2019

Experimental Design Principles to Choose the Number of Monte Carlo Replicates for Stochastic Ecological Models

M.C. Kennedy
University of Washington Tacoma, mkenn@uw.edu

Title: Experimental design principles to choose the number of Monte Carlo replicates for stochastic ecological models

Maureen C. Kennedy (corresponding author)
University of Washington, Tacoma. School of Interdisciplinary Arts and Sciences, Division of Sciences and Mathematics. 1900 Commerce St., Tacoma, WA, 98402.
mkenn@uw.edu

Abstract:

Ecologists often rely on computer models as virtual laboratories to evaluate alternative theories, make predictions, perform scenario analysis, and to aid in decision-making. The application of ecological models can have real-world consequences that drive ecological theory development and science-based decision and policy-making, so it is imperative that the conclusions drawn from ecological models have a strong, credible quantitative basis. In particular it is important to establish whether any predicted change in a model output has ecological and statistical significance. Ecological models may include stochastic components, using probability distributions to represent some modeled processes. An individual run of a stochastic ecological model is a random draw from an infinitely large population, requiring replicate simulations to estimate the distribution of model outcomes. An important consideration is the number of Monte Carlo replicates necessary to draw useful conclusions from the model analysis. A simple framework is presented that borrows from well-understood techniques for experimental design, including confidence interval estimation and sample size power analysis. The desired precision of interval estimates for model prediction, or the minimum desired detectable effect size between scenarios, is established by the researcher in the context of the model objectives and the ecological system. The number of replicates required to achieve that level of precision or detectable effect is computed given an estimate of the variability in the model outcomes of interest. If the number of replicates is computationally prohibitive, then the expected precision or detectable effect for that sample size should be reported. An example is given for a stochastic model of fire spread integrated with an eco-hydrological model.

Keywords: stochastic simulation; confidence interval; prediction interval; inference; estimation

1. Introduction

An ecological model is an abstraction of a real-world system that represents, using mathematical relationships, rules, and computer code, out best understanding of how that system functions. Even if an ecologist has no experience in developing mathematical models or writing computer code, they often use existing ecological models as virtual laboratories to evaluate alternative hypotheses, to inform experimental design, to make predictions for future states of a system, to perform scenario analysis, and to aid in decision-making for environmental and resource management. Models are increasingly used for purposes such as informing regulatory guidelines (National Research Council, 2007), for conservation and natural resource management (e.g., Fieberg and Ellner, 2001), and to predict ecological consequences of climate change (e.g., Keane et al., 2001). There is a corresponding need for defensible standards of model development, use, documentation, and interpretation of ecological model predictions (Grimm et al., 2006; Jakeman et al., 2006; Schmolke et al., 2010).

In general, ecological models are either deterministic or stochastic. For a deterministic model, replicate simulations with the same inputs and parameters give identical model predictions. In a stochastic model, probability distributions represent some modeled processes, such that replicate simulations with the same inputs and parameters give variable model predictions. In that manner stochastic simulations use probability structures to represent uncertainty in the modeled processes and input data, yielding distributions of model outputs rather than point estimates. For example, WMfire is a stochastic model of fire spread (Kennedy et al., 2017) coupled with a deterministic eco-hydrological model (RHESSys; Tague and Band 2004). With a randomly located ignition point, and spread governed by probability structures informed by the underlying landscape, replicate simulations on identical landscapes result in variable fire areas. Across multiple WMFire simulations we then can describe a distribution of fire occurrence rather than a single realization.

A consequence of implementing stochastic processes in an ecological model is that each individual simulation is a single random draw from an infinitely large population of possible outcomes. It follows that, regardless of the overarching model objective, a single run of a stochastic model is insufficient to characterize a model prediction. Suppose a single realization of WMFire estimated a mean 200 ha burned per year under baseline conditions, and a single realization predicted a mean 350 ha burned per year under a scenario of reduced precipitation. It is impossible to know whether the predicted change in mean area burned is a model response to the change in climate or if it would be expected under the random variability of WMFire.

Commonly we take a Monte Carlo approach, where for a given scenario multiple independent model replicates are simulated (N), giving a distribution of model predictions. In the above toy example, instead of single run we might perform 100 replicate simulations in each scenario (baseline, reduced precipitation) and obtain a mean value of 200 ha with a standard error of 10 ha for the baseline condition, and a mean value of 350 ha with a standard error of 15 ha for the reduced precipitation condition. In this case, given the documented variability in WMFire predictions of mean area burned per year we can conclude that WMFire predicts increased area burned with reduced precipitation. This leads inevitably to the question: how many Monte Carlo replicate simulations do I need to satisfy my modeling objectives? For example, Kennedy et al. (2017) use 500 replicate WMFire simulations to assess the model of fire spread against expected fire regimes at two different watersheds.

The choice of Kennedy et al. (2017) to use 500 Monte Carlo replicate simulations without evaluation of the underlying stochastic model variability is an example of a common *ad* *hoc* approach: choose an arbitrarily large (*sensu* Byrne 2013) number of replicate simulations, without an accompanying quantitative justification. A brief survey of recently published modeling studies (Appendix A) illustrates that this is the most common technique (Fig A.1). Alternatively, under severe computational constraints, we simulate as many replicates as possible without quantifying the uncertainty associated with a small sample size. Adapting the statistical principles of experimental design to stochastic ecological modeling may provide a more robust alternative to the current *ad hoc* approaches.

When considering the number of replicate Monte Carlo simulations, we are concerned with both estimation of mean model outputs, as well as the effect size when comparing some modeling scenario to a baseline. As with empirical studies with large sample sizes, the more Monte Carlo replicates are produced the smaller is the effect size that can be detected statistically. The fewer the number of Monte Carlo replicates the more difficult it is to distinguish actual predicted effects from random variability, an issue if the model is computationally intensive. When planning a modeling study using a stochastic ecological model, we need to determine the number of Monte Carlo replicates necessary to conclude if the mean system state is predicted to change in a way that is both meaningful (the change in mean has practical effect on the system) and significant (the change in mean is different than zero, relative to the standard error). To answer the question of how many replicate simulations, we can expand the idea of applying a design of experiments approach for modeling studies (Lorscheid et al., 2012).

The objective here is to suggest an alternative to the *ad hoc* approach in determining the number of replicate simulations of a stochastic ecological model. To that end a general framework is presented (Fig. 1) for a thoughtful quantitative analysis of the number of simulations necessary to achieve a pre-specified level of precision in stochastic model outputs, and to use that in study development. When presenting a modeling study, the reporting of mean model estimates, the variability in model estimates, and the distribution of model estimates should all be standard practice. The application of this framework is illustrated with an example using WMFire to compare fuel loading and moisture condition scenarios.

2. Methods

2.1. WMFire description

WMfire is a stochastic model of fire spread (Kennedy et al., 2017) coupled with a deterministic eco-hydrological model (RHESSys; Tague and Band 2004). The overarching objective of the coupled model is to predict and understand fire and watershed dynamics under climate change and management scenarios. A full description of WMFire can be found at Kennedy et al. (2017), here we give a brief overview. RHESSys calls WMFire once each month, sending pixel-defined values for litter loading, relative moisture deficit (calculated from the ratio of actual evapotranspiration (ET) to potential evapotranspiration (PET); 1-ET/PET), and the digital elevation model. WMFire draws a random number of ignitions from a Poisson distribution, and random ignition pixel is located uniformly on the grid for each ignition. The ignition starts a fire according to a probability determined by the litter load and relative deficit of the ignition pixel. If the fire start is successful, fire spread proceeds iteratively by testing the neighbors of newly ignited cells against a probability of spread, calculated from the litter load and relative deficit of the neighboring pixel, and the slope and wind direction between the newly burned cell and its neighbor, relative to the direction of spread. Fire spread continues until either all tests of spread fail, or the fire spans the grid. WMFire returns to RHESSys the grid with the probability of spread associated with any burned pixels. RHESSys interprets this grid to implement any fire effects on the burned pixels.

To characterize the expected variability in model outputs (Y_k) given the stochastic contribution of WMFire to RHESSys, we run WMFire in uni-directional coupling with RHESSys. This saves computation time, where WMFire receives inputs from RHESSys, but does not modify RHESSys dynamics (as in Kennedy et al. 2017). For this example modeling study we choose the Santa Fe watershed located in New Mexico, USA, with a mean ignition rate of 2/month (see Kennedy et al. 2017 for a description of the watershed and simulation structure).

2.2. Model scenario description

To illustrate how this framework can inform model application, two model scenarios are designed. The goal would be to determine if, for each scenario, model predictions change from the baseline historical condition of Kennedy et al. (2017). The first scenario is an increase of 10% in fuel loading across the landscape all years in the simulation; the second scenario is a 10% decrease in evapotranspiration across the landscape all years in the simulation (representing increased dryness). Next we give an overview of the framework illustrated in Figure 1.

2.3. Framework to determine the number of Monte Carlo Replicates

2.3.1. Define independent model replicate

In order to use standard statistical principles of experimental design, we need to identify a single independent model replicate. For example, in a time series of simulated fire spread in a fully coupled WMFire-RHESSys modeling system, the fire hazard in a given year depends on the past history of fire occurrence. Therefore each simulated year is not independent of other years in the same time series. However, a full time series of fire occurrence would be independent of replicate full time series. In the case of the Santa Fe watershed, WMFire is run from historical climate spanning the years 1941-2008. Each replicate time series repeats the conditions in this timeframe. Therefore we consider an independent model replicate to be a single WMFire time series of fire occurrence. Independent model outputs are then individual summaries of each replicate time series.

2.3.2. Identify model outputs of interest

Model outputs of interest to characterize fire regimes include measures of fire size, the time between fires, and the seasonality of wildfire. The mean annual area burned ($\bar{A}$, ha yr^{-1}) measures, for a single time series, the mean area burned in the watershed per year. The natural fire rotation represents the time it takes to burn an entire watershed of a given size, as the landscape area divided by mean annual area burned (nfr, years). The mean fire return interval is the mean number of years between successive fires at least 100 ha in size (μ_{fri}, years). Seasonality is represented by the probability June is the month with the most fires in a time series. This probability is estimated by the proportion of Monte Carlo replicate time series for which the most fires in the time series occur in June (p_{June}). For these model outputs we consider both *estimation* of mean model predictions, as well as *inference* in the comparison of model predictions among model scenarios.

Estimation is the practice of providing the best estimate of the model output (Y_k), either as a point estimate (e.g., the mean value $\bar{Y}_k$), or as an interval estimate at some level of confidence (1-α). The width or precision of this confidence interval is determined by the population variability (standard deviation, σ) and the sample size (N), where all else being equal a larger sample size gives a narrower confidence interval.

In general, *inference* is the process of rejecting or failing to reject statistical hypotheses (e.g., $\mu_1 = \mu_2$). For a given population variability, sample size for the case of inference determines our power (1-β) to determine statistically a particular effect size (change in estimated value; δ*). For a given power, a larger sample size means we can detect a smaller effect size.

2.3.3. Conduct pilot study to estimate model variability

A common pre-requisite to determine sample size requirements for both inference and estimation is to obtain a value for the population standard deviation (σ), which quantifies the variability in the population. In empirical ecological studies this is often estimated using a pilot study, or from previous measurements in similar systems. For stochastic ecological models this can be accomplished in the process of model development and assessment, or in preparation to use an existing model for a new study. As much as parameter estimation and sensitivity analysis are standard practices for model development, so should be exploratory analysis of the distribution of model outputs with Monte Carlo replicate simulations of a stochastic model. When a model is deemed adequate for application, estimates of model output variability should be included along with parameter estimates and associated uncertainty. For example, a prediction of mean annual burned of 188 ha yr^{-1} is interpreted differently if the standard deviation 52 ha yr^{-1} v. 5 ha yr^{-1}. Information about the variability in the model outputs can then be used to determine appropriate number of simulations for the application of a stochastic ecological model in a more complex factorial design. Ideally the pilot study would be completed in the process of model development, but if it hasn't been conducted then an individual model user should perform their own pilot study.

For the WMFire pilot study 10,000 Monte Carlo replicate simulations were performed at the baseline historical condition of Kennedy et al. (2017) (see Appendix B for details of pilot study), with the model outputs calculated for each replicate simulation. Table 1 gives the mean, standard deviation, and coefficient of variation for each WMFire model output across 10,000 pilot study replicates.

2.3.4. Choose margin of error and/or detectable effect size

Byrne (2013) outlines a strategy for sample size determination for stochastic cognitive models that is based on principles of confidence interval estimation (see also Driels and Shin 2004), which we adapt here. The margin of error (E) can be interpreted as the maximum likely distance between a sample mean and the population mean with some level of confidence ($1-\alpha$). The total width of a confidence interval around the mean value is 2E. A narrower confidence interval may be considered more precise. Byrne (2013) shows that for the purpose of sample size determination, if the coefficient of variation is known then the margin of error can be standardized to estimating the population mean value within some proportion (w) of its true value, without knowing the population mean value. For example, the desired precision might be $w=0.1$, that is that the sample mean value is within 10% of the population mean value. For WMFire we consider estimation within 10% ($w=0.10$) and 5% ($w=0.05$) of the population mean value.

For inference we are interested in the minimum detectable effect (δ^*), the minimum difference in mean predicted value between some baseline scenario and a treatment scenario that is considered to be ecologically significant. Consider a simple 2-sample design, where the stochastic simulation model is used to determine whether the population mean model output (μ, estimated by $\bar{Y}$) is predicted to change between a baseline simulation (control C; μ_C estimated by $\bar{Y}_c$) and a treatment scenario (treatment T; μ_T estimated by $\bar{Y}_T$). The null hypothesis is H_0: $\mu_C = \mu_T$. and δ^* is the minimum difference between population means ($|\mu_C - \mu_T|$) that we are interested in detecting. For the WMFire example, we assume a minimum detectable effect of 20 ha yr^{-1}, 5 years, 0.5 years, and 0.10 for mean annual area burned, natural fire rotation, fire return interval, and the probability that in a time series the most fires occur in June, respectively.

2.3.5a. Number of replicate simulations for estimation

There are two main requirements to use simple statistical methods to determine the number of Monte Carlo replicates. The first is that the replicate Monte Carlo simulations represent a random sample, which can be ensured by a quality random number generator. The second is that the model outputs for each Monte Carlo replicate are independent and identically distributed. This requires the modeler to choose carefully model outputs that meet the requirements (as in choosing measurements that meet these requirements in an empirical study design; see 1, above). The sampling distribution of the estimator must also be determined. In the case of the mean model output, with sufficient replicates we can use the central limit theorem and the normal distribution. That is the approach taken here.

To determine the number of Monte Carlo replicate simulations required to achieve the stated margins of error (within 10% or 5% of the population mean value), we assume through the central limit theorem that the sample mean follows a normal distribution. If your sample size is small, then this assumption may not be valid. Given a standard normal distribution and a specified level of confidence, then the standard normal critical value can be identified ($z_{\alpha/2}$; e.g., for $\alpha = 0.05$, $z_{\alpha/2}$ is 1.96). Using the results of the pilot study, we can estimate the coefficient of variation (CV) as σ/μ for each of our model outputs. Let w be the proportion of the population mean value we are interested in estimating within, then the sample size N can be determined as (Byrne 2013; see Appendix C for derivation):

$$N \geq \left(\frac{z_{\alpha/2}}{w} CV\right)^2 \qquad (1)$$

We use this relationship to determine sample size requirements to achieve a margin of error at a given proportion of the mean size (Byrne, 2013), with varying values of the CV (Figure 2a). Alternatively, for a given CV we can calculate the sample size required to achieve distances of varying proportion from the true mean value (Figure 2b):

$$w \geq \frac{z_{\alpha/2} CV}{\sqrt{N}} \qquad (2)$$

Supplement S1 gives example scripts for the R statistical program (R Core Team, 2017) to determine sample sizes for estimation. Note that Byrne (2013) also provides web-based utilities to calculate sample size requirements (http://chil.rice.edu/research/nomr/, last accessed Dec 17, 2018).

If the model prediction is a proportion, the calculation is somewhat easier to standardize. Here we define E as the maximum likely distance between the population proportion (π) and the sample proportion (p). We know that the standard deviation of the proportion is π(1-π) and the sample size is calculated as:

$$N \geq \left(\frac{z_{\alpha/2}\sqrt{\pi(1-\pi)}}{E}\right)^2 \qquad (3)$$

If π is known, then the standard deviation is known. A conservative approach is to assume π = 0.5, which maximizes the standard deviation for the proportion. Note that this may result in an overestimation of required sample size, as the sample size required to estimate a lower or higher population proportion would be smaller. If there is good prior information for the value of the population proportion then that can be used to determine a reasonable sample size. For example, assuming a proportion of 0.5 results in a sample size requirement of 97 for estimation (with E =0.1). If we assume the proportion to be 0.79, then the required sample size would drop to 64.

2.3.5b. Number of replicate simulations for scenario comparisons (inference)

To determine minimum sample size requirements for scenario comparison we need to specify the significance level (α), the desired power (1-β; the probability of detecting a true effect if one exists), the desired effect size (δ*, $|\mu_C - \mu_T|$), and the standard deviation of the output of interest (σ). For 2 samples (2-sided) and where N is small (and assuming we don't know the population standard deviation), we use the t-distribution rather than the standard normal distribution. The sample size in this scenario can be determined as:

$$N \geq 2\left[\frac{\sigma}{\delta^*}\left(t_{\alpha/2,2(N-1)} + t_{\beta(1),2(N-1)}\right)\right]^2 \qquad (4)$$

where N is the number of Monte Carlo replicates ***for each scenario***, and 2(N-1) are the degrees of freedom associated with the t-distribution for 2-samples. $t_{alpha/2}$ is the two-sided t-critical value at significant level α, and $t_{\beta\,(1)}$ is the one-sided t-critical value for power 1-β (where β = 1-power). Note that the sample size is on both sides of the equation, requiring an iterative procedure (Zar, 2010). The R statistical program (R Core Team, 2017) has a built-in function that performs the calculation for the 2-sample t-test and proportion test (Supplement S2). We can then determine the sample size required to detect a given effect size with various values of σ (Figure 2c). For a given sample size (N), we rearrange equation 6 to solve for δ*:

$$\delta^* = \sigma\sqrt{\frac{2}{N}}\left(t_{\alpha/2,2(N-1)} + t_{\beta(1),2(N-1)}\right) \qquad (5)$$

Figure 2d gives, for a given value of σ, the sample size required to detect increasing effects.

Table 1 gives the sample size required to meet each margin of error and effect size value for WMFire. For example, if we want to detect if our 10% increase in fuel loading changes mean annual area burned at least by 20 ha yr^{-1}, we should conduct at least 144 Monte Carlo replicates. If we are interested in smaller changes in mean annual area burned we would have to increase the number of Monte Carlo replicates.

2.3.6. Perform simulation study

For our WMFire simulation example we have designed two scenarios (increase fuel load 10%, decrease evapotranspiration 10%), which we will compare to our baseline condition. Note that we perform this analysis as a factorial design, simulating both the baseline condition and each of the scenarios with the same number of Monte Carlo replicates. Assume that we are interested in detecting a change in mean annual area burned of at least 20 ha yr^{-1}, a change in natural fire rotation of at least 5 years, and a change in mean fire return interval of at least 0.5 years. For each scenario we are also interested in estimating the probability the most fires in a time series occur in June within 0.1 of the true probability (rather than detecting a change). From Table 1 we see that sample size requirements differ for each target output, with the largest sample size for estimating natural fire rotation (associated with the largest coefficient of variation). We therefore choose 157 Monte Carlo replicates for all scenarios. Note that if the objective of the simulation study were to detect a change in the probability that June is the most common month for fire occurrence, then we would require 401 replicate simulations.

With a 10% increase in fuel loading, WMFire predicts mean annual area burned in the Santa Fe watershed of 303.3 ha yr^{-1}, a natural fire rotation of 25.1 years, a mean fire return interval of 4.1 years, and probability of 0.91 that June has the most fires that occur in a time series (Table 2). With a 10% decrease in evapotranspiration (corresponding to an increase in relative water deficit, or drier fuels), WMFire predicts mean annual area burned in the Santa Fe watershed of 256.8 ha yr^{-1}, a natural fire rotation of 28.8 years, a mean fire return interval of 3.9 years, and probability of 0.764 that June has the most fires that occur in a time series (Table 2). Figure 3 gives boxplots of each model prediction for each model scenario.

3. Discussion

How many replicate simulations should I conduct? There is no single numerical answer to this question (Figure 1; Table 1). As with empirical study design, design of experiments using stochastic ecological models requires thoughtful consideration of desired precision of estimation or effect sizes for scenario analysis, in the context of the overarching modeling objectives, while considering the underlying variability in the model output and any computational limitations. The basic principles of study design need to be included in the standard toolkit of stochastic model development and analysis. Large round numbers like 100 or 1000 are often accepted as sufficient (Fig. A1b), but this qualifies as arbitrarily large absent a quantitative analysis of the model variability.

3.1. More is not necessarily better

In general we have an instinct that more replicates is better. In the context of empirical ecological studies, this is often the case because we tend to exist in the realm of low statistical power. A sample size that is too small to detect meaningful effects is likely a waste of resources, with results that are difficult to interpret meaningfully. This is also true for simulations of stochastic ecological models. In the case of high computational burden, it is imperative to determine the number of replicates necessary to make meaningful comparisons and predictions.

As sample size goes to infinity, δ^* goes to 0, such that minute effects may be detectable statistically that are not meaningful for the ecological system. We desire to identify the number of replicate Monte Carlo simulations that is able to detect statistically a meaningful change in the output of interest. Larger number of replicates may be able to detect statistical differences that are not meaningful, both wasting resources and possibly leading to inappropriate conclusions where statistical significance does not imply practical significance. This is a consideration in particular for stochastic ecological models that do not suffer from high computational burdens, where a very large number of replicates is possible. This may lead the ecologist to the other extreme. Tiny effects that are not of practical significance may be detectable given a large number of Monte Carlo replicates. In this case, more is not necessarily better as the effect size itself would be of interest, not just detecting statistical differences (Steel et al., 2013).

When reporting the results of a simulation study using a stochastic ecological model, declaring that you have taken a large number of Monte Carlo replicates is meaningless absent consideration of the underlying variability in the model outputs of interest. The definition of a "large" number of simulations is relative to the variability in model outputs. There are scenarios where 100, or even 1000 replicate simulations may be inadequate (Figure 1, Byrne 2013), and some where 50 maybe sufficient. A quantitative analysis like that outlined here is required to justify choices of the number of Monte Carlo replicates.

Note also that even for an individual stochastic model, the number of simulations required will depend on the target model output (Table 1). If the modeling experiment involves multiple model outputs, the number of replicates may be chosen to meet the requirements of the most variable output. For example, in the WMFire case if all of the model outputs are results of interest, the number of replicates should be chosen for the natural fire rotation (nfr), as that is the most variable output (Table 1). If instead the priority of the modeling study is to detect a change in seasonality of wildfire (e.g., the probability that in a time series more fires occur in June than any other month), then a larger number of replicates may be required.

3.2. Interpretation of stochastic ecological model predictions

Basic statistical principles can also be applied to the interpretation of stochastic ecological model predictions, and it is important to avoid common statistical pitfalls (Steel et al., 2013) in stochastic model study design. As with empirical studies, both mean values and standard deviations should be presented with stochastic model predictions (e.g., Table 2). The pilot simulation study needed to determine the number of replicates is not sufficient for a model application under a factorial design. It is possible that the coefficient of variation does not scale with the model predictions, and it may increase or decrease depending on the scenario (Table 2). The distributions of predictions should be visualized (e.g., with boxplots; Fig. 3) to compare scenarios, and confidence intervals for the model replicates should be reported. Effect sizes should be reported (Lorscheid et al., 2012), with accompanying statistical interpretation.

In the case of the example WMFire scenarios presented here, an appropriate conclusion would be that the model predicts an increase of 115.9 ha yr^{-1} annual area burned with a 10% increase in fuel loading (Table 2; Fig. 3). This value is both statistically significant given the standard error in the model estimate, and of practical significance relative to the minimum detectable effect of 20 ha yr^{-1}. Note also that we can construct an interval estimate for the population mean model prediction of (291.0, 315.6 ha yr^{-1}) for mean annual area burned with a 10% increase in fuel loading.

An example of an inappropriate conclusion in the example WMFire scenario analysis would be that the model predicts a change in the seasonality of fire with a decrease of 10% in evapotranspiration (represented by an estimated decrease in the probability that, in a time series, more fire occur in June than any other month; Table 2, Fig. 3). Although the point estimate of the probability the most fires in a time series occur in June is lower with a 10% decrease in evapotranspiration, that change is not of statistical significance with 157 replicate simulations. It is also not of practical significance if the goal is to detect a change in the proportion of at least 0.1 (Table 1). Since we did not choose the number of replicates to detect a change in seasonality, our interpretations are limited. In contrast, if we had instead used 2000 Monte Carlo replicates with the same results, then we could have concluded that the change in seasonality was statistically significant. In this case such a simple interpretation would be misleading because while the change is statistically significant, the effect size is so small as to be of questionable ecological significance.

3.3. Considerations

The simulation pilot study (Appendix B) is an up-front computational investment used to estimate the variability in model outputs, either made in the process of model development or in simulation study design. The pilot study does not necessarily provide the true value of σ, or a value for the coefficient of variation that is robust across all possible applicable model domains. As with a pilot study in empirical study design, the goal is rather to provide a best guess to the variability and to inform the design of more complex modeling experiments with higher computational burden (e.g., a 2x3 factorial design of model scenarios, with 2 management actions and 3 temperature changes). It is possible, particularly in a scenario analysis, that the CV for a model output may be sensitive to the scenario conditions (Table 2). This is why, as in an empirical study, it is important to include estimates of the variability realized in the simulation study across modeling scenarios.

3.4. Conclusions

The guidelines presented here are not meant to be exhaustive of all model applications, but rather to establish a framework, or a set of principles, to motivate quantitative consideration of the number of Monte Carlo replicates. These guidelines can supplant the *ad hoc* approach that seems prevalent in the current literature (Appendix A), and help to set a standard for the application and interpretation of stochastic ecological models. The expected variability in important stochastic ecological model outputs is an important component of stochastic model development, and should become part of the model domain and documentation. These estimates should be updated as the model is modified and adapted for different applications.

Acknowledgements

This work was funded by the National Science Foundation Science, Engineering and Education for Sustainability Award Number 1520847. Erin Hanan, Don McKenzie, and Christina Tague provided valuable feedback that greatly improved this manuscript.

References

Byrne, M.D., 2013. How many times should a stochastic model be run? An approach based on confidence intervals. Proc. 12th Int. Conf. Cogn. Model. 445–450.

Driels, M.R., Shin, Y.S., 2004. Determining the Number of Iterations for Monte Carlo Simulations of Weapon Effectiveness. Monterey, California. Naval Postgraduate School. NPS-MAE-04-005. http://hdl.handle.net/10945/798. Last accessed Dec 17, 2018.

Fieberg, J., Ellner, S.P., 2001. Stochastic matrix models for conservation and management : a comparative review of methods. Ecol. Lett. 4, 244–266.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-custard, J., Grand, T., Heinz, S.K., Huse, G., Huth, A., Jepsen, J.U., Jørgensen, C., Mooij, W.M., Birgit, M., Pe, G., Piou, C., Railsback, S.F., Robbins, A.M., Robbins, M.M., Rossmanith, E., Nadja, R., Strand, E., Souissi, S., Stillman, R.A., Vabo, R., Visser, U., DeAngelis, D.L., 2006. A standard protocol for describing individual-based and agent-based models. Ecol. Modell. 8, 115–126. doi:10.1016/j.ecolmodel.2006.04.023

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Softw. 21, 602–614. doi:10.1016/j.envsoft.2006.01.004

Keane, R.E., Austin, M., Field, C., Huth, A., Lexer, M.J., Peters, D., Solomon, A., Wyckoff, P., 2001. Tree mortality in gap models: Application to climate change. Clim. Change 51, 509–540. doi:10.1023/A:1012539409854

Kennedy, M.C., Mckenzie, D., Tague, C., Dugger, A.L., 2017. Balancing uncertainty and complexity to incorporate fire spread in an eco-hydrological model. Int. J. Wildl. Fire 26, 706–718. doi:10.1071/WF16169

Lorscheid, I., Heine, B.O., Meyer, M., 2012. Opening the “Black Box” of Simulations: Increased Transparency and Effective Communication Through the Systematic Design of Experiments. Comput. Math. Organ. Theory 18, 22–62. doi:10.1007/s10588-011-9097-3

National Research Council, 2007. Models in Environmental Regulatory Decision Making. The National Academies Press, Washington, D.C. doi:10.17226/11972

R Core Team, 2017. R: A language and environment for statistical computing.

Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V., 2010. Ecological models supporting environmental decision making: A strategy for the future. Trends Ecol. Evol. 25, 479–486. doi:10.1016/j.tree.2010.05.001

Steel, E.A., Kennedy, M.C., Cunningham, P.G., Stanovick, J.S., 2013. Applied statistics in ecology: Common pitfalls and simple solutions. Ecosphere 4. doi:10.1890/ES13-00160.1

Tague, C., Band, L., 2004. RHESSys: Regional Hydro-Ecologic Simulation System—An Object-Oriented Approach to Spatially Distributed Modeling of Carbon, Water, and Nutrient Cycling. Earth Interact. 8, 1–42.

Zar, J.H., 2010. Biostatistical Analysis, 5th ed. Prentice Hall, Pearson, Upper Saddle River, NJ.

Tables

Table 1: Summary statistics and sample size requirements for WMFire model predictions. μ and σ give the mean and standard deviation for 10000 Monte Carlo replicate baseline simulations. CV is the coefficient of variation (σ/μ), N_E gives the sample size (per model scenario) required to estimate the mean value within 10% or 5% , δ is an effect size considered to be of practical significance for each output, and N_δ is the number of replicates required to be able to detect that effect with 90% power. All calculations assume $\alpha = 0.05$. $\bar{A}$ is the mean annual area burned per year, nfr is the natural fire rotation, μ_{fri} is the mean fire return interval between fires of at least 100 ha (years), and p_{June} is the probability that June is the month with the most fires in a time series. For the proportion estimate the margin of error (0.1 or 0.05 * μ) is simply the proportion (0.1 or 0.05). Here we assume p=0.5 for a conservative estimate of the required sample size for estimation, regardless of the point estimate.

	$\bar{A}$ (ha yr^{-1})	nfr (years)	μ_{fri} (years)	p_{June}
μ	188.4	40.6	5.2	0.792
σ	52.2	13.6	1.3	NA
CV	0.28	0.33	0.25	NA
0.1μ	18.8	4.1	0.52	0.1
N_E	31	42	25	97
0.05μ	9.4	2.0	0.026	0.05
N_E	121	168	97	385
δ	20	5	0.5	0.10
N_δ	144	157	144	401

Table 2. Summary statistics for WMFire predictions for each of the three scenarios, as well as at baseline conditions with N = 157 Monte Carlo replicates. Scenario 1 is a 10% increase in fuel load, scenario 2 is a 10% decrease in evapotranspiration (an increase in relative deficit). Mean WMFire predicted values across 157 replicate simulations (standard deviation in parentheses).

Scenario	$\bar{A}$ (ha yr^{-1})	nfr (years)	μ_{fri} (years)	p_{June}
Baseline (N=157)	187.4 (55.3)	41.3 (15.1)	5.3 (1.3)	0.783
S1	303.3 (78.7)	25.1 (8.5)	4.1 (0.75)	0.911
S2	256.8 (59.4)	28.8 (7.0)	3.9 (0.74)	0.764

Figure Captions

Figure 1. General framework for determining number of Monte Carlo replicates. Model development and assessment aggregates the many methods to develop ecological models. Once a model is deemed adequate, an independent model replicate should be defined (1), and iid (independent and identically distributed) model outputs identified (2). A pilot study of some baseline condition is performed to estimate the standard deviation (σ) and the coefficient of variation (3; Appendix B). The results of the pilot study should be included in model documentation and a repository of all model outputs generated by the pilot study maintained (to prevent future computational effort). Choose a desired margin of error (E) and/or a detectable effect size (δ) in the context of the study (4), and calculate sample size (5). If the number of replicates is computationally feasible, perform study (6). If not, determine what is feasible and calculate the expected margin of error and/or detectable effect size, and judge whether the results will be meaningful. If they are, perform study. For study results, report simulation study confidence intervals and/or effect sizes (6).

Figure 2. (a) Number of Monte Carlo replicates required to achieve a margin of error with different proportion of the mean value (w) for increasing coefficients of variation (CV). (b) for a given CV (0.25), number of replicates required to achieve a margin of error with increasing proportion of the mean value (w). (c) Number of Monte Carlo replicates required to achieve different effect sizes with increasing standard deviation (example taken from nfr from Table 1). (d) Number of Monte Carlo replicates required to detect increasing effect sizes (δ*) with 90% power, assuming σ = 14 years.

Figure 3. Boxplot of model predictions across 157 replicate simulations comparing baseline distribution to each model scenario for a) mean annual area burned; b) natural fire rotation; and

c) mean fire return interval. B is baseline, S1 is a 10% increase in fuel load compared to baseline, and S2 is a 10% decrease in evapotranspiration compared to baseline.