Oikos 000: 001–004, 2013
doi: 10.1111/j.1600-0706.2013.01073.x
Published 2013. This article is a US Government work and is in the public domain in the USA.
Oikos © 2013 Nordic Society Oikos
Subject Editor: Dustin Marshall. Accepted 11 October 2013

Ecologists should not use statistical significance tests to interpret simulation model results

J. Wilson White, Andrew Rassweiler, Jameal F. Samhouri, Adrian C. Stier and Crow White

J. W. White (whitejw@uncw.edu), Dept of Biology and Marine Biology, Univ. of North Carolina Wilmington, Wilmington, NC 28403, USA. – A. Rassweiler, Marine Science Inst., Univ. of California Santa Barbara, Santa Barbara, CA 93106, USA. – J. F. Samhouri, Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA 98112, USA. – A. C. Stier, Dept of Zoology, Univ. of British Columbia, Vancouver, BC, V6T 1Z4, Canada. – C. White, Dept of Biological Sciences, California Polytechnic State Univ., San Luis Obispo, CA 93407, USA.

Simulation models are widely used to represent the dynamics of ecological systems. A common question with such models is how changes to a parameter value or functional form in the model alter the results. Some authors have chosen to answer that question using frequentist statistical hypothesis tests (e.g. ANOVA). This is inappropriate for two reasons. First, p-values are determined by statistical power (i.e. replication), which can be arbitrarily high in a simulation context, producing minuscule p-values regardless of the effect size. Second, the null hypothesis of no difference between treatments (e.g. parameter values) is known a priori to be false, invalidating the premise of the test. Use of p-values is troublesome (rather than simply irrelevant) because small p-values lend a false sense of importance to observed differences. We argue that modelers should abandon this practice and focus on evaluating the magnitude of differences between simulations.

A growing number of authors in the ecological literature use statistical methods common to experimental ecology to analyze the output of ecological simulation models. For example, authors may use analysis of variance (ANOVA) to test whether model runs with different parameter values or different functional forms produce statistically different outputs. We view significance testing applied to simulation model output as a misuse of statistical theory. In this article we explain our reasoning with the goals of discouraging the practice, encouraging instead a focus on the magnitude of differences between simulations (i.e. effect sizes), and sparking discussion regarding when – if ever – statistical significance tests could be appropriate.

The perils of placing too much emphasis on statistical tests are well known in ecology. The past few decades have seen several essays reminding ecologists not to conflate statistical with biological significance, that p-values are essentially arbitrary thresholds, and that p-values are meaningless unless accompanied by measures of effect size and statistical power (Yoccoz 1991, Johnson 1999, Hurlbert and Lombardi 2009, Beninger et al. 2012). The use of frequentist statistical tests in a simulation model setting presents two additional issues, the first practical and the second philosophical:

1) Statistical power is determined by replication (Berkson 1938), which is a trivial notion in the era of modern computing. Therefore power, and thus p-values, are determined only by the number of simulations one chooses to run.
2) The ‘truth’ of a testable null hypothesis is assumed to be unknown. In a model context, the programmer knows the ‘truth’ (because they know the model parameters) and testing a known-to-be-false null hypothesis does not provide useful information (Savage 1957, Johnson 1999, Anderson et al. 2000).

While we contend that a focus on statistical significance is inappropriate, we also argue that quantitative evaluation of differences in effect size among model scenarios is entirely appropriate. Moreover, in some cases frequentist statistical models (e.g. regression, ANOVA) may provide accessible and simple methods for quantifying effect sizes. Though the distinction between statistical and biological significance (Yoccoz 1991) may at first appear subtle, it is critical. Below we use a few recent examples from the ecological literature to illustrate the context in which these issues arise, highlight potential pitfalls in the analysis and interpretation of simulation model output, and offer ways forward for the specific examples mentioned here and for analysis and interpretation of simulation model output in general.

MANOVA with n = 24 000

Marzloff et al. (2013) present a validation and sensitivity analysis of TRITON, a simulation model of alternative-state dynamics in temperate rocky reef communities that centers on trophic interactions between lobsters *Jasus edwardsii*, sea urchins *Centrostephanus rodgersii* and seaweeds *Ecklonia radiata* and *Phyllospora comosa*. As part of their sensitivity analysis they compared the effects of alternative formulations of the lobster predatory functional response (Holling type I, II or III). To do so they performed 8000 simulations with random initial conditions and used MANOVA to compare the effects of functional response (the fixed effect) on a multivariate index of community state. They found that the form of the functional response had a highly significant effect in the MANOVA ($p < 10^{-15}$, $F_{2,23997} = 67.5$), but conceded that of course this significance was due to having nearly 24 000 denominator degrees of freedom in the *F*-test. They therefore ignored the MANOVA results and simply compared model results visually, concluding that the predator functional response had little effect on community state despite the exceedingly low p-value.

This example nicely illustrates the two key problems we have identified. The question Marzloff et al. (2013) were attempting to answer (does the lobster functional response affect the predictions of their model regarding community state?) was not suited for null hypothesis testing. We know a priori from analytical models that different forms of the predator functional response produce distinct dynamics (Oaten and Murdoch 1975a, b), so the question is not whether the model outcomes will be different, but rather how different they will be. Even if we did not have the advantage of prior knowledge about the dynamical consequences of differences between functional responses, it would still be redundant to test for those differences using a null hypothesis framework. The null hypothesis is implicitly posed as follows: 'model results using each of the three functional responses are drawn from populations with identical distributions'. Because we know the model was programmed with different functions and parameters in each case, we also know a priori that the null hypothesis is false. Thus any failure to reject the null hypothesis is by definition a type-II error, and the test merely needs sufficient power to detect an effect and avoid that error.

The second problem with hypothesis tests like the one Marzloff et al. (2013) perform is the arbitrariness of statistical power itself in this setting. A sample size of 24,000 will generally produce a significant result ($p < 0.05$, or even much lower) regardless of the magnitude of the biological effect size. Indeed, when dealing with this type of model, one can literally choose the desired p-value by setting the number of runs. With sufficient computer time, there is no limit to how small a value can be obtained. This sample size influence occurs, of course, in empirical experiments, not just computer simulations: any effect size, no matter how small, can be found significant if one is able to obtain enough replicates (Berkson 1938). However the ease and extremely low cost of replication in a simulation model setting represents an absurd extreme for this general principle.

Marzloff et al. (2013) were correct to disregard the low p-values in their MANOVA, although we disagree with their choice to fall back on a qualitative visual inspection of the model output. A better alternative would have been to directly quantify the effect size in the MANOVA (or simply compare the distributions of community states; in their case a geometric comparison of the positions of model solutions in principal components space), having determined beforehand what magnitude of difference (perhaps a percentage difference from some baseline) would constitute a significant ecological effect in the study system. This is analogous to the argument that ecologists should focus on 'biological significance' rather than statistical significance (Berkson 1942, Yoccoz 1991).

Ecological simulation models are not the only context in which one could argue that the null hypothesis is known to be false a priori. Johnson (1999) catalogues several trivial null hypotheses that were rejected in empirical field studies (e.g. that logged and unlogged forests have the same density of trees). Johnson (1999) and others have argued that null hypotheses are usually false in observational studies, such as when comparing some variable (e.g. fish abundance) among two locations; it is unthinkable that the two locations would have precisely identical population statistics. This situation stands in contrast to the case of a controlled ecological experiment, in which experimental replicates are randomly assigned to treatments, and a null hypothesis of no difference among the treatment populations is actually a reasonable expectation (Johnson 1999). Because simulation models are in a sense controlled numerical experiments it is tempting to analyze their output as one would an empirical experiment. However, in simulation models the relationship between treatment (parameter value, functional form) and some dependent response (model output) is explicit (mechanistic) and known (or at least understood to exist); the question is simply how many replicates are needed to detect that relationship statistically. In empirical ecology, we accept that there is usually some correspondence between statistical significance and biological significance (i.e. effect size); replication is difficult and costly to obtain, so power is a limiting factor and if a statistically significant signal is detected, it is likely to be biologically meaningful as well. That correspondence breaks down with the immense replication and power available in a simulation model setting. Indeed, it is these features of models that enable their analysis to bypass significance-tests and focus on the key result of interest, effect size.

Are peer reviewers the problem?

We used Marzloff et al. (2013) as an instructive example of the folly of null hypothesis tests in part because the authors themselves admitted that the p-values they report are essentially meaningless. Indeed, the halfhearted use of MANOVA in their paper gives the sense of a test that was demanded by a reviewer during the peer review process (after all, if the test results are meaningless, why bother conducting the test at all, or taking up space in the paper to report it?). We have encountered similar pressure from reviewers in our own work, and worry that it may be a common occurrence. A non-systematic search quickly revealed many recent examples spanning the breadth of the ecological literature, from terrestrial plant ecology to fisheries to life-history theory (Marshall et al. 2008, Dauphin et al. 2009, Esther et al. 2010, Makler-Pick et al. 2011, Tam and Ang 2012). We suspect that calculating p-values for simulation model results is actually more commonplace than a literature search would reveal, because doubtless there are many instances of authors attempting this practice, but having it stripped out in review (which all of the authors of this paper have done as reviewer). We have also experienced the opposite interaction, in which reviewer advocate for the necessity of p-values. For example, Samhouri et al. (2009) analyzed seven different Ecopath with Ecosim models of marine food web dynamics to identify system-specific indicators (i.e. biomass of a particular functional group) that reveal changes to key ecosystem attributes (i.e. emergent properties such as diversity, net primary productivity, mean trophic level, etc.). For each model, they simulated a range of fishing perturbations to the model ecosystem, recorded the responses of a suite of indicators and attributes, and calculated the correlations between each indicator and attribute. In their paper, they reported only the magnitude of the correlation coefficients, not their statistical significance, noting that because of their large sample size, even small, biologically unmeaningful correlations would be statistically significant (also see Fulton et al. 2005). This omission was a point of contention during review, as a reviewer focused on the potential for reporting spurious correlations in the absence of p-values. Indeed, there are plenty of examples in the literature in which p-values are reported to support or refute the use of indicators (Travers et al. 2006). Samhouri et al. (2009) won the argument with their reviewer, but we wonder how frequently well-intentioned reviewers – particularly those with more experience with empirical data than models – insist on inappropriate hypothesis tests for simulation data. The importance of p-values and rigorous hypothesis tests is drilled into budding ecologists early and often, so the mistake is understandable. We hope that this paper helps formalize the argument for modelers, and clarify the understanding of reviewers, of the use and misuse of statistical tests in modeling studies.

Is there any use for frequentist statistics in ecological simulation models?

We have argued that frequentists tests of null hypotheses are useless when comparing model simulation output with different parameter values or functional forms. However, there are situations in which statistical tools associated with frequentist significance testing can be used productively in concert with simulation models. For example, methods such as ANOVA provide a useful and familiar framework for partitioning variance and calculating effect sizes in multifactorial simulations. It is appropriate to use ANOVA in that way, provided one ignores (and does not report) the p-values calculated along the way.

Corell et al. (2012) provide a representative example of this usage. They examined the factors affecting dispersal distances of planktonic marine larvae in the Baltic Sea. They used a numerical hydrodynamic circulation model to describe three-dimensional flow field of ocean currents in the study region, and conducted simulations in which they released simulated Lagrangian particles (virtual 'larvae') into the flow field and observed their trajectories. Larval trajectories were affected by random turbulence, so the dispersal model includes stochasticity. The authors examined the effects of multiple factors on dispersal trajectories (e.g. spawning season, larval depth, duration of the larval period, etc.) and created a factorial design to examine each of 216 individual treatment combinations, with three replicate simulations per combination. They then analyzed their results using a five-way ANOVA. However, rather than use ANOVA to test the hypothesis that different parameter values produce different dispersal patterns (which is of course true), they focused on the variance components returned by ANOVA in order to determine which factor contributed to greater absolute variation in dispersal distance. This type of usage is perfectly reasonable (see also Legendre and De Cáceres 2013). Similar applications of statistical methods could be, for example, using a Komolgorov–Smirnov test to compare two distributions generated by data; again, however, one should focus on the effect size or test statistic itself, rather than the p-value.

Another context in which hypothesis testing is usefully applied to simulation results is when one desires to simulate the empirical measurement of a system. This might be done to test alternative statistical or experimental approaches in a system with known dynamics, or to determine how the output of a simulated process compares to observed data. In the first case, models are used to simulate both process and measurement error, and model analysis focuses on determining the level of empirical replication needed to detect a process (Hoban et al. 2012) or validating a new statistical method for detecting certain phenomena (Dakos et al. 2012). These studies are essentially statistical power analyses in which the known falsehood of the null hypothesis is taken as a given.

An additional case where statistics could appropriately be applied to simulation outputs is when simulation results are being compared to observed empirical data. For example, Walker and Cyr (2007) simulated neutral community dynamics and used statistics to determine whether those dynamics matched observed species abundance distributions. Similar comparisons are common in time series applications where model forecast skill is the statistic of interest (Sugihara 1994). Although the details of comparing simulated results to real data is a separate topic that is beyond the scope of our commentary, we note that standard frequentist approaches are also perilous in that context (Waller et al. 2003), and newer methods such as Approximate Bayesian Computation are more reliable (Hoban et al. 2012).

Conclusion

Ecologists increasingly appreciate the importance of stochastic variability, spatiotemporal heterogeneity, and complex nonlinearities to the dynamics of natural systems (Comins et al. 1992, Anderson et al. 2008, Berkley et al. 2010). This realization – coupled with the availability of fast computers – is motivating increasing reliance upon large simulation models. Here we have explained why it is inappropriate to rely upon frequentist statistical hypothesis tests designed for low-replication empirical experiments when comparing highly replicated outputs of simulation models. We have also pointed out a few suitable applications of frequentist statistics to simulation model output, and there are surely others.

Nonetheless, we call upon authors (and reviewers) to avoid the temptation to analyze model output with a focus on statistical significance. The key insights to be to gleaned from simulation models, as with empirical data, must come from interpretation of biological significance.

Acknowledgements – Authorship statement: all authors designed the study; JWW wrote the manuscript with input from all authors. AR, JFS, ACS and CW contributed equally and their author order is alphabetical. JWW was supported by Florida Sea Grant PD/12-18. ACS was supported by a Killam Fellowship. We thank the Subject Editor for suggestions that improved this manuscript. We also respectfully acknowledge the authors whose works we singled out for discussion. We thank O. Shelton, M. Mangel and the Subject Editor for suggestions that improved the manuscript.

References

Anderson, C. N. K. et al. 2008. Why fishing magnifies fluctuations in fish abundance. – Nature 452: 835–839.

Anderson, D. R. et al. 2000. Null hypothesis testing: problems, prevalence and an alternative. – J. Wildlife Manage. 64: 912–923.

Beninger, P. G. et al. 2012. Strengthening statistical usage in marine ecology. – J. Exp. Mar. Biol. Ecol. 426–427: 97–108.

Berkley, H. A. et al. 2010. Turbulent dispersal promotes species coexistence. – Ecol. Lett. 13: 360–371.

Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. – J. Am. Stat. Ass. 33: 526–542.

Berkson, J. 1942. Tests of significance considered as evidence. – J. Am. Stat. Ass. 37: 325–335.

Comins, H. N. et al. 1992. The spatial dynamics of host–parasitoid systems. – J. Anim. Ecol. 61: 735–748.

Corell, H. et al. 2012. Depth distribution of larvae critically affects their dispersal and the efficiency of marine protected areas. – Mar. Ecol. Prog. Ser. 467: 29–46.

Dakos, V. et al. 2012. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. – PLoS ONE 7:e41010.

Dauphin, G. et al. 2009. Host kairomone learning and foraging success in an egg parasitoid: a simulation model. – Ecol. Entomol. 34: 193–203.

Esther, A. et al. 2010. Sensitivity of plant functional types to climate change: classification tree analysis of a simulation model. – J. Veg. Sci. 21: 447–461.

Fulton, E. A. et al. 2005. Which ecological indicators can robustly detect effects of fishing? – ICES J. Mar. Sci. 62: 540–551.

Hoban, S. et al. 2012. Computer simulations: tools for population and evolutionary genetics. – Nat. Rev. 13: 110–113.

Hurlbert, S. H. and Lombardi, C. M. 2009. Final collapse of the Neyman–Pearson decision theoretic framework and rise of the neoFisherian. – Ann. Zool. Fenn. 46: 311–349.

Johnson, D. H. 1999. The insignificance of statistical significance testing. – J. Wildlife Manage. 63: 763–772.

Legendre, P. and De Cáceres, M. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. – Ecol. Lett. 16: 951–963.

Makler-Pick, V. et al. 2011. Exploring the role of fish in a lake ecosystem (Lake Kinneret, Isreal) by coupling an individual-based fish population model to a dynamic ecosystem model. – Can. J. Fish. Aquat. Sci. 68: 1265–1284.

Marshall, D. J. et al. 2008. Offspring size variation within broods as a bet-hedging strategy in unpredictable environments. – Ecology 89: 2506–2517.

Marzloff, M. P. et al. 2013. Sensitivity analysis and pattern-oriented validation of TRITON, a model with alternative community states: insights on temperate rocky reefs dynamics. – Ecol. Modell. 258: 16–32.

Oaten, A. and Murdoch, W. W. 1975a. Functional response and stability in predator–prey systems. – Am. Nat. 109: 289–298.

Oaten, A. and Murdoch, W. W. 1975b. Switching, functional response, and stability in predator–prey systems. – Am. Nat. 109: 299–318.

Samhouri, J. F. et al. 2009. Quantitative evaluation of marine ecosystem indicator performance using food web models. – Ecosystems 12: 1283–1298.

Savage, I. R. 1957. Nonparametric statistics. – J. Am. Stat. Ass. 52: 331–344.

Sugihara, G. 1994. Nonlinear forecasting for the classification of natural time series. – Phil. Trans. R. Soc. A 348: 477–495.

Tam, T. and Ang, P. O. 2012. Object-oriented simulation of coral competition in a coral reef community. – Ecol. Modell. 245: 111–120.

Travers, M. et al. 2006. Simulating and testing the sensitivity of ecosystem-based indicators to fishing in the southern Benguela ecosystem. – Can. J. Fish. Aquat. Sci. 63: 943–956.

Walker, S. C. and Cyr, H. 2007. Testing the standard neutral model of biodiversity in lake communities. – Oikos 116: 143–155.

Waller, L. A. et al. 2003. Monte Carlo assessments of goodness-of-fit for ecological simulation models. – Ecol. Modell. 164: 49–63.

Yoccoz, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. – Bull. Ecol. Soc. Am. 72: 106–111.