esa ECOSPHERE

An efficient protocol for the global sensitivity analysis of stochastic ecological models

Thomas A. A. Prowse,[1,]† Corey J. A. Bradshaw,[1] Steven Delean,[1] Phillip Cassey,[1] Robert C. Lacy,[2] Konstans Wells,[1,5] Matthew E. Aiello-Lammens,[3,6] H. R. Akçakaya,[3] and Barry W. Brook[4]

[1]*The Environment Institute and School of Biological Sciences, The University of Adelaide, Adelaide, South Australia 5005 Australia*
[2]*Chicago Zoological Society, Brookfield, Illinois 60513 USA*
[3]*Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794 USA*
[4]*School of Biological Sciences, Private Bag 55, University of Tasmania, Hobart, Tasmania 7001 Australia*

Citation: Prowse, T. A. A., C. J. A. Bradshaw, S. Delean, P. Cassey, R. C. Lacy, K. Wells, M. E. Aiello-Lammens, H. R. Akçakaya, and B. W. Brook. 2016. An efficient protocol for the global sensitivity analysis of stochastic ecological models. Ecosphere 7(3):e01238. 10.1002/ecs2.1238

Abstract. Stochastic simulation models requiring many input parameters are widely used to inform the management of ecological systems. The interpretation of complex models is aided by global sensitivity analysis, using simulations for distinct parameter sets sampled from multidimensional space. Ecologists typically analyze such output using an "emulator"; that is, a statistical model used to approximate the relationship between parameter inputs and simulation outputs and to derive sensitivity measures. However, it is typical for *ad hoc* decisions to be made regarding: (1) trading off the number of parameter samples against the number of simulation iterations run per sample, (2) determining whether parameter sampling is sufficient, and (3) selecting an appropriate emulator. To evaluate these choices, we coupled different sensitivity-analysis designs and emulators for a stochastic, 20-parameter model that simulated the reintroduction of a threatened species subject to predation and disease, and then validated the emulators against new output generated from the simulation model. Our results lead to the following sensitivity analysis-protocol for stochastic ecological models. (1) Run a single simulation iteration per parameter sample generated, even if the focal response is a probabilistic outcome, while sampling extensively across the parameter space. In contrast to designs that invested in many model iterations (tens to thousands) per parameter sample, this approach allowed emulators to capture the input-output relationship of the simulation model more accurately and also to produce sensitivity measures that were robust to variation inherent in the parameter-sampling stage. (2) Confirm that parameter sampling is sufficient, by emulating subsamples of the sensitivity-analysis output. As the subsample size is increased, the cross-validatory performance of the emulator and sensitivity measures derived from it should exhibit asymptotic behavior. This approach can also be used to compare candidate emulators and select an appropriate interaction depth. (3) If required, conduct further simulations for additional parameter samples, and then report sensitivity measures and illustrate key response curves using the selected emulator. This protocol will generate robust sensitivity measures and facilitate the interpretation of complex ecological models, while minimizing simulation effort.

Key words: boosted regression trees; ecological model; emulator; global sensitivity analysis; metamodel; parameter uncertainty; population growth rate; probability of extinction; species interactions.

Received 9 June 2015; revised 8 September 2015; accepted 8 October 2015. Corresponding Editor: van Kooten.

[5] Present address: Environmental Futures Research Institute, Griffith University, Brisbane, Queensland 4111 Australia.
[6] Present address: Department of Environmental Studies and Science, Pace University, Pleasantville, New York 10570 USA.

† **E-mail:** thomas.prowse@adelaide.edu.au

Introduction

Following an exponential rise in the availability of cheap computing power, simulation models are increasingly used to study ecological systems that are too difficult to investigate empirically or experimentally, or as planning tools in adaptive management (Green et al. 2005, Hastings et al. 2005, Valle et al. 2009). However, an increase in model complexity—for example, by including more species or interactions—necessitates the estimation of additional input parameters about which there is often considerable uncertainty. For example, a simple population viability analysis to estimate extinction risk might only require information about the survival and fertility rates of a focal species that are readily derived from empirical studies (Boyce 1992, Brook et al. 2000). In contrast, parameters governing the strength or functional form of ecological processes or interactions (e.g., density feedbacks, predator-prey or disease-host dynamics) are typically more difficult to estimate accurately.

Although simple models are easier to interpret, complex simulation models are often required to address specific applied or theoretical questions (e.g., Lindenmayer and Possingham 1995, Bradshaw et al. 2012). Modelers are therefore faced with a trade-off between the need to simulate important ecological processes adequately, which might necessitate many parameters, and the need to construct interpretable and computationally tractable model systems (Levins 1966, Ginzburg and Jensen 2004). Sensitivity analysis is the primary tool used to determine whether simulation models produce outputs that are robust to parameter uncertainty. Sensitivity analysis is also used to assess the relative importance of input parameters to guide future research (that refines estimates for critical parameters) or to simplify the model by identifying unimportant parameters that can be fixed or removed entirely. Ecologists most commonly use "local" (or "one-at-a-time") sensitivity analyses (Conroy and Brook 2003, Naujokaitis-Lewis et al. 2009, Coutts and Yokomizo 2014), and quantify the effect of variation in single parameters on the model output, whilst all other parameters remain fixed at default values (Turyani and Rabitz 2000, Cariboni et al. 2007). In contrast, a "global" sensitivity analysis varies all parameters simultaneously and can provide robust sensitivity measures in the presence of nonlinear responses and interactions among parameters (Drechsler 1998, Sobol 2001, Wainwright et al. 2014).

Although global sensitivity analysis can facilitate the interpretation of complex models, the approach suffers from the "curse of dimensionality" because it relies on running simulations for many samples drawn from the multidimensional parameter space (Cariboni et al. 2007, Wainwright et al. 2014). The combinatorial explosion typically prohibits fully orthogonal sensitivity-analysis designs for complex models, and is usually addressed by sampling a fixed number of k parameter sets, often by generating k random values for each parameter independently (McCarthy et al. 1995), or by implementing "Latin hypercube" sampling with k divisions to guarantee better coverage of the parameter space (Fang et al. 2006). After running simulations for each parameter sample, the output can be summarized with a second descriptive model, known as an "emulator" or "meta-model". We use the former term because "meta-model" also refers to an ecological model composed of linked components such as coupled demographic-epidemiological models (Lacy et al. 2013). An emulator is a statistical model or machine-learning technique that approximates the complex function linking the model outputs to its inputs, $Y = f(\mathbf{X})$, with a simpler mathematical function $\eta(\mathbf{X}) \approx f(\mathbf{X})$ (Ratto et al. 2012, Marie and Simioni 2014). The emulator can be used to produce summary sensitivity metrics; that is, measures of the variance in the output due to variation in the parameter inputs (Storlie et al. 2009).

This approach to global sensitivity analysis appeals to ecologists, in part because emulators can distill complex models to an interpretable set of the most influential input-output response curves. Although emulators can potentially reproduce simulation outputs with great accuracy and speed (Marie and Simioni 2014), the literature currently offers little guidance with respect to optimal parameter-sampling designs and emulators for stochastic ecological models. In particular, there is a clear trade-off between the number of parameter samples taken and the number of simulation iterations that can be feasibly run per sample. A central tenet of population and conservation biology is that stochastic processes (e.g., demographic

and environmental stochasticity) can influence the fate of populations and ecosystems more generally (Shaffer 1981, Traill et al. 2007, Melbourne and Hastings 2008). Therefore, stochastic models are usually run many times for each parameter set tested to characterize mean responses and their plausible ranges (e.g., population growth rate, minimum expected population size) or to derive probabilistic outcomes (e.g., probability of extinction) (Shaffer 1981, Harris et al. 1987).

Ecologists applying global sensitivity analyses to stochastic models have typically made ad hoc decisions regarding how to balance the number of parameter samples against the number of iterations per sample (e.g., McCarthy et al. 1995, Bradshaw et al. 2012). For example, McCarthy et al. (1995) constructed a population model for the helmeted honeyeater (*Lichenostomus melanops cassidix*) that included four fertility parameters, and used global sensitivity analysis to test the influence of those parameters on the risk of population decline. They generated 500 parameter samples by random, independent sampling of the four parameters, ran 10 simulation iterations per sample (for a total of 5000 simulations), and emulated the output using logistic regression. As ecological models become more complex, however, another option is to maximize the coverage of the multidimensional parameter space by running a single simulation per sample (e.g., Prowse et al. 2013).

In this paper, we tested different sensitivity-analysis designs and emulators for a complex ecological model, to generate rules of thumb for conducting global sensitivity analysis for stochastic models. We show that maximizing the coverage of the parameter space is the best strategy, even when the response of interest is a probabilistic outcome, and that machine-learning techniques provide a convenient choice for emulation. We also introduce a standardized method for choosing an appropriate emulator and determining when parameter sampling is sufficient for that emulator to capture the input–output relationship of the simulation model.

Methods

Model details

We constructed a hypothetical ecological model using an annual time step that simulated the reintroduction of a hypothetical threatened species. The simulated reintroduction required 20 input parameters that governed the size of the founding population, the species' demography (reproduction and survival), the introduction and transmission of a disease with detrimental effects on demographic rates, the population size of a predator inhabiting the re-introduction site, and the ratio-dependent functional response of that predator. We assumed the purpose of the model was to investigate which ecological parameters or processes were most important in determining the post-introduction fate of the simulated populations. We therefore produced two different outputs from the model to be used as response variables for emulation: (1) the annual population growth rate (r) for each re-introduction averaged over a 10-yr simulation time frame, and (2) a binary output representing the final status of the re-introduction (extinct or extant). The latter output permitted consideration of a probabilistic outcome, namely the probability of population extinction. We constructed the model in the *R* computing environment (version 3.0.3; R Development Core Team 2014). All model parameters are described in Table 1, with the full *R* code provided in the Supplement to this paper.

Demography.—We initiated the simulated reintroduction with an equal number of mature males and females as dictated by the starting population size. The fate of the reintroduction was then a function of: (1) survival probabilities for two age classes, juveniles (0–1 yr) and subadult/adults (>1 yr); (2) the reproductive output of females, in turn governed by the age at maturity, probability of breeding and fertility rate (i.e., mean number of offspring per year); and (3) the sex ratio of offspring produced. We assumed a polygynous breeding system such that all mature females could potentially breed provided at least one mature male was present in the population. To incorporate demographic stochasticity, we modeled the outcome of all probabilities with binomial distributions and sampled the number of offspring produced by each reproducing female from Poisson distributions. We simulated environmental stochasticity by sampling survival rates each year from a normal distribution with mean equal to the deterministic rate and an arbitrary standard deviation of 0.15.

Table 1. Input parameters for the modeled reintroduction. Shown are default values for each parameter, as well as the integer values or uniform (*U*) ranges that we tested for sensitivity-analysis designs, using Latin hypercube or random sampling.

Parameter	Default	Sensitivity Analysis
Demography		
Starting population size (*nStart*)	50	2, 4, …, 98, 100
Age at maturity (*ageMaturity*)	2	1, 2, 3
Sex ratio (proportion males) at birth (*sr*)	0.5	*U*(0.25, 0.75)
Probability of mature females breeding (*pBreed*)	0.75	*U*(0.5, 1)
Fertility rate for breeding females (*m*)	6	*U*(2, 10)
Survival rate of 1+ yr olds (*s1plus*)	0.75	*U*(0.5, 1)
Reduction in survival rate for 0–1 yr olds relative to 1 + class (*s0.mult*)	0.75	*U*(0.5, 1)
Disease dynamics		
Probability of infection from outside source (*pIoutside*)	0.05	*U*(0, 0.1)
Probability of maternal disease transfer (*pMatTrans*)	0.5	*U*(0, 1)
Logistic parameter controlling density-dependent disease transmission (*beta*)	0.1	*U*(0, 0.2)
Probability of recovering from infection (*pRecover*)	0.75	*U*(0.5, 1)
Probability of acquiring resistance following recovery (*pResistant*)	0.5	*U*(0, 1)
Probability of losing resistance (*pLoseResistance*)	0.25	*U*(0, 0.5)
Demographic effects of disease		
Reduction in *pBreed* for infected females (*pBreed.Imult*)	0.5	*U*(0, 1)
Reduction in fertility for infected females (*m.Imult*)	0.5	*U*(0, 1)
Reduction in survival for infected 1+ yr olds (*s1plus.Imult*)	0.5	*U*(0, 1)
Reduction in survival for infected 0-1 yr olds (*s0.Imult*)	0.5	*U*(0, 1)
Predation		
Predator population size (*P*)	5	1, 2, …, 9, 10
Attack rate of predator (*a*)	0.05	*U*(0, 0.1)
Handling time for prey (*h*)	0.25	*U*(0, 0.5)

Disease dynamics and demographic effects

We used a stochastic Susceptible-Infected-Resistant-Susceptible (SIRS) model to simulate the impact of a hypothetical disease on the re-introduced population, and assumed all starting animals were susceptible to the disease. The disease could be transmitted between individuals through: (1) vertical transfer from mother to offspring according to a specified probability; and (2) density-dependent disease transmission with the annual probability of infection calculated according to a logistic function:

$$\frac{1}{1+e^{-(\alpha+\beta I)}}$$

where I is the number of infected individuals in the population, β is the slope parameter, and α is the intercept parameter that we set so this expression was equal to the specified probability of infection from an outside source (*pIoutside*, Table 1) when $I = 0$ (i.e., when no infected individuals were present within the population). The epidemiology of infected individuals was governed by three probabilities: (1) the probability of recovery; (2) the probability of acquiring resistance following recovery; and (3) the probability of eventually losing this resistance. Again, we included demographic stochasticity by sampling the outcome of all probabilities from binomial distributions. For infected individuals, we simulated the demographic effects of disease by reducing survival probabilities and female-specific probabilities of reproduction and fertility rates.

Predation.—We modeled the impact of a predator on the re-introduced population by assuming a ratio-dependent, Type III functional response such that the expected number of individuals removed by predators each year was:

$$f(N,P)=\left(\frac{aN^2}{P^2+ahN^2}\right)P$$

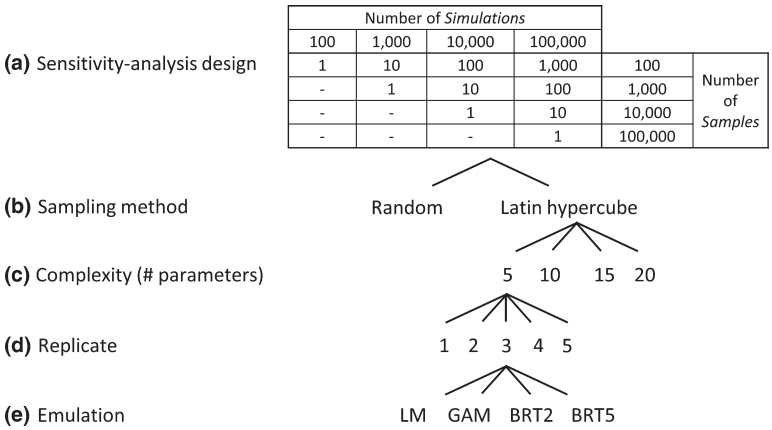

Fig. 1. The orthogonal structure used to evaluate different sensitivity-analysis designs and emulators (for clarity, only one complete branch is shown). (a) Matrix of ten different sensitivity-analysis designs tested. The numbers within each cell provide the numbers of iterations run per sample (calculated as the number of simulations divided by the number of samples taken). (b) The method for drawing parameter samples. (c) The complexity of the sensitivity analysis (number of parameters included). (d) Replicate sensitivity analyses for each complexity level. For complexities <20, each replicate represented a distinct sensitivity analysis that included a different combination of parameters randomly sampled from the available inputs. For a complexity of 20, each replicate sensitivity analysis included all parameters so only the sampling and simulation steps were distinct. (e) Sensitivity-analysis output was emulated using different statistical techniques and predictions from each emulator were then verified against validation datasets produced for each design × complexity × replicate combination. LM, linear model; GAM, generalized additive model; BRT2 and BRT5, boosted regression trees with tree complexities of 2 and 5, respectively.

where a is the attack rate, h is the handling time, N is the population size of the focal species (i.e., the prey), and P is the population size of the predator. We then calculated the probability of surviving predation by calculating $(N - f)/N$ and modified survival rates accordingly.

Sensitivity analyses

We used a hierarchical structure to evaluate different sensitivity-analysis designs and emulators (Fig. 1). We tested ten different sensitivity-analysis designs that varied: (1) the total number of *simulations*; (2) the number of *samples* (i.e., distinct parameter sets) tested; and (3) the number of *iterations* per sample (calculated as the number of simulations divided by the number of samples taken). These designs ranged from those capable of producing precise point estimates of r for few samples across the parameter space (e.g., 100,000 simulations that were allocated to 100 samples with 1000 iterations per sample) to those favoring more comprehensive coverage of the parameter space (i.e., more samples) at the expense of fewer stochastic iterations per sample (with the extreme case being 100,000 simulations allocated to 100,000 samples with 1 iteration per sample). Parameter samples were produced by either: (1) drawing random samples from a set of independent uniform distributions without taking into account the previously generated sample points (Table 1), or (2) drawing Latin hypercube samples from the same distributions using the function randomLHS in the R package lhs (Fig. 1; Carnell 2012). Latin hypercube sampling implements an a priori equal-area subdivision of the sample space and then samples randomly within each subdivision (McKay et al. 1979).

We tested the ten designs across sensitivity analyses with different levels of *complexity*, reflecting the inclusion of 5, 10, 15 or all 20 input parameters in the sensitivity analysis (Fig. 1). We assigned each parameter a default value to be used when that parameter was excluded from the sensitivity analysis (Table 1). We anticipated that, for complexities less than 20 (i.e., when

some parameters were fixed at their default values), the performance of each design might be affected by the combination of parameters selected for inclusion. We therefore produced five *replicates* for each complexity level by randomly sampling five different combinations of parameters for each (Fig. 1).

Emulators

We tested three different emulators that are commonly used in ecology: linear models (LM; actually a generalized linear model for the probability of extinction), generalized additive models (GAM), and boosted regression trees (BRT), which we fit using the R packages base, mgcv and dismo, respectively (Wood 2011, Hijmans et al. 2013). We fit LM emulators without polynomial or interaction terms, reasoning that as ecological models become more complex (e.g., >10 parameters), practitioners are unlikely to wish to fit and interpret many of these terms (e.g., Bradshaw et al. 2012). GAM provide a data-driven method of accounting for non-linear relationships between response and predictor variables by including smoothing functions of those predictors (Hastie and Tibshirani 1990, Wood 2006), and we fixed an upper limit of 3 for the degrees of freedom of the smoothing terms.

BRT combine regression trees (models that relate a response to their predictors by recursive binary splits) and boosting (a method for combining simple models to improve predictive performance), and can fit complex, non-linear relationships and automatically handle interactions between predictors (Elith et al. 2008). We fit BRT models using the function gbm.step with a learning rate of 0.01, a bag fraction of 0.75, and a tree complexity of 2 or 5 (i.e., first- or fourth-order interactions included), and optimized the number of fitted trees based on 5-fold cross-validation. In rare instances, BRT models could not be fitted using this model specification, in which cases we decreased the learning rate until the BRT fit successfully. We assumed Gaussian or binomial error distributions for emulating the population growth rate (r) or probability of extinction (e), respectively. We also used relative influence metrics from the BRT models to rank the sensitivity of these outputs to each input parameter, and examined the stability of these ranking across replicates of the different sensitivity-analysis designs. These relative influence measures are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, averaged over all trees, and scaled to sum to 100 (Elith et al. 2008). This provides a single sensitivity measure for each input parameter that incorporates the contribution of that parameter to main effects and interactions, with higher numbers indicating more important parameters.

Validation of design-emulator couplings

Our aim was to validate the ability of coupling different sensitivity-analysis designs and emulators to describe variation in the simulated population growth rate (r) or probability of extinction (e) across the multidimensional parameter space. To produce validation datasets for this purpose, we drew 200 new Latin hypercube samples within each design × complexity × replicate combination. We then ran 10,000 simulations of the stochastic model for each combination to derive 200 highly precise ("true") point estimates of r and e for those validation samples. Finally, we summarized the output of different sensitivity-analysis designs, using emulators and compared the ability of these emulators to predict the validation data sets. For the population growth rate, we calculated the validation R^2 for each emulator and sensitivity-analysis design; that is, the proportion of variance in r in the out-of-sample validation dataset that could be explained by predictions from the emulator. For the probability of extinction, we first binned predicted and validation probabilities by deciles to produce ordinal data with 10 levels (i.e., 0–10%, 10–20%, etc.), and then calculated the proportion of correctly classified cases; that is, the proportion of cases for which the validation bin was predicted correctly by the emulator.

Choosing an appropriate emulator and verifying that parameter sampling is sufficient

Given that modelers are unlikely to produce validation datasets for the purpose of choosing appropriate emulators and evaluating the adequacy of parameter sampling, we investigated an alternative (simpler) procedure for evaluating

these choices. Specifically, using BRT emulators with different interaction depths, we modeled subsamples of the sensitivity analysis output of size n, where $n \leq$ the total number of parameter samples. We then plotted changes in the cross-validation deviance from the emulation. We also calculated a measure of the "stability" of sensitivity measures between pairs of BRT models as the subsample size was increased. To achieve this, we borrowed the concept of "turnover" which is used in community ecology to quantify the similarity of proportional species abundances between pairs of sites. In our context, the "stability" of relative influence metrics p between two BRT emulators j was calculated as:

$$\text{stability of sensitivity measures} = e^{\left\{\left(\sum_{j=1}^{2}\sum_{i=1}^{S} p_{ij}\ln p_{ij}\right)/\left(\sum_{i=1}^{S} p_{i\cdot}\ln p_{i\cdot}\right)\right\}}$$

where $p_{i\cdot}$ is the relative influence of parameter i averaged across both emulators (De'ath 2012). This stability measure converges to 1 as the relative influence metrics become more similar and can therefore be used to evaluate how many parameter samples are required to produce robust sensitivity measures. We compared conclusions derived using these methods to those reached by validating emulators against the validation datasets.

Results

Sensitivity-analysis designs that favored a higher number of samples (i.e., more parameter sets and fewer stochastic iterations per set) proved the most appropriate for describing variation in the output of the reintroduction model. For simplicity, we focus on the results for sensitivity analyses that used Latin hypercube sampling here. (This approach is compared to random sampling below.) In general, sensitivity-analysis performance decreased as the complexity of the analysis (i.e., the number of parameters included) increased (Figs. 2 and 3). For sensitivity analyses with a complexity of 5, for example, the computationally least-demanding design (100 parameter samples and just a single iteration per sample) was sufficient to produce high validation scores. These validation metrics ranged from 0.891–0.934 (for population growth rate, r) and 0.729–0.789 (for the probability of extinction, e) depending on the emulator used (Figs. 2a and 3a). However, these ranges fell to 0.667–0.769 and 0.387–0.598, respectively, for the same simple design when the complexity of the sensitivity analysis was increased to 20 separate parameters (Figs. 2a and 3a). If the total number of simulations increased to 100,000 (representing a 1000-fold increase in simulation time) for the 20-parameter case, but the number of parameter samples was fixed at 100, then the validation results were only improved marginally, with ranges of 0.680–0.800 (r) and 0.603–0.623 (e) for sensitivity analyses (Figs. 2g and 3g).

As the number of Latin hypercube samples increased >100, two clear results emerged. First, BRT models (that included interactions) were the most successful at describing variation in simulation outputs across the multidimensional parameter space (Figs. 2g–j and 3g–j). The performance of the linear models and non-linear GAM were similar to each other, although the latter slightly outperformed the former (Figs. 2g–j and 3g–j). Second, for any given number of simulations in the sensitivity-analysis designs, maximizing the number of parameter samples (i.e., taking as many parameter samples as simulations and running a single iteration per sample) was clearly supported. For example, for sensitivity analyses involving 100,000 simulations, only 100 parameter samples (with 1000 stochastic iterations per sample), and the highest complexity of 20, a BRT emulator with a tree complexity of 5 produced mean validation scores of 0.680 and 0.603 for sensitivity analyses on r and e, respectively (Figs. 2g and 3g). By comparison, modifying the latter design to take 100,000 samples instead (and running 1 iteration per sample) improved these mean validation scores to 0.954 and 0.767, respectively. The corresponding ranges for these validation scores obtained from the five replicates were small (0.942–0.961 and 0.740–0.790, respectively), indicating that these results were robust to random variation in the Latin hypercube samples taken to produce the training and validation datasets.

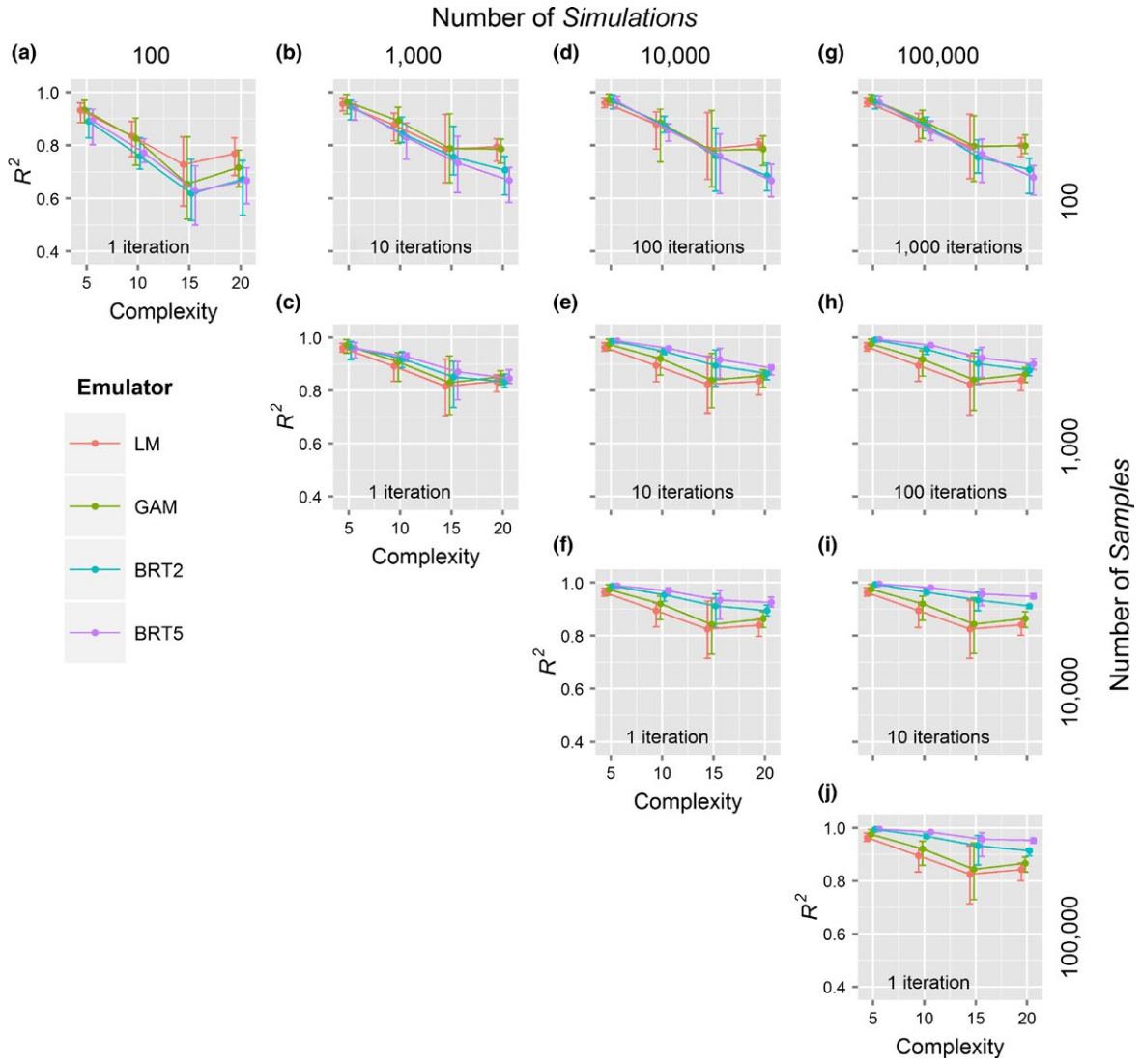

Fig. 2. Performance of different sensitivity-analysis designs that used simulated population growth rate (r) as the response variable. Points denote the mean validation R^2 and whiskers denote ranges calculated from five different parameter combinations within each complexity level (higher values represent superior performance). Abbreviations of emulators fitted to sensitivity analysis output are: LM, linear model; GAM, generalized additive model; BRT2 and BRT5, boosted regression trees with tree complexities of 2 and 5, respectively.

The improvement in validation scores for Latin hypercube sampling relative to random sampling was small and, as expected, of most importance to those sensitivity-analysis designs that relied on few parameter samples (Fig. 4). Assuming a BRT emulation of a sensitivity analysis with a complexity of 20, 100 parameter samples and 1000 iterations per sample, for example, mean validation scores were slightly higher for designs that used Latin hypercube sampling (r: 0.680 vs. 0.671; e: 0.603 vs. 0.589). However, these minor differences were no longer evident when we used 100 000 parameter samples, because at this sampling level the Latin hypercube approach converges on random (Fig. 4).

Given the strong performance of BRT fitted to the sensitivity analysis output, we used the relative influence metrics derived from these fits to quantify the sensitivity of the simulated population growth rate to parameter inputs. We then

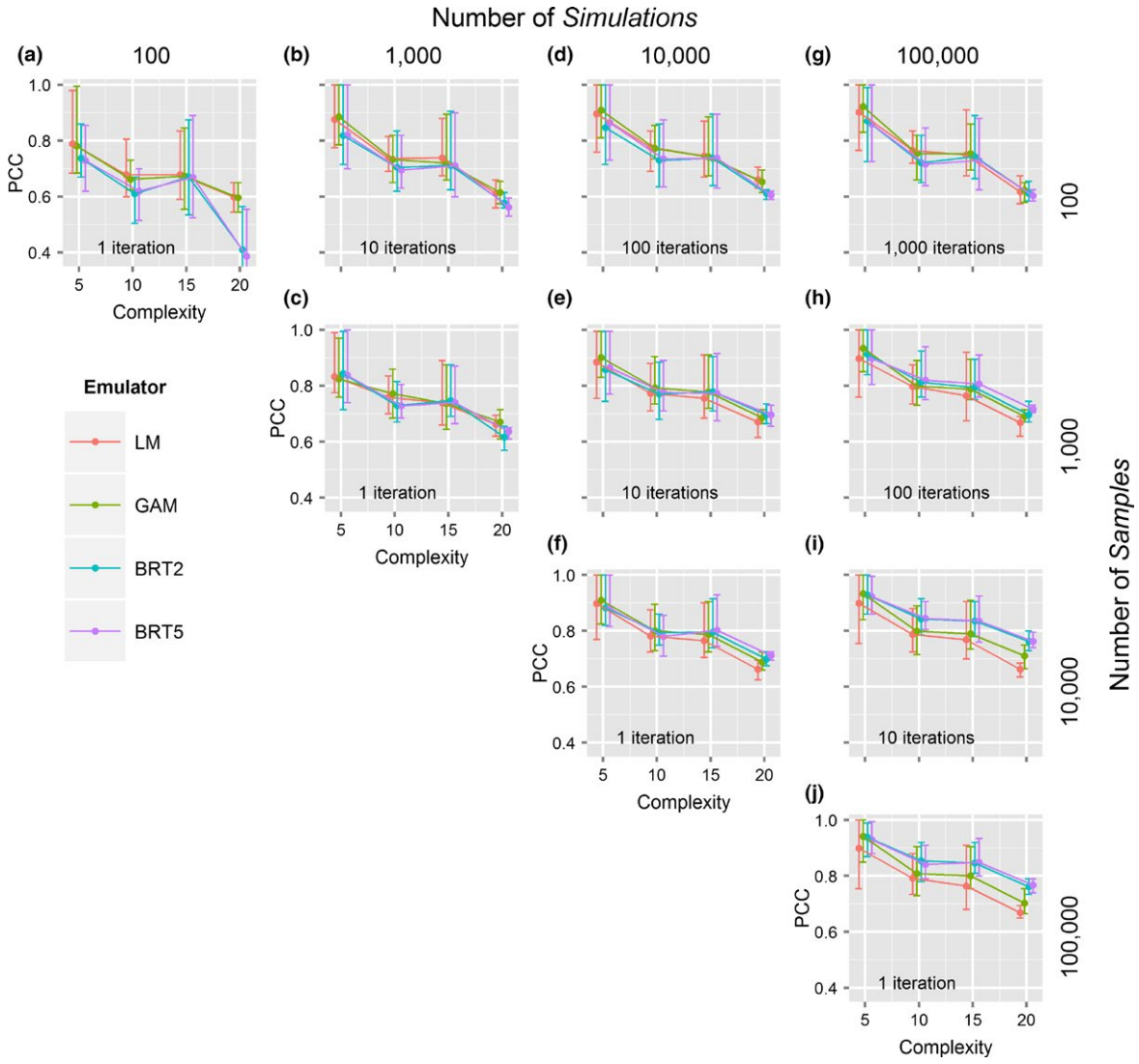

Fig. 3. Performance of different sensitivity-analysis designs that used simulated probability of extinction as the response variable. The validation metric is the proportion of correctly classified cases (PCC). All other details are as for Fig. 2.

compared the mean and variability of these metrics across sensitivity-analysis designs for models with a complexity of 20 (i.e., all parameters included). Designs limited to 100 parameter samples yielded highly variable parameter sensitivities, even when we ran many total simulations (Fig. 5). Assuming 100 samples for sensitivity analysis on *r*, for example, ranges for the relative influence of parameter *s1plus* (survival rate of uninfected individuals aged ≥1 yr) across replicates were 14.3–47.7% (100 simulations), 40.9–49.2% (1000 simulations), 40.9–53.1% (10,000 simulations), and 42.0–52.9% (100,000 simulations). In contrast, increasing the number of samples improved the reliability of these parameter sensitivities; for example, 100,000 samples and 100,000 simulations produced stable relative influence metrics for parameter *s1plus* with a range <1% (40.5–41.1%). There was also evidence that mean relative influence metrics converged as more Latin hypercube samples were taken (Fig. 5), paralleling the improvement in validation values.

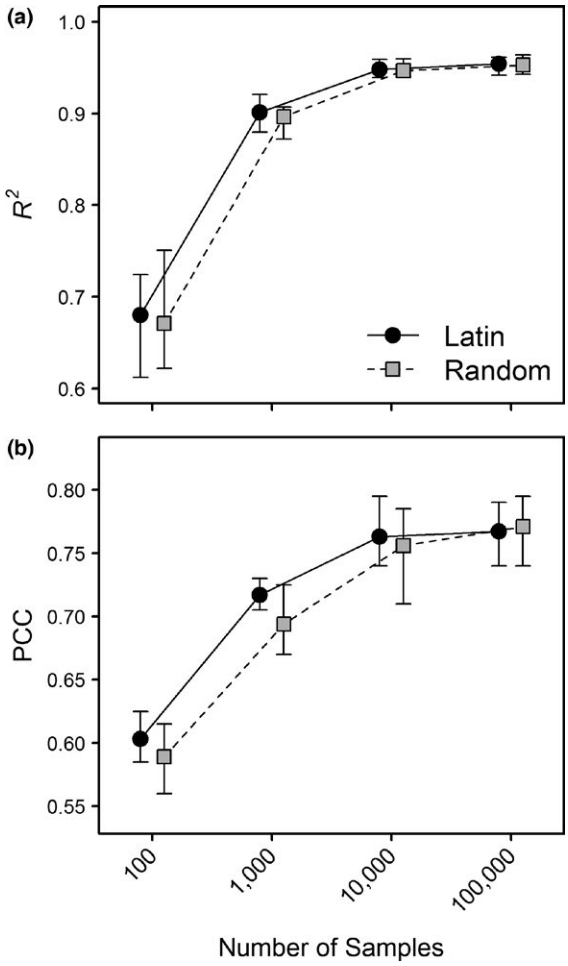

Fig. 4. Performance of sensitivity-analysis designs based on Latin hypercube or random sampling of the parameter space. (a) Validation R^2 for designs that used simulated population growth rate (r) as the response variable. (b) The proportion of correctly classified cases (PCC) for designs that used the probability of extinction (e) as the response. These results are for sensitivity-analysis designs with only a single iteration per parameter sample (i.e., total number of simulations = number of samples). The emulator is a BRT with a tree complexity of 5. All other details are as for Fig. 2.

We investigated changes in emulation-based cross-validation deviance, as well as the stability of the sensitivity measures, as parameter sampling increased (Fig. 6). Using this approach, conclusions regarding the choice of an appropriate emulator and the necessary coverage of the parameter space paralleled those derived from the more laborious approach of producing new validation datasets from the re-introduction model. Using r as the response variable, for example, the cross-validation deviance and sensitivity measures stabilized by approximately 40 000 samples, and the cross-validation results support a BRT emulator with a tree complexity of 5 (Fig. 6a, c). Similarly, validation metrics converged by around 40,000 samples and supported a BRT emulator with substantial interaction depth (Fig. 6e).

Discussion

Global sensitivity analysis is underused in ecology (Naujokaitis-Lewis et al. 2009, Coutts and Yokomizo 2014), perhaps because there is little objective guidance or general rules of thumb on how it should be applied to stochastic ecological models. The approach we have evaluated here requires modelers to: (1) choose a sensitivity-analysis design, evaluating trade-offs between the total execution time and the number of simulations run, as well as between the number of parameter samples and the number of iterations per sample; and (2) select an emulator with which to derive sensitivity measures and other descriptors of the simulated system (e.g., response curves, interaction plots). Given our results, we suggest the following protocol for applying global sensitivity analysis to stochastic ecological models.

Step 1: Run a single simulation per parameter sample

We found that global sensitivity-analysis designs that concentrated simulation effort on covering the multidimensional parameter space outperformed those that invested more computational time in numerous model iterations per parameter sample. The former strategy increased the capacity of emulators to capture the true dynamics of the simulation model, both for the mean population growth rate and probability of extinction (Figs. 2 and 3), and also stabilized emulation-based sensitivity measures (Fig. 5). To maximize coverage of the parameter space with minimum computational effort, we therefore recommend that just one model iteration should be run per sample. This result challenges the conventional wisdom that

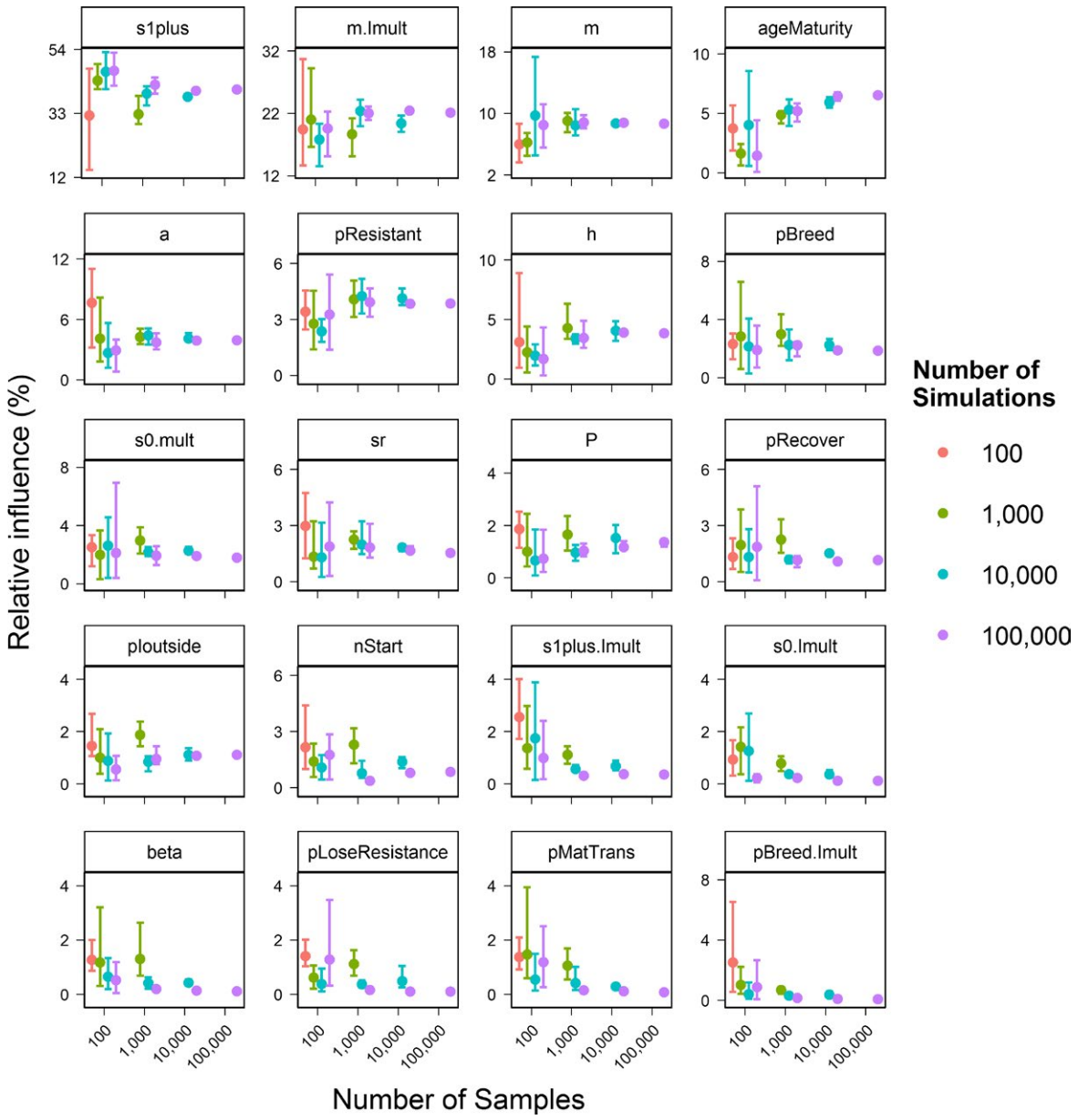

Fig. 5. Sensitivity of the simulated population growth rate to different parameter inputs. This figure plots the relative influence of model parameters derived from BRT emulators fitted to the output of different sensitivity-analysis designs. Latin hypercube sampling was used, and the emulator is a BRT with a tree complexity of 5. Points denote means and whiskers denote ranges calculated from five replicates of each design for a complexity level of 20 (i.e., all input parameters included in the sensitivity analyses). Note that the ranges plotted are obscured by points in some cases, and that these ranges reflect variation due to the sampling of five different Latin hypercubes. The *y*-axis scales are different for each panel.

multiple iterations of stochastic ecological models are required to derive probabilistic outcomes (Lacy 1993, Brook 2000, Sabo 2008). Note however that probabilistic outputs (e.g., the probability of population extinction or decline) can be emulated for sensitivity analyses that use one simulation per sample. Simply, a binary output (e.g., extinction or persistence; decline

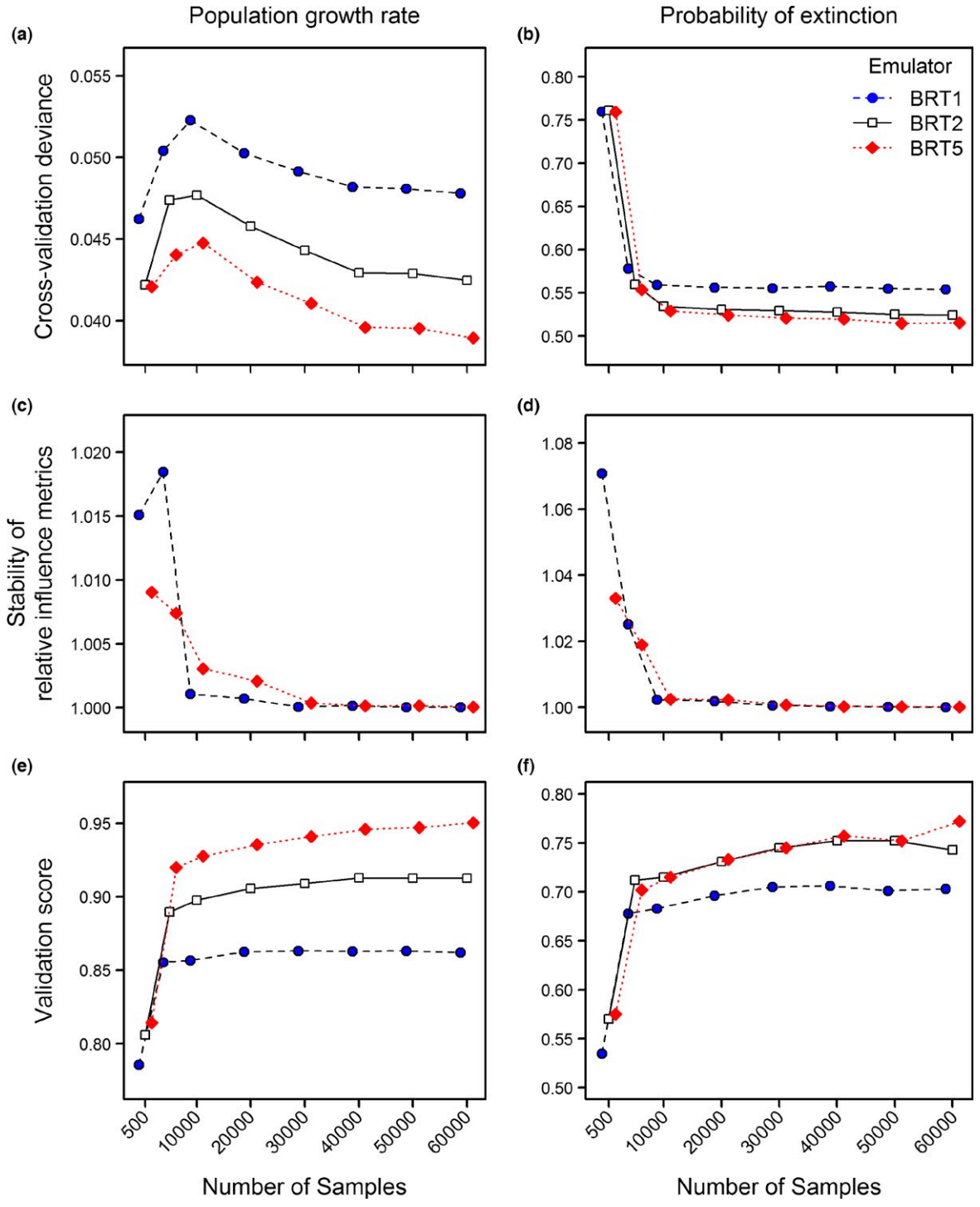

Fig. 6. Choosing an appropriate emulator. Shown are the cross-validation deviance (a, b) and the stability of sensitivity measures (c, d) derived from BRT emulators as the number of parameter samples is increased. Results are shown for emulation of the two response variables (population growth rate and probability of extinction). Validation metrics (e, f) are also shown for each response (validation R^2 and the proportion of correctly classified cases, respectively). BRT1, BRT2 and BRT5 represent boosted regression tree emulators with tree complexities of 1, 2 and 5, respectively. For clarity, validation results from the BRT2 emulator are excluded from panels c and d.

or increase) for each sample can be modeled using a binomial error distribution and probabilities on the continuous 0–1 scale can then be predicted for any combination of input parameters (Prowse et al. 2014).

Of the two commonly used sampling regimes we evaluated (Latin hypercube and random sampling), the former attempts to produce samples that are more representative of the multidimensional parameter space, but is more difficult to implement. We found that relative improvements in model emulation due to Latin hypercube sampling were small and limited to sensitivity-analysis designs that used small numbers (i.e., hundreds) of parameter samples (Fig. 4). Although Latin hypercube sampling should be favored *a priori*, we therefore anticipate that using this strategy will be most essential for computationally demanding models. In such cases, a small, feasible number of samples can be derived in the first instance and the adequacy of this sampling level then evaluated (see Step 3).

Sensitivity measures are conditional on the ranges and probability distributions from which input parameters are sampled (Cariboni et al. 2007). Our study drew parameter samples from uniform distributions with defined ranges, but other distributions could be used based on empirical studies, such as posterior distributions for parameters derived from a Bayesian analysis or expert knowledge (O'Hara et al. 2002, Tenhumberg et al. 2004, Kéry and Schaub 2012, Merow et al. 2014). As ecological models increase in complexity, however, more input parameters are required about which there is often little or imprecise empirical information. We anticipate that, for the most complex models, sampling parameters from uniform distributions with wide but plausible ranges will be an attractive option (e.g., Cassey et al. 2014).

Step 2: Emulate subsamples of the sensitivity-analysis output with a candidate set of emulators

Because sensitivity measures and other descriptors of the model are generated from the emulator, it is critical to select an appropriate emulation method. Although candidate emulators might represent different techniques (see Marie and Simioni 2014 for additional emulator options that we did not evaluate here), we anticipate that modelers will often wish to test emulators incorporating different interaction depths (Fig. 6). Flexible, machine-learning techniques are suitable for this purpose – for example, boosted regression trees (BRT) automatically fit nonlinear relationships and interactions of different depths moderated by the tree complexity specified (a tree complexity of 1 produces an additive model; Elith et al. 2008). Of course, linear models can incorporate interactions and some nonlinearity by including polynomial terms, and selecting these components within a linear modelling framework is certainly possible for simple simulation models (e.g., McCarthy et al. 1995). For complex simulations with many parameters, however, this approach is less feasible and interpretable because of the large number of terms required for emulation using linear modeling. In contrast, relative influence metrics from BRT provide a single, interpretable sensitivity measure for each input parameter that includes the contribution of that parameter to main effects and interactions.

When constructing a sensitivity analysis for complex simulations, ecologists rarely confirm that parameter sampling is sufficient for their chosen emulator to capture the input-output relationship of the simulation model. Emulation for a range of subsamples of size n, where n is less than or equal to the total number of parameter samples, can be used to plot changes in emulation-based cross-validation performance and in sensitivity measures as samples are added. This provides a standard methodology for evaluating different emulators and the coverage of the parameter space. For example, assuming the probability of population extinction as the focal output of our hypothetical model, these plots demonstrated that 10 to 20 thousand parameter samples provided sufficient coverage, while a BRT emulator incorporating first-order interactions was appropriate (Fig. 6b, d). With the population growth rate as the focal output, however, 40,000 parameter samples were required to stabilize these measures while emulation incorporating higher-order interactions was supported (Fig. 6a, c). In this latter example, a modeler would necessarily trade off the ecological interpretability of emulation limited to first-order interactions against the increase in performance afforded by greater interaction depth.

Whereas the fitting time for statistical emulators such as linear models and GAM is negligible, one potential disadvantage of machine-learning emulators like random forests and BRT it that they can be computationally expensive for large datasets. For example, assuming sensitivity analyses of our model that included all 20 parameters and drawing 50,000 parameter samples, BRT emulation for the population growth rate cost CPU times of 17 and 37 min (on a 2.7 GHz processor), for tree complexities of 1 and 5, respectively. We recognize that such computational cost could lead some practitioners to favor speedier emulators (e.g., Lehuta et al. 2010), but also note that machine-learning approaches can be tuned to individual datasets by adjusting fitting parameters. When fitting BRT using the function gbm.step in R, for example, CPU times can be reduced by adjusting parameters that affect the speed with which the optimal number of trees is selected (e.g., the number of cross-validation folds, the number of trees to add each cycle, and the threshold improvement in predictive deviance required to continue adding more trees) (Hijmans et al. 2013). BRT also require specification of a "learning rate" that is used to shrink the contribution of each fitted tree to the final model and can be optimized for different datasets (Elith et al. 2008). In our study, optimizing learning rates was not practicable because our comprehensive experimental design required 800 BRT emulations; however, such optimization is feasible for typical cases when a single sensitivity analysis is required.

Step 3: Increase parameter sampling if required and report sensitivity measures

If parameter sampling insufficient for convergence of cross-validation and sensitivity measures, additional parameter sampling and simulation can be done (i.e., repeat Steps 1 and 2). When using Latin hypercube sampling, hypercubes can be augmented while maintaining the "Latin" properties of the sampling design (Carnell 2012). Assuming emulation plots are deemed satisfactory, sensitivity measures derived by emulating the complete sensitivity analysis output can then be reported and critical and redundant parameters can then be identified. For example, results from 100,000 simulations indicated that 6 parameters from our toy model contributed a total relative influence of less than 1% (Fig. 5) and could therefore be fixed at nominal values with negligible effects on the simulation outputs. Further, since one primary advantage of global sensitivity analysis is to facilitate the interpretation of complex models, the emulator should be used to produce summary plots of key response curves and interactions (e.g., McCarthy et al. 1995, Prowse et al. 2015).

This protocol is specific to global sensitivity analyses for which mean or probabilistic simulation outputs are the primary focus. We expect this to be true for many applied ecological models; however, different sensitivity-analysis designs might be required to characterize how the variance of a simulated output changes across the parameter space. The validation-based approach we have taken here could also be used to test our protocol for sensitivity analyses on multivariate outputs (Vinatier et al. 2013). Further, we only considered sensitivity-analysis designs for a single model structure, whereas the predictions and management recommendations derived from ecological models are also sensitive to the model structure used (Hosack et al. 2008). However, we used uniform parameter ranges with a lower bound of zero for many parameters (Table 1), which provides one method of testing parameter sets that effectively exclude some processes. For example, parameter samples that specified a low probability of disease infection from an outside source (*pIoutside*) would usually produce simulations with no disease component. An alternative method to test different model structures is to construct the model such that the inclusion/exclusion of different processes is controlled by a "switch" (i.e., a binary variable) that is then included as an input parameter at the sampling stage (for an example, see Prowse et al. 2014). Emulation of such designs is usually more complicated, however, because each simulated process requires some input parameters that become redundant when that process is excluded from the model.

Conclusion

As simulation models are increasingly used to study ecological systems and inform management decisions, a challenge is to choose the

model complexity necessary to represent biological processes adequately, whilst respecting the objective of parsimony (Green et al. 2005). Global sensitivity analysis can assist with model simplification, but in the ecological context, decisions regarding the total number of simulations, parameter samples, iterations per sample and subsequent emulation have so far largely reflected the ad hoc preferences of individual researchers. Our protocol offers a standardized methodology for implementing global sensitivity analyses for complex, stochastic models in a computationally efficient manner. We tested these approaches on a simulation model designed for speed because our experimental design required many more simulations (c. 18 million) than would usually be run for an applied modelling study. However, we anticipate that this protocol will appeal particularly to practitioners using computationally demanding approaches, including individual-based models. For extremely complex models, Steps 1 and 2 of our protocol can be implemented and the adequacy of parameter sampling evaluated and reported, even if increasing parameter sampling (Step 3) is not feasible.

Acknowledgments

This research was funding by an Australian Research Council Discovery Grant to B.W.B., P.C. and C.J.A.B. (DP120101019) and an US NSF grant to R.C.L. and H.R.A. (DEB-1146198).

Literature Cited

Boyce, M. S. 1992. Population viability analysis. Annual Review of Ecology and Systematics 23:481–506.

Bradshaw, C. J. A., C. R. McMahon, P. S. Miller, R. C. Lacy, M. J. Watts, M. L. Verant, J. P. Pollak, D. A. Fordham, T. A. A. Prowse, and B. W. Brook. 2012. Novel coupling of individual-based epidemiological and demographic models predicts realistic dynamics of tuberculosis in alien buffalo. Journal of Applied Ecology 49:268–277.

Brook, B. W. 2000. Pessimistic and optimistic bias in population viability analysis. Conservation Biology 14:564–566.

Brook, B. W., J. J. O'Grady, A. P. Chapman, M. A. Burgman, H. R. Akcakaya, and R. Frankham. 2000. Predictive accuracy of population viability analysis in conservation biology. Nature 404:385–387.

Cariboni, J., D. Gatelli, R. Liska, and A. Saltelli. 2007. The role of sensitivity analysis in ecological modelling. Ecological Modelling 203:167–182.

Carnell, R. 2012. lhs: Latin Hypercube Samples. R package version 0.10. http://CRAN.R-project.org/package=lhs.

Cassey, P., T. A. A. Prowse, and T. M. Blackburn. 2014. A population model for predicting the successful establishment of introduced bird species. Oecologia 175:417–428.

Conroy, S. D. S., and B. W. Brook. 2003. Demographic sensitivity and persistence of the threatened white- and orange-bellied frogs of Western Australia. Population Ecology 45:105–114.

Coutts, S. R., and H. Yokomizo. 2014. Meta-models as a straightforward approach to the sensitivity analysis of complex models. Population Ecology 56:7–19.

De'ath, G. 2012. The multinomial diversity model: linking Shannon diversity to multiple predictors. Ecology 93:2286–2296.

Drechsler, M. 1998. Sensitivity analysis of complex models. Biological Conservation 86:401–412.

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

Fang, K.-T., R. Li, and A. Sudjianto. 2006. Design and modeling for computer experiments. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Ginzburg, L. R., and C. X. J. Jensen. 2004. Rules of thumb for judging ecological theories. Trends in Ecology and Evolution 19:121–126.

Green, J. L., et al. 2005. Complexity in ecology and conservation: mathematical, statistical, and computational challenges. BioScience 55:501–510.

Harris, R. B., L. A. Maguire, and M. L. Shaffer. 1987. Sample sizes for minimum viable population estimation. Conservation Biology 1:72–76.

Hastie, T. J., and R. J. Tibshirani. 1990. Generalized additive models. Chapman and Hall/CRC, New York, USA.

Hastings, A., P. Arzberger, B. Bolker, S. Collins, A. R. Ives, N. A. Johnson, and M. A. Palmer. 2005. Quantitative bioscience for the 21st century. BioScience 55:511–517.

Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith. 2013. dismo: Species distribution modeling. R package version 0.9-3. http://CRAN.R-project.org/package=dismo.

Hosack, G. R., K. R. Hayes, and J. M. Dambacher. 2008. Assessing model structure uncertainty through an analysis of system feedback and Bayesian networks. Ecological Applications 18:1070–1082.

Kéry, M., and M. Schaub. 2012. Bayesian population analysis using WinBUGS. Academic, Oxford, UK.

Lacy, R. C. 1993. VORTEX—a computer-simulation model for population viability analysis. Wildlife Research 20:45–65.

Lacy, R. C., P. S. Miller, P. J. Nyhus, J. P. Pollak, B. E. Raboy, and S. L. Zeigler. 2013. Metamodels for transdisciplinary analysis of wildlife population dynamics. PLoS One 8:e84211.

Lehuta, S., S. Mahevas, P. Petitgas, and D. Pelletier. 2010. Combining sensitivity and uncertainty analysis to evaluate the impact of management measures with ISIS-Fish: marine protected areas for the Bay of Biscay anchovy (*Engraulis encrasicolus*) fishery. ICES Journal of Marine Science 67:1063–1075.

Levins, R. 1966. Strategy of model building in population biology. American Scientist 54:421–431.

Lindenmayer, D. B., and H. P. Possingham. 1995. Modeling the viability of metapopulations of the endangered Leadbeater's possum in southeastern Australia. Biodiversity and Conservation 4:984–1018.

Marie, G., and G. Simioni. 2014. Extending the use of ecological models without sacrificing details: a generic and parsimonious meta-modelling approach. Methods in Ecology and Evolution 5:934–943.

McCarthy, M. A., M. A. Burgman, and S. Ferson. 1995. Sensitivity analysis for models of population viability. Biological Conservation 73:93–100.

McKay, M. D., R. J. Beckman, and W. J. Conover. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245.

Melbourne, B. A., and A. Hastings. 2008. Extinction risk depends strongly on factors contributing to stochasticity. Nature 454:100–103.

Merow, C., A. M. Latimer, A. M. Wilson, S. M. McMahon, A. G. Rebelo, and J. A. Silander. 2014. On using integral projection models to generate demographically driven predictions of species' distributions: development and validation using sparse data. Ecography 37:1167–1183.

Naujokaitis-Lewis, I. R., J. M. R. Curtis, P. Arcese, and J. Rosenfeld. 2009. Sensitivity analyses of spatial population viability analysis models for species at risk and habitat conservation planning. Conservation Biology 23:225–229.

O'Hara, R. B., E. Arjas, H. Toivonen, and I. Hanski. 2002. Bayesian analysis of metapopulation data. Ecology 83:2408–2415.

Prowse, T. A. A., C. N. Johnson, C. J. A. Bradshaw, and B. W. Brook. 2014. An ecological regime shift resulting from disrupted predator-prey interactions in Holocene Australia. Ecology 95:693–702.

Prowse, T. A. A., C. N. Johnson, P. Cassey, C. J. A. Bradshaw, and B. W. Brook. 2015. Ecological and economic benefits to cattle rangelands of restoring an apex predator. Journal of Applied Ecology 52:455–466.

Prowse, T. A. A., C. N. Johnson, R. C. Lacy, C. J. Bradshaw, J. P. Pollak, M. J. Watts, and B. W. Brook. 2013. No need for disease: testing extinction hypotheses for the thylacine using multi-species metamodels. Journal of Animal Ecology 82:355–364.

Ratto, M., A. Castelletti, and A. Pagano. 2012. Emulation techniques for the reduction and sensitivity analysis of complex environmental models. Environmental Modelling and Software 34:1–4.

R Development Core Team. 2014. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Sabo, J. L. 2008. Population viability and species interactions: life outside the single-species vacuum. Biological Conservation 141:276–286.

Shaffer, M. L. 1981. Minimum population sizes for species conservation. BioScience 31:131–134.

Sobol, I. M. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation 55:271–280.

Storlie, C. B., L. P. Swiler, J. C. Helton, and C. J. Sallaberry. 2009. Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. Reliability Engineering and System Safety 94:1735–1763.

Tenhumberg, B., A. J. Tyre, A. R. Pople, and H. P. Possingham. 2004. Do harvest refuges buffer kangaroos against evolutionary responses to selective harvesting? Ecology 85:2003–2017.

Traill, L. W., C. J. A. Bradshaw, and B. W. Brook. 2007. Minimum viable population size: a meta-analysis of 30 years of published estimates. Biological Conservation 139:159–166.

Turyani, T., and H. Rabitz. 2000. Local methods. Pages 81–100 *in* A. Saltelli, K. Chan, and E. M. Scott, editors. Sensitivity analysis: gauging the worth of scientific models. John Wiley and Sons, West Sussex, UK.

Valle, D., C. L. Staudhammer, W. P. Cropper, and P. R. van Gardingen. 2009. The importance of multimodel projections to assess uncertainty in projections from simulation models. Ecological Applications 19:1680–1692.

Vinatier, F., M. Gosme, and M. Valantin-Morison. 2013. Explaining host-parasitoid interactions at the landscape scale: a new approach for calibration and sensitivity analysis of complex spatio-temporal models. Landscape Ecology 28:217–231.

Wainwright, H. M., S. Finsterle, Y. J. Jung, Q. L. Zhou, and J. T. Birkholzer. 2014. Making sense of global

sensitivity analyses. Computers and Geosciences 65:84–94.

Wood, S. N. 2006. Generalized additive models: an introduction with R. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. Journal of the Royal Statistical Society B 73:3–36.

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.1238/supinfo