A Guide to Bayesian Model Checking for Ecologists

Paul B. Conn[1,5], Devin S. Johnson[1], Perry J. Williams[2,3], Sharon R. Melin[1], and Mevin B. Hooten[4,2,3]

[1]*Marine Mammal Laboratory, NOAA, National Marine Fisheries Service, Alaska Fisheries Science Center, 7600 Sand Point Way NE, Seattle, WA 98115 USA*

[2]*Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523 USA*

[3]*Department of Statistics, Colorado State University, Fort Collins, CO 80523 USA*

[4]*U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, CO 80523 USA*

Abstract. Checking that models adequately represent data is an essential component of applied statistical inference. Ecologists increasingly use hierarchical Bayesian statistical models in their research. The appeal of this modeling paradigm is undeniable, as researchers can build and fit models that embody complex ecological processes while simultaneously accounting for observation error. However, ecologists tend to be less focused on checking model assumptions and assessing potential lack of fit when applying Bayesian methods than when applying more traditional modes of inference such as maximum likelihood. There are also multiple ways of assessing the fit of Bayesian models, each of which has strengths and weaknesses. For instance, Bayesian p-values are relatively easy to compute, but are well known to be conservative, producing p-values biased toward 0.5. Alternatively, lesser known approaches to model checking, such as prior predictive

[5]Email: paul.conn@noaa.gov

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/ecm.1314

checks, cross-validation probability integral transforms, and pivot discrepancy measures may produce more accurate characterizations of goodness-of-fit but are not as well known to ecologists. In addition, a suite of visual and targeted diagnostics can be used to examine violations of different model assumptions and lack of fit at different levels of the modeling hierarchy, and to check for residual temporal or spatial autocorrelation. In this review, we synthesize existing literature to guide ecologists through the many available options for Bayesian model checking. We illustrate methods and procedures with several ecological case studies, including i) analysis of simulated spatio-temporal count data, (ii) N-mixture models for estimating abundance of sea otters from an aircraft, and (iii) hidden Markov modeling to describe attendance patterns of California sea lion mothers on a rookery. We find that commonly used procedures based on posterior predictive p-values detect extreme model inadequacy, but often do not detect more subtle cases of lack of fit. Tests based on cross-validation and pivot discrepancy measures (including the "sampled predictive p-value") appear to be better suited to model checking and to have better overall statistical performance. We conclude that model checking is necessary to ensure that scientific inference is well founded. As an essential component of scientific discovery, it should accompany most Bayesian analyses presented in the literature.

Key words: *Bayesian model checking, Bayesian p-value, goodness-of-fit, hierarchical model, model diagnostics, posterior checks*

Introduction

Ecologists increasingly use Bayesian methods to analyze complex hierarchical models for natural systems (Hobbs and Hooten 2015). There are clear advantages of adopting a

Bayesian mode of inference, as one can entertain models that were previously intractable using common modes of statistical inference (e.g., maximum likelihood). Ecologists use Bayesian inference to fit rich classes of models to their datasets, allowing them to separate measurement error from process error, and to model features such as temporal or spatial autocorrelation, individual level random effects, and hidden states (Link et al. 2002, Clark and Bjørnstad 2004, Cressie et al. 2009). Applying Bayesian calculus also results in posterior probability distributions for parameters of interest; used together with posterior model probabilities, these can provide the basis for mathematically coherent decision and risk analysis (Link and Barker 2006, Berger 2013, Williams and Hooten 2016).

Ultimately, the reliability of inference from a fitted model (Bayesian or otherwise) depends on how well the model approximates reality. There are multiple ways of assessing a model's performance in representing the system being studied. A first step is often to examine diagnostics that compare observed data to model output to pinpoint if and where any systematic differences occur. This process, which we term *model checking*, is a critical part of statistical inference because it helps diagnose assumption violations and illuminate places where a model might be amended to more faithfully represent gathered data. Following this step, one might proceed to compare the performance of alternative models embodying different hypotheses using any number of model comparison or out-of-sample predictive performance metrics (see Hooten and Hobbs 2015, for a review) to gauge the support for alternative hypotheses or optimize predictive ability (Fig. 1).

Non-Bayesian statistical software often include a suite of goodness-of-fit diagnostics that examine different types of lack of fit (Table 1). For instance, when fitting generalized linear (McCullagh and Nelder 1989) or additive (Wood 2006) models in the R programming environment (R Development Core Team 2017), one can easily access

diagnostics such as quantile-quantile, residual, and leverage plots. These diagnostics allow one to assess the assumed probability model, to examine whether there is evidence of heteroskedasticity, and to pinpoint outliers. Likewise, in capture-recapture analysis, there are established procedures for assessing overall fit and departures from specific model assumptions that are coded in user-friendly software such as U-CARE (Choquet et al. 2009). Results of such goodness-of-fit tests are routinely reported when publishing analyses in the ecological literature.

The implicit requirement that one conduct model checking exercises is not often adhered to when reporting results of Bayesian analyses. For instance, a search of *Ecology* articles published in 2014 indicated that only 25% of articles employing Bayesian analysis on real datasets reported any model checking or goodness-of-fit testing (Fig. 2). There are several reasons why Bayesian model checking (hereafter, BMC) is uncommon. First, it likely has to do with inertia; the lack of precedent in ecological literature may lead some authors looking for templates on how to publish Bayesian analyses to conclude that model checking is unnecessary. Second, when researchers seek to publish new statistical methods, applications may be presented more as proof-of-concept exhibits than as definitive analyses that can stand up to scrutiny on their own. In such studies, topics like goodness-of-fit and model checking are often reserved for future research, presumably in journals with less impact. Third, all of the articles we examined did a commendable job in reporting convergence diagnostics to support their contention that MCMC chains had reached their stationary distribution. Perhaps there is a mistaken belief among authors and reviewers that convergence to a stationary distribution, combined with a lack of prior sensitivity, implies that a model fits the data. In reality, convergence diagnostics such as trace plots only allow us to check the algorithm for fitting the model, not the model itself. Finally, it

may just be a case of fatigue: it takes considerable effort to envision and code complex hierarchical models of ecological systems, and the extra step of model checking may seem burdensome. Regardless of the reason for not reporting BMC, it is concerning because poorly specified models can lead to incorrect scientific inference.

If we accept the premise that Bayesian models should be routinely checked for compatibility with data, a logical next question is how best to conduct such checks. Unfortunately, there is no single best answer. Most texts in ecology (e.g., King et al. 2009, Link and Barker 2010, Kéry and Schaub 2012) focus on posterior predictive checks, as pioneered by Guttman (1967), Rubin (1981, 1984), and Gelman et al. (1996) (among others). These procedures are also the main focus of popular Bayesian analysis texts (e.g., Cressie and Wikle 2011, Gelman et al. 2014) and are based on the intuitive notion that data simulated from the posterior distribution should be similar to the data one is analyzing. However, "Bayesian p-values" generated from these tests tend to be conservative (biased toward 0.5) because the data are used twice (once to fit the model and once to test the model; Bayarri and Berger 2000, Robins et al. 2000). Depending on the data, the conservatism of Bayesian p-values can be considerable (Zhang 2014) and can be accompanied by an inability to detect lack of fit (Yuan and Johnson 2012, Zhang 2014). By contrast, other less familiar approaches (such as prior predictive checks, sampled posterior p-values, cross-validated probability integral transforms, and pivot discrepancy measures) may produce more accurate characterizations of model fit.

In this monograph, we collate relevant statistical literature with the goal of providing ecologists with a practical guide to BMC. We start by defining a consistent notation that we use throughout the paper. Next, we inventory a number of BMC procedures, providing pros and cons for each approach. We illustrate BMC with several examples; code to

implement these examples are available in an accompanying R package, `HierarchicalGOF` (Conn et al. 2018). In the first example, we use simulation to study the properties of a variety of BMC procedures applied to spatial models for count data. In the second example, we apply BMC procedures to check the closure assumption (i.e., that the population being sampled is closed with respect to births, deaths, and movement) of N-mixture models, using both simulated data and data from northern sea otters (*Enhydra lutris kenyoni*) in Glacier Bay, Alaska, U.S.A. Finally, we apply BMC to examine attendance patterns of California sea lions (CSL; *Zalophus californianus*) using capture-recapture data from a rookery on San Miguel Island, California, U.S.A. We conclude with several recommendations on how model checking results should be presented in the ecological literature.

Background and notation

Before describing specific model checking procedures, we first establish common notation. Bayesian inference seeks to describe the posterior distribution, $[\boldsymbol{\theta}|\mathbf{y}]$, of model parameters, $\boldsymbol{\theta}$, given data, $\mathbf{y}$. Throughout the paper, we use bold lowercase symbols to denote vectors. Matrices are represented with bold, uppercase symbols, while Roman (unbolded) characters are used for scalars. The bracket notation '$[\ldots]$' denotes a probability distribution or mass function, and a bracket with a vertical bar '|' denotes that it is a conditional probability distribution (Gelfand and Smith 1990).

The posterior distribution is often written as

$$[\boldsymbol{\theta}|\mathbf{y}] = \frac{[\mathbf{y}|\boldsymbol{\theta}][\boldsymbol{\theta}]}{[\mathbf{y}]}, \quad (1)$$

where $[\mathbf{y}|\boldsymbol{\theta}]$ is the assumed probability model for the data, given parameters (i.e., the likelihood), $[\boldsymbol{\theta}]$ denotes the joint prior distribution for parameters, and $[\mathbf{y}]$ is the marginal distribution of the data. In Bayesian computation, the denominator $[\mathbf{y}]$ is frequently ignored because it is a fixed constant that does not affect inference (although it is needed when computing Bayes factors for model comparison and averaging; Link and Barker 2006). The exact mechanics of Bayesian inference are well reviewed elsewhere (e.g., King et al. 2009, Link and Barker 2010, Hobbs and Hooten 2015), and we do not attempt to provide a detailed description here. For the remainder of this treatment, we assume that the reader has familiarity with the basics of Bayesian inference, including Markov Chain Monte Carlo (MCMC) as a versatile tool for sampling from $[\boldsymbol{\theta}|\mathbf{y}]$.

In describing different model checking procedures, we often refer to data simulated under an assumed model. We use $\mathbf{y}_i^{rep}$ to denote the ith simulated dataset under the model that is being checked. In some situations, we may indicate that the dataset was simulated using a specific parameter vector, $\boldsymbol{\theta}_i$; in this case, denote the simulated dataset as $\mathbf{y}_i^{rep}|\boldsymbol{\theta}_i$. We use the notation $T(\mathbf{y}, \boldsymbol{\theta})$ to denote a discrepancy function that is dependent upon data and possibly the parameters $\boldsymbol{\theta}$. For instance, we might compare the discrepancy $T(\mathbf{y}, \boldsymbol{\theta})$ calculated with observed data to a distribution obtained by applying $T(\mathbf{y}^{rep}, \boldsymbol{\theta})$ to multiple replicated data sets. Examples of candidate discrepancy functions are provided in Table 2.

Model checking procedures

Our goal in this section is to review relevant BMC procedures for typical models in ecology, with the requirement that such procedures be accessible to statistically-minded ecologists. As such, we omit several approaches that have good statistical properties but have been

criticized (e.g., Johnson 2007*b*, Zhang 2014) as too computationally intensive, conceptually difficult, or problem-specific. For instance, we omit consideration of double-sampling methods that may increase the computational burden of a Bayesian analysis by an order of magnitude (Johnson 2007*b*), including "partial posterior" and "conditional predictive" p-values (e.g., Bayarri and Berger 1999, Robins et al. 2000, Bayarri and Castellanos 2007). A brief summary of the model checking procedures we consider is provided in Table 3; we now describe each of these approaches in greater depth.

Prior predictive checks

Box (1980) argued that the hypothetico-deductive process of scientific learning can be embodied through successive rounds of model formulation and testing. According to his view, models are built to represent current theory and an investigator's knowledge of the system under study; data are then collected to evaluate how well the existing theory (i.e., model) matches up with reality. If necessary, the model under consideration can be amended, and the process repeats itself.

From a Bayesian standpoint, such successive rounds of *estimation* and *criticism* can be embodied through posterior inference and model checking, respectively (Box 1980). If one views a model, complete with its assumptions and prior beliefs, as a working model of reality, then data simulated under a model should look similar to data gathered in the real world. This notion can be formalized through a prior predictive check, where replicate data $\mathbf{y}^{rep}$ are simulated via

$$\begin{aligned} \boldsymbol{\theta}^{rep} &\sim [\boldsymbol{\theta}] \\ \mathbf{y}^{rep} &\sim [\mathbf{y}|\boldsymbol{\theta}^{rep}] \end{aligned} \tag{2}$$

and then compared to observed data $\mathbf{y}$ via a discrepancy function (Appendix S1, Alg. 1).

When the prior distribution $[\boldsymbol{\theta}]$ is proper (i.e., integrates to 1.0), p-values from prior predictive checks are uniformly distributed under the null model (Bayarri and Berger 2000). The main problem with this approach is that prior distributions must be able to predict the likely range of data values; therefore, they require substantial expert opinion or data from previous studies. In our experience, when Bayesian inference is employed in ecological applications, this is not often the case. Still, prior predictive checks may be useful for Bayesian models that serve as an embodiment of current theory about a study system (e.g., population or ecosystem dynamics models). Alternatively, a subset of data (test data) can be withheld when fitting a model, and the posterior distribution $[\boldsymbol{\theta}|\mathbf{y}]$ can be substituted for $[\boldsymbol{\theta}]$ in Eq. 2. If used in this manner, prior predictive checks can be viewed as a form of cross-validation, a subject we examine in a later subsection (see *cross-validation tests*).

Prior predictive checks appear to have found little use in applied Bayesian analysis (but see Dey et al. 1998), at least in the original form proposed by Box (1980). However, they are important as historical precursors of modern day approaches to Bayesian model checking. Further, several researchers have recently used discrepancy measures calculated on prior predictive data sets to help calibrate posterior predictive (e.g., Hjort et al. 2006) or joint pivot discrepancy (Johnson 2007*a*) p-values so that they have a uniform null distribution. These calibration exercises are not conceptually difficult but do have a high computational burden (Yuan and Johnson 2012). The properties (e.g., type I error probabilities, power) of p-values produced with these methods also depend critically on the similarity of the real world data-generating process with the prior distributions used for calibration (Zhang 2014).

Posterior predictive checks

Posterior predictive checks are the dominant form of Bayesian model checking advanced in statistical texts read by ecologists (e.g., King et al. 2009, Link and Barker 2010, Kéry and Schaub 2012, Gelman et al. 2014). Although sample size was small ($n = 25$), a survey of recent *Ecology* volumes indicated that posterior predictive checks are also the dominant form of BMC being reported in ecological literature (Fig. 2). Posterior predictive checks are based on the intuition that data simulated under a fitted model should be comparable to the real-world data the model was fitted to. If observed data differ from simulated data in a systematic fashion (e.g., excess zeros, increased skew, increased variance, lower kurtosis), it indicates that model assumptions are not being met.

Posterior predictive checks can be used to look at differences between observed and simulated data graphically, or can be used to calculate "Bayesian p-values" (Appendix S1, Alg. 2). Bayesian p-values necessarily involve application of a discrepancy function, $T(\mathbf{y}, \boldsymbol{\theta})$, for comparing observations to simulated data. Omnibus discrepancy functions help diagnose global lack of fit, while targeted discrepancy functions can be used to look for systematic differences in specific data features (Table 2). Posterior predictive checks involve cyclically drawing parameter values from the posterior distribution (i.e., $\boldsymbol{\theta}_i \sim [\boldsymbol{\theta}|\mathbf{y}]$) and then generating a replicate dataset for each i, $\mathbf{y}_i^{rep} \sim [\mathbf{y}|\boldsymbol{\theta}_i]$, to compute the reference distribution for the discrepancy test statistic (Gelman et al. 2014, Appendix S1, Alg. 2).

Posterior predictive checks are straightforward to implement. Unfortunately, Bayesian p-values based on these checks tend to be conservative in the sense that the distribution of p-values calculated under a null model (i.e., when the data generating model and estimation model are the same) is often dome-shaped instead of the uniform distribution

expected of frequentist p-values (Robins et al. 2000). This feature arises because data are used twice: once to approximate the posterior distribution and to simulate the reference distribution for the discrepancy measure, and a second time to calculate the tail probability (Bayarri and Berger 2000). As such, the power of posterior predictive Bayesian p-values to detect significant differences in the discrepancy measure is low. Evidently, the degree of conservatism can vary across data, models, and discrepancy functions, making it difficult to interpret or compare Bayesian p-values across models. In an extreme example, Zhang (2014) found that posterior predictive p-values almost never rejected a model, even when the model used to fit the data differed considerably from the model used to generate it.

Another possible criticism of posterior predictive checks is that they rely solely on properties of simulated and observed data. Given that a lack of fit is observed, it may be difficult to diagnose where misspecification has occurred within the modeling hierarchy (e.g., priors, mean structure, choice of error distribution). Further, a poorly specified mean structure (e.g., missing important covariates) may still result in reasonable fit if the model is made sufficiently flexible (e.g., via random effects or covariance).

These cautions do not imply that posterior predictive checks are devoid of value. Indeed, given that tests are conservative, small (e.g., < 0.05) or very large (e.g., > 0.95) p-values strongly suggest lack of fit. Further, graphical displays (see *Graphical techniques*) and targeted discrepancies (Table 2) may help pinpoint common assumption violations (e.g., lack of independence, zero inflation, overdispersion). However, it is often less clear how to interpret p-values and discrepancies that indicate no (or minor) lack of fit. In these cases, it seems necessary to conduct simulation-based exercises to determine the range of p-values that should be regarded as extreme, and to possibly calibrate the observed p-value with those obtained in simulation exercises (e.g., Dey et al. 1998, Hjort et al. 2006).

Some practical suggestions may help to reduce the degree of conservatism of posterior predictive p-values. Lunn et al. (2013) suggest that the level of conservatism depends on the discrepancy function used; discrepancy functions that are solely a function of simulated and observed data (e.g., proportion of zeros, distribution of quantiles) may be less conservative than those that also depend on model parameters (e.g., summed Pearson residuals). Similarly, Marshall and Spiegelhalter (2003) suggest reducing the impact of the double use of data by iteratively simulating random effects when generating posterior predictions for each data point, a procedure they term a "mixed predictive check" (also called "ghosting"). For instance, rather than basing a posterior prediction directly on random effect realizations available from MCMC sampling, we could instead simulate random effects from a leave-one-out distribution. For an example of this latter approach, see *Spatial regression simulations*.

Sampled posterior p-values

Posterior predictive checks rely on multiple draws from a posterior distribution. Alternatively, one can simulate a single parameter vector from the posterior, $\tilde{\boldsymbol{\theta}} \sim [\boldsymbol{\theta}|\mathbf{y}]$, and then generate replicate datasets conditional on this parameter vector alone (i.e., $\mathbf{y}_i^{rep} \sim [\mathbf{y}|\tilde{\boldsymbol{\theta}}]$), otherwise calculating the p-value in the same manner. This choice may seem strange because the resulting p-value can vary depending upon the posterior sample, $\tilde{\boldsymbol{\theta}}$, but a variety of theoretical arguments (e.g., Johnson 2004; 2007*a*, Yuan and Johnson 2012, Gosselin 2011) and several simulation studies (e.g., Gosselin 2011, Zhang 2014) suggest that it may be a preferable choice, both in terms of Type I error control and power to detect lack of fit. In fact, sampled posterior p-values are guaranteed to at least have an asymptotic uniform distribution under the null (Gosselin 2011). Sampled posterior p-values

can also be calculated using pivotal discrepancy measures, reducing computational burden (i.e., eliminating the requirement that replicate datasets be generated). We describe an example of this approach in *Spatial regression simulations*.

Pivotal discrepancy measures (PDMs)

In addition to overstated power to detect model lack of fit, posterior predictive p-values are limited to examining systematic differences between observed data and data simulated under a hypothesized model. As such, there is little ability to examine lack of fit at higher levels of modeling hierarchy. One approach to conducting goodness-of-fit tests at multiple levels of the model is to use discrepancy functions based on pivotal quantities (Johnson 2004, Yuan and Johnson 2012). Pivotal quantities are random variables that can be functions of data, parameters, or both, and that have known probability distributions that are independent of parameters (e.g., Casella and Berger 1990, section 9.2.2). For instance, if

$$y \sim \mathcal{N}(\mu, \sigma^2)$$

then $z = \frac{y-\mu}{\sigma}$ has a standard $\mathcal{N}(0,1)$ distribution. Thus, z is a pivotal quantity in that it has a known distribution independent of μ or σ.

This suggests a potential strategy for assessing goodness-of-fit; for instance, in a Bayesian regression model

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}), \tag{3}$$

where $\mathbf{X}$ represents a design matrix, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\mathbf{I}$ is an

identity matrix, we might keep track of

$$z_{ij} = \frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}_j}{\sigma_j} \quad (4)$$

for each of $j \in 1, 2, \ldots, n$ samples from the posterior distribution (i.e., drawing each $(\boldsymbol{\beta}_j, \sigma_j)$ pair from $[\boldsymbol{\theta}|\mathbf{y}]$). Systematic departures of z_{ij} from the theoretical $\mathcal{N}(0, 1)$ distribution can point to model misspecification. Although we have focused on the data model in Eq. 3, note that the same approach could be used at higher levels of the modeling hierarchy.

The advantage of using PDMs is that the reference distribution is known and does not necessarily involve simulation of replicated datasets, $\mathbf{y}^{rep}$. In practice, there are several difficulties with using pivotal quantities as discrepancy measures in BMC. First, as with the sampled predictive p-value, p-values using PDMs are only guaranteed to be uniform under the null if calculated with respect to a single posterior parameter draw, $\tilde{\boldsymbol{\theta}} \sim [\boldsymbol{\theta}|\mathbf{y}]$. The joint distribution of PDMs calculated across $i \in 1, 2, \ldots, n$ samples from the posterior distribution are not independent because they depend on the same observed data, $\mathbf{y}$ (Johnson 2004). As with the Bayesian p-value calculated using a posterior predictive check, this latter problem can result in p-values that are conservative. Yuan and Johnson (2012) suggest comparing histograms of a pivotal discrepancy function $T(\mathbf{y}, \boldsymbol{\theta}_i)$ to its theoretical distribution, f, to diagnose obvious examples of model misspecification.

A second problem is that, to apply these techniques, one must first define a pivotal quantity and ascertain its reference distribution. Normality assessment is relatively straightforward using standardized residuals (e.g., Eq. 4), but pivotal quantities are not necessarily available for other distributions (e.g., Poisson). However, Yuan and Johnson (2012), building upon work of Johnson (2004), proposed an algorithm based on cumulative

distribution functions (CDFs) that can apply to any distribution, and at any level of a hierarchical model (Appendix S1, Alg. 3). For continuous distributions, this algorithm works by defining a quantity $w_{ij} = g(y_{ij}, \boldsymbol{\theta})$ (this can simply be $w_{ij} = y_{ij}$) with a known CDF, F. Then, according to the probability integral transformation, $F(w)$ will be uniformly distributed if the the modeled distribution function is appropriate. Similarly, for discrete distributions, we can apply a randomization scheme (Smith 1985, Yuan and Johnson 2012) to transform discrete variables into continuously distributed uniform variates. For example, when y_{ij} has integer valued support, we can define

$$w_{ij} \quad = \quad F(y_{ij} - 1|\boldsymbol{\theta}) + u_{ij} f(y_{ij}|\boldsymbol{\theta}),$$

where u_{ij} is a continuosly uniform random variable on (0,1) and $F()$ and $f()$ are the cumulative mass and probability mass functions associated with $[\mathbf{y}|\boldsymbol{\theta}]$, respectively. In this case, w_{ij} will be uniformly and continuously distributed on (0,1) if the assumed distribution is reasonable; deviation from uniformity can point to model misspecification.

We have written the PDM algorithm in terms of the data distribution $[\mathbf{y}|\boldsymbol{\theta}]$ (Appendix S1), but the algorithm can be applied to any level of a hierarchical model. Further, the algorithm can be applied separately to different categories of mean response (e.g., low, medium, or high levels of predicted responses). These advantages are extremely appealing in that one can more thoroughly test distributional assumptions and look for places where lack of fit may be occurring, something that can be difficult to do with posterior predictive checks. We apply this algorithm in *Spatial regression simulations* and provide R code for applying this approach to generic MCMC data in the R package `HierarchicalGOF` accompanying this paper (see *Software* for more information).

Cross-validation tests

Cross-validation consists of leaving out one or more data points, conducting an analysis, and checking how model predictions match up with actual observations. This process is often repeated sequentially for different partitions of the data. It is most often used to examine the relative predictive performance of different models (i.e., for model selection, Arlot and Celisse 2010). However, one can also use cross-validation to examine model fit and determine outliers. The primary advantage of conducting tests in this fashion is that there is no duplicate use of data as with posterior predictive tests or those based on joint PDMs. However, cross-validation can be computationally intensive (sometimes prohibitively so) for complicated Bayesian hierarchical models.

One approach to checking models using cross-validation is the cross-validated probability integral transform (PIT) test, which has long been exploited to examine the adequacy of probabilistic forecasts (e.g., Dawid 1984, Früiiwirth-Schnatter 1996, Gneiting et al. 2007, Czado et al. 2009). These tests work by simulating data at a set of times or locations, and evaluating the CDF of the predictions at a set of realized data (where realized data are not used to fit the model). This can be accomplished in a sequential fashion for time series data, or by withholding data (as with leave-one-out cross-validation). In either case, if the distribution of the CDF values from the realized data diverge from a Uniform(0,1) distribution it is indicative of model deficiency. In particular, a U-shape suggests an underdispersed model, a dome-shape suggests an overdispersed model, and skew (i.e., mean not centered at 0.5) suggests bias. Congdon (2014) provided an algorithm for computing PIT diagnostic histograms for both continuous and discrete data in Bayesian applications (see Appendix S1, Alg. 4).

Cross-validation can also be useful for diagnosing outliers in spatial modeling applications. For instance, Stern and Cressie (2000) and Marshall and Spiegelhalter (2003) use it to identify regions that have inconsistent behavior relative to the model. Such outliers can indicate that the model does not sufficiently explain variation in responses, that there are legitimate "hot spots" worthy of additional investigation (Marshall and Spiegelhalter 2003), or both.

For certain types of data and models it is possible to approximate leave-one-out cross-validation tests with a single sample from the posterior distribution. For instance, in random effects models, importance weighting and resampling can be used to approximate the leave-one-out distribution (Stern and Cressie 2000, Qiu et al. 2016). Similarly, Marshall and Spiegelhalter (2007) use a "ghosting" procedure to resample random effects and thereby approximate the leave-one-out distribution. When applicable, such approaches have well known properties (i.e., a uniform distribution of p-values under the null; Qiu et al. 2016).

Residual tests

Lunn et al. (2013) suggest several informal tests based on distributions of Pearson and deviance residuals. These tests are necessarily informal in Bayesian applications because residuals all depend on $\boldsymbol{\theta}$ and are thus not truly independent as required in unbiased application of goodness-of-fit tests. Nevertheless, several rules of thumb can be used to screen residuals for obvious assumption violations. For example, standardized Pearson

residuals for continuous data,

$$r_i = \frac{y_i - E(y_i|\boldsymbol{\theta})}{\sqrt{\mathrm{Var}(y_i|\boldsymbol{\theta})}},$$

should generally take on values between -2.0 and 2.0. Values far from this range represent outliers. Similarly, for the Poisson and binomial distributions, a rule of thumb is that the mean saturated deviance should approximately equal sample size for a well-fitting model (Lunn et al. 2013).

For time series, spatial, and spatio-temporal models, failure to account for autocorrelation can result in bias and overstated precision (Lichstein et al. 2002). For this reason, it is important to look for evidence of residual spatio-temporal autocorrelation in analyses where data have a spatio-temporal index. There are a variety of metrics to quantify autocorrelation, depending upon the ecological question and types of data available (e.g., Perry et al. 2002). For Bayesian regression models, one versatile approach is to compute a posterior density associated with a statistic such as Moran's I (Moran 1950) or Getis-Ord G* (Getis and Ord 1992) on residuals. For example, calculating Moran's I for each posterior sample j relative to posterior residuals $\mathbf{y} - \mathrm{E}(\mathbf{y}|\boldsymbol{\theta}_j)$, a histogram of I_j values can be constructed; substantial overlap with zero suggests little evidence of residual spatial autocorrelation. Moran's I is dependent upon a pre-specified distance-weighting scheme, thus investigators can simulate a posterior sample of Moran's I at several different choices of weights or neighborhoods to evaluate residual spatial autocorrelation at different scales.

Graphical techniques

Many of the previously described tests require discrepancy functions, and it may be difficult to formulate such functions for different types of lack of fit (e.g., Table 1). Displaying model checking information graphically may lead to more rapid intuition about where models do or do not fit the data. Alternative plots can be made for each type of model checking procedure (e.g., posterior predictive checks, sampled predictive checks, or even PDMs). For instance, Ver Hoef and Frost (2003) plotted posterior predictive χ^2 discrepancy values for different sites where harbor seal counts had been performed. Models accounting for overdispersion clearly resulted in improved fit at a majority of sites. The consistency of predictions was clear in this case, whereas a single p-value one would not effectively communicate where and how predictions were inaccurate.

Gelman et al. (2014) argued that residual and binned residual plots are instructive for revealing patterns of model misspecification. In spatial problems, maps of residuals can be helpful in detecting whether lack of fit is spatially clustered. The types of plots that are possible are many and varied, so it is difficult to provide a comprehensive list in this space. However, we illustrate several types of diagnostic plots in the following examples.

Computing

We conduct all subsequent analyses using a combination of R (R Development Core Team 2017) and JAGS (Plummer 2003). We used R to simulate data and to conduct model testing procedures; JAGS was used to conduct MCMC inference and produce posterior predictions. We developed an R package, `HierarchicalGOF`, that contains all of our code. This package is publicly available at

`https://github.com/pconn/HierarchicalGOF/releases`, and has been archived on Zenodo (Conn et al. 2018). The code is predominantly model-specific; however, it can be used as a template for ecologists conducting their own model checking exercises.

Examples

Spatial regression simulations

We examined alternative model checking procedures for spatially explicit regression models applied to simulated count data. Such models are often used to describe variation in animal or plant abundance over space and time, and can be used to map abundance distributions or examine trends in abundance (e.g., Sauer and Link 2011, Conn et al. 2014). A common question when modeling count data is whether there is overdispersion relative to the commonly chosen Poisson distribution. In ecological data, several sources of overdispersion are often present, including a greater number of zero counts than expected under the Poisson (zero inflation; Agarwal et al. 2002), and heavier tails than predicted by the Poisson (Potts and Elith 2006, Ver Hoef and Boveng 2007). Another important question is whether there is residual spatial autocorrelation that needs to be taken into account for proper inference (Legendre 1993, Lichstein et al. 2002).

In this simulation study, we generated count data under a Poisson distribution where the true mean response is a function of a hypothetical covariate, spatially autocorrelated error, and additional Gaussian noise. Data simulated in this manner arise from a spatially autocorrelated Poisson-normal mixture, and can be expected to be overdispersed relative to the Poisson, in much the same way as a negative binomial distribution (a Poisson-gamma mixture). We then examined the effectiveness of alternative model checking procedures for

diagnosing incorrect model specification, such as when spatial independence is assumed. We also studied properties of model checking procedures when the correct estimation model is specified.

For a total of 1000 simulation replicates, this study consisted of the following steps:

1. Locate $n = 200$ points at random in a square study area $\mathcal{A}_1$, where $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \mathbb{R}^2$. Call the set of $n = 200$ points $\mathcal{S}$.

2. Generate a hypothetical, spatially autocorrelated covariate $\mathbf{x}$ using a Matérn cluster process on $\mathcal{A}_2$ (see Appendix S2).

3. Generate expected abundance for all $s \in \mathcal{S}$ as $\boldsymbol{\mu} = \exp(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\epsilon})$ where $\mathbf{X}$ is an $(n \times 2)$ design matrix, $\boldsymbol{\beta}$ are regression coefficients, $\boldsymbol{\eta}$ are spatially autocorrelated random effects (see Appendix S2), and $\boldsymbol{\epsilon}$ are iid Gaussian errors. The first column of $\mathbf{X}$ is a vector of all 1s, and the second column consists of $\mathbf{x}$.

4. Simulate count data, $y_i|\mu_i \sim \text{Poisson}(\mu_i)$, at each of the $i \in \{1, 2, \ldots, 200\}$ points.

5. Fit a sequence of three models to each data set according to the following naming convention:

   - `Pois0`: Poisson model with no overdispersion

$$Y_i \sim \text{Poisson}(\exp(\mathbf{x}_i'\boldsymbol{\beta}))$$

- `PoisMix`: A Poisson-normal mixture with iid error

$$\begin{aligned} Y_i &\sim \text{Poisson}(\exp(\nu_i)) \\ \nu_i &\sim \text{Normal}(\mathbf{x}_i'\boldsymbol{\beta}, \tau_\epsilon^{-1}), \end{aligned}$$

  where τ_ϵ^{-1} is the error variance

- `PoisMixSp`: The data-generating model, consisting of a Poisson-normal mixture with both independent and spatially autocorrelated errors induced by a predictive process (cf. Banerjee et al. 2008):

$$\begin{aligned} Y_i &\sim \text{Poisson}(\exp(\nu_i)) \\ \nu_i &\sim \text{Normal}(\mathbf{x}_i'\boldsymbol{\beta} + \eta_i, \tau_\epsilon^{-1}) \\ \eta_i &= \mathbf{w}_i'\tilde{\boldsymbol{\eta}} \\ \tilde{\boldsymbol{\eta}} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \end{aligned}$$

6. Finally, a number of model checking procedures were employed on each simulated dataset.

A depiction of the data-generating algorithm (i.e., steps 1-4) is provided in Fig. 3; mathematical details of this procedure, together with a description of Bayesian analysis methods used in step 5 are provided in Appendix S2. We now describe model checking procedures (step 6) in greater detail.

Posterior predictive p-values

For each dataset and statistical model, we calculated several posterior predictive p-values with different discrepancy measures. These included χ^2, Freeman-Tukey, and deviance-based omnibus p-values, as well as directed p-values examining tail probabilities (Table 2). Tail probabilities were examined by comparing the 95% quantile of simulated and estimated data.

For the `Pois0` model, calculation of posterior predictive p-values was straightforward; posterior predictions ($\mathbf{y}^{rep}$) were simulated from a Poisson distribution, with an expectation that depends on posterior samples of $[\boldsymbol{\beta}|\mathbf{y}]$. For the other two models (i.e., `PoisMix` and `PoisMixSp`), it was less obvious how best to calculate posterior predictions. For instance, we identified at least three ways to simulate replicated data, $\mathbf{y}^{rep}$ for `PoisMixSp` (Fig. 4). Initial explorations suggested similar performance of predictions generated via the schematics in Figs. 4A-B, but the approach in Fig. 4B was used in reported results. We also examined the relative performance of a "mixed predictive check" (Marshall and Spiegelhalter 2007, Fig. 4C) for the `PoisMixSp` model.

To calculate some of the omnibus discrepancy checks (Table 2), one must also specify a method for calculating the expectation, $E(y_i|\boldsymbol{\theta})$. As with posterior predictions, this calculation depends on what one admits to being a parameter (e.g., are the latent $\boldsymbol{\nu}$ variables part of the parameter set, $\boldsymbol{\theta}$?). We opted to start with the lowest level parameters possible. For instance, for `PoisMix` we calculate the expectation relative to the parameter set $\boldsymbol{\theta} \equiv \{\boldsymbol{\beta}, \tau_\epsilon\}$; as such, the lognormal expectation is $E(y_i|\boldsymbol{\theta}) = \exp(\mathbf{x}_i\boldsymbol{\beta} + 0.5\tau_\epsilon^{-1})$. For `PoisMixSp`, we compute the expectation relative to $\boldsymbol{\theta} \equiv \{\boldsymbol{\beta}, \tau_\epsilon, \tau_\eta\}$, so that $E(y_i|\boldsymbol{\theta}) = \exp(\mathbf{x}_i\boldsymbol{\beta} + 0.5(\tau_\epsilon^{-1} + \tau_\eta^{-1}))$.

Pivotal discrepancy measures

We used Alg. 3 (Appendix S1) to conduct PDM tests on each simulated data set and model type. For all models, we assessed fit of the Poisson stage; for the `PoisMix` and `PoisMixSp` models, we also applied PDM tests on the Gaussian stage (see e.g., Fig. 5). These tests produce a collection of p-values for each fitted model; one for each posterior parameter sample (i.e., one for each MCMC iteration). We used the median p-value from this collection to summarize overall PDM goodness-of-fit.

Sampled predictive p-values

In addition to the median p-value from applying PDM tests, we also sampled a single PDM p-value at random from each MCMC run. This p-value was used as the sampled predictive p-value for each fitted model.

K-fold cross-validation

We used a cross-validation procedure to estimate an omnibus p-value for the `PoisMix` model, but did not attempt to apply it to the `Pois0` or `PoisMixSp` models owing to high computational cost. To improve computational efficiency, we modified Alg. 4 (Appendix S1) to use k-fold cross-validation instead of leave-one-out cross-validation. For each simulated dataset, we partitioned data into $k = 40$ "folds" of $m = 5$ observations each. We then fit the `PoisMix` model to each unique combination of 39 of these groups, systematically leaving out a single fold for testing (each observation was left out of the analysis exactly once). We then calculated an empirical CDF value for each omitted

observation i as

$$u_i \;=\; n^{-1}\sum_{j=1}^{n} I(y_{ij}^{rep} < y_i) + 0.5I(y_{ij}^{rep} = y_i).$$

Here, $I(y_{ij}^{rep} < y_i)$ is a binary indicator function taking on the value 1.0 if the posterior prediction of observation i at MCMC sample j (y_{ij}^{rep}) is less than the observed data at i. The binary indicator function $I(y_{ij}^{rep} = y_i)$ takes on the value 1.0 if $y_{ij}^{rep} = y_i$.

According to PIT theory, the u_i values should be uniformly distributed on $(0, 1)$ if the model being tested does a reasonable job of predicting the data. For each simulated dataset, we used a χ^2 test (with ten equally space bins) to test for uniformity; the associated p-value was used as an omnibus cross-validation p-value.

Posterior Moran's I for spatial autocorrelation

To test for residual spatial autocorrelation, we calculated a posterior distribution for the Moran's I statistic on residuals for each model fitted to simulated data. For each of $j \in 1, 2, \ldots, n$ samples from the posterior distribution (e.g., for each MCMC sample), Moran's I was calculated using the residuals $\mathbf{y} - E(\mathbf{y}|\theta_j)$. For `Pois0`, we set $E(\mathbf{y}|\theta_j) = \exp(\mathbf{X}\boldsymbol{\beta})$; for `PoisMix` and `PoisMixSp`, we set $E(\mathbf{y}|\theta_j) = \exp(\boldsymbol{\nu})$.

Spatial regression simulation results

Posterior predictive p-values were extremely conservative, with p-values highly clustered near 0.5 under the null case where the data generating model and estimation model were the same (Fig. 7). In contrast, an unbiased test should generate an approximately uniform distribution of p-values under the null. Tests using the median p-value associated with

PDMs were also conservative, as were mixed predictive checks and those calculated relative to posterior Moran's I statistics. At least in this example, the mixed predictive check actually appeared slightly more conservative than posterior predictive checks. Posterior predictive checks that depended on parameters in the discrepancy function (e.g, χ^2, deviance-based discrepancies) appeared to be slightly more conservative than those that depended solely on observed and simulated data properties (e.g., the 'tail' discrepancy comparing upper quantiles). In fact, the only p-values that appeared to have good nominal properties were sampled predictive p-values and cross-validation p-values. We did not explicitly quantify null properties of cross-validation p-values, but these should be uniform under the null because the data used to fit and test the model were truly independent in this case.

For the `Pois0` model, the mean directed posterior predictive p-value examining tail probabilities was 0.09 over all simulated data sets; the means of all other p-values (posterior predictive and otherwise) were < 0.01. As such, all model checking procedures had high power to appropriately detect the inadequacy of the basic Poisson model. Examining a representative plot of over- and under-predictions shows the inadequacy of the `Pois0` model: overdispersion is clearly present, and residuals are spatially clustered (Fig. 6).

For the `PoisMix` model, only the cross-validation test, the Moran I test, and tests based on PDMs of the Gaussian portion of the model had any power to detect model inadequacy (Fig. 7). Of these, the sampled predictive p-value had higher power than the p-value based on the median PDM. The remaining model checking approaches (notably including those based on posterior predictive checks) had no power to detect model inadequacy (Fig. 7).

The need for closure: N-mixture models

N-mixture models are a class of hierarchical models that use count data collected from repeated visits to multiple sites to estimate abundance in the presence of an unknown detection probability (Royle 2004). That is, counts y_{ij} are collected during sampling visits $j = 1, \ldots, J$, at sites $i = 1, \ldots, n$, and are assumed to be independent binomial random variables, conditional on constant abundance N_i and detection probability p; $y_{ij} \sim \text{Binomial}(N_i, p)$. Additionally, N_i is assumed to be an independent random variable with probability mass function $[N_i|\boldsymbol{\theta}]$ (e.g., Poisson, negative binomial, Conway-Maxwell Poisson). The assumption of constant abundance $N_{ij} = N_i \; \forall j$ is critical for accurate estimates of N_i and p (Barker et al. 2017). In practice, this assumption implies that a population at site i is closed with respect to births, deaths, immigration, and emigration, for all replicate temporal surveys at the site. Violation of this assumption can lead to non-identifiability of the N and p parameters, or worse, posterior distributions that converge, but result in N_i being biased high and p being biased low (Kéry and Royle 2016, Appendix S3).

The appropriateness of the closure assumption has often been determined by judgment of the investigators, who assess whether time between replicate surveys is short relative to the dynamics of the system, and whether individual movement is small, compared to the size of sample plots (e.g., Efford and Dawson 2012; but see Dail and Madsen 2011, for a frequentist test of this assumption using a model selection approach). As an alternative, we consider the utility of BMC to assess the closure assumption for N-mixture models. We first consider a brief simulated example where truth is known. We then examine real data consisting of counts of sea otters from aerial photographs taken in Glacier Bay National

Park, southeastern Alaska. For additional model checking examples and other violations of assumptions of the N-mixture model, including zero-inflation, extra-Poisson dispersion, extra-binomial dispersion, unmodeled site covariates, and unmodeled detection covariates, see Kéry and Royle (2016, section 6.8).

Simulation

We examined the most common form of N-mixture model for ecological data,

$$\begin{aligned} y_{ij} &\sim \text{Binomial}(N_i, p_i), \\ N_i &\sim \text{Poisson}(\lambda_i), \\ \log(\lambda_i) &= \mathbf{x}_i'\boldsymbol{\beta}, \\ \text{logit}(p_i) &= \mathbf{w}_i'\boldsymbol{\alpha}, \end{aligned} \tag{5}$$

where p_i and the expected abundance λ_i depend on covariates $\mathbf{w}_i$ and $\mathbf{x}_i$, respectively. We used Eq. (5) to simulate data, with one additional step to induce violation of the closure assumption. We examined a series of eight cases where the closure assumption was increasingly violated by letting

$$N_{ij} \sim \text{Discrete-Uniform}(N_{i,j-1}(1-c), N_{i,j-1}(1+c)),$$

for $j = 2, \ldots, J$, and $c = \{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$, where c can be interpreted as the maximum proportion of the population that could move in or out of a site between $j-1$ and j. When c equals zero, $N_{i,j} = N_{i,j-1}$, and thus, $N_{i,j} = N_i$, and the closure assumption is met. For all values of c, we set $\boldsymbol{\beta} = (4, 1)'$ and $\boldsymbol{\alpha} = (1, -1)'$, $i = 1, \ldots, n = 300$, $j = 1, \ldots, J = 5$. The covariate matrices $\mathbf{X}$ and $\mathbf{W}$ each had

dimensions 300×2, where the first column was all ones, and the second column was generated by sampling from a Bernoulli distribution with probability 0.5 for all i. We then fit Eq. 5 to the generated data using a MCMC algorithm written in R. Using the fitted model, we assessed the effectiveness of posterior predictive and sampled predictive p-values for diagnosing the closure assumption. When $c = 0$, the model used to generate the data was the same as the model used to fit the data, and our model checking procedures should indicate no lack of model fit. In all other cases, the closure assumption was violated, with the degree of violation proportional to the value of c. Annotated `R` code, results, and figures from the simulation are provided in Appendix S3.

N-mixture results

When the closure assumption was met ($c = 0$), the estimated posterior distributions recovered true parameter values well, which was expected (Table 4, Appendix S3). The posterior predictive p-value was 0.48, and the sampled predictive p-value was 0.27, suggesting no lack of model fit from either model checking proceedure (Table 4).

When the closure assumption was violated (i.e., $c > 0$), MCMC chains appeared to converge (Appendix S3), and convergence was often supported by Gelman-Rubin diagnostics (Table 4). However, abundance was always overestimated when the closure assumption was violated, and the true abundance value used to simulate the data was always outside estimated 95% credible intervals (Table 4). The posterior predictive p-values did not suggest lack of model fit when $c < 0.10$, and suggested lack of model fit otherwise (Table 4). The sampled predictive p-value correctly identified violation in the closure assumption (assuming a type I error rate of 0.05) for all values of c, for this simulation (Table 4). The effective sample sizes of the MCMC chains were small due to the

autocorrelation between abundance and detection probability in the N-mixture model (Table 4). Mean abundance estimates erroneously increased, with increased violation in the closure assumption, and confidence intervals failed to cover the true abundance value by allowing just 5% of the population to move in or out of a site between surveys.

We note that assessing the closure assumption of N-mixture models using posterior predictive p-values and sampled predictive p-values may be challenging in some areas of the parameter space, because the biased parameter estimates obtained from fitting data from an open population can produce data $\mathbf{y}_i^{rep} \sim [\mathbf{y}|\boldsymbol{\theta}_{biased}]$ that are almost indistinguishable (i.e., similar first and second moments) from the open population data. Further, other scientifically plausible models where N_i (or λ_i) are not identifiable also lead to data that are indistinguishable from data generated under an N-mixture model (Barker et al. 2017). Thus, model-checking is an important step in evaluating a model, but is not a replacement for proper study design.

Estimating sea otter detection probability from aerial photographs

Williams et al. (2017) describe a framework for using aerial photograph data to fit N-mixture models, where photographs are taken such that a subset of images overlap in space. The subset of overlapping images provides temporal replication of counts of individuals at spatial locations that can be used to estimate p in the N-mixture modeling framework. To assess the utility of their approach, Williams et al. (2017) conducted an aerial survey in Glacier Bay National Park, southeastern Alaska. During the survey, they identified groups of sea otters at the surface of the ocean and then flew over the groups of sea otters multiple times. They captured an image of the group of sea otters during each flight over the group. In their study, a primary observer operated the camera, and a

secondary observer watched the groups of sea otters to ensure the closure assumption of N-mixture models was met. That is, whether sea otters dispersed out of, or into, the footprint of the photograph among temporal replicates. According to observer notes, 20 of the 21 groups of sea otters that were photographed multiple times did not appear to violate the closure assumption. For analysis, Williams et al. (2017) omitted the one site that appeared to violate the closure assumption. Here, we use Bayesian model checking as a formal method for assessing the closure assumption of two data sets that are used to fit the N-mixture model. The first data set is the complete set of 21 observations initially collected for Williams et al. (2017). The second data set is the data provided in Table 1 of Williams et al. (2017) which omits the problematic site. The full data set is provided in the R package `HierarchicalGOF` (Conn et al. 2018). As in our N-mixture model simulation study above, we used Bayesian p-values and sampled posterior predictive values to check our model. We used each data set to fit the model

$$
\begin{aligned}
y_{ij} &\sim \text{Binomial}(N_i, p), \\
N_i &\sim \text{Poisson}(\lambda_i), \\
\lambda_i &\sim \text{Gamma}(0.001, 0.001), \\
p &\sim \text{Beta}(1, 1),
\end{aligned}
$$

using an MCMC algorithm written in R (Appendix S3). The Bayesian p-value for the full data set (21 sites) was 0.048 and the sampled posterior predictive value was 0.059, suggesting potential lack of model fit. The Bayesian p-value for the restricted data set used in Williams et al. (2017) was 0.5630 and the sampled posterior predictive value was 0.823, suggesting no lack of model fit. Thus, model checking proceedures can provide a formal

method for examining the closure assumption of N-mixture models for our example, and corroborates the auxillary information collected by the observers. We note that successful identification of violation of the closure assumption in the sea otter case should not be taken as evidence that these will be readily detected in other cases.

Should I stay or should I go? Hidden Markov Models

In this example, we present another assessment of goodness-of-fit for a model that is quickly becoming popular within the ecological community, the Hidden Markov Model (HMM; Zucchini and MacDonald 2009). HMMs are a general class of models for time series data that describe the dynamics of a process in terms of potentially unobserverable (latent) states that generate observable data according to state-dependent distributions. Using HMMs, ecologists can construct models that make inference to biologically relevant 'states' (e.g., infection status, foraging/not foraging) even when data consist solely of cues (e.g., field observations, locations of satellite tags).

One implicit (and seldom tested) assumption of HMM models is that the amount of time spent within a state (the *residence time*) is geometrically distributed. The geometric distribution implies a strictly decreasing distribution of residence times, and may not be realistic for certain ecological time series. For instance, if a hidden state corresponds to "foraging," one might expect a dome-shaped distribution of residence times.

In this section, we use BMC to assess the assumption of geometrically distributed residence times in HMMs applied to California sea lion (CSL) rookery attendance patterns. We do this by comparing the fit of a Bayesian HMM, as well as the fit of an alternative Bayesian hidden *semi*-Markov model (HSMM) that allows more flexible residence time distributions.

The HMM is formed by considering a time series of categorical variables, $Z_1, \ldots, Z_T$ that represent the hidden states. For each t, $Z_t \in \{1, \ldots, S\}$, where S is the number of latent states. The Z_t process follows a Markov chain with transition matrix $\mathbf{\Gamma}_t$ in which the j, k entry is $\Gamma_{tjk} = [Z_t = k | Z_{t-1} = j]$. The state process is hidden (at least partially), so, the researcher is only able to make observation y_t with distribution $[y_t|Z_t]$ and observations are independent given the hidden states. For n independent individual replications, the complete likelihood is

$$[\mathbf{y}, \mathbf{Z}|\boldsymbol{\psi}, \mathbf{\Gamma}] \;\; = \;\; \prod_{i=1}^{n} \prod_{t=1}^{T} [y_{it}|Z_{it}, \boldsymbol{\psi}_t] \; [Z_{it}|Z_{i,t-1}, \mathbf{\Gamma}_t],$$

where $\boldsymbol{\psi}_t$ is a parameter vector for the observation process. For Bayesian inference within an MCMC algorithm, we make use of the forward algorithm (see Zucchini and MacDonald 2009) to integrate over the missing state process and evaluate the integrated likelihood $[\mathbf{y}|\boldsymbol{\psi}, \mathbf{\Gamma}]$, thus we can generate a posterior sample without having to sample $\boldsymbol{Z}$ in the process.

The CSL data are composed of a time series or capture-history of 66 females on San Miguel I., California over the course of 2 months (61 days) during the pupping season. It was noted whether or not a previously marked CSL female was seen on a particular day (i.e., $y_{it} = 1, 0$, respectively, $i = 1, \ldots, 66$ and $t = 1, \ldots, 61$). The probability of observing a particular CSL female on a given day depends on her unobserved reproductive state: (1) pre-birth, (2) neonatal, (3) at-sea foraging, and (4) on-land nursing. The detection probability for CSL females in the pre-birth state is likely to be low because without a pup they are not attached to the rookery and can come and go as they please. In the neonatal state the female remains on shore for approximately 5–7 days to nurse the newborn pup.

After this period, the female begins foraging trips where it feeds for several days and returns to nurse the pup. While the CSL female is at-sea it has a detection probability of 0.0. For females that have just given birth, or are returning from a foraging trip, they will be tending to their pups and are more available to be detected.

To make inference on the attendance patterns of the CSL we used an HMM with the state transition matrix

$$\mathbf{\Gamma}_t = \mathbf{\Gamma} = \begin{bmatrix} \gamma_1 & 1-\gamma_1 & 0 & 0 \\ 0 & \gamma_2 & 1-\gamma_2 & 0 \\ 0 & 0 & \gamma_3 & 1-\gamma_3 \\ 0 & 0 & 1-\gamma_4 & \gamma_4 \end{bmatrix}.$$

This allows the process to pass from each state to the next in the reproductive schedule with alternating (3) at-sea and (4) on-land states. Conditioning on the reproductive state, the observation model is

$$[y_{it}|Z_{it}] \quad = \quad \text{Bernoulli}(\psi(Z_{it})),$$

where the detection parameters are constrained as $\psi(1) = \psi_1$, $\psi(3) = 0$, and $\psi(2) = \psi(4) = \psi_2$. The parameters ψ_1 and ψ_2 represent pre-birth and after-birth detection probability.

To assess model fit, we used the Freeman-Tukey fit statistic

$$T(\mathbf{y}; \boldsymbol{\psi}, \mathbf{\Gamma}) \quad = \quad \sum_t \left(\sqrt{d_t} - \sqrt{E[d_t]}\right)^2,$$

where d_t is the number of observed detections on occasion t and $E[d_t]$ is the expected number of detections given by the HMM model. The Freeman-Tukey statistic is less sensitive to small expected values than other discrepancy functions (e.g., χ^2), which is important in this example since the expected number of detections is small in early summer. For day t, the expected number of detections is

$$E[d_t] \quad = \quad n\boldsymbol{\delta}'\boldsymbol{\Gamma}^{t-1}\boldsymbol{\psi},$$

were $\boldsymbol{\delta} = (1, 0, 0, 0)'$, as all animals start in the pre-birth state, and $\boldsymbol{\psi} = (\psi_1, \psi_2, 0, \psi_2)'$

Two versions of the HMM model were fit to the data, one in which ψ_1 and ψ_2 were constant through time and one in which they were allowed to vary with each occasion (shared additive time effect). For variable time ψ models, detection was parameterized logit (ψ_{lt}) = logit $(\psi_l) + \epsilon_t$ for $l = 1, 2$, $t = 1, \ldots, 61$, and $\epsilon_1 = 0$ for identifiability. We used the following prior distributions in this analysis:

- [logit (γ_k)] $\propto 1$
- $[\psi_l] = U(0, 1);\ l = 1, 2$
- $[\epsilon_t] \propto \exp\{-|\epsilon_t|/2\};\ t = 2, \ldots, 61.$

The Laplace prior for ϵ_t was used to shrink unnecessary deviations to zero.

A collapsed MCMC sampler using the forward algorithm to calculate $[\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\gamma}]$ was used so that the Z_{it} process did not have to be sampled. Each sampler was run for 50,000 iterations following burn-in. To calculate the reference distribution for the discrepancy function, replicated data were simulated at every 10th iteration. After fitting, the posterior predictive p-value for both models was ≈ 0, which strongly implies lack of fit. Although

individual detection heterogeneity might be the source of fit issues, examination of Figure 8 suggests a systematic positive bias in the initial days and a negative bias in the middle of season, indicating possible issues with basic model structure.

The Markov assumption of the latent state process implies that, after landing in state k, the amount of time spent there is geometrically distributed with parameter $1 - \gamma_k$. Further, this implies that the most common (i.e., modal) amount of time spent is one time step. As γ_k approaches 1, this distribution flattens out, but retains a mode of 1. An alternative model that relaxes this assumption is the HSMM. In the HSMM, the residence time is explicitly modeled and at the end of the residence period a transition is made to another state with probability $\tilde{\Gamma}_{jk}$. For an HSMM, $\tilde{\Gamma}_{kk} = 0$ because remaining in a state is governed by the residence time model. This extra generality comes at a computational cost; however, Langrock and Zucchini (2011) provide a method for calculating an HSMM likelihood with an HMM algorithm, such that the forward algorithm can still be used for inference.

In terms of the CSL analysis, the off-diagonal elements of the HSMM transition matrix occur at the same locations as in the HMM but are all equal to 1 because after the residence time has expired, the animal immediately moves to the next stage in the reproductive schedule (alternating between at-sea and on-land at the end). The residence time was modeled using a shifted Poisson(λ_k) distribution; that is, residence time minus 1 is Poisson distributed. We set prior distributions for residence time parameters as $[\log \lambda_k] \propto 1$. Prior distributions for the detection parameters remained the same as before. Using the "HSMM as HMM" technique of Langrock and Zucchini (2011), we sampled the posterior distributions using the same MCMC algorithm as in the HMM case.

The p-value for the Tukey fit statistic under the constant time model was 0.09, so, it

was an improvement over the HMM models, but still low enough to cause concern. However, for the time varying ψ HSMM model, the p-value was 0.82, indicating a substantial improvement in fit. By reducing the probability that an animal would transition from pre-birth to birth states immediately after the start of the study, the HSMM model was able to accommodate a similar average residence time to the HMM without maintaining a mode of 1 (Figure 8), producing a more biologically realistic model.

Discussion

Ecologists increasingly use hierarchical Bayesian models to analyze their data. Such models are powerful, allowing researchers to represent complex, and often dynamic, ecological processes. Under the Bayesian calculus, ecologists can partition observation error from process error, produce detailed predictions, and properly carry through uncertainty when making inferences. The ability to build complex models is exciting, but does not absolve us of the need to check whether models fit our data. If anything, complicated models should be subject to *more* scrutiny than simple models, as there are more places where things can go wrong.

One way to ensure a model fits the data is simply to build a sufficiently flexible model. To take an extreme example, a saturated model (one where there is a separate parameter for each datum) fits the data perfectly. No one would actually do this in practice; science proceeds by establishing generalities, and there is no generality implicit in such a model. Further, there is no way to predict future outcomes. Indeed, models with high complexity can fit the data well, but may have poorer predictive ability and inferential value than a model of lower complexity (Burnham and Anderson 2002, Hooten and Hobbs 2015).

When unsure of the desirable level of complexity or number of predictive covariates to include in a model, one approach is to fit a number of different models and to average among the models according to some criterion (e.g., Green 1995, Hoeting et al. 1999, Link and Barker 2006). Still, unless one conducts model checking exercises, there is no assurance that *any* of the models fit the data. Further, there are costs to model averaging, especially in Bayesian applications where considerable effort is needed to implement an appropriate algorithm. In such cases, it may make more sense to iterate on a single model (Ver Hoef and Boveng 2015), and thus, model checking becomes even more important.

We have described a wide variety of Bayesian model checking procedures with the aim of providing ecologists an overview of possible approaches, including strengths and limitations. Our intention is not to be prescriptive, but to guide ecologists into making an appropriate choice. For instance, using simulation, we showed that the popular posterior predictive p-value (and several other metrics) can have a larger than nominal α value, so that our ability to "reject" the null hypothesis that data arose from the model is overstated. In the spatial regression example, the Bayesian p-value often failed to reject models without spatial structure even when data were simulated with considerable spatial autocorrelation. The overstated probability of rejection is due to the double use of data, which are used both to fit the model and also to calculate a tail probability. However, as shown in the sea otter and California sea lion examples, the posterior predictive p-value can be useful in diagnosing obvious cases of lack of fit and in producing more biologically realistic models. Other choices, such as those based on cross-validation, have better stated properties and would be preferable on theoretical grounds, but may be more difficult to implement. Regardless of the approach(es) chosen, ecologists can start incorporating BMC as a standard part of their analysis workflow (e.g., Fig. 1). As in the case of 'p-hacking'

(Head et al. 2015), care should be taken to choose appropriate goodness of fit measures without first peeking at results (i.e. employing multiple discrepancy measures and only reporting those that indicate adequate fit).

In ecology, simplistic processes are rare: we often expect heterogeneity among individuals, patchy responses, and variation that is partially unexplained by gathered covariates. Therein lies an apparent contradiction: we expect lack of fit in our models, but still want to minimize biases attributable to poor modeling assumptions. From our perspective, the goal of model checking should not be to develop a model that fits the data perfectly, but rather to probe models for assumption violations that result in systematic errors. For instance, employing an underdispersed model (e.g., the Poisson instead of the negative binomial or the normal instead of the t distribution) will often lead to estimates that are too precise and to predictions that are less extreme than real world observations. Basing inference on such a model could have real world implications if used to inform environmental policy, as it would tend to make decision makers overly confident in their projections. In the case of basic science, such over-confidence can have real ramifications for confirmation or refutation of existing theory. It is therefore vital that we do a better job of conducting and reporting the results of model checks when publishing ecological research.

So far, we have viewed Bayesian model checking primarily as a confirmatory procedure used to validate (or more precisely, invalidate) modeling assumptions. However, model checking can be an important tool for scientific discovery. If a model fails to fit the data, the inquisitive ecologist will want to ask "why?" In some cases, deviations in model predictions from observations may actually suggest alternative scientific hypotheses worthy of additional investigation. We thus urge ecologists to view model checking not as a box to check off on their way to publication, but as an important tool to learn from their data and

hasten the process of scientific discovery.

Acknowledgments

We thank B. Brost, A. Ellison, T. Ergon, J. Ver Hoef, and an anonymous reviewer for comments on previous versions of our manuscript, and J. Laake for initial ideas on modeling detectability when estimaing CSL attendance patterns. The findings and conclusions in the paper of the NOAA authors do not necessarily represent the views of the reviewers nor the National Marine Fisheries Service, NOAA. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Agarwal, D. K., A. E. Gelfand, and S. Citron-Pousty. 2002. Zero-inflated models with application to spatial count data. Environmental and Ecological Statistics **9**:341–355.

Arlot, S., and A. Celisse. 2010. A survey of cross-validation procedures for model selection. Statistics Surveys **4**:40–79.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang. 2008. Stationary process approximation for the analysis of large spatial datasets. Journal of the Royal Statistical Society B **70**:825–848.

Barker, R. J., M. R. Schofield, W. A. Link, and J. R. Sauer. 2017. On the reliability of N-mixture models for count data. Biometrics **00**:10.1111/biom.12734.

Bayarri, M., and J. O. Berger. 2000. P values for composite null models. Journal of the American Statistical Association **95**:1127–1142.

Bayarri, M., and M. Castellanos. 2007. Bayesian checking of the second levels of hierarchical models. Statistical Science **22**:322–343.

Bayarri, M. J., and J. O. Berger, 1999. Quantifying surprise in the data and model verification. Pages 53–82 *in* J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors. Bayesian Statistics 6. Oxford University Press, London, U.K.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. Genetics **162**:2025–2035.

Berger, J. O. 2013. Statistical decision theory and Bayesian analysis. Springer Science & Business Media, New York.

Box, G. E. 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society. Series A (General) **143**:383–430.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd Edition. Springer-Verlag, New York.

Casella, G., and R. L. Berger. 1990. Statistical Inference. Duxbury Press, Belmont, California.

Choquet, R., J.-D. Lebreton, O. Gimenez, A.-M. Reboulet, and R. Pradel. 2009. U-CARE: Utilities for performing goodness of fit tests and manipulating CApture–REcapture data. Ecography **32**:1071–1074.

Clark, J. S., and O. N. Bjørnstad. 2004. Population time series: process variability, observation errors, missing values, lags, and hidden states. Ecology **85**:3140–3150.

Congdon, P. 2014. Applied Bayesian modelling. John Wiley & Sons, Hoboken, New Jersey.

Conn, P. B., D. S. Johnson, and P. J. Williams. 2018. HierarchicalGOF: 1.0.1. Zenodo. http://doi.org/10.5281/zenodo.1231501 .

Conn, P. B., J. M. Ver Hoef, B. T. McClintock, E. E. Moreland, J. M. London, M. F. Cameron, S. P. Dahle, and P. L. Boveng. 2014. Estimating multi-species abundance using automated detection systems: ice-associated seals in the eastern Bering Sea. Methods in Ecology and Evolution **5**:1280–1293.

Cressie, N., C. Calder, J. Clark, J. Ver Hoef, and C. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications **19**:553–570.

Cressie, N., and C. K. Wikle. 2011. Statistics for spatio-temporal data. Wiley, Hoboken, New Jersey.

Czado, C., T. Gneiting, and L. Held. 2009. Predictive model assessment for count data. Biometrics **65**:1254–1261.

Dail, D., and L. Madsen. 2011. Models for estimating abundance from repeated counts of an open metapopulation. Biometrics **67**:577–587.

Dawid, A. P. 1984. Statistical theory: the prequential approach. Journal of the Royal Statistical Society. Series A (General) **147**:278–292.

Dey, D. K., A. E. Gelfand, T. B. Swartz, and P. K. Vlachos. 1998. A simulation-intensive approach for checking hierarchical models. Test **7**:325–346.

Efford, M. G., and D. K. Dawson. 2012. Occupancy in continuous habitat. Ecosphere **3**:1–15.

Früiiwirth-Schnatter, S. 1996. Recursive residuals and model diagnostics for normal and non-normal state space models. Environmental and Ecological Statistics **3**:291–309.

Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association **85**:398–409.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2014. Bayesian data analysis, third edition. CRC Press, Boca Raton, Florida.

Gelman, A., X.-L. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica **6**:733–760.

Getis, A., and J. K. Ord. 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis **24**:189–206.

Gneiting, T., F. Balabdaoui, and A. E. Raftery. 2007. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69**:243–268.

Gosselin, F. 2011. A new calibrated Bayesian internal goodness-of-fit method: sampled posterior p-values as simple and general p-values that allow double use of the data. PloS One **6**:e14770.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**:711–732.

Guttman, I. 1967. The use of the concept of a future observation in goodness-of-fit problems. Journal of the Royal Statistical Society. Series B (Methodological) **29**:83–100.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. 2015. The extent and consequences of p-hacking in science. PLoS Biology **13**:e1002106.

Hjort, N. L., F. A. Dahl, and G. H. Steinbakk. 2006. Post-processing posterior predictive p values. Journal of the American Statistical Association **101**:1157–1174.

Hobbs, N. T., and M. B. Hooten. 2015. Bayesian Models: A Statistical Primer for Ecologists. Princeton University Press, Princeton, New Jersey.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: a tutorial. Statistical Science **14**:382–417.

Hooten, M. B., and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs **85**:3–28.

Johnson, V. E. 2004. A Bayesian χ^2 test for goodness-of-fit. Annals of Statistics **32**:2361–2384.

Johnson, V. E. 2007*a*. Bayesian model assessment using pivotal quantities. Bayesian Analysis **2**:719–734.

Johnson, V. E. 2007*b*. Comment: Bayesian checking of the second levels of hierarchical models. Statistical Science **22**:353–358.

Kéry, M., and J. A. Royle. 2016. Applied hierarchical modeling in ecology. Elsevier, London, U.K.

Kéry, M., and M. Schaub. 2012. Bayesian population analysis using WinBUGS: a hierarchical perspective. Academic Press, London, U.K.

King, R., B. Morgan, O. Gimenez, and S. Brooks. 2009. Bayesian analysis for population ecology. CRC Press, Boca Raton, Florida.

Langrock, R., and W. Zucchini. 2011. Hidden Markov models with arbitrary state dwell-time distributions. Computational Statistics & Data Analysis **55**:715–724.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology **74**:1659–1673.

Lichstein, J., T. Simons, S. Shriner, and K. E. Franzreb. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs **72**:445–463.

Link, W., and R. Barker. 2010. Bayesian inference with ecological applications. Academic Press, London, U.K.

Link, W., E. Cam, J. Nichols, and E. Cooch. 2002. Of BUGS and birds: Markov chain Monte Carlo for hierarchical modeling in wildlife research. Journal of Wildlife Management **66**:277–291.

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

Lunn, D., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. 2013. The BUGS book: a practical introduction to Bayesian analysis. Chapman & Hall/CRC, Boca Raton, Florida.

Marshall, E. C., and D. J. Spiegelhalter. 2003. Approximate cross-validatory predictive checks in disease mapping models. Statistics in Medicine **22**:1649–1660.

Marshall, E. C., and D. J. Spiegelhalter. 2007. Identifying outliers in Bayesian hierarchical models: a simulation-based approach. Bayesian Analysis **2**:409–444.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Chapman and Hall, New York.

Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. Biometrika **37**:17–23.

Perry, J., A. Liebhold, M. Rosenberg, J. Dungan, M. Miriti, A. Jakomulska, and S. Citron-Pousty. 2002. Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. Ecography **25**:578–600.

Plummer, M., 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Page 125 *in* Proceedings of the 3rd international workshop on distributed statistical computing, volume 124. Technische Universit at Wien Wien, Austria.

Potts, J. M., and J. Elith. 2006. Comparing species abundance models. Ecological Modelling **199**:153–163.

Qiu, S., C. X. Feng, and L. Li. 2016. Approximating cross-validatory predictive p-values with integrated IS for disease mapping models. arXiv **1603.07668**.

R Development Core Team, 2017. R: a Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

Robins, J. M., A. van der Vaart, and V. Ventura. 2000. Asymptotic distribution of P values in composite null models. Journal of the American Statistical Association **95**:1143–1156.

Royle, J. 2004. N-mixture models for estimating population size from spatially replicated counts. Biometrics **60**:108–115.

Rubin, D. B. 1981. Estimation in parallel randomized experiments. Journal of Educational and Behavioral Statistics **6**:377–401.

Rubin, D. B., et al. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics **12**:1151–1172.

Sauer, J. R., and W. A. Link. 2011. Analysis of the North American breeding bird survey using hierarchical models. Auk **128**:87–98.

Smith, J. Q. 1985. Diagnostic checks of non-standard time series models. Journal of Forecasting **4**:283–291.

Stern, H. S., and N. Cressie. 2000. Posterior predictive model checks for disease mapping models. Statistics in Medicine **19**:2377–2397.

Ver Hoef, J. M., and P. L. Boveng. 2007. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology **88**:2766–2772.

Ver Hoef, J. M., and P. L. Boveng. 2015. Iterating on a single model is a viable alternative to multimodel inference. Journal of Wildlife Management **79**:719–729.

Ver Hoef, J. M., and K. J. Frost. 2003. A Bayesian hierarchical model for monitoring harbor seal changes in Prince William Sound, Alaska. Environmental and Ecological Statistics **10**:201–219.

Williams, P. J., and M. B. Hooten. 2016. Combining statistical inference and decisions in ecology. Ecological Applications **26**:1930–1942.

Williams, P. J., M. B. Hooten, J. N. Womble, and M. R. Bower. 2017. Estimating

occupancy and abundance using aerial images with imperfect detection. Methods in Ecology and Evolution **8**:1679–1689.

Wood, S. N. 2006. Generalized additive models. Chapman & Hall/CRC, Boca Raton, Florida.

Yuan, Y., and V. E. Johnson. 2012. Goodness-of-fit diagnostics for Bayesian hierarchical models. Biometrics **68**:156–164.

Zhang, J. L. 2014. Comparative investigation of three Bayesian p values. Computational Statistics & Data Analysis **79**:277–291.

Zucchini, W., and I. L. MacDonald. 2009. Hidden Markov models for time series: an introduction using R. CRC Press, Boca Raton, Florida.

TABLES

TABLE 1. Types and causes of lack of fit in statistical models

Concept	Description
Dependent responses	Many statistical models assume independent response variables. Lack of independence can have multiple causes, including behavioral coupling and unmodeled explanatory variables, with the latter often inducing residual spatial or temporal autocorrelation. The usual result is inflated sample size, underestimated variance, and overdispersion relative to the assumed model.
Kurtosis	The sharpness of the peak of a probability distribution. Assuming a probability distribution with too high a kurtosis can increase the impact of outliers on an analysis.
Nonidentical distribution	Statistical models often assume that responses are identically distributed (i.e., have the same underlying probability distribution). However, this need not be the case. For instance, *Heteroskedasticity* refers to the case in which variance changes as a function of the magnitude of the response.
Outliers	Outliers consist of observations that are surprisingly different than those predicted by a statistical model. They can arise because of measurement error, or because of model misspecification (particularly with regard to kurtosis). Outliers can often have undue influence on the results of an analysis (i.e., high leverage), and it may be advantageous to choose models that are robust to the presence of outliers.
Over-parameterization	A model is overparameterized whenever two or more combinations of parameters give the same, optimal solution given the data and assumed model. If overparameterization is a function of the model only (i.e., could not be resolved by collection of more data), a particular parameter set is said to be *non-identifiable*. If it is overparameterized because data are too sparse to discriminate between alternative solutions, a particular parameter set is said to be *non-estimable*. Overparameterization can be studied analytically or (perhaps more commonly) through numerical techniques such as singular value decomposition. It can be difficult to diagnose in Bayesian applications because it typically results in a multimodal posterior distribution, and it can be difficult to discern whether all the modes have been reached.
Overdispersion	A condition where the statistical model is incapable of reproducing the amount of variation observed in a data set. Three common types of overdispersion in ecological data are (i) unmodeled heterogeneity, (ii) dependent responses, and (iii) zero-inflation.
Skewness	The amount of asymmetry of an assumed probability density about its mean.

TABLE 2. Discrepancy functions and pivotal quantities useful for Bayesian model checking.

Name	Definition	Comments
A. Omnibus discrepancy functions		
χ^2	$T(\mathbf{y}, \boldsymbol{\theta}) = \sum_i \frac{(y_i - E(y_i\|\boldsymbol{\theta}))^2}{E(y_i\|\boldsymbol{\theta})}$	Often used for count data; suggested by Gelman et al. (2014) (among others).
Deviance (D)	$T(\mathbf{y}, \boldsymbol{\theta}) = -2\log[\mathbf{y}\|\boldsymbol{\theta}]$	Used by King et al. (2009)
Likelihood ratio statistic	$T(\mathbf{y}, \boldsymbol{\theta}) = 2\sum_i y_i \log(\frac{y_i}{E(y_i\|\boldsymbol{\theta})})$	Used by Lunn et al. (2013)
Freeman-Tukey Statistic	$T(\mathbf{y}, \boldsymbol{\theta}) = \sum_i (\sqrt{y_i} - \sqrt{\mathrm{E}(y_i\|\boldsymbol{\theta})})^2$	Less sensitive to small expected values than χ^2; suggested by Kéry and Royle (2016) for count data.
B. Targeted discrepancy functions		
Proportion of zeros	$T(\mathbf{y}) = \sum_i I(y_i = 0)$	Zero inflation check for count data
Kurtosis checks	$T(\mathbf{y}) = y_p$	Using the pth quantile can be useful for checking for proper tail behavior.
C. Pivotal quantities		
$Y \sim$ Exponential(λ)	$\lambda\bar{Y} \sim$ Gamma(n, n)	Note n is sample size
$Y \sim \mathcal{N}(\mu, \sigma^2)$ (Gaussian)	$\frac{Y-\mu}{\sigma} \sim \mathcal{N}(0, 1)$	For mean μ and standard deviation σ
$Y \sim$ Weibull(α, β)	$\beta Y^\alpha \sim$ Exponential(1)	
Y from *any* distribution	$Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$	For large sample size (n), Z converges in distribution to a standard normal (Slutsky's theorem) and Z is termed an "asymptotically pivotal quantity."

TABLE 3. A summary of Bayesian model checking approaches. For each method, we describe whether each method (1) tends to be "conservative" (i.e., an overstated α value), (2) whether all levels of the modeling hierarchy can be evaluated ("all levels"), (3) whether out-of-sample data are used to assess lack of fit ("out of sample"), and (4) computing cost ("cost").

Method	conservative	all levels	out of sample	cost
Pivotal discrepancy	Yes	Yes	No	medium
Posterior predictive check	Yes	No	No	low
Prior predictive check	No	Yes	No	low
Predictive PIT tests	No	No	Yes	high
Sampled predictive p-value	No	Maybe	No	low
Graphical	Maybe	Maybe	No	low

TABLE 4. Results of one simulation for examining the effect of the closure assumption on model fit in the sea otter example. The notation c represents the maximum proportion of the population that could move in or out of a site between $j-1$ and j, p-value is the posterior predicitive p-value using a χ-squared goodness-of-fit statistic, sppv is the sampled predictive p-value using the sum of variance test statistic, Abundance is the mean of the marginal posterior distribution for total abundance at the 300 sites, the 95% CRI are the 95% credible intervals, GR is the multi-variate Gelman-Rubin convergence diagnostic, and ESS is the effective sample size of 1,000,000 MCMC iterations.

c	p-value	sppv	Abundance (truth=50,989)	95% CRI	GR	ESS
0.00	0.48	0.27	51,200	(49,295, 53,481)	1.00	3,420
0.05	0.40	1.00	60,047	(56,605, 63,868)	1.00	3,260
0.10	0.00	1.00	81,299	(75,223, 89,601)	1.01	3,194
0.15	0.00	1.00	97,066	(89,149, 104,360)	1.13	3,199
0.20	0.00	0.02	117,624	(108,825, 127,007)	1.03	3,184
0.25	0.00	0.01	119,397	(110,477, 125,992)	1.06	3,206
0.30	0.00	0.00	133,797	(124,194, 141,117)	1.10	3,195
0.35	0.00	0.00	139,951	(133,351, 147,086)	1.00	3,213

FIGURE CAPTIONS

FIGURE 1. A decision diagram describing the steps to adopt when reporting the results of Bayesian analyses in the literature, particularly when results will be used for conservation and management or to inform ecological theory. The first step is to formulate reasonable ecological models, ensuring that the model(s) and associated software is free of errors and that convergence to the posterior distribution can be achieved (using Markov chain Monte Carlo, for instance). Following this step, models should be checked against observed data to diagnose possible model misspecification (the subject of this article). Assuming no obvious inadequacies, various model comparison or averaging techniques can be used to compare the predictive performance of alternative models that embody different ecological hypotheses. Finally, we suggest conducting robustness analyses (prior sensitivity analyses, simulation analyses where model assumptions are violated) to gauge the importance of implicit parametric assumptions on ecological inference.

FIGURE 2. Type of model checking procedures used in $n = 31$ articles published in the journal Ecology during 2014 and 2015. Articles were found via a Web of Science for articles including the topic "Bayesian" (search conducted 10/1/2015). Six articles were determined to be non-applicable (N/A) because they either (1) were simulation studies, or (2) used approximate Bayesian computation, which is conceptually different than traditional Bayesian inference (e.g., Beaumont et al. 2002). Of the remaining 25, 20 did not report any model checking procedures. Five articles reported specific model checking procedures, which included a combination of Bayesian cross-validation (*Cross.val*,), frequentist software (*Non-Bayes*), posterior predictive p-values (*Pp.pval*), and posterior predictive graphical checks (*Pp.gc*). Some articles also investigated prior sensitivity which can be regarded as a form of model checking, but we do not report prior sensitivity checks here.

FIGURE 3. A depiction of how simulated count data are generated. First, a spatially autocorrelated covariate is generated using a Matérn cluster process (A) over a region $\mathcal{A}_2$. Second, a spatially autocorrelated random effect is simulated according to a predictive process formulation (B), where the parent process occurs at a knot level (C; open circles). The covariate and spatial random effect values combine on the log scale to generate expected abundance (C). Sampling locations (C; small points) are randomly placed over a subregion, $\mathcal{A}_1$ of the study area, where $\mathcal{A}_1$ is defined by the inner box of knot values. Finally, counts are simulated according to a Poisson distribution (D). Note that counts are simulated in $\mathcal{A}_1 \subset \mathcal{A}_2$ to eliminate possible edge effects.

FIGURE 4. Three possible ways of simulating replicate data to calculate posterior predictive p-values for the spatial regression simulation study. Solid boxes indicate parameters or latent variables that occur in the directed graph for observed counts, while dashed boxes indicate posterior predictions. In (A), replicate data (y_i^{rep}) for a given observation i depend only upon the latent variable ν_i, posterior samples of which are available directly from MCMC sampling. In (B), replicate values of ν_i are simulated (ν_i^{rep}) prior to generating posterior predictions. In (C), an example of a "mixed predictive check," spatially autocorrelated random effects are also resimulated (η_i^{rep}), conditional on the values of random effects at other sites, $\boldsymbol{\eta}_{-i}$, and parameters describing spatial autocorrelation (i.e., precision τ_η and exponential decay ϕ).

FIGURE 5. Example computation of a χ^2 discrepancy test using a CDF pivot for a single posterior sample of a Normal-Poisson mixture model (without spatial autocorrelation) fit to simulated count data. In this case, the test focuses on the fit of the the latent variable $\boldsymbol{\nu}$ to a Gaussian distribution with mean given by the linear predictor (i.e., $\mathbf{X}\boldsymbol{\beta}$) and precision τ as specified in the `PoisMix` model. The test we employed

partitions the linear predictor based on 20%, 40%, 60%, and 80% quantiles (solid lines), and assesses whether the Gaussian CDF in these ranges is uniformly distributed within five bins. If modeling assumptions are met, there should be a roughly equal number of observations in each bin. For the data presented here, there appears to underpredictions at low and high values of the linear predictor.

FIGURE 6. A plot of over- and under-predictions obtained from fitting a basic Poisson regression model (`Pois0`) to count data generated from an overdispersed Poisson model subject to spatial autocorrelation. We plot the spatial location of each data point, and indicate whether each datum was below ("Under"), above ("Over") or within ("Neither") a 95% posterior predictive credible interval (the limits of which were calculated as the 2.5% and 97.5% quantiles of predicted counts from the `Pois0` model). Data are clearly overdispersed relative to the basic Poisson model, with only 80 of 200 observations within posterior predictive intervals (we would expect an average of 190 of 200 observations to be within credible intervals if the model fitted the data well). There is also evidence of considerable spatial clustering not accounted for with the single explanatory covariate, with overpredictions occurring in the southwest and underpredictions occurring more often in the northeast.

FIGURE 7. Histogram bin heights showing the relative frequency of 1000 p-values as obtained in the spatial regression simulation study (histograms have 10 bins). The dashed line represents the case where the simulation and estimation model were the same (`PoisMixSp`). An unbiased test should have a roughly uniform distribution in this case, whereas concave distributions indicate that the test is conservative. A greater frequency of low p-values (e.g., < 0.1) under `PoisMix` (solid lines) indicate a higher power of rejecting the `PoisMix` model, a model that incorrectly omits the possibility of residual spatial

autocorrelation. The following types of p-values were calculated: k-fold cross-validation ('Cross.val'; `PoisMix` model only), a mixed predictive p-value using the Freeman-Tukey discrepancy ('Mixed.FT'; `PoisMixSp` model only), posterior Moran's I ('Moran'), median pivot discrepancy on the Gaussian ('Pivot.Gauss') and Poisson ('Pivot.Pois') parts of the model, a posterior predictive p-value with a χ^2 discrepancy function ('PP.ChiSq'), posterior predictive p-values using a deviance-based discrepancy calculated relative to the Poisson ('PP.Dev.Pois') and Gaussian ('PP.Dev.Gauss') portions of the likelihood, a posterior predictive p-value calculated with the Freeman-Tukey discrepancy ('PP.FT'), a posterior predictive p-value using a 95th quantile discrepancy ('PP.Tail'), and sampled predictive p-values relative the Gaussian ('Sampled.Gauss') and Poisson ('Sampled.Pois') parts of the model.

FIGURE 8. Observed and expected values for the number of detected animals that were previously marked. Light and dark blue envelopes represent the 50 and 90th highest probability density interval for the expected number of detections under the HMM model, respectively. The red envelopes represent the equivalent intervals for the HSMM model with shifted Poisson residence time distributions for each state. The gaps in the envelopes represent days in which resighting did not occur and detection probabilities were fixed to 0.

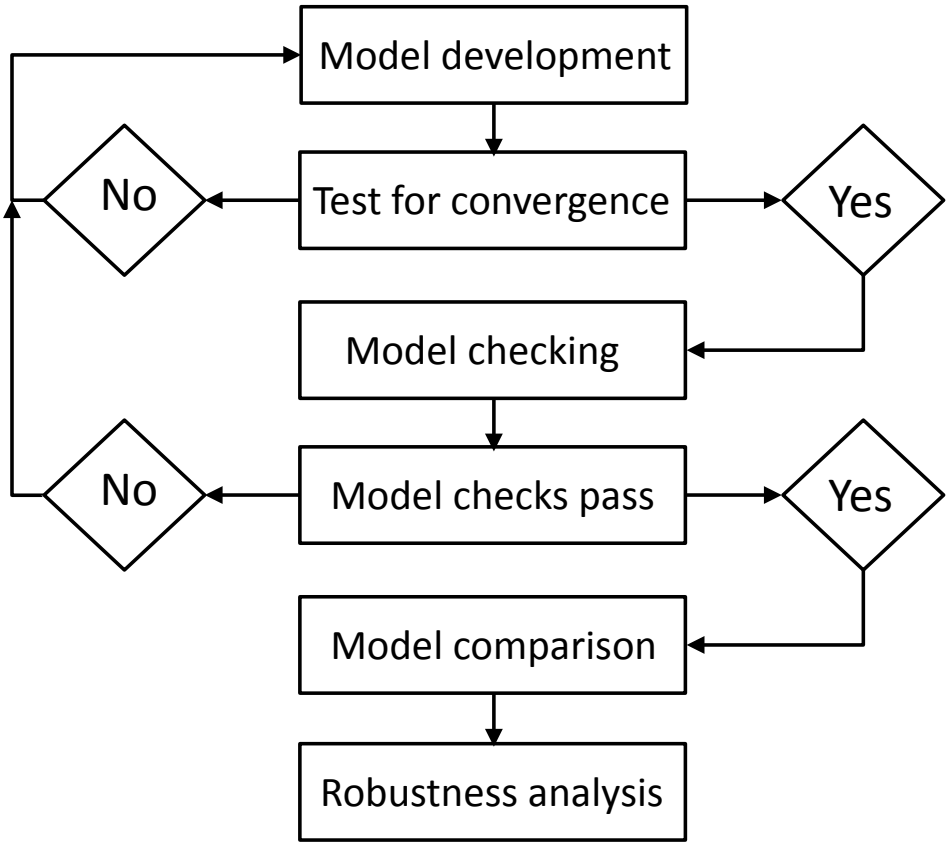

FIG 1

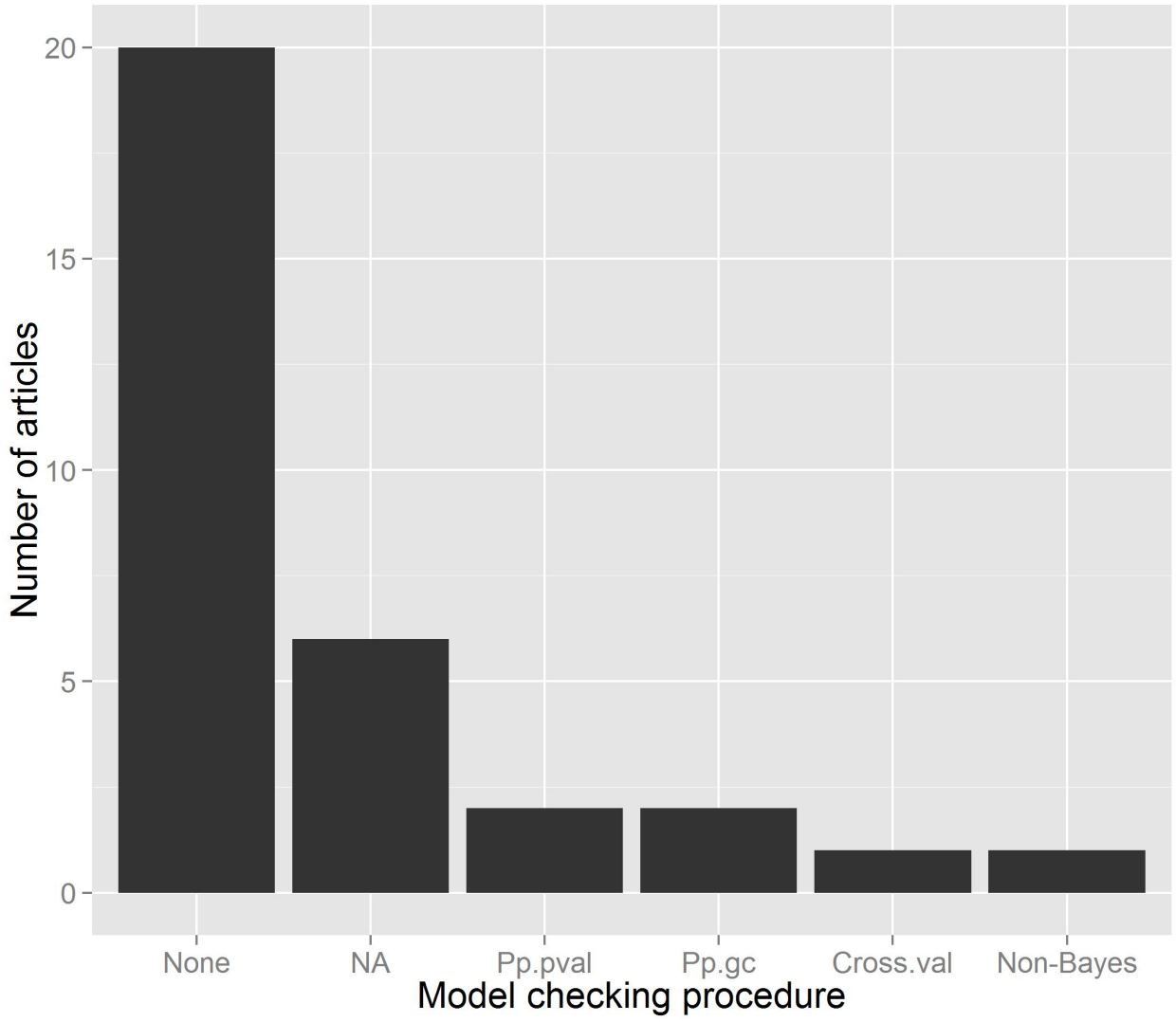

FIG 2

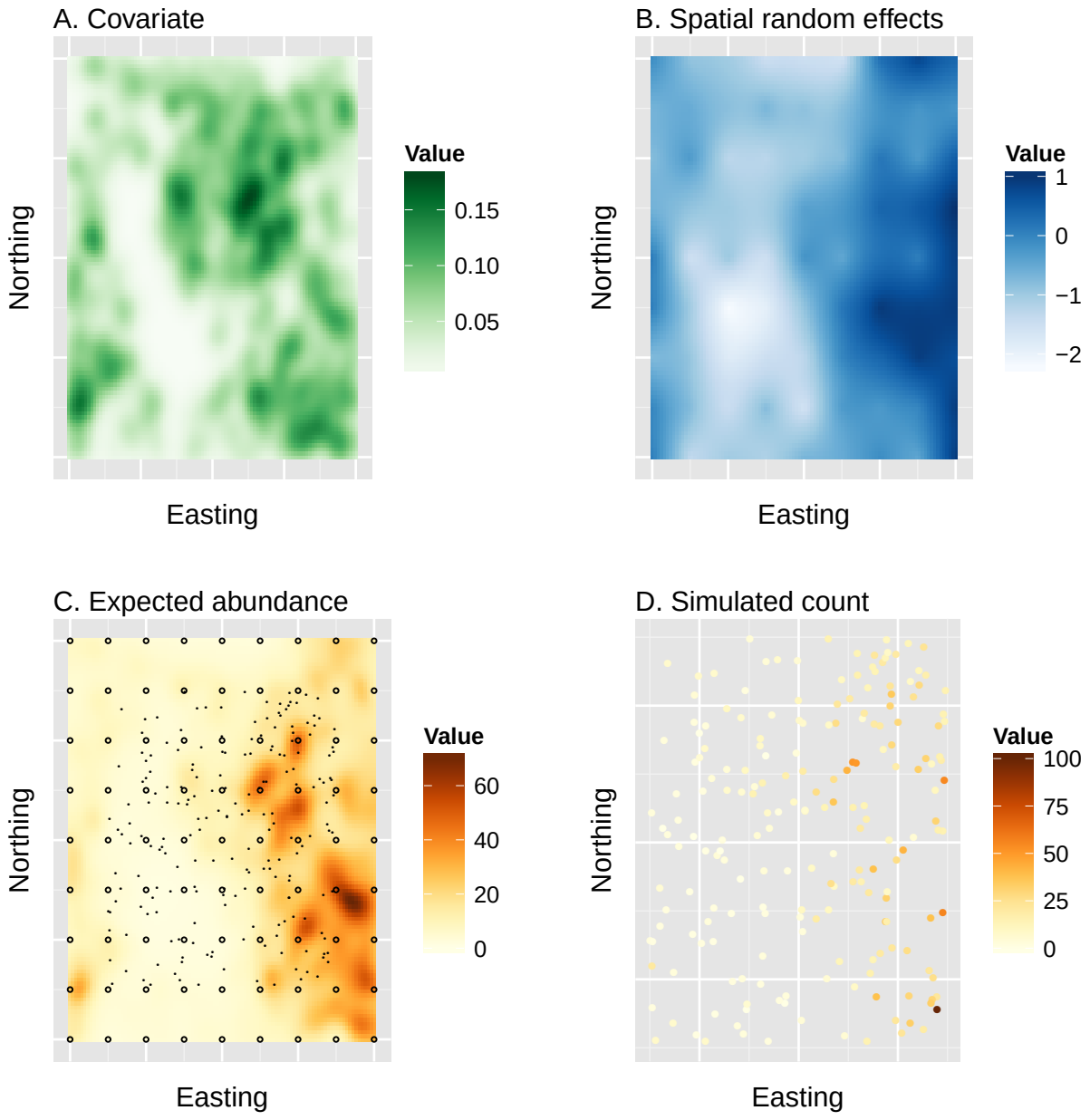

FIG 3

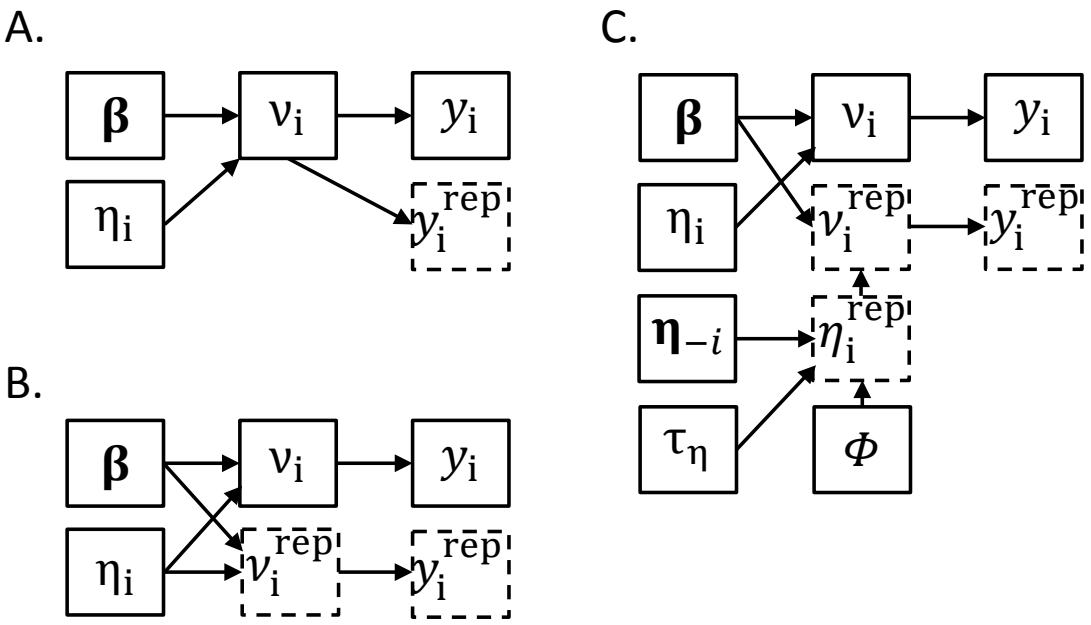

FIG 4

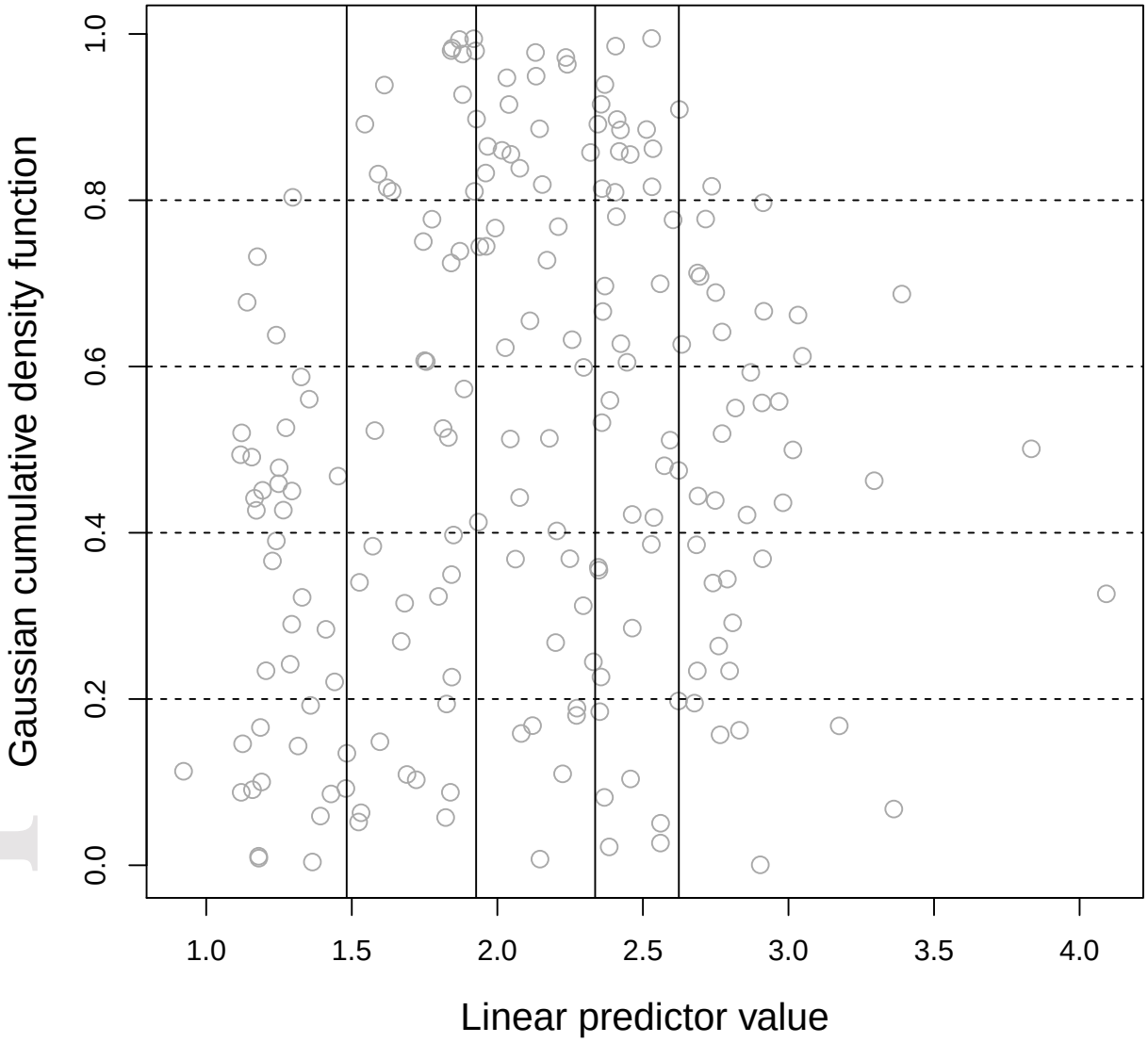

FIG 5

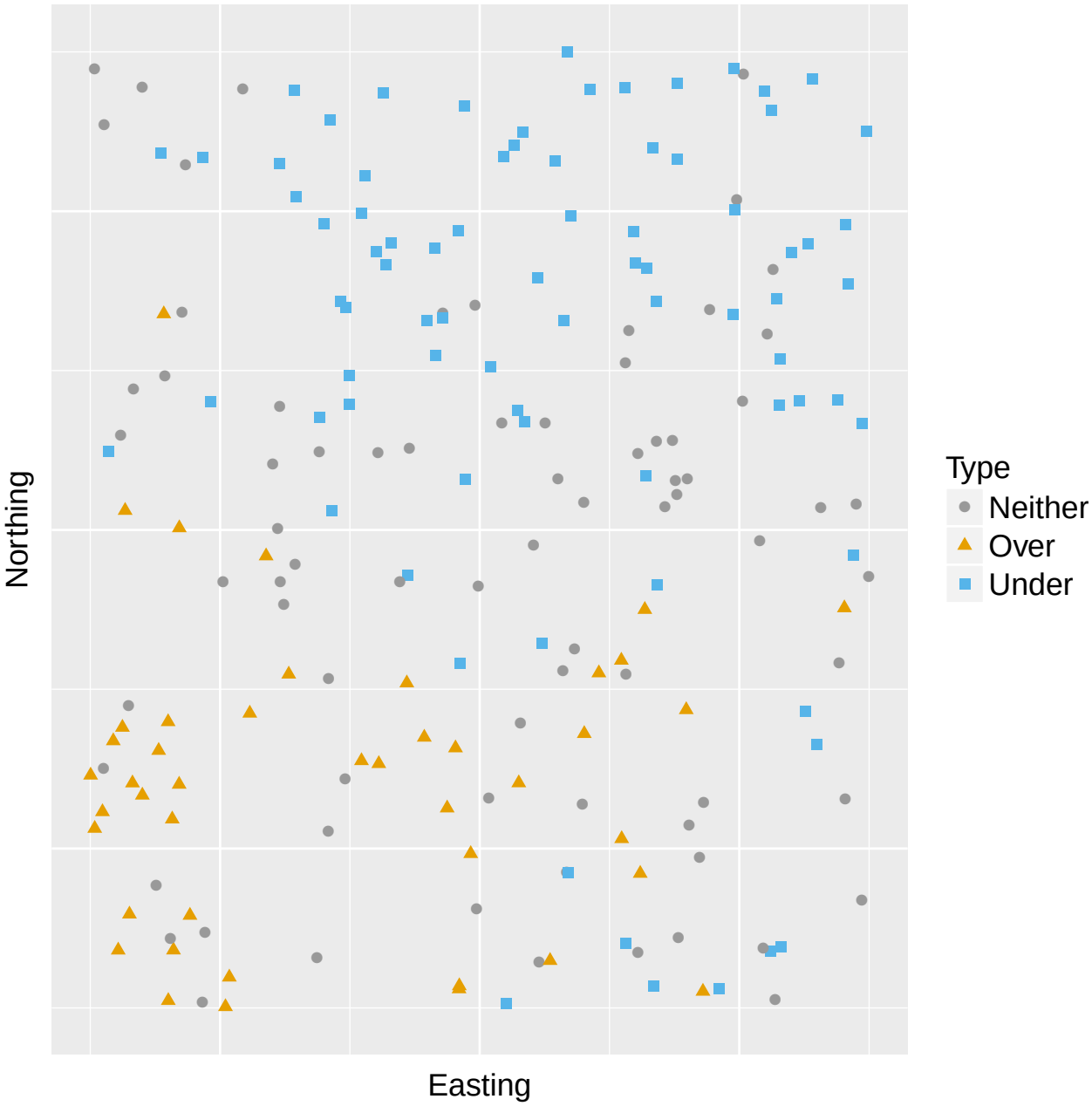

FIG 6

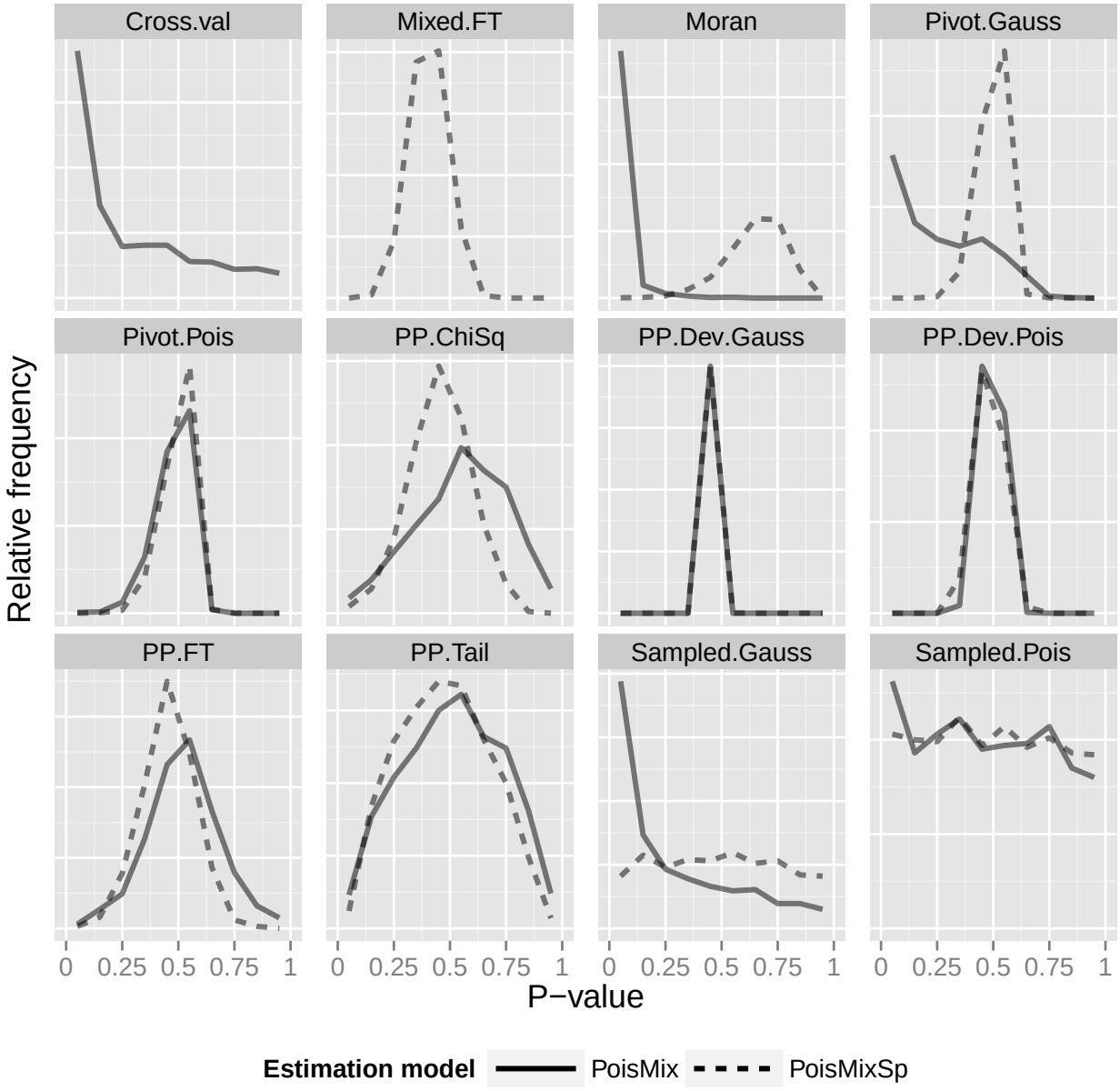

FIG 7

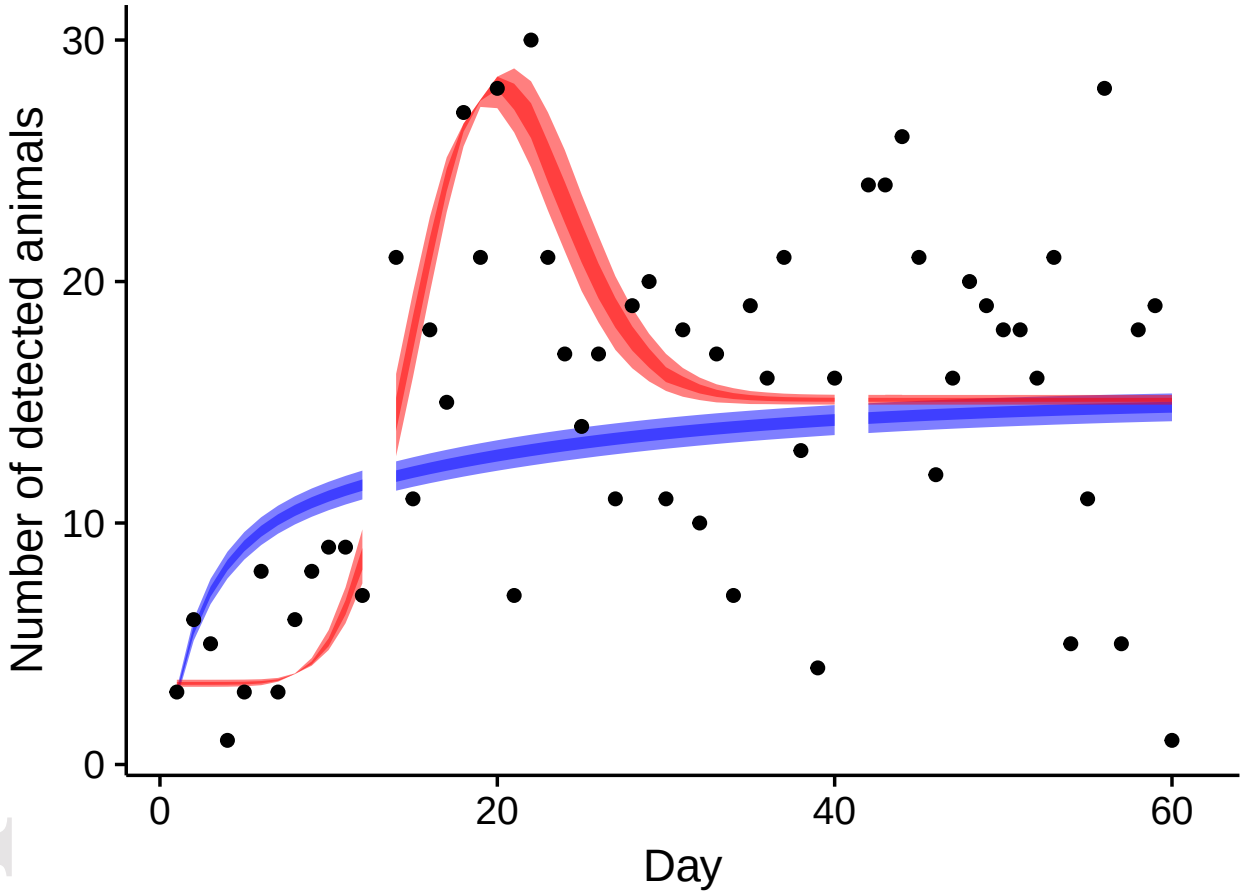

Fig 8