models often provide a single, static prediction of a species distribution in these human-driven environmental scenarios. In contrast, process-based (and hybrid) models are often used to predict dynamic features of species distributions, such as invasion rate (Kearney *et al.*, 2009a), succession and the influence of disturbance, land use and management measures on species persistence (Schumacher & Bugmann, 2006; Jeltsch *et al.*, 2011). In general, correlative models are essentially static and without access to the description of non-equilibrium, periodic, chaotic or alternative stable states (but see claims by De Marco *et al.*, 2008).

THIRTEEN FEATURES COMMON TO OR DIFFERENT BETWEEN CORRELATIVE AND PROCESS-BASED MODELS

1. Assumptions

Both purely correlative models and fitted process-based models have the same statistical analysis assumptions: error structure assumptions (such as independence of data, homogeneity and stationarity of variance), homogeneity of sampling effort and constant observation error. Both approaches can be adjusted to accommodate violations of these assumptions (e.g. spatial autocorrelation, Dormann *et al.*, 2007; non-stationarity, Hothorn *et al.*, 2011; detection probability and observer error, Royle *et al.*, 2005), but this is rarely done (for example see Latimer *et al.*, 2006; Bierman *et al.*, 2010). Both approaches assume that the relevant mechanisms influencing a species' distribution are captured. Usually, the assumption is made that the functional forms of the relationships between species occurrence and environmental variables are correct. In contrast, forward process-based models do not use distribution data for their development, and therefore such data can be used to validate forward-process models (Chuine & Beaubien, 2001).

Correlative models require species to be in equilibrium with their environment, i.e. occurring throughout the suitable *environmental* space (although not to fill the *geographic* space completely). Process-based models can abolish the equilibrium assumption and use data from the non-equilibrium trajectory to fit the model (for potential bias resulting from transient phase data see Moilanen, 2000). When the equilibrium assumption is removed it is possible to assess range dynamics

Table 1 Comparative aspects of correlative and process-based models as discussed in this paper.

Topic	Issues discussed
Assumptions	Error structure, structure of functional relationships, relevant processes/predictors, equilibrium with environment
Information required	Distribution data, environmental data, ecological and biological knowledge
Determination of model structure	Variable selection, alternative functional relationships, submodels
Verification	Technical correctness, model diagnostics
Validation	Cross-validation, external validation, parameter validation, sensitivity, specificity
Sources of uncertainty in model predictions	Input data, model misspecification, regression dilution, stochasticity
Equifinality	Over-parameterization, collinearity, non-identifiability
Extrapolation	Model domain, (micro-)evolution, constancy of limiting factors and interactions
Transferability to other species, sites and times	Functional types, correlation structure
When to stop: accuracy versus complexity	Deployment time, re-parameterization, sensitivity analysis
Communicability/transparency of the model	Documentation, open source code/software
Knowledge potentially gleaned from the model	Surprise, emergence
Common errors and misuses	Lack of uncertainty analysis, use beyond purpose, overconfidence in communication

under climate change (Kearney *et al.*, 2008; Keith *et al.*, 2008) or commercial exploitation (Cabral *et al.*, 2011).

2. Information required

Ecological knowledge of a given target species is used in both correlative and process-based models. In correlative models, it guides the pre-selection of explanatory variables. For example, when they are correlated, direct variables are preferable over indirect (e.g. temperature has a potential direct effect on thermal regulation, while elevation serves as a less specific proxy; Austin, 2002; Guisan & Thuiller, 2005). In process-based models, ecological knowledge guides the selection and formulation of processes represented in the model. Relevant ecological knowledge can be derived from experimental results or process observations (e.g. seed dispersal, phenology, growth). Correlative models require data on species distribution and relevant environmental factors for deriving correlations, while fitted process-based models require these same data to calibrate their unknown parameters. Data on species distribution and environmental factors are also used in model validation (see below).

Processes that occur at smaller spatial or temporal scales than environmental data used in model building can be difficult to capture. For instance, a species' range may be limited by a few days of frost not captured in commonly used monthly temperature averages. A small refuge may result in a species' presence in a 100-km^2 cell of unsuitable habitat, but land-cover data might not be at a sufficient resolution to capture this important microhabitat (Trivedi *et al.*, 2008). The effect of temperature on many organisms is complex and cannot be represented by interpolated monthly mean air temperature from weather station records. The latter problem has been dealt with by combining models of microclimate with models of behavioural thermoregulation (Kearney *et al.*, 2009b). Finding additional ways to represent subscale heterogeneity for the analysis at larger scales is an open challenge for both correlative and mechanistic models.

3. Determination of model structure

At first glance, it would seem that correlative and process-based models have little in common when it comes to the selection of which variables/processes are incorporated in the model. If, however, a likelihood can be formulated for the process-based model, information-theoretical model selection (Burnham & Anderson, 2002; Johnson & Omland, 2004) can proceed similarly for both approaches (see O'Hara & Sillanpää, 2009 for a Bayesian perspective). With Bayesian approaches becoming more widespread, more similarities with respect to the determination of model structure between correlative and process-based models may emerge, informing the analyst on processes relevant for the given data (e.g. Van Oijen *et al.*, 2005).

A more crucial distinction is *how* processes are included. In a correlative approach, allowing for nonlinearity, functional relationships are derived by fitting species occurrences or abundance to environmental data. In process-based models, choices about the specific process structure have to be made based on theory or observation, with potentially large ramifications for the model output even from seemingly small choices (such as between frequency- and density-dependence of disease transmission: Wasserberg *et al.*, 2009). Evaluation of alternative choices is rare (or rarely published), however, and Smith *et al.*'s (2008) study on different density-dependence schemes for cormorant population dynamics is a rare exception. Effectively, this means that in addition to the validation of the complete model all its components need to be validated as well (e.g. LaDeau, 2010).

4. Verification

Verification refers to testing the technically correct implementation of the model, i.e. that the model does what it was specified to do (Schmolke *et al.*, 2010). The use of this technical term is somewhat unfortunate, because, philosophically, verification of models is impossible (Oreskes *et al.*, 1994), but we use it in line with other publications. Verification of a process-based model is usually carried out by running the model using settings for which the outcome is known, or can be derived analytically, and by comparing the model output with the expected result. Also, dimension analysis is a crucial ingredient, i.e. checking that the units of the right- and left-hand sides of model equations are identical. Essentially, the aim of verification is to try hard to find a flaw in the implementation by producing inconsistent results. In correlative models, verification comprises double-checking of settings, assumptions (error distributions) and pre-processing steps. In that sense, model diagnostics (i.e. distribution of residuals, check for spatial autocorrelation) are the most common steps in the 'verification' of statistical models. For process-based models, beside the consistency with fundamental physical laws such as conservation of mass and energy, reproduction of analytical results or simulations using, for example, the virtual ecologist approach (Zurell *et al.*, 2010) are options. Model verification is more complex than these lines suggest, and we thus refer to Starfield *et al.* (1990) and Grimm & Railsback (2005) for further details.

5. Validation

Validation refers to the assessment of the correctness of model predictions using data not used for the building or calibration of the model. When independent data are available (ideally from another time *and* region; Lebreton *et al.*, 1992; Schröder & Richter, 1999; Araújo *et al.*, 2005), both modelling approaches can be validated externally. The commonly used cross-validation (also called internal validation) of correlative models is intrinsically optimistic compared with external validation, because it 'only' validates the model for data from the same region and time. The generality of the model hence remains unassessed. For fitted process-based models, external validation can also be carried out by comparing their

parameter estimates with independent parameter estimates (Cabral & Schurr, 2010; Hartig *et al.*, unpublished). In contrast to tuned process-based and correlative models, external validation is the rule in forward process-based models, where processes are usually parameterized on separate data sets. The absence of independent parameter estimates in the literature gives an important feedback to empiricists for improving the knowledge of the species' biology.

In addition to validation, both model approaches should also be assessed for their sensitivity and specificity (in the statistical sense). First, by using a simulated species, the ability of the model to recover the known true distribution/parameters can be assessed (i.e. the sensitivity of a method; Reineking & Schröder, 2006). Second, by randomizing the validation data, the model's tendency to find patterns in data where there are none can be gauged (specificity). These tests seem to be more established (but not necessarily published; Grimm & Railsback, 2005) for process-based models than for correlative models (but see Dormann *et al.*, 2008a).

6. Sources of uncertainty in model predictions

Generally, model uncertainty is poorly quantified (Clark *et al.*, 2001). Five sources of reducible (or epistemic) uncertainty pertain to both modelling approaches (Beven & Freer, 2001; Barry & Elith, 2006; Refsgaard *et al.*, 2007): input data uncertainty, model misspecification, equifinality (see next section), parameter uncertainty, model stochasticity and regression dilution. Even worse, these errors may be non-independent, thus amplifying their effects rather than outbalancing them. Obviously, incorrect input data used for parameterization of processes or fitting of the correlative model will bias predictions. For correlative models it has been shown that bias in presence–absence data (e.g. due to the so-called botanist effect; Applequist *et al.*, 2007; Pautasso & McKinney, 2007) is more serious than undersampling per se (Dennis *et al.*, 1999; Royle *et al.*, 2005). Similarly, an incorrectly specified model, for example one where a nonlinear process is represented by a linear relationship or a relevant predictor/process is absent, can distort the output. In correlative and fitted process-based models the distortion may be difficult to detect, as it may be compensated by altered values of other fitted parameters, leading to good model fits. In forward process-based models, the incorrect representation of processes is likely to yield stronger bias in model predictions because there is no room for such compensation. Because correlative models are typically more flexible than process-based models, the latter tend to be more biased (Hartig *et al.*, unpublished). However, this also implies that a fitted model may be giving the right results for the wrong reasons (see next section 'Equifinality'). The inclusion of stochastic processes in process-based models (e.g. dispersal, mortality) or the use of randomization steps in correlative models (cross-validation, bootstrap aggregation) will yield different model outcomes despite identical initial conditions/data. Thus, even attempts to include more and more processes in order to make a model more realistic are ultimately confronted with such stochastic and irreducible (= aleatory) variability, defining the fundamental limit of a model's accuracy.

A large, but in its effects difficult to quantify, uncertainty derives from the lack of representation of small-scale processes in large-scale data. Animals can avoid microclimatically adverse conditions so that the climate encountered by an organism is often different from the regional average. The difficulty of deciding whether to include a microscale process into a process model is conceptually similar to having an indirect explanatory variable in a regression model. In addition to this scale problem, coarse environmental data usually have large errors, which leads to 'regression dilution' and hence underestimation of the strength of a relationship (McInerny & Purves, 2011). (Note that error in the response variable does *not* cause a bias in ordinary regression, while error in the predictor variable does; Draper & Smith, 1998.)

7. Equifinality

For a given data set, several parameterizations may exist that equally fit the data ('non-identifiability'). This equifinality (Beven & Freer, 2001) is the consequence of a statistically ill-posed problem, where the information content in the calibration data set is insufficient to filter out a single parameter set from all possible sets. One consequence is that we cannot identify a single best parameter set that is likely to produce the right results for the *right* reasons (Kirchner, 2006). The causes of equifinality differ between correlative and process-based models (collinearity and over-parameterization, respectively), but the problem of resulting prediction uncertainty is the same. In ecology, this problem has not received much attention (but see Penteriani, 2008; Luo *et al.*, 2009), mainly because ecological models are complex and data are sparse, and hence fitting models other than very simple models is rare (Schulz *et al.*, 2001; Lele *et al.*, 2010). Even though we can use the full set of equifinal solutions for averaged prediction (both in a Bayesian as well as in a frequentist setting; Link & Barker, 2006; Dormann *et al.*, 2008b), we do not learn much about our system from this fitting exercise.

8. Extrapolation

When using distribution models for prediction beyond the data range (extrapolation in geographical space or in time, where new environmental conditions occur), more assumptions become relevant for both approaches. So far, studies have commonly considered *stationarity*, i.e. that model parameter estimates remained constant through space and time (but see Kearney *et al.*, 2009a; Hothorn *et al.*, 2011). Specifically, this means that the environmental niche of the species does not change (e.g. through microevolution, genetic drift or acclimation; Aitken *et al.*, 2008). Process-based models can alleviate this problem by trying to explicitly represent microevolutionary environmental niche shifts in the model (Kearney *et al.*, 2009a; Chevin *et al.*, 2010).

Furthermore, both correlative and process-based approaches assume that the way variables/processes interact will be the same in the extrapolated case as they were with the original data. For correlative models this means that the correlations found when the model was built will remain the same in the (far) future. For process-based models this means, for example, that the functional forms of the processes and parameter values stay the same. This may be quite likely for some processes (e.g. those depicting thermodynamic laws, such as body temperature or water balance), but less likely for others (e.g. a dispersal function or biotic interactions). The palaeoecological record indicates clearly that plant species in the past have reacted idiosyncratically to climatic changes (Huntley, 1991). Furthermore, the extent to which it is reasonable to extrapolate also depends on whether a process has been described empirically or from the structure and assumptions of a general theory. For correlative models, the model should not be extended outside the conditions under which the measurements were performed (e.g. elevated CO_2). For process-based models, it seems reasonable to extrapolate to conditions under which the general theory is supposed to hold.

9. Transferability to other species, sites and times

For correlative models, several studies have explored the transferability of a model to other species (Peterson *et al.*, 1999; Schröder & Richter, 1999; Bonn & Schröder, 2001; Hein *et al.*, 2007), other sites (Randin *et al.*, 2006; Broennimann *et al.*, 2007; Pearman *et al.*, 2007) and other times (Araújo *et al.*, 2005; Araújo & Rahbek, 2006; Giesecke *et al.*, 2006). Overall, generality was found to be very low, indicating that the models are tailored to species (or data) idiosyncrasies rather than to general features of the ecology, although they are sometimes interpreted as species responses (e.g. niche shifts: Broennimann *et al.*, 2007). Three possible reasons for low generality are: (1) the change of correlation structure of predictors in space and time, violation of assumptions of the statistical model (Bahn & McGill, 2007; Currie, 2007); (2) incorrect identification of relevant processes (Beale *et al.*, 2008); and (3) environmental factors that limit a species' distribution changing in time or space.

We are not aware of many comparative studies of this type for process-based models (but see Bugmann & Solomon, 1995, 2000; Bugmann, 1996), possibly because the choice of parameters is usually tailored to a certain species (see 'Information required', above).

10. When to stop: accuracy versus complexity

Correlative models can be fitted to data in a matter of minutes to hours. In fact, preparations of environmental and occurrence data usually take much longer than the actual statistical modelling process itself. This fast deployment time is probably the main cause of the proliferation of statistical methods in our data-rich times.

Process-based models commonly take a long time to develop, as they often simulate nonlinear dynamics and hence have to deal with issues such as numerical diffusion and time stepping (Press *et al.*, 2007). Furthermore, they are usually very sensitive to initial conditions and need burn-in periods to achieve a reproducible, stable steady state. This can take considerable computation time and hence slow down the developmental cycle even more. Even the use of an existing process model for a new species can take considerable time and effort, as the parameterization requires either collection of experimental or observational data with respect to the phenology and physiology of the species or model calibration to an existing distribution data set.

Scaling-up and sensitivity analyses of complex process-based models can be time-consuming (Bolker *et al.*, 1998; Pagel *et al.*, 2008). Due to the computational demand of dynamic process-based models, this cannot always be achieved using automated tools. In fitted process-based models, an accuracy–complexity return curve (depicting gain in accuracy over model complexity) is likely to be similar to that of a correlative model, levelling off fast once a 'sufficient' level of complexity is reached. For correlative models this is described by the 'variance–bias trade-off' (Hastie *et al.*, 2009), but for process-based models we are not aware of any modelling study systematically investigating the accuracy–complexity curve (but see the studies of Cox *et al.*, 2006; Crout *et al.*, 2009; and Martínez *et al.*, 2011, for systematic exploration of simplified versions of their process-based models).

11. Communicability/transparency of the model

Communication of a model requires: (1) a precise documentation of the steps/processes included, and (2) sufficient scientific background of both writer and reader to be able to judge their appropriateness. Model documentation is traditionally poor in process-based models and many efforts have been made to improve this situation (reviewed in Schmolke *et al.*, 2010). Also the reluctance of many ecological modellers to make their code publicly available contributes to low reproducibility of all but the simplest of models. Reasons for closed code include inelegant coding or insufficient documentation (Barnes, 2010) as well as the wish of the scientist to prevent others from using the model inappropriately or so as to diminish the modeller's own publishing prospects. For models of moderate complexity, re-implementation is actually a good way of testing the implementation, because potential errors are unlikely to be reproduced identically.

For correlative models the *de facto* standard statistical tool is R (R Development Core Team, 2010), and hence analyses can be transparently communicated through the exchange of software code (which is similarly true for other code-based software such as Python, Sage, Matlab or Mathematica). Software tools that are configured through a graphical user interface have the disadvantage that they often do not record the choices made by the user and hence require special care by

the user to record and communicate all the chosen settings. Unfortunately, this is rarely carried out and therefore efficient logging of the chosen options by the program should be the standard (as it is for example in MAXENT: Phillips & Dudík, 2008). Given the many choices available in correlative models, an analysis of the sensitivity of the results to alternative choices would be desirable, as stated above for process-based models.

12. Knowledge potentially gleaned from the model

Any useful model should be able to reproduce expected outcomes (see section 4), but it may also yield counterintuitive results, which are actually one of the most useful outcomes of modelling. They may identify new connections between processes and should generate new hypotheses that can be confirmed by experiment or other data. Surprising results obtained using a correlative model, such as an unexpected correlation between a species' distribution and a particular environmental variable, may in fact lead to the discovery of a new process, while surprises resulting from the use of process-based models usually relate to unexpected, emergent patterns as a result of nonlinear interactions between processes that are already in the model (for an example of this see Eisinger & Thulke, 2008).

We speculate that, in general, correlative models used in the exploratory sense are more likely to result in discoveries of new processes or process interactions than process-based models, where the processes and interactions have to be defined a priori. Formally comparing forward process-based models with data may detect process deficits, but will not necessarily identify the missing processes.

13. Common errors and misuses

The most common 'error' of any modeller is to 'believe' a model. Models are abstractions of reality and their correctness of abstraction has to be demonstrated (Krakauer *et al.*, 2011). No ecological model can be right a priori, because fundamental laws do not exist in ecology (Lawton, 1999), and, with respect to species distributions, there is no 'quantum biogeography'. From this first error in 'attitude' follow three common misuses. Firstly, either model type – correlative or process – is often stretched beyond (sometimes far beyond) the range of data underlying it. For example, constructing a model correlating the abundance of a freshwater fish species in temperate Europe with environmental data cannot be expected to 'work' when taking it to the Mediterranean. This is not because of the distance involved, but because the wet season is the cold season in the Mediterranean, while it is the warm season in Central Europe. We would extrapolate the parameter estimates to combinations of temperature and precipitation never encountered in the region for which the model was developed. Climate change predictions using correlative models often fall into this category (Ohlemüller *et al.*, 2006).

The second misuse is to employ the model for an application for which it was not developed, without due validation/justification. If an ecologist builds a model to understand home-range size of a passerine bird and includes landscape composition as a parameter, then this model does not automatically qualify as an assessment tool for landscape structure with respect to bird abundance. The reason is simply that for his or her initial purpose the ecologist may not have looked at abundance at all, instead inferring it as a by-product of home-range packing. But where in the model does it state that each of these 'virtual home ranges' must be occupied?

The final common misuse follows from overconfident communication of a model's predictions (even within the parameter range). A simple diagnostic is whether uncertainty was quantified or discussed: if it was not, the user/modeller is likely to be overconfident in the model's prediction. Not an error, but a missed opportunity for any model, is to omit to specify a set of predictions to sites with environmental conditions not encountered when assembling the model. Collecting data in exactly these conditions would then serve as a critical test.

CRITICAL ISSUES FOR SPECIES DISTRIBUTION MODELS

Critical issues for correlative models

Data! Everything depends on the amount, quality and appropriateness of data. Many statistical papers have developed fixes for biased sampling, missing values, unbalanced designs and so forth, but when the relevant ecological driver has not been quantified, no amount of data will be able to generate ecologically interesting hypotheses.

The causality of detected correlations is a critical issue for the use of correlative models, where the input variables are often correlated among themselves. For example, the observation that the occurrence of a species is correlated with mean annual temperature does not necessarily imply that temperature is itself a direct limiting factor, it could also be solar radiation, which is usually not represented in the data sets as accurately as temperature, or the presence of a competitor that is itself limited by temperature. If the temperature correlation is used to make a prediction of the species distribution under climate change, this could lead to incorrect results, as climate change causes temperatures to increase, but not solar radiation.

Problems can arise when extrapolating in space, as the correlations between input parameters may be different in different places and hence a non-causal correlation found in one place could lead to incorrect results when extrapolated to another place. An additional problem when extrapolating in time is that the increasing atmospheric CO_2 concentrations are likely to have an impact on plant species ranges due to the alleviation of water stress (Farquhar, 1997). This effect is unlikely to be detected in the present data because the spatial variability of CO_2 is negligible. For this reason, cross-validation using present data sets that were not used for the derivation of the model may shed some light on the

uncertainty of the model predictions when extrapolating in space, but it does not necessarily serve as an indication of the uncertainty when extrapolating in time, particularly under climate change.

The use of historical records for validation (e.g. pollen records in combination with historical climate, where available), however, could give some indication of the uncertainty when extrapolating in time, but this is rarely carried out (e.g. Pearman *et al.*, 2008). Also, with genetic information becoming increasingly available, genetic structure may reflect historical developments and hence provide additional opportunity for validation.

Critical issues for process-based models

Fitted process-based models rely on the implicit assumption that the model structure and process formulations are correct and that the unknown parameter values can be obtained by inverse modelling or available observations. Because the model parameters are fitted to reproduce observations, the same observations cannot be used to test for the correctness of the model structure and process formulations. The accuracy to which a tuned model reproduces the data is not an indication of correct process representations, as any data stream can be reproduced to an arbitrary level of accuracy using, for example, a polynomial function with sufficient degrees of freedom. On the other hand, even if a model was a true representation of the relevant processes, there is no guarantee that the correct parameter values can be obtained through inverse modelling, as the available data may not be sufficient to allow identification of an unambiguous set of parameters that best reproduces the data. In fact, many different parameter sets can reproduce the data equally well (see Equifinality). The different, apparently equally valid parameter sets can yield very different predictions when used under changed conditions (e.g. Schulz *et al.*, 2001).

Forward process-based models often rely on empirical parameterizations of the processes considered. This again introduces problems akin to those described above for correlative models, as the causality of observed correlations is not necessarily assured. If, for example, the observed correlation between mean daily temperatures and the onset of leafing or flowering was due to a cross-correlation between solar irradiance and temperatures, while solar radiation was the directly responsible variable, the model could lead to a wrong prediction under climate change, where temperatures increase but not solar radiation. Furthermore, neither empirical process-based nor correlative models would capture the adaptation of a species' phenology to changed climate.

CONCLUSIONS AND OUTLOOK

Our review of the similarities and differences between correlative and process-based species distribution models emphasizes that they sit on a continuum defined by the extent to which processes are explicitly represented. When these two broad types of models are fitted to observed data, there is considerable overlap in their assumptions, validation challenges and reproducibility problems. Although representing two very different conceptual starting points for species distribution modelling, they may well converge onto the same problems with respect to prediction of environmental and management change. Neither approach warrants the inference that reproduction of observations is indicative of the model being 'true' ('right for the wrong reason'; Judd, 2003). Both the causality of correlations found using a correlative model and the interplay of mechanisms proposed in a process-based model should be considered as hypotheses. However, in the former, the model itself and the data cannot be used to test the hypotheses, as they have already been used to generate the hypotheses. In a forward process model, on the other hand, mechanisms are proposed based on theoretical grounds or independent data, and hence, in theory, they can be tested using the match between model results and observations. However, in practice, most process-based models have a large number of adjustable parameters that need to be calibrated against observations. This precludes the use of the same data for hypothesis testing and reduces the use of the model to an extrapolation tool.

The future development of both correlative and process-based approaches is likely to see a mixing of their strengths (Mokany & Ferrier, 2010): data-driven implementation conveys trustworthiness because it is based on 'real data'; modelling of actual processes emanates scientific rigour and mechanistic understanding. The key test for either approach, however, is its usefulness for the question at hand. Can either of the two approaches identify previously unknown mechanisms thus generating knowledge? Are models accurately predicting suitable sites as confirmed by transplant experiments? Are uncertainties small enough to allow selection between different management scenarios?

We would like to highlight three avenues for research on species distributions, as follows.

1. Bayesian fitting of process-based models. To understand how certain the knowledge we put into the model actually is, we can fit the model to observed data sets ('model inversion'). Allowing the uncertainty of model parameters to enter the fitting process as priors, estimated distributions of model parameters are indicative of the statistical support of the data for this specific parameter. Note, however, that such Bayesian process modelling is still being developed and that we may 'fit models that are far beyond our ability to understand them' (Hodges, 2010, p. 3497). If a parameter's posterior distribution largely overlaps with 0, we would conclude that under this model there is no evidence for the process in this data set. For a given system, generic models could thus be tailored and simplified. Model inversion could hence be used in an inferential way.

2. 'Forward' process-based models. To avoid the need for parameter fitting, unknown parameters in process-based models can be determined using detailed observations or

experiments, such as in PHENOFIT (Chuine & Beaubien, 2001) or Niche Mapper (Kearney *et al.*, 2008). Alternatively, models can be formulated that simulate natural selection from a randomly generated pool of virtual species resembling the species of interest, akin to the JeDi model (Reu *et al.*, 2011). Forward process-based models avoid the problem of equifinality and can be used for hypothesis testing as they are much less likely to produce the right result for the wrong reasons. Forward process-based models, if based on first principles, may also lead to more reliable predictions of species distributions under environmental change, as their probability of matching species distributions under new conditions should be compared with their probability of matching them at present (first principles do not change).

3. Combined workflow. Although we juxtaposed correlative and process-based models, it may actually be fruitful to join them in a combined workflow (Mokany & Ferrier, 2010; Peng *et al.*, 2011). Scientific understanding of nature starts with *observations*, i.e. descriptive data. *Correlative models* efficiently sift through such data, thereby *generating hypotheses* on potentially underlying processes. These can then be taken up, along with ecological theory and experimental evidence, by *process-based models*, based on ecological theory and experimental evidence. Unknown parameters in process-based models could guide experimental and theoretical research to gather relevant knowledge for their quantification. The resulting process-based models can then generate predictions specifically designed for a *formal test on independent data*. In such a comprehensive approach, researchers with different interests, expertise and focus can synergistically progress the field in a way neither correlative nor process-based approaches can do by themselves.

In conclusion, we find no reason why a proponent of either of the two extremes of correlative and process-based species distribution modelling should hold the moral high ground. 'Correlationists' should be humble: their model's success may be due to spurious correlations. 'Mechanists' should be unassertive about their approach, because they will only find effects of processes that they included. Either approach must comply with nature, statistically or mechanistically, and be aware of the kinds of questions they are best suited to answer.

ACKNOWLEDGEMENTS

We are grateful to the following colleagues whose comments helped us to improve the clarity and focus of this publication: Rampal Etienne, Lee Hannah, Thomas Hickler, Steven I. Higgins, Bob O'Hara, Peter Linder, Greg McInerny, Frank Schurr, Ralf Seppelt and Konstans Wels, as well as Jens-Christian Svenning, Robert Whittaker and two anonymous referees. The work was initiated through a workshop 'The ecological niche as a window to biodiversity', organized by Christine Römermann, Bob O'Hara and Steven Higgins and funded by the LOEWE- BiK-F Biodiversity and the Climate Research Centre Frankfurt. Funding to C.F.D. by the Helmholtz Association (VH-NG 247) and the German Research Foundation DFG (DO 686/5-1), to S.J.S. by the Max Planck Society and to C.R. by the DFG (RO 3842/1-1) is gratefully acknowledged.

REFERENCES

Aitken, S.N., Yeaman, S., Holliday, J.A., Wang, T. & Curtis-McLane, S. (2008) Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, **1,** 95–111.

Angert, A.L. & Schemske, D.W. (2005) The evolution of species' distributions: reciprocal transplants across the elevation ranges of *Mimulus cardinalis* and *M. lewisii*. *Evolution*, **59,** 1671–1684.

Applequist, W.L., McGlinn, D.J., Miller, M., Long, Q.G. & Miller, J.S. (2007) How well do herbarium data predict the location of present populations? A test using *Echinacea* species in Missouri. *Biodiversity and Conservation*, **16,** 1397–1407.

Araújo, M.B. & Rahbek, C. (2006) How does climate change affect biodiversity? *Science*, **313,** 1396–1397.

Araújo, 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.

Arhonditsis, G.B., Qian, S.S., Stow, C.A., Lamon, E.C. & Reckhow, K.H. (2007) Eutrophication risk assessment using Bayesian calibration of process-based models: application to a mesotrophic lake. *Ecological Modelling*, **208,** 215–229.

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

Bahn, V. & McGill, B.J. (2007) Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, **16,** 733–742.

Barnes, N. (2010) Publish your computer code: it is good enough. *Nature*, **467,** 753.

Barry, S. & Elith, J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43,** 413–423.

Beale, C.M., Lennon, J.J. & Gimona, A. (2008) Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences USA*, **105,** 14908–14912.

Beven, K.J. & Freer, J. (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, **249,** 11–29.

Bierman, S.M., Butler, A., Marion, G. & Kühn, I. (2010) Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. *Ecography*, **33,** 451–460.

Bolker, B.M., Pacala, S.W. & Parton, W.J. (1998) Linear analysis of soil decomposition: insights from the Century model. *Ecological Applications*, **8,** 425–439.

Bonn, A. & Schröder, B. (2001) Habitat models and their transfer for single and multi species groups: a case study of carabids in an alluvial forest. *Ecography*, **24,** 483–496.

Broennimann, O., Treier, U.A., Müller-Schärer, H., Thuiller, W., Peterson, A.T. & Guisan, A. (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters*, **10,** 701–709.

Buckley, L.B. (2008) Linking traits to energetics and population dynamics to predict lizard ranges in changing environments. *The American Naturalist*, **171,** E1–E19.

Bugmann, H.K.M. (1996) A simplified forest model to study species composition along climate gradient. *Ecology*, **77,** 2055–2074.

Bugmann, H.K.M. & Solomon, A.M. (1995) The use of a European forest model in North America: a study of ecosystem response to climate gradients. *Journal of Biogeography*, **22,** 477–484.

Bugmann, H.K.M. & Solomon, A.M. (2000) Explaining forest composition and biomass across multiple biogeographical regions. *Ecological Applications*, **10,** 95–114.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretical approach*. Springer, Berlin.

Cabral, J.S. & Schurr, F.M. (2010) Estimating demographic models for the range dynamics of plant species. *Global Ecology and Biogeography*, **19,** 85–97.

Cabral, J.S., Bond, W.J., Midgley, G.F., Rebelo, A.G., Thuiller, W. & Schurr, F.M. (2011) Effects of harvesting flowers from shrubs on the persistence and abundance of wild shrub populations at multiple spatial extents. *Conservation Biology*, **25,** 73–84.

Chevin, L.-M., Lande, R. & Mace, G.M. (2010) Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. *PLoS Biology*, **8,** e1000357.

Chuine, I. (2000) A unified model for budburst of trees. *Journal of Theoretical Biology*, **207,** 337–347.

Chuine, I. & Beaubien, E.G. (2001) Phenology is a major determinant of temperate tree range. *Ecology Letters*, **4,** 500–510.

Clark, J.S., Schlesinger, W.H., Pascual, M., Carpenter, S.R., Barber, M., Collins, S., Dobson, A., Foley, J.A., Lodge, D.M., Pielke, R., Pizer, W., Pringle, C., Reid, W.V., Rose, K.A., Sala, O., Wall, D.H. & Wear, D. (2001) Ecological forecasts: an emerging imperative. *Science*, **293,** 657–660.

Cox, G.M., Gibbons, J.M., Wood, A.T.A., Craigon, J., Ramsden, S.J. & Crout, N.M.J. (2006) Towards the systematic simplification of mechanistic models. *Ecological Modelling*, **198,** 240–246.

Crout, N., Tarsitano, D. & Wood, A. (2009) Is my model too complex? Evaluating model formulation using model reduction. *Environmental Modelling and Software*, **24,** 1–7.

Currie, D.J. (2007) Disentangling the roles of environment and space in ecology. *Journal of Biogeography*, **34,** 2009–2011.

Davis, A.J., Jenkinson, L.S., Lawton, J.H., Shorrocks, B. & Wood, S. (1998) Making mistakes when predicting shifts in species ranges in response to global warming. *Nature*, **391,** 783–786.

De Marco, P., Jr, Diniz-Filho, J.A.F. & Bini, L.M. (2008) Spatial analysis improves species distribution modelling during range expansion. *Global Change Biology*, **4,** 577–580.

Dennis, R.L.H., Sparks, T.H. & Hardy, P.B. (1999) Bias in butterfly distribution maps: the effects of sampling effort. *Journal of Insect Conservation*, **3,** 33–42.

Doak, D.F. & Morris, W.F. (2010) Demographic compensation and tipping points in climate-induced range shifts. *Nature*, **467,** 959–962.

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

Dormann, C.F., Purschke, O., García Márquez, J.R., Lautenbach, S. & Schröder, B. (2008a) Components of uncertainty in species distribution analysis: a case study of the great grey shrike. *Ecology*, **89,** 3371–3386.

Dormann, C.F., Schweiger, O., Arens, P. *et al.* (2008b) Prediction uncertainty of environmental change effects on temperate European biodiversity. *Ecology Letters*, **11,** 235–244.

Draper, N.R. & Smith, H. (1998) *Applied regression analysis*. Wiley, New York.

Eisinger, D. & Thulke, H.H. (2008) Spatial pattern formation facilitates eradication of infectious diseases. *Journal of Applied Ecology*, **45,** 415–423.

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

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

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

Farquhar, G.D. (1997) Carbon dioxide and vegetation. *Science*, **278,** 1411.

Giesecke, T., Hickler, T., Kunkel, T., Sykes, M.T. & Bradshaw, R.H.W. (2006) Towards an understanding of the Holocene distribution of *Fagus sylvatica* L. *Journal of Biogeography*, **34,** 118–131.

Grimm, V. & Railsback, S.F. (2005) *Individual-based modeling and ecology*. Princeton University Press, Princeton, NJ.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8,** 993–1009.

Hastie, T., Tibshirani, R.J. & Friedman, J.H. (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer, Berlin.

Hein, S., Binzenhöfer, B., Poethke, H., Biedermann, R., Settele, J. & Schröder, B. (2007) The generality of habitat suitability models: a practical test with two insect groups. *Basic and Applied Ecology*, **8,** 310–320.

Heisey, D.M., Osnas, E.E., Cross, P.C., Joly, D.O., Langenberg, J.A. & Miller, M.W. (2010) Linking process to pattern: estimating spatiotemporal dynamics of a wildlife epidemic from cross-sectional data. *Ecological Monographs*, **80,** 221–240.

Hijmans, R.J. & Graham, C.H. (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, **12,** 2272–2281.

Hodges, J.S. (2010) Are exercises like this a good use of anybody's time? *Ecology*, **91,** 3496–3500.

Hothorn, T., Müller, J., Kneib, T., Schröder, B. & Brandl, R. (2011) Decomposing environmental, spatial and spatiotemporal components of species distributions. *Ecological Monographs*, **81,** 329–347.

Huntley, B. (1991) How plants respond to climate change: migration rates, individualism and the consequences for plant communities. *Annals of Botany*, **67**(Suppl. 1)**,** 15–22.

Jeltsch, F., Moloney, K.A., Schwager, M., Körner, K. & Blaum, N. (2011) Consequences of correlations between climatic and landscape changes for species survival. *Agriculture, Ecosystems and Environment*, **145,** 49–58.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19,** 101–108.

Judd, K. (2003) Bayesian reconstruction of chaotic times series: right results for the wrong reasons. *Physical Review E*, **67,** 026212.

Kearney, M., Phillips, B.L., Tracy, C.R., Christian, K.A., Betts, G. & Porter, W.P. (2008) Modelling species distributions without using species distributions: the cane toad in Australia under current and future climates. *Ecography*, **31,** 423–434.

Kearney, M., Shine, R. & Porter, W.P. (2009b) The potential for behavioral thermoregulation to buffer ''cold-blooded'' animals against climate warming. *Proceedings of the National Academy of Sciences USA*, **106,** 3835–3840.

Kearney, M.R. & Porter, W.P. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12,** 334–350.

Kearney, M.R., Porter, W.P., Williams, C., Ritchie, S. & Hoffmann, A.A. (2009a) Integrating biophysical models and evolutionary theory to predict climatic impacts on species' ranges: the dengue mosquito *Aedes aegypti* in Australia. *Functional Ecology*, **23,** 528–538.

Kearney, M.R., Wintle, B.A. & Porter, W.P. (2010) Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. *Conservation Letters*, **3,** 203–213.

Keenan, T., Maria Serra, J., Lloret, F., Ninyerola, M. & Sabate, S. (2011) Predicting the future of forests in the Mediterranean under climate change, with niche- and process-based models: CO_2 matters! *Global Change Biology*, **17,** 565–579.

Keith, D.A., Akcakaya, H.R., Thuiller, W., Midgley, G.F., Pearson, R.G., Phillips, S.J., Regan, H.M., Araújo, M.B., Rebelo, T.G. & Akçakaya, H.R. (2008) Predicting extinction risks under climate change: coupling stochastic population models with dynamic bioclimatic habitat models. *Biology Letters*, **4,** 560–563.

Kirchner, J.W. (2006) Getting the right answer for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, **42,** W03S04.

Kleidon, A. & Mooney, H.A. (2000) A global distribution of biodiversity inferred from climatic constraints: results from a process-based modelling study. *Global Change Biology*, **6,** 507–523.

Krakauer, D.C., Collins, J.P., Erwin, D., Flack, J.C., Fontana, W., Laubichler, M.D., Prohaska, S.J., West, G.B. & Stadler, P.F. (2011) The challenges and scope of theoretical biology. *Journal of Theoretical Biology*, **276,** 269–276.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S. & Prentice, I.C. (2005) A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, **19,** GB1050.

LaDeau, S.L. (2010) Advances in modeling highlight a tension between analytical accuracy and accessibility. *Ecology*, **91,** 3488–3492.

Latimer, A.M., Wu, S., Gelfand, A.E. & Silander, J.A., Jr (2006) Building statistical models to analyze species distributions. *Ecological Applications*, **16,** 33–50.

Lawton, J.H. (1999) Are there general laws in ecology? *Oikos*, **84,** 177–192.

Lebreton, J.-D., Burnham, K.P., Clobert, J. & Anderson, D.R. (1992) Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, **62,** 67–118.

Lele, S.R. (2010) Model complexity and information in the data: could it be a house built on sand? *Ecology*, **91,** 3493–3496.

Lele, S.R., Nadeem, K. & Schmuland, B. (2010) Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, **105,** 1617–1625.

Link, W.A. & Barker, R.J. (2006) Model weights and the foundations of multimodel inference. *Ecology*, **87,** 2626–2635.

Loehle, C. & Leblanc, D. (1996) Model-based assessments of climate change effects on forests: a critical review. *Ecological Modelling*, **90,** 1–31.

Luo, Y., Weng, E., Wu, X., Gao, C., Zhou, X. & Zhang, L. (2009) Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. *Ecological Applications*, **19,** 571–574.

Martínez, I., Wiegand, T., Camarero, J.J., Batllori, E. & Gutiérrez, E. (2011) Disentangling the formation of contrasting tree-line physiognomies combining model selection and Bayesian parameterization for simulation models. *The American Naturalist*, **177,** E136–E152.

McInerny, G.J. & Purves, D.W. (2011) Fine-scale environmental variation in species distribution modelling: regres-

sion dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2,** 248–257.

Moilanen, A. (2000) The equilibrium assumption in estimating the parameters of metapopulation models. *Journal of Animal Ecology*, **69,** 143–153.

Mokany, K. & Ferrier, S. (2010) Predicting impacts of climate change on biodiversity: a role for semi-mechanistic community-level modelling. *Diversity and Distributions*, **17,** 374–380.

Morin, X. & Thuiller, W. (2009) Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change. *Ecology*, **90,** 1301–1313.

Morin, X., Augspurger, C. & Chuine, I. (2007) Process-based modeling of species' distributions: what limits temperate tree species' range boundaries? *Ecology*, **88,** 2280–2291.

O'Hara, R.B. & Sillanpää, M.J. (2009) A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, **4,** 85–118.

Ohlemüller, R., Gritti, E.S., Sykes, M.T. & Thomas, C.D. (2006) Towards European climate risk surfaces: the extent and distribution of analogous and non-analogous climates 1931–2100. *Global Ecology and Biogeography*, **15,** 395–405.

Oreskes, N., Shrader-Frechette, K. & Belitz, K. (1994) Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, **263,** 641–646.

Pagel, J. & Schurr, F. (2011) Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, doi: 10.1111/j.1466-8238.2011.00663.x.

Pagel, J., Fritzsch, K., Biedermann, R. & Schröder, B. (2008) Annual plants under cyclic disturbance regimes – better understanding through model aggregation. *Ecological Applications*, **18,** 2000–2015.

Pautasso, M. & McKinney, M.L. (2007) The botanist effect revisited: plant species richness, county area, and human population size in the United States. *Conservation Biology*, **21,** 1333–1340.

Pearman, P.B., Guisan, A., Broennimann, O. & Randin, C.F. (2007) Niche dynamics in space in time. *Trends in Ecology and Evolution*, **23,** 149–158.

Pearman, P.B., Randin, C.F., Broennimann, O., Vittoz, P., van der Knaap, W.O., Engler, R., Le Lay, G., Zimmermann, N.E. & Guisan, A. (2008) Prediction of plant species distributions across six millennia. *Ecology Letters*, **11,** 357–369.

Peng, C., Guiot, J., Wu, H., Jiang, H. & Luo, Y. (2011) Integrating models with data in ecology and palaeoecology: advances towards a model-data fusion approach. *Ecology Letters*, **14,** 522–536.

Penteriani, V. (2008) When similar ecological patterns in time emerge from different initial conditions: equifinality in the breeding performance of animal populations. *Ecological Complexity*, **5,** 66–68.

Perry, G. & Millington, J. (2008) Spatial modelling of succession-disturbance dynamics in forest ecosystems: concepts and examples. *Perspectives in Plant Ecology, Evolution and Systematics*, **9,** 191–210.

Peterson, A.T., Soberón, J. & Sánchez-Cordero, V. (1999) Conservatism of ecological niches in evolutionary time. *Science*, **285,** 1265–1267.

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

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (2007) *Numerical recipes: the art of scientific computing*. Cambridge University Press, Cambridge.

R Development Core Team (2010) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org.

Randin, C.F., Dirnbock, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33,** 1689–1703.

Refsgaard, J.C., van der Sluijs, J.P., Hojberg, A.L. & Vanrolleghem, P.A. (2007) Uncertainty in the environmental modelling process – a framework and guidance. *Environmental Modelling and Software*, **22,** 1543–1556.

Reineking, B. & Schröder, B. (2006) Constrain to perform: regularization of habitat models. *Ecological Modelling*, **193,** 675–690.

Reu, B., Proulx, R., Bohn, K., Dyke, J.G., Kleidon, A., Pavlick, R. & Schmidtlein, S. (2011) The role of climate and plant functional trade-offs in shaping global biome and biodiversity patterns. *Global Ecology and Biogeography*, **20,** 570–581.

Rickebusch, S., Thuiller, W., Hiernaux, P., Araújo, M.B., Sykes, M.T., Schweiger, O. & Lafourcade, B. (2008) Incorporating the effects of changes in vegetation functioning and CO_2 on water availability in plant habitat models. *Biology Letters*, **4,** 556–559.

Royle, J.A., Nichols, J.D. & Kery, M. (2005) Modelling occurrence and abundance of species when detection is imperfect. *Oikos*, **110,** 353–359.

Schmolke, A., Thorbek, P., DeAngelis, D.L. & Grimm, V. (2010) Ecological models supporting environmental decision making: a strategy for the future. *Trends in Ecology and Evolution*, **25,** 479–486.

Schröder, B. & Richter, O. (1999) Are habitat models transferable in space and time? *Zeitschrift für Ökologie und Naturschutz*, **8,** 195–204.

Schulz, K., Jarvis, A., Beven, K. & Soegaard, H. (2001) The predictive uncertainty of land surface fluxes in response to increasing ambient carbon dioxide. *Journal of Climate*, **14,** 2551–2562.

Schumacher, S. & Bugmann, H.K.M. (2006) The relative importance of climatic effects, wildfires and management for future forest landscape dynamics in the Swiss Alps. *Global Change Biology*, **12,** 1435–1450.

Sitch, S., Smith, B., Prentice, I.C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J.O., Levis, S., Lucht, W., Sykes, M.T.,

Thonicke, K. & Venevsky, S. (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, **9,** 161–185.

Smith, B., Prentice, I.C. & Sykes, M.T. (2001) Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. *Global Ecology and Biogeography*, **10,** 621–637.

Smith, G.C., Parrott, D. & Robertson, A. (2008) Managing wildlife populations with uncertainty: cormorants *Phalacrocorax carbo*. *Journal of Applied Ecology*, **45,** 1675–1682.

Starfield, A.M., Smith, K.A. & Bleloch, A.L. (1990) *How to model it*. McGraw-Hill, New York.

Sutherst, R.W. & Maywald, G.F. (1985) A computerised system for matching climates in ecology. *Agriculture, Ecosystems and Environment*, **13,** 281–299.

Thuiller, W., Midgley, G.F., Hughes, G.O. & Rebelo, A.G. (2006) Migration rate limitations on climate change-induced range shifts in Cape Proteaceae. *Diversity and Distributions*, **12,** 555–562.

Trivedi, M.R., Berry, P.M., Morecroft, M.D. & Dawson, T.P. (2008) Spatial scale affects bioclimate model projections of climate change impacts on mountain plants. *Global Change Biology*, **14,** 1089–1103.

Van Oijen, M., Rougier, J. & Smith, R. (2005) Bayesian calibration of process-based forest models: bridging the gap between models and data. *Tree Physiology*, **25,** 915–927.

Vaughan, I.P. & Ormerod, S.J. (2003) Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, **17,** 1601–1611.

Wasserberg, G., Osnas, E.E., Rolley, R.E. & Samuel, M.D. (2009) Host culling as an adaptive management tool for chronic wasting disease in white-tailed deer: a modelling study. *Journal of Applied Ecology*, **46,** 457–466.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B. & Grimm, V. (2010) The virtual ecologist approach: simulating data and observers. *Oikos*, **119,** 622–635.

SUPPORTING INFORMATION

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

Appendix S1 Bibliography of process-based, hybrid and correlative species distribution models.

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

BIOSKETCHES

Carsten Dormann is a statistical ecologist with an interest in extending correlative approaches into process-based modelling. His fields of research comprise non-predictive areas of species distribution modelling, experimental community ecology and ecological networks.

Stan Schymanski is an ecohydrological process-based modeller. He seeks common thermodynamic principles behind the organization and growth of vegetation. Together with their co-authors they represent a diverse background and attitude towards the correlative–process continuum of species distribution models.

Author contributions: C.F.D. and S.J.S. led the discussion and wrote the first draft. All authors structured the study and co-wrote the final manuscript.

Editor: Jens-Christian Svenning

The papers in this Special Issue arose from two workshops entitled 'The ecological niche as a window to biodiversity' held on 26–30 July 2010 and 24–27 January 2011 in Arnoldshain near Frankfurt, Germany. The workshops combined recent advances in our empirical and theoretical understanding of the niche with advances in statistical modelling, with the aim of developing a more mechanistic theory of the niche. Funding for the workshops was provided by the Biodiversity and Climate Research Centre (BiK-F), which is part of the LOEWE programme 'Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz' of Hesse's Ministry of Higher Education, Research and the Arts.