A guide to Bayesian model checking for ecologists

Paul B. Conn[1,5], Mevin B. Hooten[2,3,4], and Devin S. Johnson[1]

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

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

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

[4]*Department of Statistics, 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 controlling for potential biases arising from sampling artifacts. However, ecologists tend to be less focused on checking model assumptions and assessing potential lack-of-fit when applying Bayesian methods than when they applying more traditional modes of inference such as maximum likelihood. There are also multiple ways of assessing goodness-of-fit for Bayesian models, each of which has strengths and weaknesses. For instance, in ecological applications, the posterior predictive p-value is probably the most widely used approach for assessing lack of fit in Bayesian models. Such p-values are relatively easy to compute, but they are well known to be conservative, producing p-values

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

biased toward 0.5. Alternatively, lesser known approaches to model checking, such as prior predictive 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 in order to guide ecologists to the many available options for Bayesian model checking. We illustrate methods and procedures with several ecological case studies, including i) explaining variation in simulated spatio-temporal count data, and (ii) modeling survival and residence times of fur seal mothers on a rookery, and (iii) using N-mixture models to model XXX. We find that commonly used procedures based on posterior predictive p-values have high power to detect extreme model inadequacy, but low power to 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 much better suited to this task and to have better overall statistical performance. We conclude that model checking is an essential component of scientific discovery and learning that should accompany most Bayesian analyses presented in the literature.

Bayesian p-value, count data, goodness-of-fit diagnostic check, hierarchical model, model checking, N-mixture model, pivot discrepancy, posterior predictive check, probability interval transform, sampled predictive p-value, spatio-temporal model

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 analyses (Link and Barker 2006, Berger 2013, Williams and Hooten In Press).

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, as 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). Note that

scientific inference can still proceed if models do not fit the data well, but conclusions need to be tempered; one approach in such situations is to estimate a variance inflation factor to adjust precision levels downward (e.g., Cox and Snell 1989, McCullagh and Nelder 1989).

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 2015), one can easily access diagnostics such as quantile-quantile, residual, and leverage plots. These diagnostics allow one to assess the reasonability of 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 as well as departures from specific model assumptions which are codified 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 in the ecological literature. For instance, a search of recent volumes of Ecology 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 is uncommon. First, it likely has to do with momentum; 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 Markov chains from MCMC output 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? Finally, it may just be that those publishing Bayesian analyses in ecological literature ". . . like artists, have the bad habit of falling in love with their models" (to borrow a quote attributed to G.E.P. Box and referenced by Link and Barker (2010) with regard to model checking). However, models can be poor at returning our affection; indeed this monograph can be viewed as a partial atonement for unrequited love.

If we accept the premise that Bayesian models in ecology 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 low power to detect lack-of-fit (Yuan and Johnson 2012, Zhang 2014). By

contrast, other approaches less familiar to ecologists (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 have collated relevant statistical literature with the goal of providing ecologists with a practical guide to Bayesian model checking. We start by defining a consistent notation that we use throughout the paper. Next, we work to compile a bestiary of Bayesian model checking procedures, providing positives and negatives associated with each approach. We illustrate Bayesian model checking using several simulation studies (including spatial regression and ...), as well as a case study involving capture-recapture sampling of adult female fur seals (*Callorhinus ursinus*) on a rookery in Alaska, 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.

The posterior distribution is often written as

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

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 a single, 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 3.

Model checking procedures

Our goal in this section is to review relevant Bayesian model checking procedures for typical models in ecology, with the requirement that such procedures be accessible to statistics-savvy 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 to be of relevant use in common applications. 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 (see e.g., Bayarri and Berger 1999, Robins et al. 2000, Bayarri and Castellanos 2007).

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 all its set of 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 A, Alg. 1).

When the prior distribution(s) $[\boldsymbol{\theta}]$ are proper statistical distributions, p-values from prior predictive checks are uniformly distributed under the null model and have properly stated frequentist properties. The main problem with this approach is that the models being considered need to have considerable historical investment and proper prior distributions informed by 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 hierarchical 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 shall 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 precursor to 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 ($n = 25$), our survey of recent *Ecology* volumes indicated that posterior predictive checks are also the dominant form of Bayesian model checking being reported in ecological literature (if any checking is reported at all; 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, lower kurtosis), it is good indication 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 A, Alg. 2). Bayesian p-values necessarily involve application of a discrepancy function, $T(\mathbf{y}, \boldsymbol{\theta})$, for comparing observed and simulated data. There are several omnibus discrepancy measures that can be employed to examine overall lack-of-fit, and targeted discrepancy measures can be used to look for specific data features that systematically differ between simulated and observed data (Table 3).

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) tends to be 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 a simulation study with two different model types, 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 is occurring within the modeling hierarchy (e.g., poorly specified priors, errant mean structure, underdispersed error distribution). Further, a poorly specified mean structure may still result in reasonable fit of the model if the model is made sufficiently flexible through inclusion of random effects.

These cautions do not imply that posterior predictive checks are completely devoid of value. Indeed, given that tests are conservative, small (e.g., < 0.05) or very large (e.g., > 0.95) p-values are strongly suggestive of lack-of-fit. Further, graphical displays (see *Graphical techniques*) and targeted discrepancies (Table 3) 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 that indicate no (or little) lack-of-fit. P-values close to 0.15 or 0.2 are especially problematic. Hjort et al. (2006) developed approaches for calibrating Bayesian p-values so that they are asymptotically uniform under the null model. However, their approach can be computationally intensive and/or difficult to implement.

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 resimulating random effects when generating posterior predictions for each data point (a procedure they term a "mixed predictive check"). For an example of this latter approach, see *Spatial models for count data*.

Sampled posterior p-values

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 a discrepancy test statistic (Gelman et al. 2004, ; Appendix A, Alg. 2). 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 for $[\tilde{\boldsymbol{\theta}}]$, 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) (i.e., when the model fit to the data is the "true" model). Sampled posterior p-values can also be calculated using pivotal discrepancy measures at a single posterior sample of parameter values, reducing computational burden (i.e., eliminating the requirement that replicate datasets be generated). We describe an example of this approach in *Spatial models for count data*.

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 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, that have known probability distributions that are independent of parameters (see e.g., Casella and Berger 1990, section 9.2.2). For instance, consider a simple normal (Gaussian) model

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

Recall from introductory statistics classes that $z = \frac{y-\mu}{\sigma}$ has a standard $f = \mathcal{N}(0,1)$ 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

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

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 $N(0,1)$ distribution can point to model misspecification. Note that, although we have focused on the data model in Eq. 3, 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, $\mathbf{y}^{rep}$. However, in practice, there are several difficulties with using pivotal quantities as discrepancy measures in Bayesian model checking. 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. If an omnibus Bayesian p-value is desired, a test can be implemented by appealing to limiting distributions of order statistics (Johnson 2004), but these tests are conservative and have low power to detect lack of fit.

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 A, 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, $F(\mathbf{w})$ should 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 deviate 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 A), but the algorithm can be applied (without loss of generality) 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 *Examples* 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, running an analysis, and seeing 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; see e.g. Arlot and Celisse 2010). However, it also possible to use cross-validation techniques to examine model fit and diagnose outlier behavior. The major 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). The major disadvantage is that it can be computationally challenging for complicated 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 computing the CDF of the predictions evaluated at the 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, divergence from a Uniform(0,1) distribution is indicative of a 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) provides an algorithm for computing PIT diagnostic histograms for both continuous and discrete data in Bayesian applications (see Appendix A, 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 either 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.

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, as 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,

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

should generally take on values between -2.0 and 2.0. Values very far out of this range represent outliers. Similarly, for the Poisson and binomial distributions, an approximate 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 $\mathbf{Y} - \text{E}(\mathbf{Y}|\boldsymbol{\theta}_j)$, a histogram of I_j values can be constructed; substantial overlap with zero suggests little evidence of residual spatial autocorrelation. As calculation of Moran's I is dependent upon a a pre-specified distance weighting scheme, investigators might simulate a posterior sample of Moran's I at several different choices of weights or neighborhoods to evaluate residual spatial autocorrelation at different scales.

Just build a bigger model! Tradeoffs between fit and prediction

One way to ensure a model fits the data is simply to build a model high complexity. To take an extreme example, one could simply start with a saturated model (one where there is a separate parameter for each datum) so that the model fits the data perfectly. No one would actually do this in practice; science proceeds be establishing generalities, and there is no generality implicit in such a model. Further, there is no way to predict future outcomes with such a model. Indeed, models with high complexity can fit the data well, but may have poorer predictive ability 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 (see, 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 using this approach, 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.

Graphical techniques

Many of the tests described previously require discrepancy functions, and it may be difficult to formulate such functions for different types of lack-of-fit (e.g., Table 1). Many scientists are visual learners, and displaying model checking information graphically can

lead to more rapid intuition about where models fit 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, Gelman et al. (2014) argues that residual and binned residual plots can be 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 conducted all analyses using a combination of R (R Development Core Team 2015) 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 will be published to a permanent repository following manuscript acceptance. The code is predominantly model-specific; however, we hope it can be used as a template for ecologists conducting their own model checking exercises.

Simulation studies

We conducted several simulation studies to illustrate application of alternative model checking procedures when trying to detect departures from model assumptions. Simulation is extremely useful for illustrating concepts as truth is known, and we can examine the

large-scale properties of model testing procedures, including the important case when the same model is used to fit the data as is used to generate them. For example, we can describe the empirical null distribution of Bayesian p-values, which must be uniformly distributed on (0,1) to provide an unbiased test.

Simulation has been previously used to assess the properties of alternative Bayesian model checking procedures (e.g., Gosselin 2011, Yuan and Johnson 2012, Zhang 2014), but the problems studied have often been simplistic relative to the types of Bayesian models used in real world ecological applications. In this section, we study the properties of different model checking methods when applied to simulated data sets more typical for ecological inference. First, we examine spatial regression models applied to simulated count data. In this case, we are interested in detecting residual spatial autocorrelation and overdispersion relative the Poisson distribution. In our second example, we investigate N-mixture models.

Spatial models for count data

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 generate 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 can be viewed as arising from a spatially autocorrelated Poisson-normal mixture, and can be expected to be overdispersed relative to the Poisson, in much the same way that a negative binomial distribution (a Poisson-gamma mixture) is. We then examine the effectiveness of alternative model checking procedures for diagnosing incorrect model specification, such as when spatial independence is assumed. We also study 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 $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \mathbb{R}^2$, and $\mathcal{A}_1$ and $\mathcal{A}_2$ are subsets of $\mathbb{R}^2$. 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 B).

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 a two-column design matrix specifying a linear effect of $\mathbf{x}$, $\boldsymbol{\eta}$ are spatially autocorrelated random effects, and $\boldsymbol{\epsilon}$ are iid Gaussian errors.

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

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

- `PoisMixSp`: The data-generating model, consisting of a Poisson-normal mixture with iid 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 B. As it is the main focus of the paper, we next describe model checking procedures (step 6) in greater detail.

Posterior predictive p-values

For each dataset and estimation model, we computed a number of posterior predictive p-values. These included χ^2, Freeman-Tukey, and deviance-based omnibus p-values, as well as directed p-values examining tail probabilities (Table 3). 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 simply 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 3), 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 $E(y_i|\boldsymbol{\theta}) = \exp(\mathbf{x}_i\boldsymbol{\beta} + 0.5\tau_\epsilon^{-1})$. 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 A) 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 sample of parameters (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 A) 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

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

Here, $I(y_{ij}^{rep} < y_i)$ is a binary indicator function taking on the value 1.0 if and only 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 and only 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. 6). By 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, there did not appear to be much reason to go to the extra effort of computing a mixed predictive check, as they actually appeared slightly more conservative than their posterior predictive counterparts. 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 are 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 for the `Pois0` model. As such, all model checking procedures had high power to appropriately detect the inadequacy of the basic Poisson model.

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. 6). 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. 6).

Examples

N-mixture

Fur seal capture-recapture

Discussion

Previous articles in the ecological literature that use Bayesian analysis have tended to focus on prior sensitivity, convergence diagnostics and sometimes model comparison (e.g., DIC or cross-validation) - not as much focus on GOF.

GOF on most general model, then model selection/comparison/averaging (Burnham and Anderson 2002).

Standard regression analysis software In capture-recapture analysis and other areas of ecological statistics, there has been considerable focus on developing procedures to assess goodness-of-fit (e.g., Choquet et al. 2009)

Mean structure vs. dispersion - not always obvious where misspecification occurs.

The purpose of this paper was to provide a general overview of common approaches to Bayesian model checking and the strengths and weaknesses of each. We do not, however, claim to be entirely comprehensive. For instance, we limited our discussion to approaches that are relatively straightforward to implement. There are a variety of options to producing p-values with good statistical properties provided one has the time and technical acumen to implement them (e.g., "partial posterior" and "conditional predictive" p-values; see Bayarri and Berger 1999, Robins et al. 2000, Bayarri and Castellanos 2007). Another possibility is to use data sets simulated from the prior predictive distribution to study the

realized Bayesian p-values under the null model and calibrate p-values accordingly (Hjort et al. 2006). For example, if p-values exhibit a dome shaped pattern (c.f., dashed lines in Fig. 6), we might adjust p-value cut off values to be the 5th and 95th quantiles of the realized values. Other approaches to model checking may be useful in more specialized areas of ecology. For instance, Shipley (2009) introduced directional-separation tests for assessing the path structure of directed, acyclic graphs. These tests can be useful for assessing the graph structure of ecological networks.

Acknowledgments

Views and conclusions in this article represent the views of the authors and the U.S. Geological Survey but do not necessarily represent findings or policy of the U.S. National Oceanic and Atmospheric Administration. 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. Statistcs Surveys **4**:40–79.

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

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.

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.

Box, G. E. 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society. Series A (General) pages 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, CA.

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.

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

626 Cox, D. R., and E. J. Snell. 1989. Analysis of binary data. CRC Press.

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

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

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

634 Dawid, A. P. 1984. Present position and potential developments: Some personal views:
635 Statistical theory: The prequential approach. Journal of the Royal Statistical Society.
636 Series A (General) pages 278–292.

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

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

641 Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis, 2nd
642 Edition. Chapman and Hall, Boca Raton.

643 Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2014. Bayesian data analysis,
644 Third edition. Taylor & Francis.

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

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

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

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

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

657 Guttman, I. 1967. The use of the concept of a future observation in goodness-of-fit
658 problems. Journal of the Royal Statistical Society. Series B (Methodological) pages
659 83–100.

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

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

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

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

668 Johnson, V. E. 2004. A Bayesian χ^2 test for goodness-of-fit. Annals of Statistics
669 **32**:2361–2384.

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

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

674 Kéry, M., and J. A. Royle. 2016. Applied Hierarchical Modeling in Ecology. Elsevier,
675 London.

676 Kéry, M., and M. Schaub. 2012. Bayesian population analysis using WinBUGS: a
677 hierarchical perspective. Academic Press.

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

680 Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology
681 **74**:1659–1673.

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

684 Link, W., and R. Barker. 2010. Bayesian Inference with Ecological Applications. Academic
685 Press, London, U.K.

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

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

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

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

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

698 McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models. Chapman and Hall,
699 New York.

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

701 Perry, J., A. Liebhold, M. Rosenberg, J. Dungan, M. Miriti, A. Jakomulska, and
702 S. Citron-Pousty. 2002. Illustrations and guidelines for selecting statistical methods for
703 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.

R Development Core Team, 2015. 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.

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.

Shipley, B. 2009. Confirmatory path analysis in a generalized multilevel context. Ecology **90**:363–368.

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

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

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

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

731 Williams, P. J., and M. B. Hooten. In Press. Combining statistical inference and decisions
732 in ecology. Ecological Applications .

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

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

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

TABLES

FIGURE CAPTIONS

FIGURE 1. A decision diagram describing the steps we suggest ecologists 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 (see 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, $\mathcal{A}_\infty$ of the study area, where $\mathcal{A}_\infty$ 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., τ_η and 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. 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.

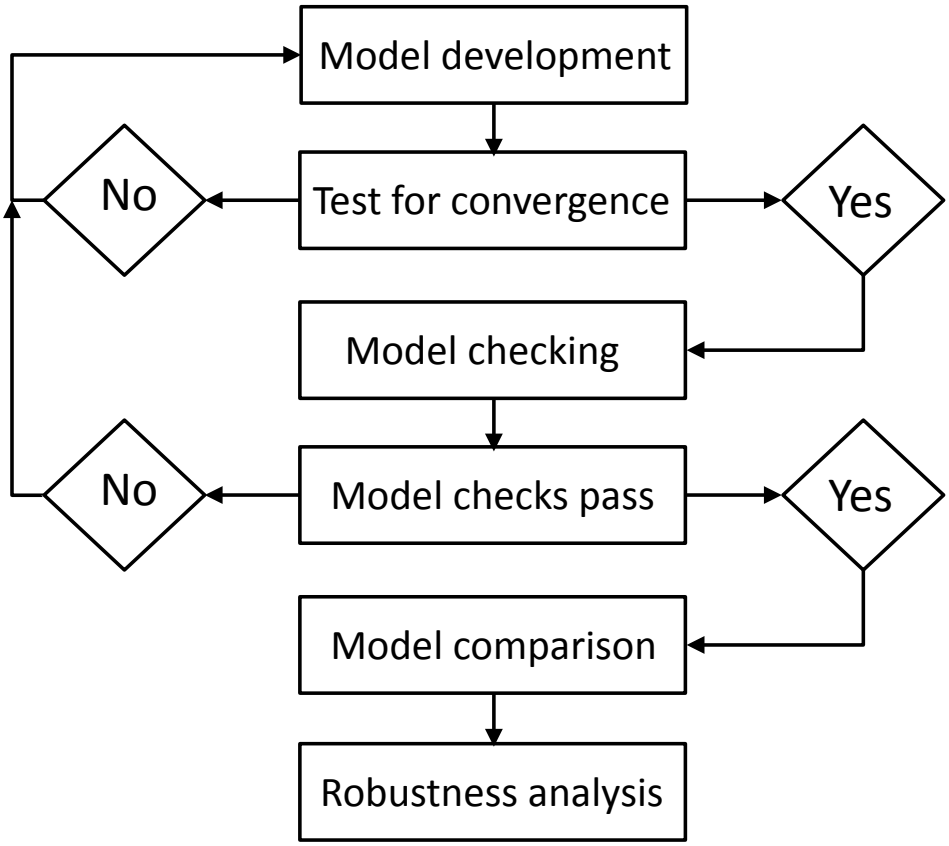
Model development
Test for convergence
No
Yes
Model checking
Model checks pass
No
Yes
Model comparison
Robustness analysis

Fig 1

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.
Overdispersion	Although dependent responses can certainly induce it, the term overdispersion is more a symptom of lack-of-fit, namely that the statistical model is incapable of reproducing the amount of variation seen in a data set. Three common types of overdispersion in ecological data are (i) unmodeled heterogeneity, (ii) zero inflation in count data (more zero observations are obtained than expected under canonical models such as the Poisson), and (iii) heavy tails (more extreme observations than predicted under the assumed model). The latter is often a result of kurtosis misspecification (see *Higher moments* below).
Higher moments	Overdispersion refers to a misspecification (underestimate) of variance, which is defined as a second moment when studied in terms of moment generating functions. However, higher moments may also be misspecified. For instance, *skewness* refers to the third moment and depicts the amount of asymmetry of an assumed probability density about its mean; *kurtosis* refers to the fourth moment and to the tail behavior of the distribution.
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.
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 increases as a function of the magnitude of the response.
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.

TABLE 2. Discrepancy functions and pivotal quantities useful for hierarchical 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 $p\%$ quantile can be useful for checking for proper tail behavior.
C. Pivotal quantities		
$Y \sim \text{Exponential}(\lambda)$	$\lambda\bar{Y} \sim \text{Gamma}(n, n)$	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
$Y \sim \text{Weibull}(\alpha, \beta)$	$\beta Y^\alpha \sim \text{Exponential}(1)$	
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 allows for (1) computation of an overall p-value ("p-value?"), (2) whether the method tends to be conservative (i.e., has overstated power to detect goodness-of-fit; "conservative?"), (3) whether all levels of the modeling hierarchy can be evaluated ("all levels?"), (4) whether out-of-sample data are used to assess lack-of-fit ("out-of-sample?"), and (5) computing cost ("cost").

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

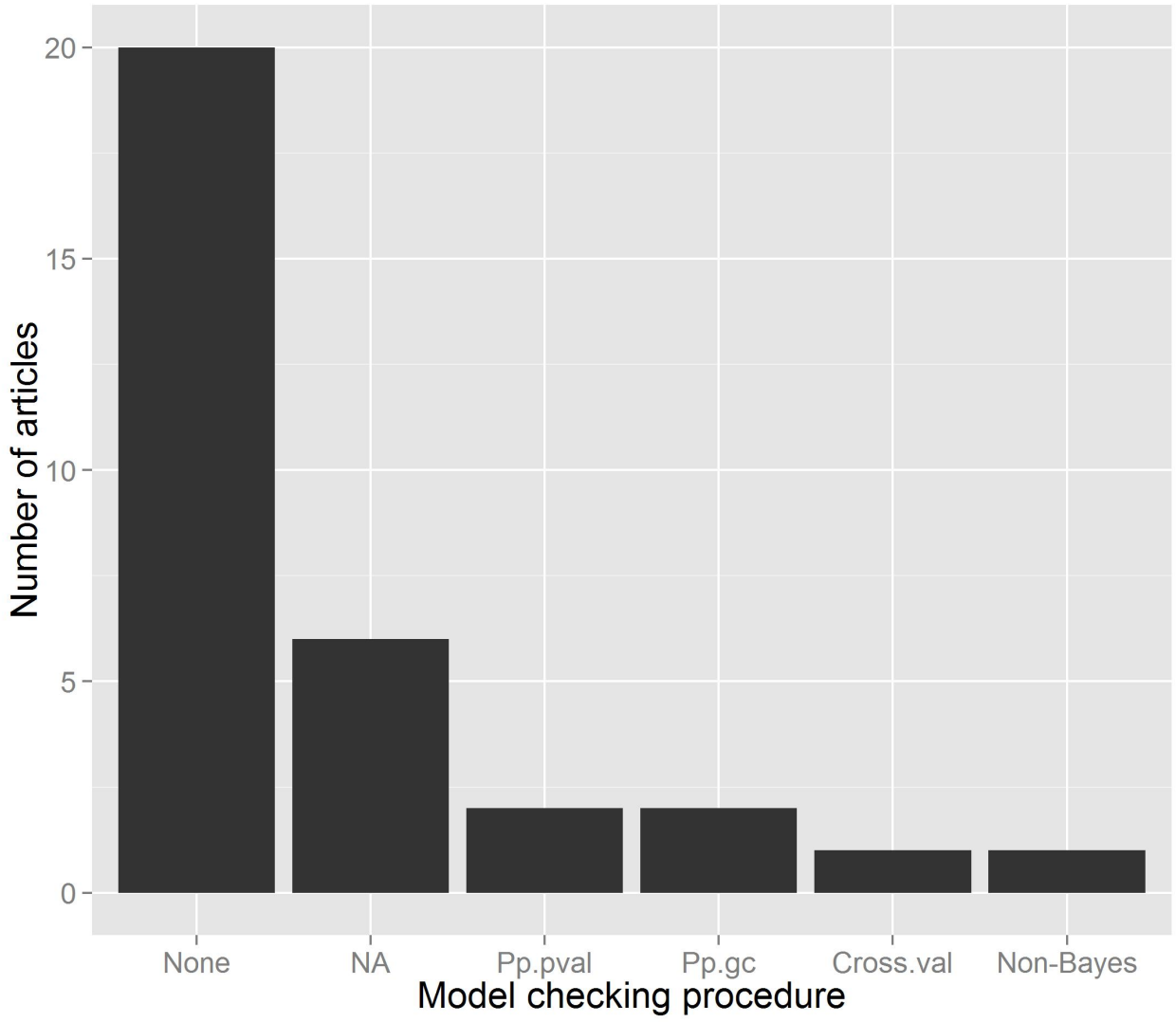
20
15
10
5
0
Number of articles
None
NA
Pp.pval
Pp.gc
Cross.val
Non-Bayes
Model checking procedure

FIG 2

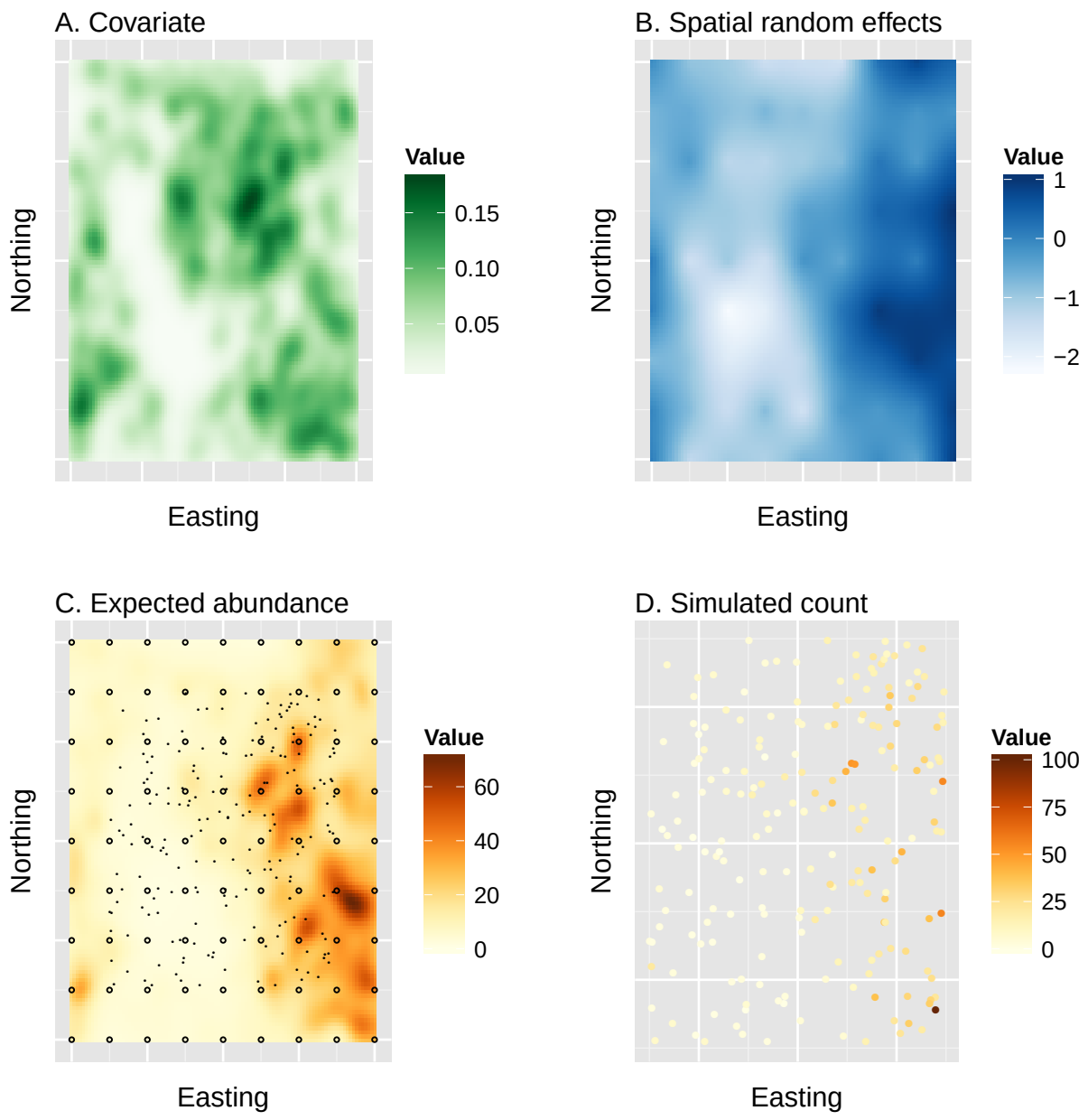
A. Covariate
Northing
Easting
Value
0.15
0.10
0.05
B. Spatial random effects
Northing
Easting
Value
1
0
−1
−2
C. Expected abundance
Northing
Easting
Value
60
40
20
0
D. Simulated count
Northing
Easting
Value
100
75
50
25
0

FIG 3

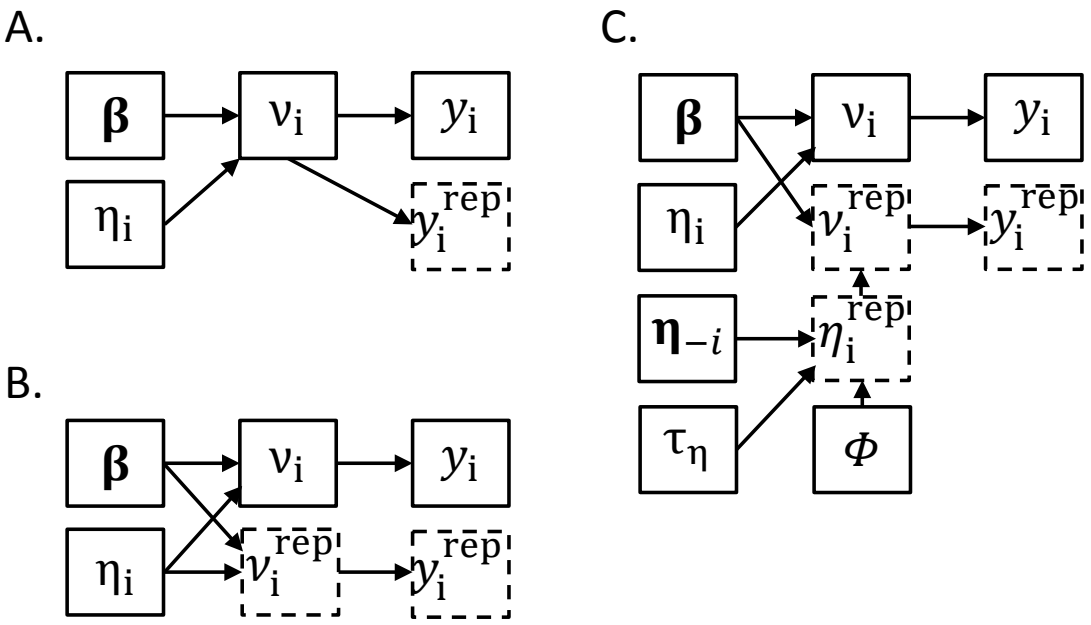
A.
β
ν_i
y_i
η_i
y_i^rep
B.
β
ν_i
y_i
η_i
ν_i^rep
y_i^rep
C.
β
ν_i
y_i
η_i
ν_i^rep
y_i^rep
η_{−i}
η_i^rep
τ_η
Φ

Fig 4

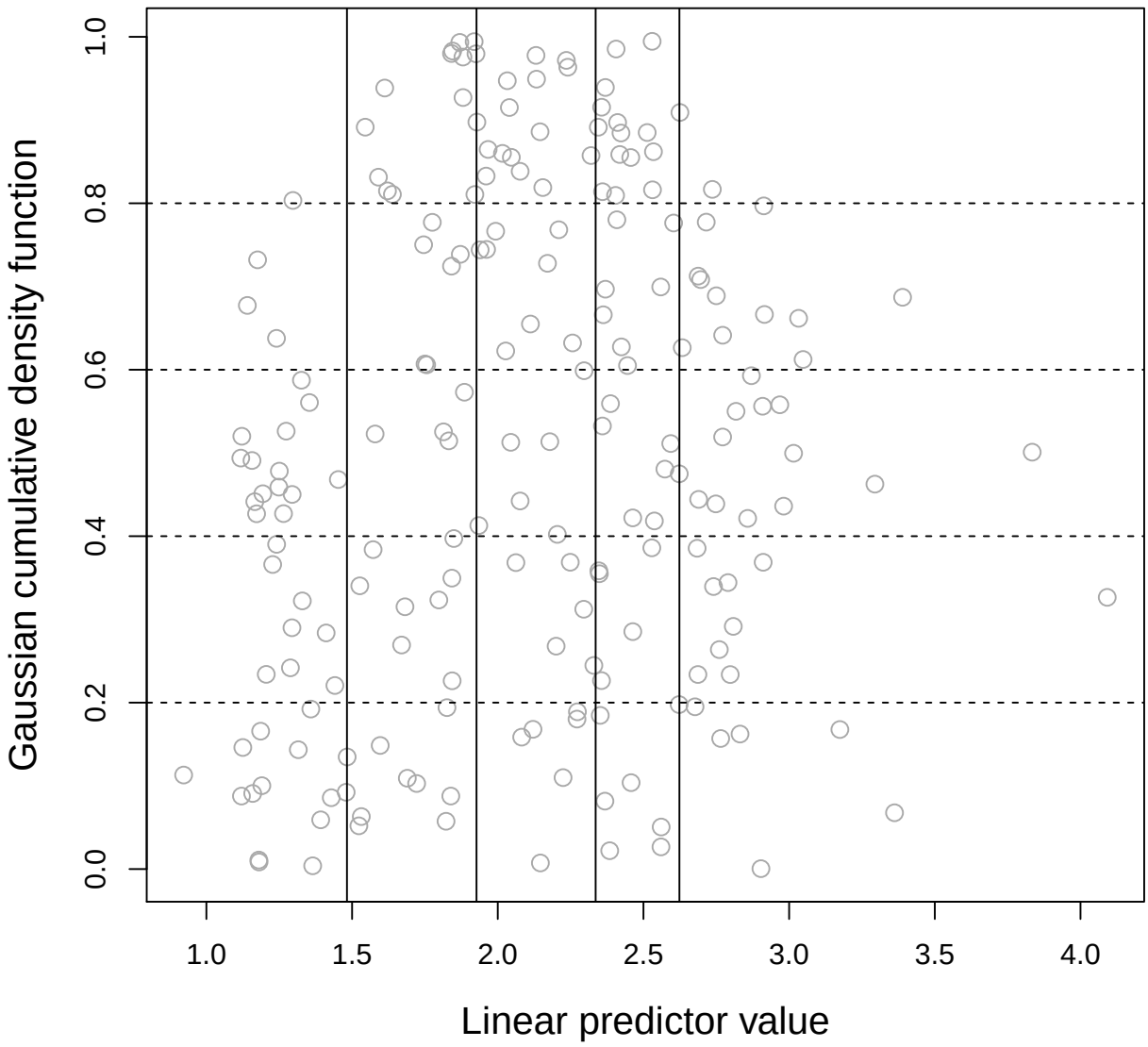
Gaussian cumulative density function
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Linear predictor value

FIG 5

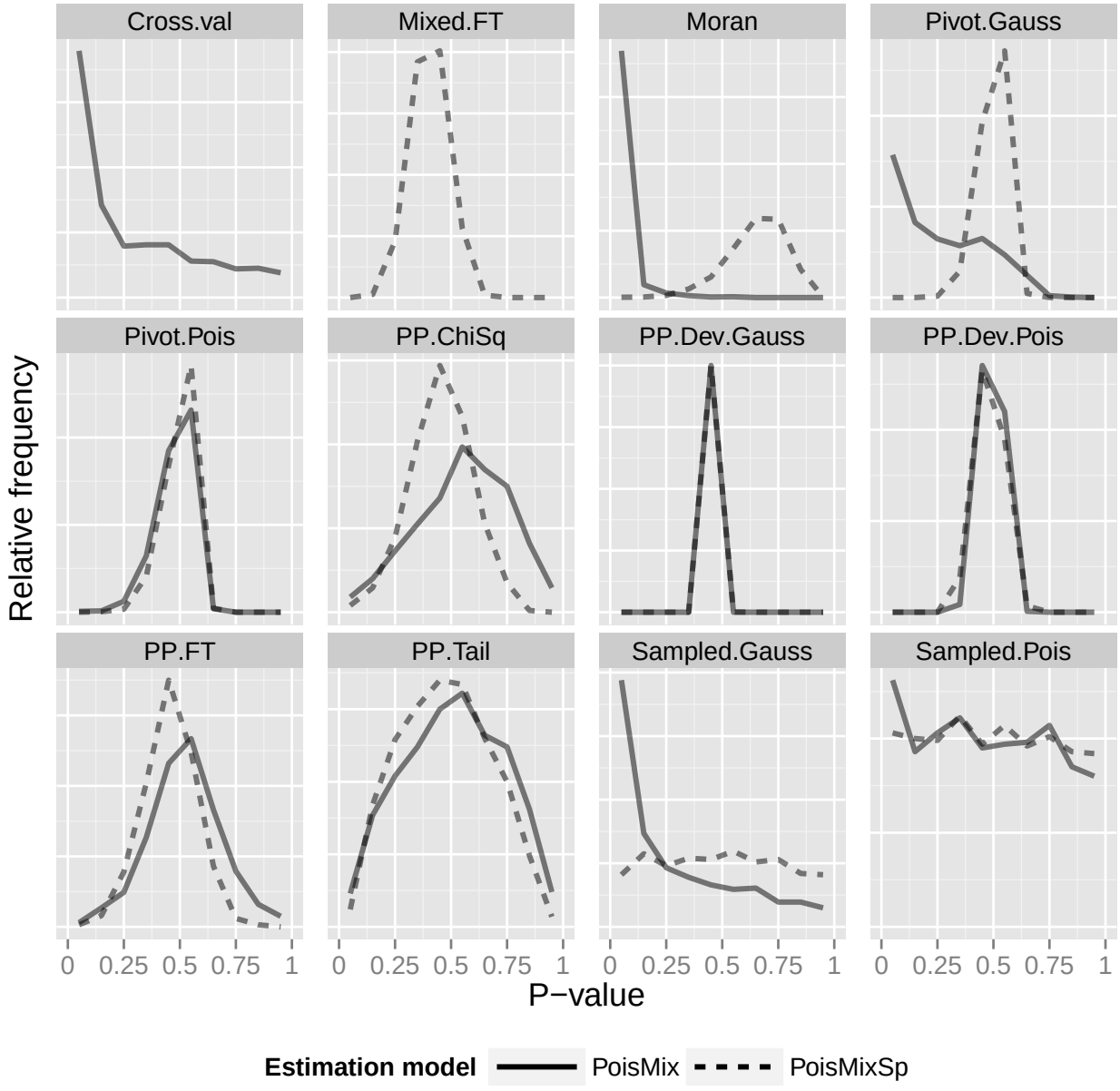
Cross.val
Mixed.FT
Moran
Pivot.Gauss
Pivot.Pois
PP.ChiSq
PP.Dev.Gauss
PP.Dev.Pois
PP.FT
PP.Tail
Sampled.Gauss
Sampled.Pois
Relative frequency
0
0.25
0.5
0.75
1
P−value
Estimation model
PoisMix
PoisMixSp

FIG 6