Appendix 3. Description of simulation analyses, and goodness-of-fit evaluation for empirical model.

Simulation description

We conducted a series of simulations to evaluate whether the statistical model could accurately recover regional population trends, given noisy and incomplete datasets. We conducted 250 independent simulations, each 20 years in duration, wherein we generated random log-linear slopes from $Uniform(-0.1,0.1)$ for each of two hypothetical strata. The relative abundance in each stratum was set to 1 in the initial year of the simulations.

Simulations were intended to approximate the data collection process during pre-breeding migration in our empirical study. Simulations assumed there were 13 migration monitoring stations that counted birds in each year of study, and stations operated for the same durations as in the empirical analysis described in the main text. For each station, we selected model parameters that would lead to "realistic" annual counts and temporal variation in counts.

In each simulation, we randomly assigned migration parameters ($\rho_{j,s}$) to each of 13 migration monitoring stations by drawing from a uniform distribution on the log-scale between log(0.01) and log(3), resulting in a wide but realistic range of "capture rates" and numbers of seasonal migrants among stations; most $\rho_{j,s}$ in the empirical analysis were less than 0.5, though several stations had values larger than 1. For the remainder of model parameters, we used median parameter estimates from the empirical model fit to pre-breeding migration, to ensure that simulated datasets contained realistic values of process and observation variance. However, unlike the empirical analysis, we assumed that no stations were known *a priori* to exclusively monitor a single stratum. These simulations therefore represent an especially difficult monitoring situation in which regional trends must be estimated based only on noisy mixtures of birds from multiple regions at all stations simultaneously. In each simulation, we also assumed the data collection process was identical to that in the empirical analysis; several stations only operated for a limited number of years and days across the study duration, and feather isotopes were only collected at a subset of monitoring sites (also in a limited number of years at each station), corresponding to those in the empirical analysis. We used these parameters and data constraints to simulate new datasets of observed counts on each day of the season, at each monitoring station.

For each simulation, we fit the Bayesian statistical model to the simulated datasets and calculated estimates of regional trends to compare to the "true" (simulated) trends that generated the data. Critically, when we fit the Bayesian statistical model to the simulated data, we did not enter the

simulated values of regional abundance ($X_{j,y}$) or migration parameters ($\rho_{j,s}$) as data. Instead, the model estimated those parameters using only a small sample of simulated birds with "known" breeding origins at a subset of sites and years, combined with simulated daily migration counts at each station. For converged models (those with R-hat statistics less than 1.1 for all monitored parameters), we calculated the mean bias in estimates of stratum-level trends and coverage of the 95% credible intervals. If the model can recover stratum-level trends, we expected that bias would be minimal on average and the 95% credible intervals would contain the true (i.e., simulated) stratum-level trends in 95% of simulations.

Results of simulations confirmed that the model could accurately recover stratum-level population trends under a wide range of hypothetical population trends and migration scenarios. Models converged for 249 of 250 simulations. Estimates of regional trends were highly correlated with true (i.e., simulated) trends, bias in estimated trends was extremely small (mean = -0.0008), and 95% credible intervals overlapped with the true trends for 95.7% of estimates (477 of the 498 stratum-level trend estimates from the 249 converged model runs; Fig. A3.1).

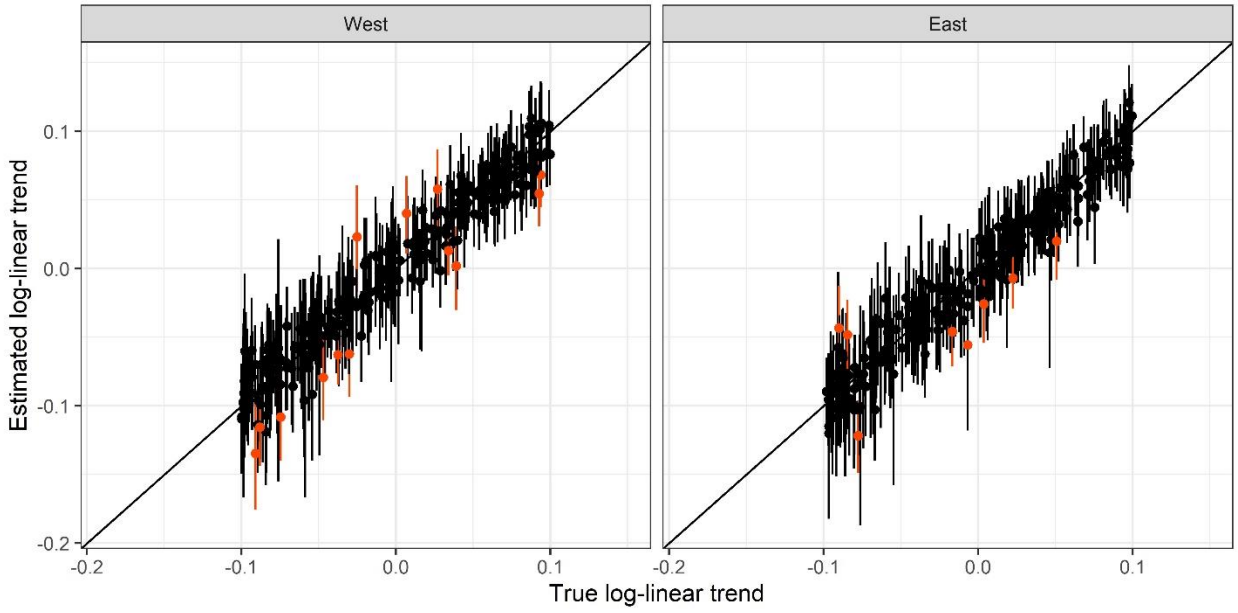

Fig. A3.1. Results of simulation analysis, comparing true (i.e., simulated) log-linear trends to estimated trends for each of two hypothetical strata. Red dots/whiskers indicate simulations in which the 95% credible intervals on estimates did not contain true trends.

Posterior Predictive Checks

We used posterior predictive checks to confirm that the empirical datasets were "similar" to simulated datasets based on the fitted models. The logic of this exercise is discussed more fully in Kery and Royle 2016; pp 192-198) and is a common approach for evaluating goodness-of-fit for Bayesian hierarchical models with multi-level error structures.

In brief, for each iteration of the MCMC fitting algorithm, we calculated an expected value for each datapoint in the empirical dataset, and additionally, we generated an entire simulated dataset based on the estimated model parameters and variance components (i.e., a new dataset that was "perfectly consistent" with the fitted model). We then calculated a measure of the discrepancy between the expected values from the fitted model and 1) the empirical data, and 2) the simulated data. A mis-calibrated model can be diagnosed when the discrepancy between the observed data and the model is consistently higher or lower than the discrepancy between the simulated data and the model. The proportion of simulated datasets with lower discrepancy measures than the observed data is called the "Bayesian p-value", and a well-calibrated model will have a Bayesian p-value close to 0.5, while p-values close to either 0 or 1 (and far from 0.50) indicate model miscalibration. This method therefore determines whether the empirical dataset "looks" as if it was simulated from the fitted model.

We calculated expected values for each datapoint as:

$$Expected_i = \exp(\log(T_{s,y}) + \log(f(d, \mu_s, \sigma_s)) + 0.5\omega_s^2).$$

Discrepancy measures were calculated as chi-squared statistics, where:

$$\chi_i^2 = \frac{(Count_i - Expected_i)^2}{Expected_i}.$$

To provide a measure of model "fit" for each station in each year of study, we summed these measures across all days of the season, for each monitoring station, within each year. We then calculated the proportion of simulated datasets with lower discrepancy statistics than the observed datasets for each station-year combination, within each season. Results for pre-breeding and post-breeding migration are depicted in Figures A3.2 and A3.3.

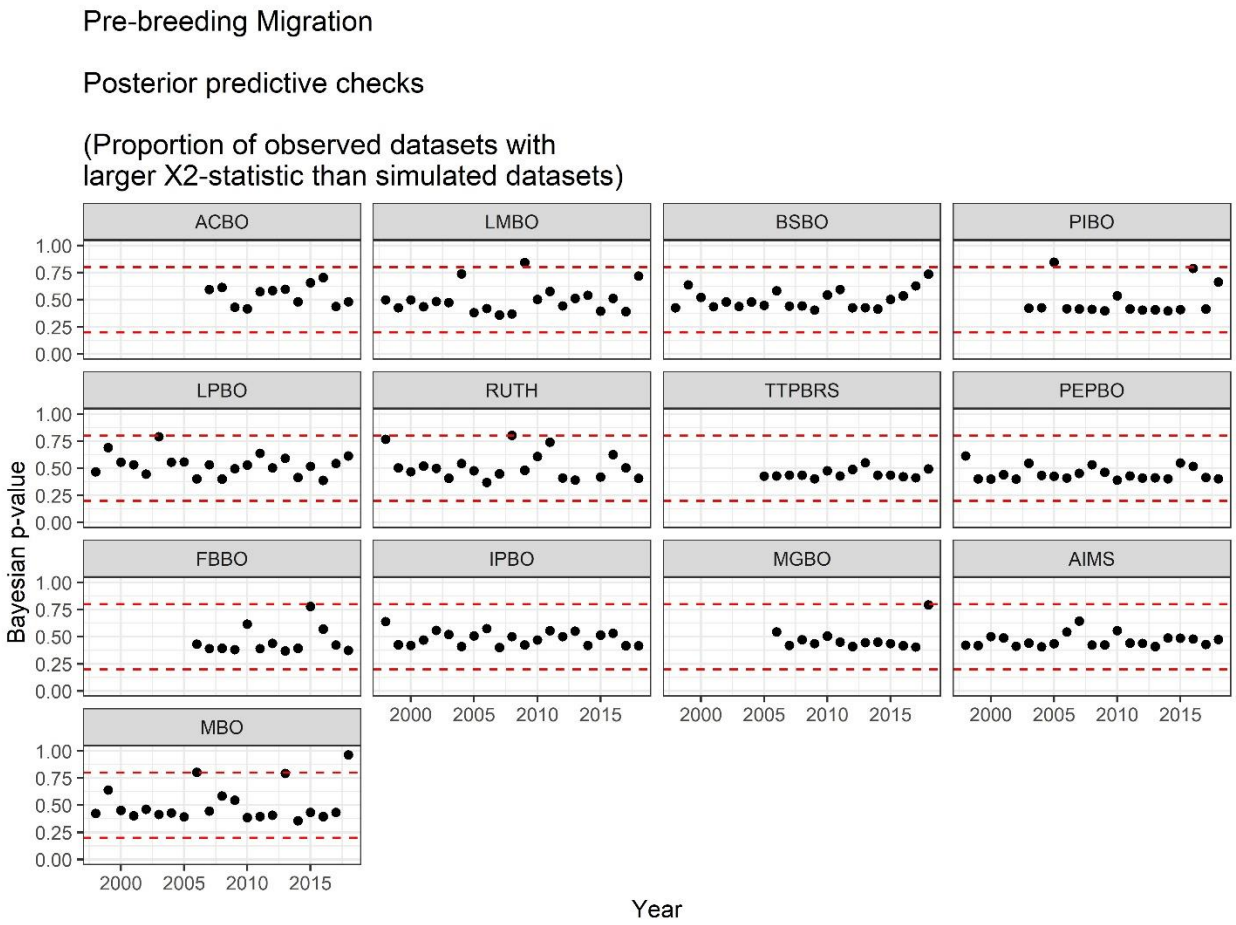

Fig. A3.2. Comparison of discrepancy measures for observed counts and simulated counts at each MCMC iteration, summed across all days of the season for each year, at each monitoring station for pre-breeding migration. Discrepancy measures is X-squared residual between observed counts and predicted count. Discrepancy for simulated count is summed squared residual between simulated counts and predicted count. P-values close to 0.50 indicate that the model is reproducing the observed distribution of data; values close to 0 or 1 indicate discrepancy is very different between simulated and observed data. Red dashed lines are positioned at p-values of 0.2 and 0.8 to help visualize instances where Bayesian p-values are indicating a potential lack of fit. Plots arranged in order from farthest west to farthest east.

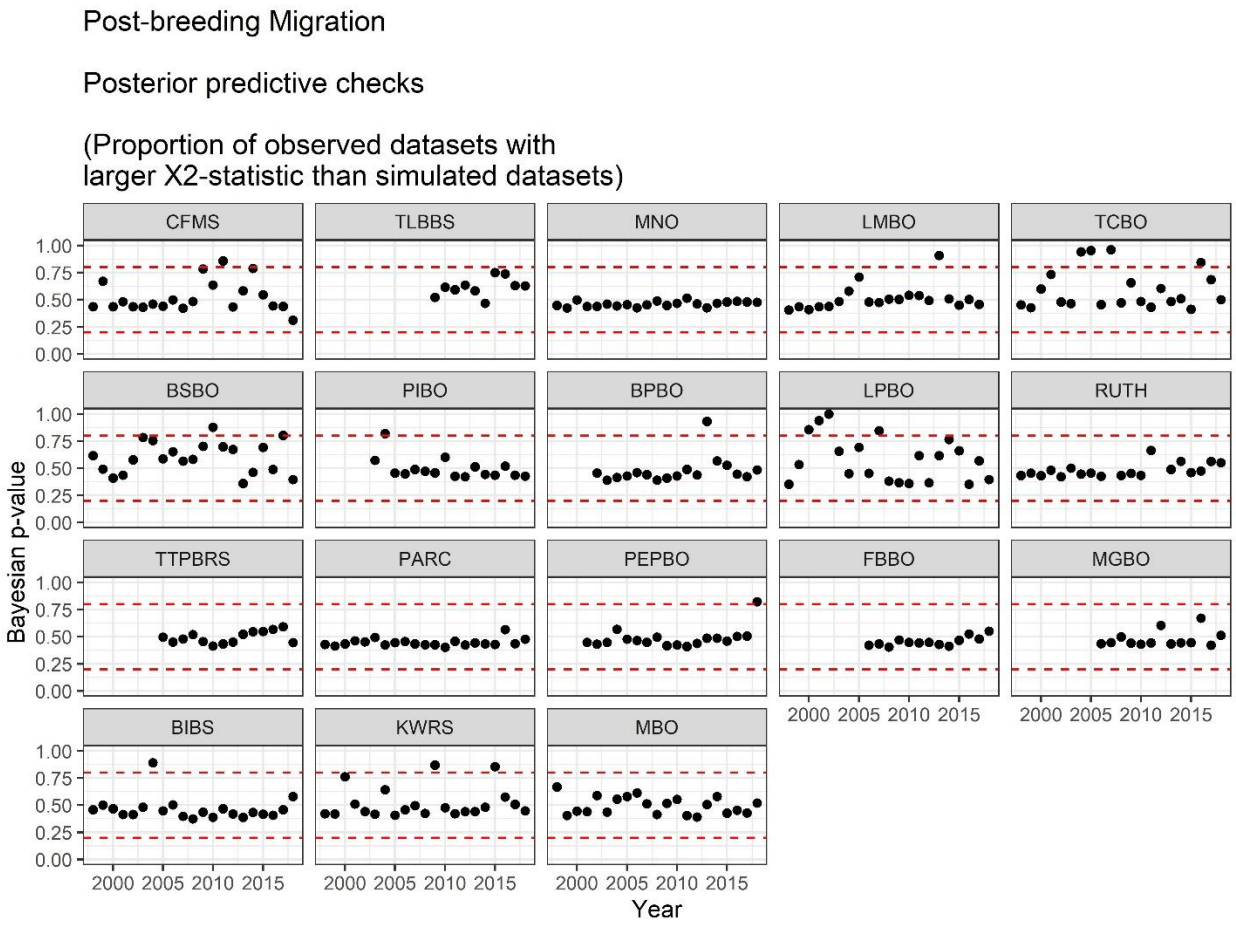

Fig. A3.3. Comparison of discrepancy measures for observed counts and simulated counts at each MCMC iteration, summed across all days of the season for each year, at each monitoring station for post-breeding migration. Discrepancy measures is X-squared residual between observed counts and predicted count. Discrepancy for simulated count is summed squared residual between simulated counts and predicted count. P-values close to 0.50 indicate that the model is reproducing the observed distribution of data; values close to 0 or 1 indicate discrepancy is very different between simulated and observed data. Red dashed lines are positioned at p-values of 0.2 and 0.8 to help visualize instances where Bayesian p-values are indicating a potential lack of fit. Plots arranged in order from farthest west to farthest east.

We also plotted observed total counts versus expected total counts for each station, within each year of study. Annual expected counts were calculated as $\sum_{d=1}^{365}\left(\exp\left(log(T_{s,y}) + log\left(\frac{1}{\sigma_s\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{d-\mu_S}{\sigma_s}\right)}\right) + offset_{d,s,y} + 0.5\omega_s^2\right)\right)$. These comparisons are illustrated in Figures A3.4 and A3.5, and mainly illustrate that the variance components in the model allow for accurate characterizations of temporal variation in seasonal totals at most stations, in most years.

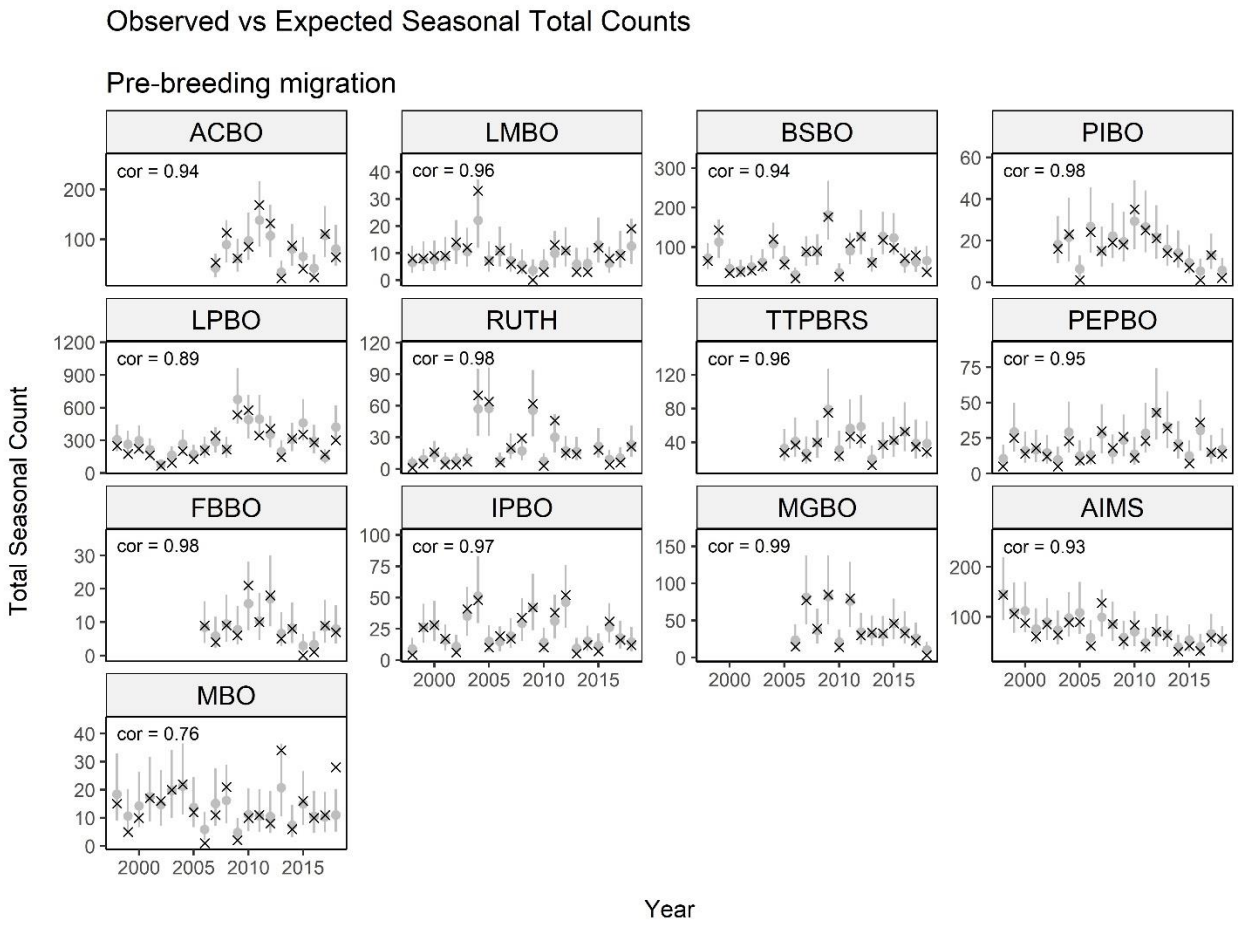

Fig. A3.4. Comparison of observed (black "x") and expected (gray dots and whiskers) total seasonal counts in each season, at each monitoring station. Whiskers are 95% prediction intervals (i.e., 95% of observations are expected to fall inside those bounds). Pearson correlation between observed and expected (mean of posterior) counts are written in top left of each facet. Plots arranged in order from farthest west to farthest east.

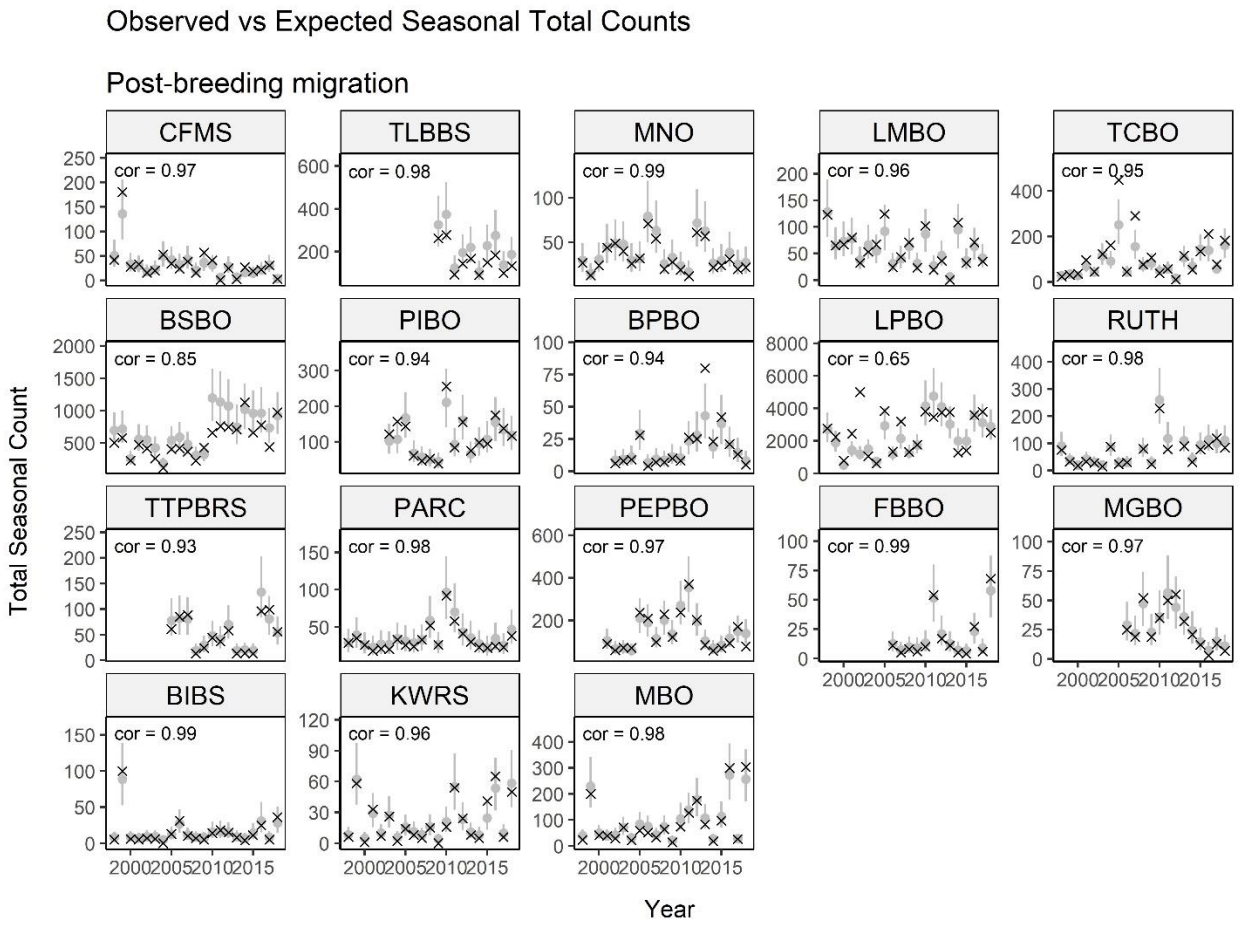

Fig. A3.5. Comparison of observed (black "x") and expected (gray dots and whiskers) total seasonal counts in each season, at each monitoring station. Whiskers are 95% prediction intervals (i.e., 95% of observations are expected to fall inside those bounds). Pearson correlation between observed and expected (mean of posterior) counts are written in top left of each facet. Plots arranged in order from farthest west to farthest east.

Literature Cited

Kéry, M., and Royle, J. A. (2015). Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models. Academic Press.