Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2016) **25**, 238–249

A Journal of Macroecology

Global Ecology and Biogeography

MACROECOLOGICAL METHODS

Cross-scale integration of knowledge for predicting species ranges: a metamodelling framework

Matthew V. Talluto[1,2,3,4*], Isabelle Boulangeat[1,2], Aitor Ameztegui[5], Isabelle Aubin[6], Dominique Berteaux[1,2,7], Alyssa Butler[1,2], Frédérik Doyon[8,9], C. Ronnie Drever[10], Marie-Josée Fortin[11], Tony Franceschini[1], Jean Liénard[12], Dan McKenney[6], Kevin A. Solarik[2,5], Nikolay Strigul[12], Wilfried Thuiller[3,4] and Dominique Gravel[1,2]

[1]*Biogeography and Metacommunity Ecology Lab, Département de biologie, Université du Québec à Rimouski, Rimouski, Quebec, Canada,* [2]*Quebec Centre for Biodiversity Science, Montreal, Quebec, Canada,* [3]*Université Grenoble Alpes, Laboratoire d'Ecologie Alpine (LECA), F-38000 Grenoble, France,* [4]*CNRS, Laboratoire d'Ecologie Alpine (LECA), F-38000 Grenoble, France,* [5]*Centre d'Étude de la Forêt, Département des Sciences Biologiques, Université du Québec à Montréal, Montreal, Quebec, Canada,* [6]*Great Lakes Forestry Centre, Canadian Forest Service, Natural Resources Canada, Sault Ste Marie, Ontario, Canada,* [7]*Centre for Northern Studies, Université du Québec à Rimouski, Rimouski, Quebec, Canada,* [8]*Université du Québec en Outaouais, Gatineau, Quebec, Canada,* [9]*Institut des Sciences de la Forêt Tempérée (ISFORT), Ripon, Quebec, Canada,* [10]*The Nature Conservancy Canada, Ottawa, Ontario, Canada,* [11]*Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada,* [12]*Department of Mathematics, Washington State University, Vancouver, Washington, USA*

*Correspondence: Matthew Talluto, LECA, BP 53, 2233 Rue De La Piscine 38041 Grenoble Cedex 9, France.
E-mail: mtalluto@gmail.com

ABSTRACT

Aim Current interest in forecasting changes to species ranges has resulted in a multitude of approaches to species distribution models (SDMs). However, most approaches include only a small subset of the available information, and many ignore smaller-scale processes such as growth, fecundity and dispersal. Furthermore, different approaches often produce divergent predictions with no simple method to reconcile them. Here, we present a flexible framework for integrating models at multiple scales using hierarchical Bayesian methods.

Location Eastern North America (as an example).

Methods Our framework builds a metamodel that is constrained by the results of multiple sub-models and provides probabilistic estimates of species presence. We applied our approach to a simulated dataset to demonstrate the integration of a correlative SDM with a theoretical model. In a second example, we built an integrated model combining the results of a physiological model with presence–absence data for sugar maple (*Acer saccharum*), an abundant tree native to eastern North America.

Results For both examples, the integrated models successfully included information from all data sources and substantially improved the characterization of uncertainty. For the second example, the integrated model outperformed the source models with respect to uncertainty when modelling the present range of the species. When projecting into the future, the model provided a consensus view of two models that differed substantially in their predictions. Uncertainty was reduced where the models agreed and was greater where they diverged, providing a more realistic view of the state of knowledge than either source model.

Main conclusions We conclude by discussing the potential applications of our method and its accessibility to applied ecologists. In ideal cases, our framework can be easily implemented using off-the-shelf software. The framework has wide potential for use in species distribution modelling and can drive better integration of multi-source and multi-scale data into ecological decision-making.

Keywords
Climate change, decision making, patterns and processes, range dynamics, scaling, spatial ecology, species distribution modelling, uncertainty.

DOI: 10.1111/geb.12395

INTRODUCTION

Models of species range limits have wide applications, particularly in conservation biology where they can be used as decision-support tools in biodiversity management (Guisan *et al.*, 2013). Due to large temporal and spatial scales as well as the complex and nonlinear nature of ecosystem dynamics, it is often impossible to construct experiments that adequately explore the processes generating species range limits (Wu & Loucks, 1995; Levin, 1998). Hence, range models are essential tools that have been applied to a large number of ecological subfields, including biogeography (Schurr *et al.*, 2012), invasion biology (Catterall *et al.*, 2012; Gallien *et al.*, 2012), hybrid zone dynamics (Engler *et al.*, 2013) and the impacts of climate change on species distributions (Blois *et al.*, 2013; Thuiller *et al.*, 2014b).

Despite the recognized potential of these models, it can be difficult to produce species distribution models (SDMs) with acceptable levels of precision and bias (Guisan *et al.*, 2013). For mechanistic models, two important constraints can be problematic: (1) having the appropriate ecological theory needed to link data to modelling objectives, and (2) having sufficient data over a range of conditions to maintain coherence between the spatial and temporal scales of data and theory. In recent decades, however, modelling techniques have proliferated to take advantage of the increased number of datasets available. A growing body of theory, reflecting the diversity of processes generating species ranges, has also contributed to model diversification (Boulangeat *et al.*, 2012). Of these model types, fine-scale mechanistic models often capture important ecological processes quite well, but may perform poorly when applied at the scale of species ranges. For instance, biotic interactions are usually not modelled mechanistically at regional or continental scales because they are poorly known or unrecorded, despite being considered a key determinant of range limits (Pigot & Tobias, 2013). In contrast to mechanistic models, more correlative approaches that statistically relate species occurrences to other variables have the advantage of indirectly accounting for underlying processes (Guisan & Zimmermann, 2000). However, their predictions rely on the stationarity of the relationships between occurrences and explanatory variables in time and space, implying that the selected variables are related to the processes limiting species ranges and that their correlations are constant for calibration and projection ranges (Dormann, 2007). Extrapolating beyond the scope of the original data (e.g. predicting ranges based on future climate) is therefore problematic, because nonlinear responses to novel combinations of explanatory variables cannot be accommodated in models that do not simulate the underlying processes.

Clearly, an approach is needed to unify the strength of different modelling approaches that can also incorporate multiple data sources. To this end, we present an application of hierarchical Bayesian methods that uses outputs from multiple models to inform the results of the final model. Techniques for multimodel inference have proliferated in recent years. For example, hybrid models that allow for combinations between mechanistic and phenomenological sub-models are commonly employed in SDMs (Gallien *et al.*, 2010; Thuiller *et al.*, 2013; Boulangeat *et al.*, 2014). Within the hybrid framework, a correlative model might be used to account for abiotic variables that limit species distributions (Guisan & Thuiller, 2005), while a more mechanistic approach could include biotic interactions and space–time dynamics (Smolik *et al.*, 2010). However, the link between different sub-models is based on assumptions about the scaling of ecological processes that are poorly known and difficult to test (Gallien *et al.*, 2010) and, as such, uncertainties are approximated and difficult to attribute to different sources. An alternative to hybrid models is the direct combination of predictions, allowing models operating at the same spatio-temporal scales to be combined (e.g. model averaging, ensemble forecasting; Araújo & New, 2007). However, because uncertainty is approximated and may be poorly understood, it is not possible to evaluate the effects of convergent predictions on the total uncertainty of the outcomes, despite its potential importance in a prediction context.

We propose an alternative to these approaches using a hierarchical Bayesian framework. This approach provides a number of advantages, including: (1) the ability to incorporate multiple modes of inference (e.g. mechanistic, correlative models) (Van Oijen *et al.*, 2005; Clark & Gelfand, 2006; Hobbs & Ogle, 2011; Hartig *et al.*, 2012), (2) an easy mechanism to include multiple data sources at various scales (Levin, 1992; Peters *et al.*, 2004), and (3) an intuitive and comprehensive reporting of uncertainty in model predictions that reflects variation at all levels of organization (Cressie *et al.*, 2009; Hobbs & Ogle, 2011). Unlike hybrid methods, the aim is not to link different sub-models into a single model, but to condition the predictions of a metamodel at the target scale (e.g. an entire species' range) with information from independent sub-models at a variety of spatial scales, allowing for more flexibility regarding the type of information included. By integrating all available knowledge into a single prediction, our approach potentially mitigates the limitations inherent in each individual model, contributing to more robust predictions (Guisan & Thuiller, 2005; Araújo & Guisan, 2006). Moreover, a more comprehensive understanding of uncertainty can guide biodiversity management and prioritize future data collection by identifying parameters that make the greatest contribution to variance in the model predictions (McMahon *et al.*, 2011).

We illustrate our framework with two examples. We begin with a hypothetical example using simulated data, where we define the framework and demonstrate its application to multiple sources of information from different scales. In a second example, we apply the framework to combine presence–absence information with phenological data to improve uncertainty estimates and reduce bias when predicting changes to the range of sugar maple (*Acer saccharum* Marsh.), a widespread and dominant tree species from eastern North America, in response to climate change. We provide a more formal mathematical presentation in Appendix S1 in the Supporting Information, along with complete code and data for both examples in Appendix S2.

Example 1: adding experimental evidence for the fundamental niche to an SDM

The key idea of our approach is to formulate a metamodel that integrates data at the same ecological scale as the desired predictions, and to constrain the parameters of this model using the output of one or more sub-models. In this hypothetical example, we build a metamodel relating the distribution of an annual plant to coarse-scale climate with complementary information originating from a fine-scale experiment manipulating the precipitation regime. The metamodel attempts to capture the realized distribution of a species; as a correlative model, it implicitly captures the major physiological constraints and ecological processes constraining the distribution of the target species. However, for the purposes of forecasting, we would like to disentangle the fundamental response of a species to environmental variation from other processes in order to map the climatic envelope of where a species may be found in a natural setting. Thus, we use additional information on the physiological constraints affecting species distribution. Because these data are often collected at a finer scale than that the rangewide occurrence data, we apply a simple scaling function, drawing on ecological theory, to compute the likelihood of a set of metamodel parameters given both the occurrence and the physiological data. (See Appendix S1 for procedural details and scripts for executing the model.)

We consider data collected from a species' historical distribution, where the goal is to predict the distribution following a substantial reduction in precipitation. For the metamodel-scale data, we simulate a relatively high-quality presence–absence dataset covering a variety of ecological conditions, which we term X_M, where the subscript M indicates that the data were collected at the same scale as the metamodel (Fig. 1). We desire to model the species' distribution as a function of temperature, T_M, and precipitation, P_M. An initial version of the metamodel (θ_M) that has no constraints from other datasets will be referred to as the naive model. This naive model uses a simple logistic regression to estimate the probability of occurrence (ψ_N) as a function of temperature and precipitation:

$$\begin{aligned}\psi_N &= f(\theta_M, T_M, P_M)\\ &= p(X_M = 1|\theta_M, T_M, P_M)\\ &= \text{logit}^{-1}(\theta_M D_M)\end{aligned} \quad (1)$$

where θ_M is the parameter vector of the model, D_M is the covariate matrix (i.e. T_M, P_M), and logit^{-1} is the inverse of the logit function. We estimate parameters using a Metropolis-Hastings algorithm within a Markov chain Monte Carlo (MCMC) scheme using the proportional form of the Bayes theorem:

$$p(\theta_M|X_M, T_M, P_M) \propto p(X_M|\theta_M, T_M, P_M)\, p(\theta_M) \quad (2)$$

where $p(X_M \mid \theta_M, T_M, P_M)$ is often referred to as the *likelihood* of the data (X_M) given the model (θ_M), $p(\theta_M)$ is often referred to as the *prior distribution* of θ_M, and the goal of modelling is to estimate $p(\theta_M \mid X_M, T_M, P_M)$, the *posterior distribution* of θ_M, which gives the probability that θ_M takes particular values, given the observed data.

Thus far, we have considered only a single source of information to fit this model, and therefore the prior distribution $p(\theta_M)$ from equation 2 is uninformative. As a secondary source of information, we will consider an experiment relating the population growth rate of the plant to manipulations to the precipitation regime, with results (but no raw data) available from the literature (Fig. 1b). Furthermore, no information is available regarding the temperature regime for the experiment. Transplant experiments that evaluate performance beyond the range of a species are common and represent a plausible scenario for model integration (Hargreaves *et al.*, 2014). According to niche theory (Holt, 2009), the fundamental niche corresponds to the set of environmental conditions where the per capita intrinsic growth rate r is positive. This concept gives us a reasonable model to fit a scaling function for our sub-model (Appendix S1). If we hypothesize that the errors from Fig. 1b are normally

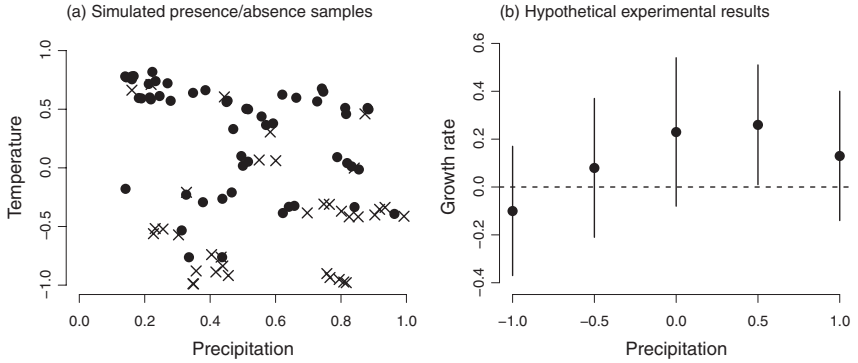

Figure 1 Two simulated datasets used to illustrate the model integration framework. (a) Presences (circles) or absences (crosses) of the species in ecological space, where the range of precipitation values sampled was 0.1–1. (b) Growth rate (r) as a function of manipulations to the precipitation regime (whiskers show ± 1 SE), with a larger range for precipitation (i.e. –1.0 to 1.0). The dashed line shows the threshold above which the species net growth rate is positive (implying presence). Axis scales for temperature and precipitation are arbitrary, but note the different scales on the horizontal axes.

with standard deviation σ_s (where the subscript S indicates information pertaining to the sub-model), then for an observation i we can interpret the probability of presence ($\psi_{S,i}$) as the probability that the observed growth rate $X_{S,i}$ is positive:

$$\psi_{S,i} = \int_0^{\infty} N(X_{S,i}, \sigma_{S,i}) \tag{3}$$

where N is the normal density function. We can then estimate the posterior distribution for the sub-model by fitting the relationship between ψ_S and precipitation (P_S) using Bayesian beta regression (Ferrari & Cribari-Neto, 2004):

$$p(\theta_S|\psi_S, P_S) \propto p(\psi_S|\theta_S, P_S)\, p(\theta_S). \tag{4}$$

Although the two datasets were collected at considerably different scales, we have sub-model predictions arising from a fine-scale experiment that are relevant at the scale of the metamodel (i.e. the probability of presence at a given precipitation regime). The scaling treats the fundamental niche as the only driver of species distribution, and only considers a single dimension of the niche. As such, it would be unwise to expect predictions from this model alone to resemble the actual distribution of the species; as a mechanistic model, it is simply too incomplete to predict distribution. However, the information from this sub-model, when applied as a constraint on the metamodel, can result in improved predictions that incorporate the information within each model.

We accomplish model integration by treating ψ_S, the posterior predictions of the sub-model θ_S, as prior information about some of the parameters of θ_M (i.e. parameters related to precipitation), expanding equation 2 to incorporate the new information from the sub-model:

$$\overbrace{p(\theta_M|X_M, T_M, P_M, \theta_S, \psi_S)}^{\text{integrated posterior}} \propto \overbrace{p(\psi_S|\theta_M, P_M)}^{\substack{\text{new information}\\ \text{from sub-model}}} \overbrace{p(X_M|\theta_M, T_M, P_M)\, p(\theta_M)}^{\substack{\text{naive metamodel}\\ \text{posterior}}} \overbrace{p(\theta_S)}^{\substack{\text{prior for}\\ \text{sub-model}}}. \tag{5}$$

As before, the metamodel θ_M can be used to predict probability of occurrence (ψ_I; where the subscript I refers to the integrated model). However, these predictions now reflect the presence–absence data X_M as well as the information from θ_S, including all of the data sources used to produce this sub-model. Finally, we note the presence of marginal distributions for both models [i.e. $p(\theta_M)$ and $p(\theta_S)$]. These can be informative (e.g. incorporating further prior information or the predictions of additional sub-models), semi-informative (e.g. to provide greater weight to more informative models) or uninformative. For the purposes of this example, we applied prior weights of 1 and 0.05 to the correlative and mechanistic models, respectively, reflecting the increased generality and much larger sample size of the correlative data. This procedure has the effect of increasing the variance of the model and prevents biasing the parameter estimation in favour of the mechanistic model.

When comparing the three models (naive metamodel, mechanistic sub-model and integrated metamodel) we observed extreme uncertainty in the first model when projecting beyond the range of the original data (Figs 2a & 3a, b). The sub-model was highly precise with respect to precipitation, thus providing a fairly strong constraint when producing the integrated model (Fig. 2b). The result was an integrated prediction that reflected the shapes of both models and showed considerably reduced uncertainty (Fig. 2). At the scale of the metamodel, considering both temperature and precipitation, we observed similar results, with reduced uncertainty in the predictions over the domain not covered by the presence–absence data (Fig. 3).

Example 2: constraining a SDM using phenological information

For the second example, we consider the problem of forecasting a species' potential distribution following climate change. There is considerable interest in comparing correlative and mechanistic projections with respect to climate change (Morin & Thuiller, 2009), and the correct characterization of uncertainty is a critical aspect of this problem (Cheaib *et al.*, 2012). Despite being a relatively common application of SDMs (Guisan & Thuiller, 2005), projecting models parameterized with modern climate data to future climate scenarios remains problematic (Araújo &

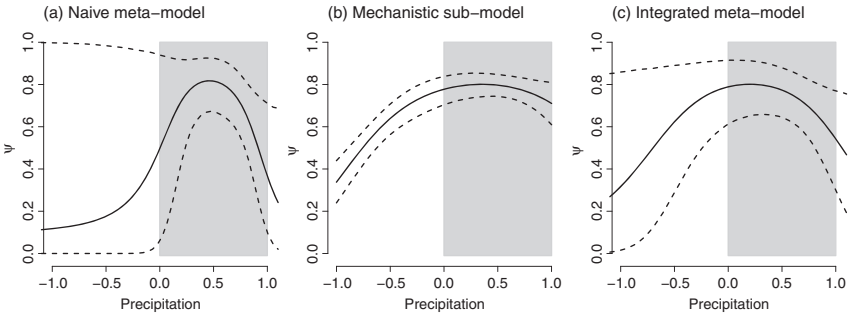

Figure 2 Comparison of the naive model, mechanistic sub-model and the integrated model showing the probability of presence (ψ) as a function of precipitation. Uncertainty is represented as dashed lines, showing the limits of 90% Bayesian credible intervals. The shaded region shows the calibration range for the naive model. (a) Naive model, using only presence–absence data. Uncertainty increases dramatically when attempting to project beyond the scope of the source data. (b) Mechanistic model, using observations of an experiment to infer probability of presence. (c) Integrated model, showing predictions that are intermediate between the two sub-models and uncertainty that is reduced compared with (a).

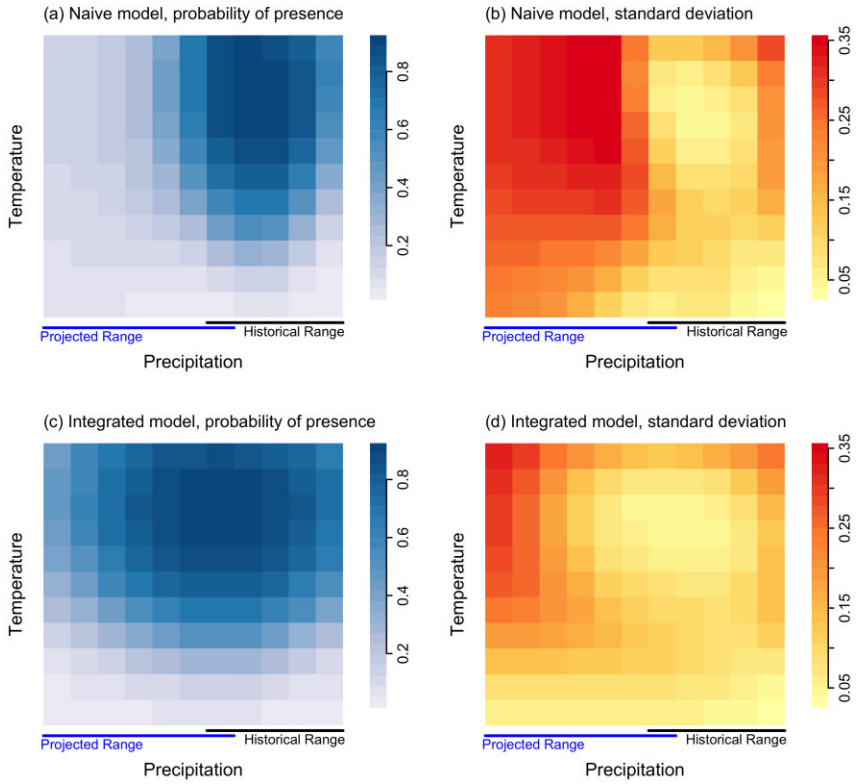

Figure 3 Maps showing the predicted probability of presence (ψ; a and c) and the standard deviation of ψ (b and d) for the naive and integrated models. Historical (i.e. where presence–absence samples were available) and predicted future precipitation regimes are shown below the horizontal axes.

Guisan, 2006). We used our framework to constrain a climate-based SDM with information obtained from Phenofit, a mechanistic model that predicts a species' probability of presence as a function of the suitability of the environment given the species' phenology (Chuine & Beaubien, 2001; Morin & Thuiller, 2009). Here we describe briefly the dataset and methods and the results of the analysis (see Appendix S1 for details of implementation and Appendix S2 for data and scripts to reproduce the analysis).

We obtained climate variables, occurrence data and Phenofit projections at 0.5° resolution for both the present and for 2100 for sugar maple, an economically and ecologically important species occurring in eastern North America (Morin & Thuiller, 2009). These data defined the metamodel scale. To these data, we added 4903 recorded presences and 21,701 absences derived from permanent forest sample plots located in the United States and Canada (see Appendix S1 for a map of plot locations). We reserved a third of this dataset for evaluation and used the remaining records to calibrate the models. We constructed the naive model by using a binomial generalized linear model (GLM) to relate the presence–absence dataset to three climate variables: the number of degree days (ddeg), mean annual precipitation (an_prcp) and the ratio of annual precipitation to potential evapotranspiration (pToPET). These variables were selected from an initial set of seven variables (see Appendix S1 and Morin & Thuiller, 2009, for details on the climate variables). We selected a GLM for its simplicity and interpretability because our focus was on demonstrating the framework, but more complex methods (e.g. generalized additive models) are compatible with the framework.

To perform the integration, we constrained the estimates of the naive model with the additional information from Phenofit while considering two different modelling objectives. The first is improving our model of the present range of the species. The use of both datasets to develop a range model for the species has a number of advantages. Assuming we have chosen climate variables that well represent the constraints on the species, including Phenofit in our model is likely to reduce bias in our estimate of the fundamental niche. The posterior predictions of the model (that is, the probability of presence in geographic space) will incorporate uncertainty from all sources. This can provide a much more accurate estimate of the uncertainty of our predictions. Thus, for our first integration we combined the naive model with the Phenofit predictions for the present; we refer to this model as 'Integrated-Present'. We also evaluated this model by constructing calibration curves and by computing the area under the receiver operating curve (AUC; see Appendix S1), which evaluates classification ability with 1 indicating perfect classification and 0.5 indicating no difference from a random model (Swets, 1988). The second modelling goal is to project changes to the range of sugar maple following climate change. Process-based and correlative models often differ substantially when projecting beyond the range of the original data (see Example 1). Thus, we used the predictions for 2100 from Phenofit (Morin & Thuiller, 2009) to condition the metamodel given the naive model predictions under future climate.

This procedure provided a consensus view of the future range of the species; we refer to this model as 'Integrated-Future'. In both cases, it was necessary to scale the Phenofit predictions, which were probabilistic, to make them compatible with the naive model, which was fitted using occurrence data. We used a latent variable approach, which posits an unobserved 'true' presence–absence dataset from which the Phenofit probabilities are derived. Similar to any other unknown parameter, we can generate a posterior distribution of this dataset by drawing samples during the MCMC procedure; thus, at each iteration, a dataset similar to the one used for the naive model was generated using the Phenofit probabilities (see Appendix S1 for a full statistical presentation of the model). This procedure expresses the information from Phenofit in a way that is compatible with the naive model, and also propagates the uncertainty in the Phenofit predictions. Additionally, it becomes possible to address future distribution via simulated future occurrences (which, by nature, are unobservable).

Model integration resulted in substantial reduction in posterior uncertainty in the parameters, and, for the Integrated-Future model, a large revision in the estimate of the response of sugar maple to temperature (i.e. the variable ddeg) (Table 1, Fig. 4). When projecting beyond the calibration range for the naive model, the greater coverage provided by the integrated model produced substantial reductions in uncertainty (Fig. 4). When considering the present species distribution, the naive and Integrated-Present models made very similar predictions (Fig. 5). Furthermore, both models performed well when evaluated against reserved data, with median AUC values of 0.802 and 0.797 for the naive and integrated models, respectively (see Appendix S1). The small difference between the models indicates that both models adequately predict the probability of the presence of sugar maple. The major advantage to integration is the improved understanding of uncertainty in the predictions, with greater uncertainty in southern portions of the range (Fig. 5). It is important to note that the naive model was the basis for the integrated model; thus the increased uncertainty present in the integrated model is not the result of a 'worse' model, but rather should be viewed as a correction to overfitting in the naive model that incorporates uncertainty arising from the processes included in the metamodel. Phenofit predicts fitness based on how climate affects phenological timings, frost injury, reproduction and survival (Chuine & Beaubien, 2001; Morin & Thuiller, 2009). Thus, climatic factors that ultimately limit species distribution might be quite different between the two source models, as illustrated by the differences in the potential future distributions predicted by Phenofit and the naive model.

The Integrated-Future model presents a different interpretation of the response of sugar maple to warmer temperatures. The naive model predicted a substantial northward migration; in other words, the expectation under the naive model is that the present estimate of the realized niche (obtained using occurrence data) is an unbiased reflection of the fundamental niche. Thus the species should track temperature northward as the climate warms. The Integrated-Future model, in contrast, predicted substantially more tolerance to warmer temperatures, reflecting similar predictions from Phenofit (Figs 4 & 6). This is because Phenofit estimates different aspects of the realized niche. Although both models predicted a northward range shift, the change under Phenofit was limited to approximately 200 km north of the present range limit of the species, compared with more than 900 km for the naive model. Phenofit also predicted little change in the southern range limit of the species, while the naive model projected loss of the species over much of the southern portion of the range (Fig. 6). The metamodel thus presents a consensus view of the niche of the species with respect to the macroclimatic variables included in the model (Fig. 5), incorporating the present range of the species (using the occurrence data) and information from Phenofit on what conditions will be tolerable in the future.

Table 1 Parameter estimates and 95% credible intervals for the naive model (i.e. a binomial GLM relating sugar maple presence–absence data to climatic variables), and two integrated models combining the naive model and either the present or future predictions from the mechanistic model Phenofit. All models were fitted on predictor variables standardized to mean = 0 and unit variance, and all estimates are on the logit scale. Climate variables included the number of degree days (ddeg), mean annual precipitation (an_prcp) and the ratio of annual precipitation to potential evapotranspiration (pToPET). Area under the receiver operating curve (AUC) values measure model classification ability, with values of 0.5 indicating no improvement over random classification and 1 indicating perfect classification.

	Naive	Integrated-Present	Integrated-Future
intercept	−0.886	−0.103	−1.037
	(−1.12, −0.66)	(−0.23, 0.026)	(−1.16, −0.92)
ddeg	2.904	3.701	6.431
	(2.34, 3.45)	(3.37, 4.03)	(6.07, 6.79)
$ddeg^2$	−6.697	−6.216	−5.241
	(−7.04, 6.35)	(−6.43, −6.00)	(−5.43, −5.05)
$ddeg^3$	1.669	1.454	0.893
	(1.52, 1.79)	(1.37, 1.53)	(0.85, 0.94)
an_prcp	0.358	0.612	1.412
	(−0.26, 1.02)	(0.28, 0.94)	(1.11, 1.71)
an_prcp^2	−0.571	−0.848	−0.975
	(−0.76, −0.40)	(−0.95, −0.75)	(−1.06, −0.90)
pToPET	2.960	2.637	2.093
	(2.29, 3.61)	(2.28, 2.99)	(1.70, 2.48)
$pToPET^2$	−0.557	−0.093	0.0064
	(−0.74, −0.37)	(−0.15, −0.03)	(−0.070, −0.073)
AUC	0.802	0.797	
	(0.78, 0.82)	(0.78, 0.81)	–*

*AUC is unavailable for the Integrated-Future model because independent validation data are not available for future predictions.

DISCUSSION

Comparison with other methods

The methods provided here expand upon the motivation of hybrid models to develop more robust approaches to using eco-

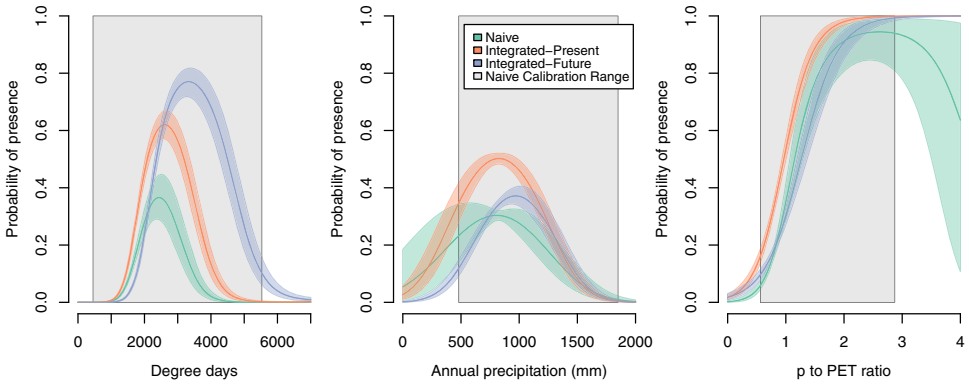

Figure 4 Response curves for each environmental variable for the three models. Predictions are broadly similar for the three models, with an increase in the optimal temperature regime predicted by the Integrated-Future model (left panel). Integration reduced prediction uncertainty for all three variables, particularly for domains outside the naive calibration range. Single-variable predictions were computed with the other variables set to their medians. Uncertainty is represented by coloured/shaded regions showing 95% Bayesian credible intervals. The grey shaded region shows the calibration range for the naive model.

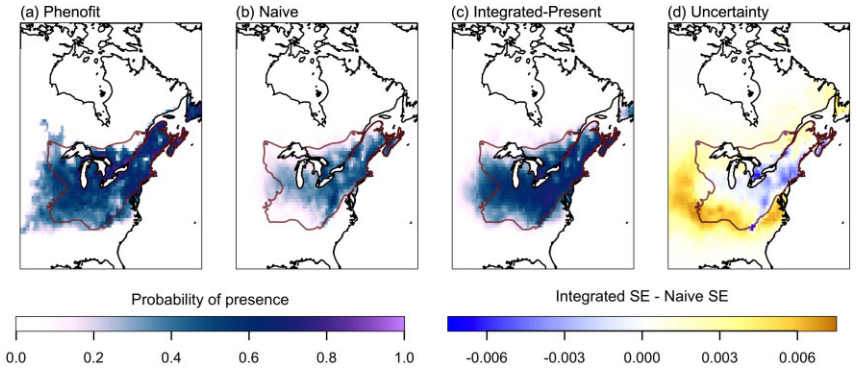

Figure 5 Range predictions for present climate for Phenofit (a), the naive model (b), the Integrated-Present model (c) and the difference between posterior prediction standard errors (SE) of the Integrated-Present and naive models (d; Integrated SE – Naive SE). Model predictions were quite similar, with reduced uncertainty where the models were in strongest agreement, and increased uncertainty near the range boundary where the models disagreed. For reference, the present range of the species is outlined in red (Little, 1971).

logical models for prediction while overcoming some limitations characteristic of other integrated approaches. In particular, it is often difficult in a hybrid model to identify parameters that can be used to connect different modelling frameworks and produce a meaningful response (Thuiller *et al.*, 2013). The difficulty of including information from experimental studies or ecological processes at lower scales can be a possible drawback of hybrid models (Smolik *et al.*, 2010; Thuiller *et al.*, 2014a). Bayesian methods provide a natural framework for the incorporation of multiple sources of information, making them an attractive alternative to SDMs. Hierarchical models in particular have the potential to capture many of the intricacies necessary for implementing hybrid models (Latimer *et al.*, 2006). Pagel & Schurr (2012) developed a hybrid approach to species distributions via a dynamic range model. Similar to our approach, their model integrated demographic information, abundance and presence/absence data within a hierarchical Bayesian framework to predict species ranges. However, their approach explicitly links the modelled processes to occurrences/abundances via a detailed demographic model, requiring data that may not be available for many species. In contrast, our approach allows for the inclusion of less complete datasets because integration is performed via the separate predictions of each model (following the application of the scaling function where necessary). Because the metamodel is expressed as a series of conditional probabilities (see Appendix S1), this information can be easily included if the probability of the metamodel can be expressed mathematically. Furthermore, Bayesian methods produce posterior distributions of parameters and predictions rather than point estimates, allowing for a comprehensive understanding of uncertainty. Finally, Bayesian methods inherently allow for feedbacks or interactions between sub-models, which may be a more realistic representation of ecological dynamics when many factors may simultaneously influence the system.

Our approach is a logical extension of other Bayesian approaches developed to deal with processes that occur at multiple scales while using several models simultaneously. In particular, it has certain similarities with Bayesian model averaging and

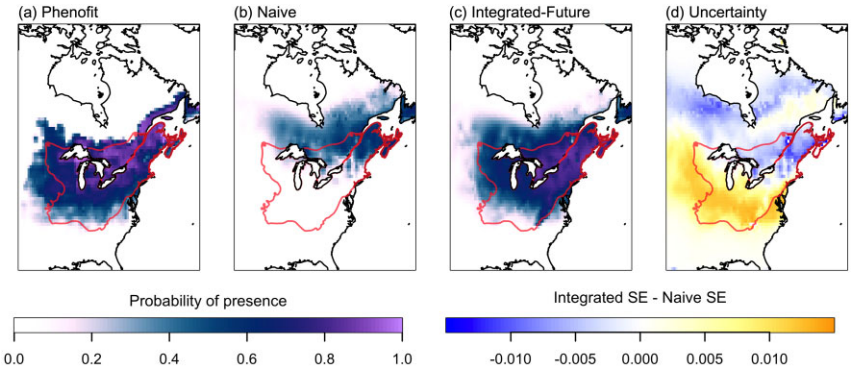

Figure 6 Range predictions for future climate for Phenofit (a), the naive model (b), the Integrated-Future model (c) and the difference between posterior prediction standard errors (SE) of the Integrated-Future and naive models (d; Integrated SE – Naive SE). The mechanistic sub-model Phenofit (a) predicted small shifts in the sugar maple range. In contrast, the naive model (b) projected a large northward change in suitable habitat. Model integration (c) produced predictions that were intermediate between the two models. Uncertainty decreased for the northern portion of the present range (red/black outline) where the models were in agreement, while it increased in the southern portion of the range where the models were in strong disagreement.

Bayesian calibration of process-based models. Bayesian model averaging aims to combine several alternative models that operate at the same scale to obtain better predictions while taking into account parameter uncertainties (Hoeting *et al.*, 1999). This is of particular interest in ecology, where the mechanisms underlying complex phenomena are often unknown (e.g. Link & Barker, 2006). Bayesian calibration of process-based models focuses on uncertainty of the parameter values, in this case the values of the parameters are calibrated by the model output (Van Oijen *et al.*, 2005; Hartig *et al.*, 2012). In contrast with these methods, our approach handles data and models operating at different hierarchical scales and uses process-based models to constrain the shape of the metamodel.

Advantages of model integration

Species distribution models are important tools that are increasingly being used by land managers for science-based decision-making (Guisan *et al.*, 2013). However, the possibility that diverse approaches will provide contrasting answers as a result of different assumptions and methodologies can create confusion and mistrust of the models, and some managers may be discouraged from incorporating their results in management plans. Integrated approaches have gained momentum in recent years, with integrative science being featured as a central theme for several science-based governmental organizations around the world (e.g. Bernier *et al.*, 2013). Incorporating information from multiple sources, particularly with respect to uncertainty, fosters a connection between scientifically generated knowledge and policy, and is therefore an important tool for adaptive management (Rehme *et al.*, 2011, Fig. 7). Such approaches are needed in designing management plans for vulnerable species and ecosystems to avoid basing decisions on too-narrow subsets of the available information (Dawson *et al.*, 2011). However, the successful use of approaches such as ours will always remain dependent on an intimate understanding of the decision-making process, emphasizing the importance of close collaboration between modellers and practitioners at all stages of model development (Guisan *et al.*, 2013).

Model uncertainty is another key factor affecting applicability of model outputs (Addison *et al.*, 2013). One of the main strengths of our approach is that it allows for a transparent identification of uncertainties and how they propagate through the models. Transparency in uncertainty can be considered as a sort of sensitivity analysis, whereby the greatest sources of uncertainty can be detected and further research directed accordingly (e.g. Example 1, Adding experimental evidence for the fundamental niche to a species distribution model; Figs 1 & 2). The new knowledge resulting from this research can then be readily incorporated into the metamodel and the model predictions updated to account for the new information. The ease of incorporating new knowledge to the modelling framework allows for a rapid adjustment of the predictions and the incorporation of the most recently available knowledge into management plans (Keith *et al.*, 2011). Furthermore, the use of linked sub-models allows for clear specification of desired model outputs (via the specification of the metamodel) while easily retaining important ecological objectives (via specification of sub-models). Transparency in the model-building process must be accompanied by a clearly documented workflow. We suggest using the sub-models as a natural proxy for specifying specific objectives, and using this as the basis for developing workflows describing the process of model integration to ensure reproducibility and applicability (Fig. 7). Adaptive approaches such as the one presented here are often highlighted as a pressing need in order to develop strategies to promote ecosystems that are both feasible and resilient (Seastedt *et al.*, 2008).

In many cases, both the data and theory needed to apply our approach already exist, and all that is needed is the development of sub-models and their integration into a metamodel. For example, climatic gradients may mediate competitive interactions (Kunstler *et al.*, 2011), which means that simple correlative

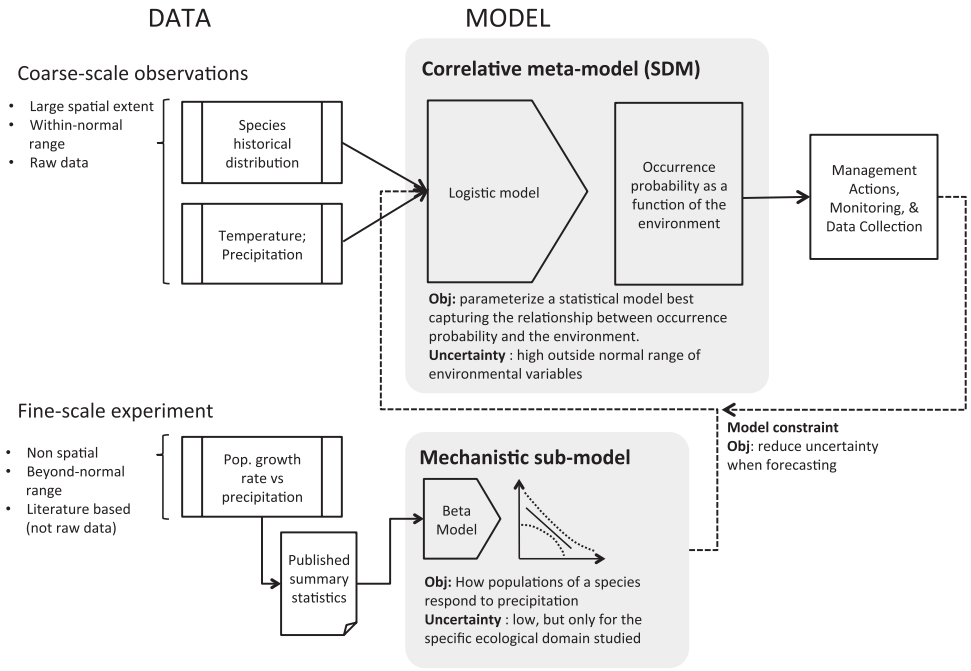

Figure 7 Sample workflow for applying the models presented in Example 1 in a management context. Critical steps include specifying the metamodel, identifying additional sources of information to be used as constraints on the metamodel and using the integrated prediction for decision-making. As additional information becomes available from monitoring the results of management, this information can be incorporated in additional sub-models to further refine the metamodel.

models that fail to account for competition may be wrong if the climate–competition association changes in the future. In North America, the US Forest Service maintains a long-term Forest Inventory Analysis database that could be utilized to parameterize a competition model. Such a model need not explicitly predict occurrence limits; rather it could be integrated with a larger-scale model to include information about competition in a distribution model. The phenological information needed to parameterize the sub-model for Example 2 (Constraining a SDM using phenological information) is similarly available for a wide range of species (Morin & Thuiller, 2009). There are also a number of networks collecting high-quality data with good temporal and spatial coverage [e.g. the National Ecological Observatory Network (NEON) and Long-Term Ecological Research sites (LTER)]. There is great potential for these kinds of data to be used in fitting sub-models of the kind used in Example 1. In other cases, efforts have already been made to compare models qualitatively (Morin & Thuiller, 2009; Cheaib *et al*., 2012). Our framework could be used post hoc on the outputs of these models to quantify uncertainty resulting from model disagreement.

Challenges

Although our approach is highly flexible and can be applied in a number of situations, there are some challenges to successfully using the framework. Data quality and availability can present a significant constraint on the number and type of models that can be implemented in our framework. One obstacle is a lack of adequate and unbiased coverage of explanatory variables; exploratory analyses can be a significant aid in understanding how data coverage affects the resulting predictions (McKenney *et al*., 2002). Integration can solve these issues to some extent by using supplemental information (and conceptual advances) in additional sub-models where coverage is weak (e.g. Example 1; Figs 2 & 3). A strength of our approach is that it can be used without the full suite of data that would be required to run a fully mechanistic model. Given that the metamodel is correlative, it can be effectively implemented with, for example, only presence–absence data, or, in the case where true absences are difficult to obtain, with presences and pseudo absences (provided sufficient care is used in interpreting the results of such a model). Consequently, any additional mechanistic data that become available will enhance predictions by constraining outputs of the metamodel.

Determining the functions to use to express the likelihood of the sub-models given the metamodel (i.e. equation 5) is a critical point. The challenge is three-fold: (1) determining which spatial and temporal scales (i.e. which processes), are to be considered; (2) selecting how to build and scale the sub-models to be consistent with the metamodels; and (3) understanding how error and uncertainty propagate from the sub-models to the metamodel. Although we argue that our proposed framework is able to easily deal with different scales and that the Bayesian framework allows for an efficient integration of uncertainty across all scales considered, the building of scaling functions is an object of investigation on its own. It is likely that the modelling process may include multiple functions operating at different scales when taking all

known processes and models into account. Indeed, if species distributions are a function of, for example, population growth rate (Guisan & Zimmermann, 2000), they will involve processes at the individual (e.g. competition) or cellular (e.g. photosynthesis) scales. Such very large differences in spatial scales would require more sophisticated upscaling methods than the simple functions we have used here. Our framework is still applicable whatever the chosen upscaling approach and is able to propagate uncertainties from sub-models to the metamodel. Indeed, if a sub-model provides poor information (due, for example, to cross-scale nonlinearities in the response to the environment), the resulting metamodel predictions may be worse than the pre-integration naive model. In general, we advise users of this framework to carefully consider the scaling in their models with respect to the biology of the organism studied, and to use prior model weights to downweight the scaled models when there is uncertainty about the applicability of a sub-model at the metamodel scale. For instance, in Example 1 we applied a model weight to decrease the influence of the mechanistic model on the metamodel (while still retaining some information contained therein), recognizing that such a simplistic model may not scale well. Finally, as always when modelling ecological systems, we urge humility in the interpretation of model results and suggest the use of model evaluation and validation tools whenever possible. We provide additional discussion on the implementation of model weights in Appendix S1.

The implementation of the model itself can present an obstacle when model complexity increases. In many cases, off-the-shelf software can adequately express the model likelihoods with minimal programming, but more complicated models will require the development of custom programs. Developing such customized code requires careful model specification, understanding of applied Bayesian methods and, in some cases, extensive programming. However, the flexibility of our approach and its transparency with respect to the propagation of uncertainty will often outweigh the implementation challenges.

Finally, this approach is not just a new methodological tool but a framework for forecasting species distributions that is fundamentally designed to make the link between modellers and practitioners while correctly estimating uncertainties and being updatable with new data and theoretical advances. A scientific approach such as that presented here is particularly adapted to synthesize available information and provide robust species distribution forecasts based on information known to be the best available scientific knowledge. It can incorporate large databases, valuing the efforts of data collection, and include models based on the latest theoretical advances, which is essential to decrease errors due to model specification (Austin, 2007). In addition, it provides practitioners and decision-makers with the best possible estimation of uncertainties, with direct applications to risk assessment or to guide the choice when investigating new research and accumulating new data. Finally, we argue that the adaptability of our approach is particularly appropriate in a world where collected data and theoretical knowledge are changing as quickly as climate, and conservation practices must be adjusted accordingly.

Data accessibility

All data, as well as all code required to repeat the analyses, have been uploaded as online supporting information in Appendix S2.

ACKNOWLEDGEMENTS

We acknowledge funding from the Quebec Centre for Biodiversity Science as well as NSERC strategic grant 430393-12. We thank editors David Currie and Arndt Hampe, as well as two anonymous referees, whose feedback greatly improved the manuscript. Additionally, comments from Charles Canham, William Godsoe, and Tamara Münkemüller improved a previous version of this manuscript. W.T. received support funding from the European Research Council under the European Community's Seven Framework Programme FP7/2007-2013 grant agreement no. 281422 (TEEMBIO).

REFERENCES

Addison, P.F.E., Rumpff, L., Bau, S.S., Carey, J.M., Chee, Y.E., Jarrad, F.C., McBride, M.F. & Burgman, M.A. (2013) Practical solutions for making models indispensable in conservation decision-making. *Diversity and Distributions*, **19**, 490–502.

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

Araújo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology and Evolution*, **22**, 42–47.

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

Bernier, P., Kurz, W.A., Lemprière, T. & Ste-Marie, C. (2013) *A blueprint for forest carbon science in Canada, 2012–2020*. Natural Resources Canada, Canadian Forest Service, Ottawa, ON.

Blois, J.L., Zarnetske, P.L., Fitzpatrick, M.C. & Finnegan, S. (2013) Climate change and the past, present, and future of biotic interactions. *Science*, **314**, 499–504.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012) Disentangling the underlying mechanisms of species abundance distribution using a comprehensive and nested modeling framework. *Ecology Letters*, **15**, 584–593.

Boulangeat, I., Georges, D., Dentant, C., Bonet, R., Van Es, J., Abdulhak, S., Zimmermann, N.E. & Thuiller, W. (2014) Anticipating the spatio-temporal response of plant diversity and vegetation structure to climate and land use change in a protected area. *Ecography*, **37**, 1230–1239.

Catterall, S., Cook, A.R., Marion, G., Butler, A. & Hulme, P.E. (2012) Accounting for uncertainty in colonisation times: a novel approach to modelling the spatio-temporal dynamics of alien invasions using distribution data. *Ecography*, **35**, 901–911.

Cheaib, A., Badeau, V., Boe, J., Chuine, I., Delire, C., Dufrêne, E., François, C., Gritti, E.S., Legay, M., Pagé, C., Thuiller, W.,

Viovy, N. & Leadley, P. (2012) Climate change impacts on tree ranges: model intercomparison facilitates understanding and quantification of uncertainty. *Ecology Letters*, **15**, 533–544.

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

Clark, J.S. & Gelfand, A.E. (2006) A future for models and data in environmental science. *Trends in Ecology and Evolution*, **21**, 375–380.

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

Dawson, T.P., Jackson, S.T., House, J.I., Prentice, I.C. & Mace, G.M. (2011) Beyond predictions: biodiversity conservation in a changing climate. *Science*, **332**, 53–58.

Dormann, C.F. (2007) Promising the future? Global change projections of species distributions. *Basic and Applied Ecology*, **8**, 387–397.

Engler, J.O., Rödder, D., Elle, O., Hochkirch, A. & Secondi, J. (2013) Species distribution models contribute to determine the effect of climate and interspecific interactions in moving hybrid zones. *Journal of Evolutionary Biology*, **26**, 2487–2496.

Ferrari, S. & Cribari-Neto, F. (2004) Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799–815.

Gallien, L., Münkemüller, T., Albert, C.H., Boulangeat, I. & Thuiller, W. (2010) Predicting potential distributions of invasive species: where to go from here? *Diversity and Distributions*, **16**, 331–342.

Gallien, L., Douzet, R., Pratte, S., Zimmermann, N.E. & Thuiller, W. (2012) Invasive species distribution models – how violating the equilibrium assumption can create new insights. *Global Ecology and Biogeography*, **21**, 1126–1136.

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

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

Guisan, A., Tingley, R., Baumgartner, J.B. *et al.* (2013) Predicting species distributions for conservation decisions. *Ecology Letters*, **16**, 1424–1435.

Hargreaves, A.L., Samis, K.E. & Eckert, C.G. (2014) Are species' range limits simply niche limits writ large? a review of transplant experiments beyond the range. *The American Naturalist*, **183**, 157–173.

Hartig, F., Dyke, J., Hickler, T., Higgins, S.I., O'Hara, R.B., Scheiter, S. & Huth, A. (2012) Connecting dynamic vegetation models to data – an inverse perspective. *Journal of Biogeography*, **39**, 2240–2252.

Hobbs, N.T. & Ogle, K. (2011) Introducing data–model assimilation to students of ecology. *Ecological Applications*, **21**, 1537–1545.

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. (1999) Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382–401.

Holt, R.D. (2009) Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences USA*, **106**, 19659–19665.

Keith, D.A., Martin, T.G., McDonald-Madden, E. & Walters, C. (2011) Uncertainty and adaptive management for biodiversity conservation. *Biological Conservation*, **144**, 1175–1178.

Kunstler, G., Albert, C.H., Courbaud, B., Lavergne, S., Thuiller, W., Vieilledent, G., Zimmermann, N.E. & Coomes, D.A. (2011) Effects of competition on tree radial-growth vary in importance but not in intensity along climatic gradients. *Journal of Ecology*, **99**, 300–312.

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

Levin, S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.

Levin, S.A. (1998) Ecosystems and the biosphere as complex adaptive systems. *Ecosystems*, **1**, 431–436.

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

Little, E.L., Jr (1971) *Atlas of United States trees: volume 1, conifers and important hardwoods*. US Department of Agriculture Miscellaneous Publication 1146. US Department of Agriculture, Washington, DC.

McKenney, D.A., Venier, L.A., Heerdegen, A. & McCarthy, M.A. (2002) A Monte Carlo experiment for species mapping problems. *Predicting species occurrences: issues of accuracy and scale* (ed. by J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall and F.B. Samson), pp. 377–381. Island Press, Washington, DC.

McMahon, S.M., Harrison, S.P., Armbruster, W.S., Bartlein, P.J., Beale, C.M., Edwards, M.E., Kattge, J., Midgley, G., Morin, X. & Prentice, I.C. (2011) Improving assessment and modelling of climate change impacts on global terrestrial biodiversity. *Trends in Ecology and Evolution*, **26**, 249–259.

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.

Pagel, J. & Schurr, F.M. (2012) Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, **21**, 293–304.

Peters, D.P.C., Pielke, R.A., Bestelmeyer, B.T., Allen, C.D., Munson-McGee, S. & Havstad, K.M. (2004) Cross-scale interactions, nonlinearities, and forecasting catastrophic events. *Proceedings of the National Academy of Sciences USA*, **101**, 15130–15135.

Pigot, A.L. & Tobias, J.A. (2013) Species interactions constrain geographic range expansion over evolutionary time. *Ecology Letters*, **16**, 330–338.

Rehme, S.E., Powell, L.A. & Allen, C.R. (2011) Multimodel inference and adaptive management. *Journal of Environmental Management*, **92**, 1360–1364.

Schurr, F.M., Pagel, J., Cabral, J.S., Groeneveld, J., Bykova, O., O'Hara, R.B., Hartig, F., Kissling, W.D., Linder, H.P., Midgley,

G.F., Schröder, B., Singer, A. & Zimmermann, N.E. (2012) How to understand species' niches and range dynamics: a demographic research agenda for biogeography. *Journal of Biogeography*, **39**, 2146–2162.

Seastedt, T.R., Hobbs, R.J. & Suding, K.N. (2008) Management of novel ecosystems: are novel approaches required? *Frontiers in Ecology and the Environment*, **6**, 547–553.

Smolik, M.G., Dullinger, S., Essl, F., Kleinbauer, I., Leitner, M., Peterseil, J., Stadler, L.M. & Vogl, G. (2010) Integrating species distribution models and interacting particle systems to predict the spread of an invasive alien plant. *Journal of Biogeography*, **37**, 411–422.

Swets, J. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Thuiller, W., Münkemüller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffers, K. & Gravel, D. (2013) A road map for integrating eco-evolutionary processes into biodiversity models. *Ecology Letters*, **16**, 94–105.

Thuiller, W., Münkemüller, T., Schiffers, K.H., Georges, D., Dullinger, S., Eckhart, V.M., Edwards, T.C., Gravel, D., Kunstler, G., Merow, C., Moore, K., Piedallu, C., Vissault, S., Zimmermann, N.E., Zurell, D. & Schurr, F.M. (2014a) Does probability of occurrence relate to population dynamics? *Ecography*, **37**, 1155–1166.

Thuiller, W., Pironon, S., Psomas, A., Barbet-Massin, M., Jiguet, F., Lavergne, S., Pearman, P.B., Renaud, J., Zupan, L. & Zimmermann, N.E. (2014b) The European functional tree of bird life in the face of global change. *Nature Communications*, **5**, 3118.

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.

Wu, J. & Loucks, O.L. (1995) From balance of nature to hierarchical patch dynamics: a paradigm shift in ecology. *Quarterly Review of Biology*, **70**, 439–466.

SUPPORTING INFORMATION

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

Appendix S1 Model description and methodology.
Appendix S2 Model code and data.

BIOSKETCH

This paper was led by members of the Biogeography and Metacommunity Ecology Lab at the University of Quebec at Rimouski. The lab focuses on interactions between species distributions, community structure and ecosystem functioning. We apply principles of spatial ecology to a variety of organisms and systems, from bacteria to entire forests. We also use theoretical and simulation models to develop hypotheses and extend our work beyond the technical limitations of empirical studies.

Editor: Arndt Hampe