

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](mailto:mkenn@uw.edu)

Follow this and additional works at: [https://digitalcommons.tacoma.uw.edu/ias\\_pub](https://digitalcommons.tacoma.uw.edu/ias_pub)

---

## Recommended Citation

Kennedy, M.C., "Experimental Design Principles to Choose the Number of Monte Carlo Replicates for Stochastic Ecological Models" (2019). *SIAS Faculty Publications*. 1033.

[https://digitalcommons.tacoma.uw.edu/ias\\_pub/1033](https://digitalcommons.tacoma.uw.edu/ias_pub/1033)

This Article is brought to you for free and open access by the School of Interdisciplinary Arts and Sciences at UW Tacoma Digital Commons. It has been accepted for inclusion in SIAS Faculty Publications by an authorized administrator of UW Tacoma Digital Commons.

1 Title: Experimental design principles to choose the number of Monte Carlo replicates for  
2 stochastic ecological models  
3 Maureen C. Kennedy (corresponding author)  
4 University of Washington, Tacoma. School of Interdisciplinary Arts and Sciences, Division of  
5 Sciences and Mathematics. 1900 Commerce St., Tacoma, WA, 98402.  
6 mkenn@uw.edu  
7

8 **Abstract:**

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

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

30

## 31 **1. Introduction**

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

44 In general, ecological models are either deterministic or stochastic. For a deterministic  
45 model, replicate simulations with the same inputs and parameters give identical model  
46 predictions. In a stochastic model, probability distributions represent some modeled processes,  
47 such that replicate simulations with the same inputs and parameters give variable model  
48 predictions. In that manner stochastic simulations use probability structures to represent  
49 uncertainty in the modeled processes and input data, yielding distributions of model outputs  
50 rather than point estimates. For example, WMfire is a stochastic model of fire spread (Kennedy  
51 et al., 2017) coupled with a deterministic eco-hydrological model (RHESSys; Tague and Band  
52 2004). With a randomly located ignition point, and spread governed by probability structures  
53 informed by the underlying landscape, replicate simulations on identical landscapes result in

54 variable fire areas. Across multiple WMFire simulations we then can describe a distribution of  
55 fire occurrence rather than a single realization.

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

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

75 The choice of Kennedy et al. (2017) to use 500 Monte Carlo replicate simulations  
76 without evaluation of the underlying stochastic model variability is an example of a common *ad*

77 *hoc* approach: choose an arbitrarily large (*sensu* Byrne 2013) number of replicate simulations,  
78 without an accompanying quantitative justification. A brief survey of recently published  
79 modeling studies (Appendix A) illustrates that this is the most common technique (Fig A.1).  
80 Alternatively, under severe computational constraints, we simulate as many replicates as possible  
81 without quantifying the uncertainty associated with a small sample size. Adapting the statistical  
82 principles of experimental design to stochastic ecological modeling may provide a more robust  
83 alternative to the current *ad hoc* approaches.

84         When considering the number of replicate Monte Carlo simulations, we are concerned  
85 with both estimation of mean model outputs, as well as the effect size when comparing some  
86 modeling scenario to a baseline. As with empirical studies with large sample sizes, the more  
87 Monte Carlo replicates are produced the smaller is the effect size that can be detected  
88 statistically. The fewer the number of Monte Carlo replicates the more difficult it is to  
89 distinguish actual predicted effects from random variability, an issue if the model is  
90 computationally intensive. When planning a modeling study using a stochastic ecological model,  
91 we need to determine the number of Monte Carlo replicates necessary to conclude if the mean  
92 system state is predicted to change in a way that is both meaningful (the change in mean has  
93 practical effect on the system) and significant (the change in mean is different than zero, relative  
94 to the standard error). To answer the question of how many replicate simulations, we can expand  
95 the idea of applying a design of experiments approach for modeling studies (Lorscheid et al.,  
96 2012).

97         The objective here is to suggest an alternative to the *ad hoc* approach in determining the  
98 number of replicate simulations of a stochastic ecological model. To that end a general  
99 framework is presented (Fig. 1) for a thoughtful quantitative analysis of the number of

100 simulations necessary to achieve a pre-specified level of precision in stochastic model outputs,  
101 and to use that in study development. When presenting a modeling study, the reporting of mean  
102 model estimates, the variability in model estimates, and the distribution of model estimates  
103 should all be standard practice. The application of this framework is illustrated with an example  
104 using WMFire to compare fuel loading and moisture condition scenarios.

## 105 **2. Methods**

### 106 *2.1. WMFire description*

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

122 probability of spread associated with any burned pixels. RHESSys interprets this grid to  
123 implement any fire effects on the burned pixels.

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

## 130 *2.2. Model scenario description*

131 To illustrate how this framework can inform model application, two model scenarios are  
132 designed. The goal would be to determine if, for each scenario, model predictions change from  
133 the baseline historical condition of Kennedy et al. (2017). The first scenario is an increase of  
134 10% in fuel loading across the landscape all years in the simulation; the second scenario is a 10%  
135 decrease in evapotranspiration across the landscape all years in the simulation (representing  
136 increased dryness). Next we give an overview of the framework illustrated in Figure 1.

## 137 *2.3. Framework to determine the number of Monte Carlo Replicates*

### 138 *2.3.1. Define independent model replicate*

139 In order to use standard statistical principles of experimental design, we need to identify a  
140 single independent model replicate. For example, in a time series of simulated fire spread in a  
141 fully coupled WMFire-RHESSys modeling system, the fire hazard in a given year depends on  
142 the past history of fire occurrence. Therefore each simulated year is not independent of other  
143 years in the same time series. However, a full time series of fire occurrence would be  
144 independent of replicate full time series. In the case of the Santa Fe watershed, WMFire is run



145 from historical climate spanning the years 1941-2008. Each replicate time series repeats the  
146 conditions in this timeframe. Therefore we consider an independent model replicate to be a  
147 single WMFire time series of fire occurrence. Independent model outputs are then individual  
148 summaries of each replicate time series.

### 149 2.3.2. Identify model outputs of interest

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

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

166 In general, *inference* is the process of rejecting or failing to reject statistical hypotheses  
167 (e.g.,  $\mu_1 = \mu_2$ ). For a given population variability, sample size for the case of inference

168 determines our power ( $1-\beta$ ) to determine statistically a particular effect size (change in estimated  
169 value;  $\delta^*$ ). For a given power, a larger sample size means we can detect a smaller effect size.

### 170 *2.3.3. Conduct pilot study to estimate model variability*

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

187 For the WMFire pilot study 10,000 Monte Carlo replicate simulations were performed at  
188 the baseline historical condition of Kennedy et al. (2017) (see Appendix B for details of pilot  
189 study), with the model outputs calculated for each replicate simulation. Table 1 gives the mean,

190 standard deviation, and coefficient of variation for each WMFire model output across 10,000  
191 pilot study replicates.

#### 192 *2.3.4. Choose margin of error and/or detectable effect size*

193 Byrne (2013) outlines a strategy for sample size determination for stochastic cognitive  
194 models that is based on principles of confidence interval estimation (see also Driels and Shin  
195 2004), which we adapt here. The margin of error (E) can be interpreted as the maximum likely  
196 distance between a sample mean and the population mean with some level of confidence (1- $\alpha$ ).  
197 The total width of a confidence interval around the mean value is 2E. A narrower confidence  
198 interval may be considered more precise. Byrne (2013) shows that for the purpose of sample size  
199 determination, if the coefficient of variation is known then the margin of error can be  
200 standardized to estimating the population mean value within some proportion (w) of its true  
201 value, without knowing the population mean value. For example, the desired precision might be  
202 w=0.1, that is that the sample mean value is within 10% of the population mean value. For  
203 WMFire we consider estimation within 10% (w=0.10) and 5% (w=0.05) of the population mean  
204 value.

205 For inference we are interested in the minimum detectable effect ( $\delta^*$ ), the minimum  
206 difference in mean predicted value between some baseline scenario and a treatment scenario that  
207 is considered to be ecologically significant. Consider a simple 2-sample design, where the  
208 stochastic simulation model is used to determine whether the population mean model output ( $\mu$ ,  
209 estimated by  $\bar{Y}$ ) is predicted to change between a baseline simulation (control C;  $\mu_C$  estimated by  
210  $\bar{Y}_C$ ) and a treatment scenario (treatment T;  $\mu_T$  estimated by  $\bar{Y}_T$ ). The null hypothesis is  $H_0: \mu_C =$   
211  $\mu_T$ . and  $\delta^*$  is the minimum difference between population means ( $|\mu_C - \mu_T|$ ) that we are interested  
212 in detecting. For the WMFire example, we assume a minimum detectable effect of 20 ha yr<sup>-1</sup>, 5

213 years, 0.5 years, and 0.10 for mean annual area burned, natural fire rotation, fire return interval,  
214 and the probability that in a time series the most fires occur in June, respectively.

### 215 *2.3.5a. Number of replicate simulations for estimation*

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

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

$$234 \quad N \geq \left( \frac{z_{\alpha/2}}{w} CV \right)^2 \quad (1)$$

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

$$239 \quad w \geq \frac{z_{\alpha/2} CV}{\sqrt{N}} \quad (2)$$

240 Supplement S1 gives example scripts for the R statistical program (R Core Team, 2017) to  
241 determine sample sizes for estimation. Note that Byrne (2013) also provides web-based utilities  
242 to calculate sample size requirements (<http://chil.rice.edu/research/nomr/>, last accessed Dec 17,  
243 2018).

244 If the model prediction is a proportion, the calculation is somewhat easier to standardize.  
245 Here we define E as the maximum likely distance between the population proportion ( $\pi$ ) and the  
246 sample proportion (p). We know that the standard deviation of the proportion is  $\sqrt{\pi(1-\pi)}$  and the  
247 sample size is calculated as:

$$248 \quad N \geq \left( \frac{z_{\alpha/2} \sqrt{\pi(1-\pi)}}{E} \right)^2 \quad (3)$$

249 If  $\pi$  is known, then the standard deviation is known. A conservative approach is to assume  $\pi =$   
250 0.5, which maximizes the standard deviation for the proportion. Note that this may result in an  
251 overestimation of required sample size, as the sample size required to estimate a lower or higher  
252 population proportion would be smaller. If there is good prior information for the value of the  
253 population proportion then that can be used to determine a reasonable sample size. For example,  
254 assuming a proportion of 0.5 results in a sample size requirement of 97 for estimation (with E  
255 =0.1). If we assume the proportion to be 0.79, then the required sample size would drop to 64.

256 *2.3.5b. Number of replicate simulations for scenario comparisons (inference)*

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

$$263 \quad N \geq 2 \left[ \frac{\sigma}{\delta^*} (t_{\alpha/2, 2(N-1)} + t_{\beta(1), 2(N-1)}) \right]^2 \quad (4)$$

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

$$272 \quad \delta^* = \sigma \sqrt{\frac{2}{N}} (t_{\alpha/2, 2(N-1)} + t_{\beta(1), 2(N-1)}) \quad (5)$$

273 Figure 2d gives, for a given value of  $\sigma$ , the sample size required to detect increasing effects.

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

### 279 2.3.6. Perform simulation study

280 For our WMFire simulation example we have designed two scenarios (increase fuel load  
281 10%, decrease evapotranspiration 10%), which we will compare to our baseline condition. Note  
282 that we perform this analysis as a factorial design, simulating both the baseline condition and  
283 each of the scenarios with the same number of Monte Carlo replicates. Assume that we are  
284 interested in detecting a change in mean annual area burned of at least  $20 \text{ ha yr}^{-1}$ , a change in  
285 natural fire rotation of at least 5 years, and a change in mean fire return interval of at least 0.5  
286 years. For each scenario we are also interested in estimating the probability the most fires in a  
287 time series occur in June within 0.1 of the true probability (rather than detecting a change). From  
288 Table 1 we see that sample size requirements differ for each target output, with the largest  
289 sample size for estimating natural fire rotation (associated with the largest coefficient of  
290 variation). We therefore choose 157 Monte Carlo replicates for all scenarios. Note that if the  
291 objective of the simulation study were to detect a change in the probability that June is the most  
292 common month for fire occurrence, then we would require 401 replicate simulations.

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

### 301 **3. Discussion**

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

### 311 *3.1. More is not necessarily better*

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

318           As sample size goes to infinity,  $\delta^*$  goes to 0, such that minute effects may be detectable  
319 statistically that are not meaningful for the ecological system. We desire to identify the number  
320 of replicate Monte Carlo simulations that is able to detect statistically a meaningful change in the  
321 output of interest. Larger number of replicates may be able to detect statistical differences that  
322 are not meaningful, both wasting resources and possibly leading to inappropriate conclusions  
323 where statistical significance does not imply practical significance. This is a consideration in  
324 particular for stochastic ecological models that do not suffer from high computational burdens,



325 where a very large number of replicates is possible. This may lead the ecologist to the other  
326 extreme. Tiny effects that are not of practical significance may be detectable given a large  
327 number of Monte Carlo replicates. In this case, more is not necessarily better as the effect size  
328 itself would be of interest, not just detecting statistical differences (Steel et al., 2013).

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

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

### 344 *3.2. Interpretation of stochastic ecological model predictions*

345         Basic statistical principles can also be applied to the interpretation of stochastic  
346 ecological model predictions, and it is important to avoid common statistical pitfalls (Steel et al.,  
347 2013) in stochastic model study design. As with empirical studies, both mean values and

348 standard deviations should be presented with stochastic model predictions (e.g., Table 2). The  
349 pilot simulation study needed to determine the number of replicates is not sufficient for a model  
350 application under a factorial design. It is possible that the coefficient of variation does not scale  
351 with the model predictions, and it may increase or decrease depending on the scenario (Table 2).  
352 The distributions of predictions should be visualized (e.g., with boxplots; Fig. 3) to compare  
353 scenarios, and confidence intervals for the model replicates should be reported. Effect sizes  
354 should be reported (Lorscheid et al., 2012), with accompanying statistical interpretation.

355         In the case of the example WMFire scenarios presented here, an appropriate conclusion  
356 would be that the model predicts an increase of 115.9 ha yr<sup>-1</sup> annual area burned with a 10%  
357 increase in fuel loading (Table 2; Fig. 3). This value is both statistically significant given the  
358 standard error in the model estimate, and of practical significance relative to the minimum  
359 detectable effect of 20 ha yr<sup>-1</sup>. Note also that we can construct an interval estimate for the  
360 population mean model prediction of (291.0, 315.6 ha yr<sup>-1</sup>) for mean annual area burned with a  
361 10% increase in fuel loading.

362         An example of an inappropriate conclusion in the example WMFire scenario analysis  
363 would be that the model predicts a change in the seasonality of fire with a decrease of 10% in  
364 evapotranspiration (represented by an estimated decrease in the probability that, in a time series,  
365 more fire occur in June than any other month; Table 2, Fig. 3). Although the point estimate of the  
366 probability the most fires in a time series occur in June is lower with a 10% decrease in  
367 evapotranspiration, that change is not of statistical significance with 157 replicate simulations. It  
368 is also not of practical significance if the goal is to detect a change in the proportion of at least  
369 0.1 (Table 1). Since we did not choose the number of replicates to detect a change in seasonality,  
370 our interpretations are limited. In contrast, if we had instead used 2000 Monte Carlo replicates

371 with the same results, then we could have concluded that the change in seasonality was  
372 statistically significant. In this case such a simple interpretation would be misleading because  
373 while the change is statistically significant, the effect size is so small as to be of questionable  
374 ecological significance.

### 375 *3.3. Considerations*

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

### 387 *3.4. Conclusions*

388 The guidelines presented here are not meant to be exhaustive of all model applications,  
389 but rather to establish a framework, or a set of principles, to motivate quantitative consideration  
390 of the number of Monte Carlo replicates. These guidelines can supplant the *ad hoc* approach that  
391 seems prevalent in the current literature (Appendix A), and help to set a standard for the  
392 application and interpretation of stochastic ecological models. The expected variability in  
393 important stochastic ecological model outputs is an important component of stochastic model

394 development, and should become part of the model domain and documentation. These estimates  
395 should be updated as the model is modified and adapted for different applications.

396 **Acknowledgements**

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

400 **References**

- 401 Byrne, M.D., 2013. How many times should a stochastic model be run? An approach based on  
402 confidence intervals. *Proc. 12th Int. Conf. Cogn. Model.* 445–450.
- 403 Driels, M.R., Shin, Y.S., 2004. Determining the Number of Iterations for Monte Carlo  
404 Simulations of Weapon Effectiveness. Monterey, California. Naval Postgraduate School.  
405 NPS-MAE-04-005. <http://hdl.handle.net/10945/798>. Last accessed Dec 17, 2018.
- 406 Fieberg, J., Ellner, S.P., 2001. Stochastic matrix models for conservation and management : a  
407 comparative review of methods. *Ecol. Lett.* 4, 244–266.
- 408 Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-custard, J., Grand,  
409 T., Heinz, S.K., Huse, G., Huth, A., Jepsen, J.U., Jørgensen, C., Mooij, W.M., Birgit, M.,  
410 Pe, G., Piou, C., Railsback, S.F., Robbins, A.M., Robbins, M.M., Rossmanith, E., Nadja, R.,  
411 Strand, E., Souissi, S., Stillman, R.A., Vabo, R., Visser, U., DeAngelis, D.L., 2006. A  
412 standard protocol for describing individual-based and agent-based models. *Ecol. Modell.* 8,  
413 115–126. doi:10.1016/j.ecolmodel.2006.04.023
- 414 Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and  
415 evaluation of environmental models. *Environ. Model. Softw.* 21, 602–614.  
416 doi:10.1016/j.envsoft.2006.01.004
- 417 Keane, R.E., Austin, M., Field, C., Huth, A., Lexer, M.J., Peters, D., Solomon, A., Wyckoff, P.,  
418 2001. Tree mortality in gap models: Application to climate change. *Clim. Change* 51, 509–  
419 540. doi:10.1023/A:1012539409854
- 420 Kennedy, M.C., Mckenzie, D., Tague, C., Dugger, A.L., 2017. Balancing uncertainty and  
421 complexity to incorporate fire spread in an eco-hydrological model. *Int. J. Wildl. Fire* 26,  
422 706–718. doi:10.1071/WF16169

423 Lorscheid, I., Heine, B.O., Meyer, M., 2012. Opening the “Black Box” of Simulations: Increased  
424 Transparency and Effective Communication Through the Systematic Design of  
425 Experiments. *Comput. Math. Organ. Theory* 18, 22–62. doi:10.1007/s10588-011-9097-3  
426 National Research Council, 2007. *Models in Environmental Regulatory Decision Making*. The  
427 National Academies Press, Washington, D.C. doi:10.17226/11972  
428 R Core Team, 2017. *R: A language and environment for statistical computing*.  
429 Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V., 2010. Ecological models supporting  
430 environmental decision making: A strategy for the future. *Trends Ecol. Evol.* 25, 479–486.  
431 doi:10.1016/j.tree.2010.05.001  
432 Steel, E.A., Kennedy, M.C., Cunningham, P.G., Stanovick, J.S., 2013. Applied statistics in  
433 ecology: Common pitfalls and simple solutions. *Ecosphere* 4. doi:10.1890/ES13-00160.1  
434 Tague, C., Band, L., 2004. RHESSys: Regional Hydro-Ecologic Simulation System—An  
435 Object-Oriented Approach to Spatially Distributed Modeling of Carbon, Water, and  
436 Nutrient Cycling. *Earth Interact.* 8, 1–42.  
437 Zar, J.H., 2010. *Biostatistical Analysis*, 5th ed. Prentice Hall, Pearson, Upper Saddle River, NJ.  
438  
439

440 **Tables**

441 **Table 1:** Summary statistics and sample size requirements for WMFire model predictions.  $\mu$  and  
 442  $\sigma$  give the mean and standard deviation for 10000 Monte Carlo replicate baseline simulations.  
 443 CV is the coefficient of variation ( $\sigma/\mu$ ),  $N_E$  gives the sample size (per model scenario) required to  
 444 estimate the mean value within 10% or 5% ,  $\delta$  is an effect size considered to be of practical  
 445 significance for each output, and  $N_\delta$  is the number of replicates required to be able to detect that  
 446 effect with 90% power. All calculations assume  $\alpha = 0.05$ .  $\bar{A}$  is the mean annual area burned per  
 447 year, nfr is the natural fire rotation,  $\mu_{fri}$  is the mean fire return interval between fires of at least  
 448 100 ha (years), and  $p_{June}$  is the probability that June is the month with the most fires in a time  
 449 series. For the proportion estimate the margin of error (0.1 or 0.05 \*  $\mu$ ) is simply the proportion  
 450 (0.1 or 0.05). Here we assume  $p=0.5$  for a conservative estimate of the required sample size for  
 451 estimation, regardless of the point estimate.

	$\bar{A}$ (ha yr <sup>-1</sup> )	nfr (years)	$\mu_{fri}$ (years)	$p_{June}$
$\mu$	188.4	40.6	5.2	0.792
$\sigma$	52.2	13.6	1.3	NA
CV	0.28	0.33	0.25	NA
0.1 $\mu$	18.8	4.1	0.52	0.1
$N_E$	31	42	25	97
0.05 $\mu$	9.4	2.0	0.026	0.05
$N_E$	121	168	97	385
$\delta$	20	5	0.5	0.10
$N_\delta$	144	157	144	401

452

453

454 **Table 2.** Summary statistics for WMFire predictions for each of the three scenarios, as well as at  
 455 baseline conditions with N = 157 Monte Carlo replicates. Scenario 1 is a 10% increase in fuel  
 456 load, scenario 2 is a 10% decrease in evapotranspiration (an increase in relative deficit). Mean  
 457 WMFire predicted values across 157 replicate simulations (standard deviation in parentheses).  
 458

Scenario	$\bar{A}$ (ha yr <sup>-1</sup> )	nfr (years)	$\mu_{\text{fri}}$ (years)	p <sub>June</sub>
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

459

460



461 **Figure Captions**

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

475 Figure 2. (a) Number of Monte Carlo replicates required to achieve a margin of error with  
476 different proportion of the mean value ( $w$ ) for increasing coefficients of variation (CV). (b) for a  
477 given CV (0.25), number of replicates required to achieve a margin of error with increasing  
478 proportion of the mean value ( $w$ ). (c) Number of Monte Carlo replicates required to achieve  
479 different effect sizes with increasing standard deviation (example taken from nfr from Table 1).  
480 (d) Number of Monte Carlo replicates required to detect increasing effect sizes ( $\delta^*$ ) with 90%  
481 power, assuming  $\sigma = 14$  years.

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

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