Appendix 2 – Goodness-of-fit evaluation for statistical model

We evaluated goodness-of-fit of our statistical model using several approaches. First, we visually compared predicted and observed counts for each colony, which are shown in Figures 2 and 3 in the main text of the manuscript. Second, we conducted posterior predictive checks, which evaluate whether the fitted statistical model can produce simulated datasets that “look like” the observed data. The degree to which a simulated dataset “looks like” the observed data is summarized using a discrepancy measure (see discussion in Kery and Royle 2016; pp 192-198). This approach entails generating an entirely new simulated dataset at each iteration of the MCMC fitting algorithm. Next, the simulated data is compared to the fitted values at each step of the MCMC algorithm using a discrepancy measure (in this study, we used root mean squared error; RMSE). The same discrepancy measure (RMSE) is also calculated between the observed data and the fitted values. The calculated discrepancy measures for the simulated datasets and observed dataset are then compared to evaluate whether the observed data has consistently higher or lower discrepancy with the fitted values than simulated datasets (which are, by definition, perfectly consistent with the fitted model).

Model misspecification can be detected if the discrepancy between the observed data and the fitted model is consistently higher or lower than the discrepancy between the simulated data and the fitted model. The proportion of simulated datasets with higher discrepancy measures than the observed data are summarized using a “Bayesian p-value”, where a Bayesian p-value close to 0.5 implies the observed data are highly consistent with the fitted model, while p-values close to either 0 or 1 (and far from 0.50) indicate the observed data are not consistent with the fitted model.

Figures S2.1 and S2.2 illustrate the results of posterior predictive checks for storm-petrels and puffins, respectively. The Bayesian p-value was 0.72 for the petrel model, and 0.43 for the puffin model. Both measures indicate that the observed data are consistent with the fitted models.

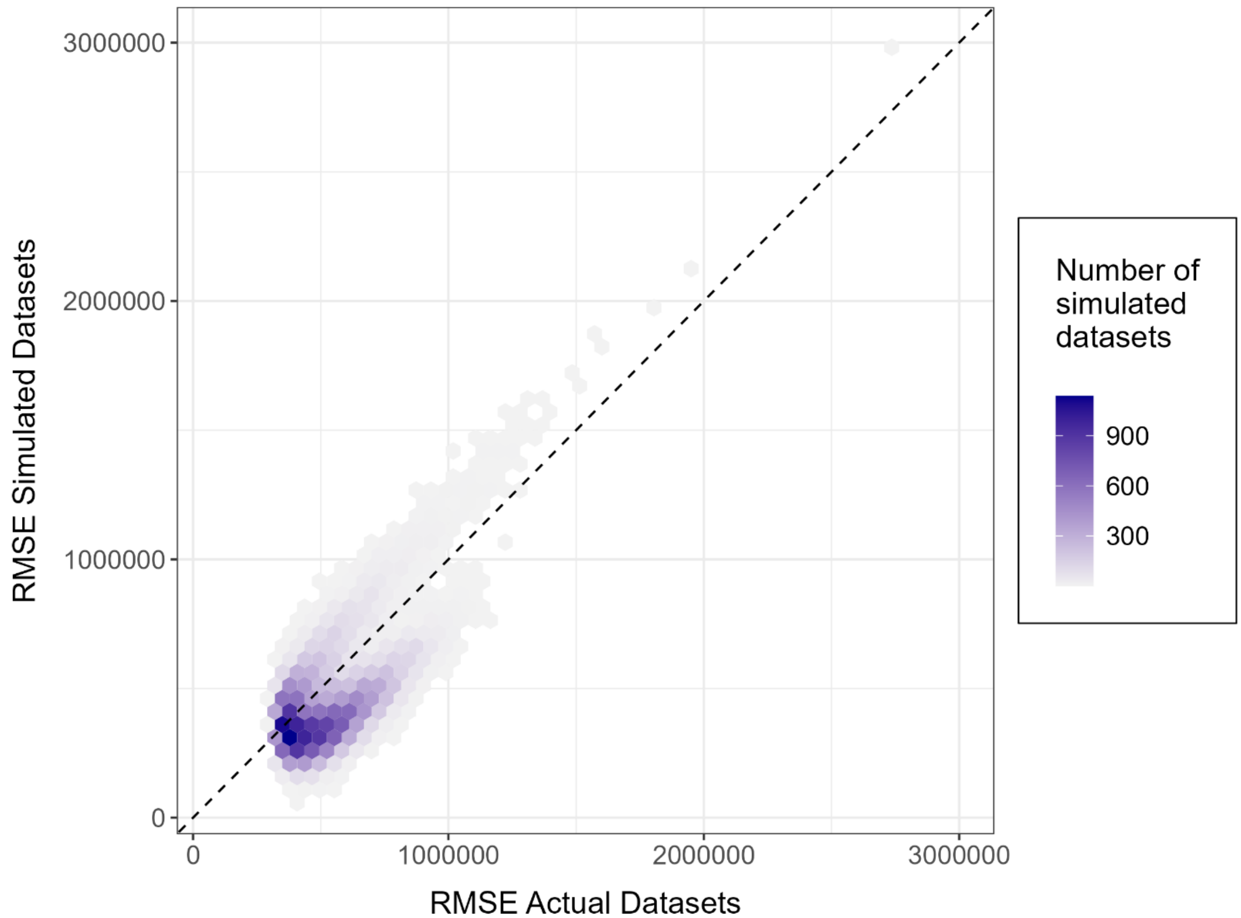

Figure S2.1. Posterior predictive check for the storm-petrel model, comparing discrepancy measures (root mean squared error; RMSE) between the observed data and fitted model (x axis) to discrepancy measures between simulated data and fitted model (y axis). Correspondence along the 1:1 line indicates that simulated datasets have similar discrepancy to fitted model as the observed data, suggesting adequate model fit.

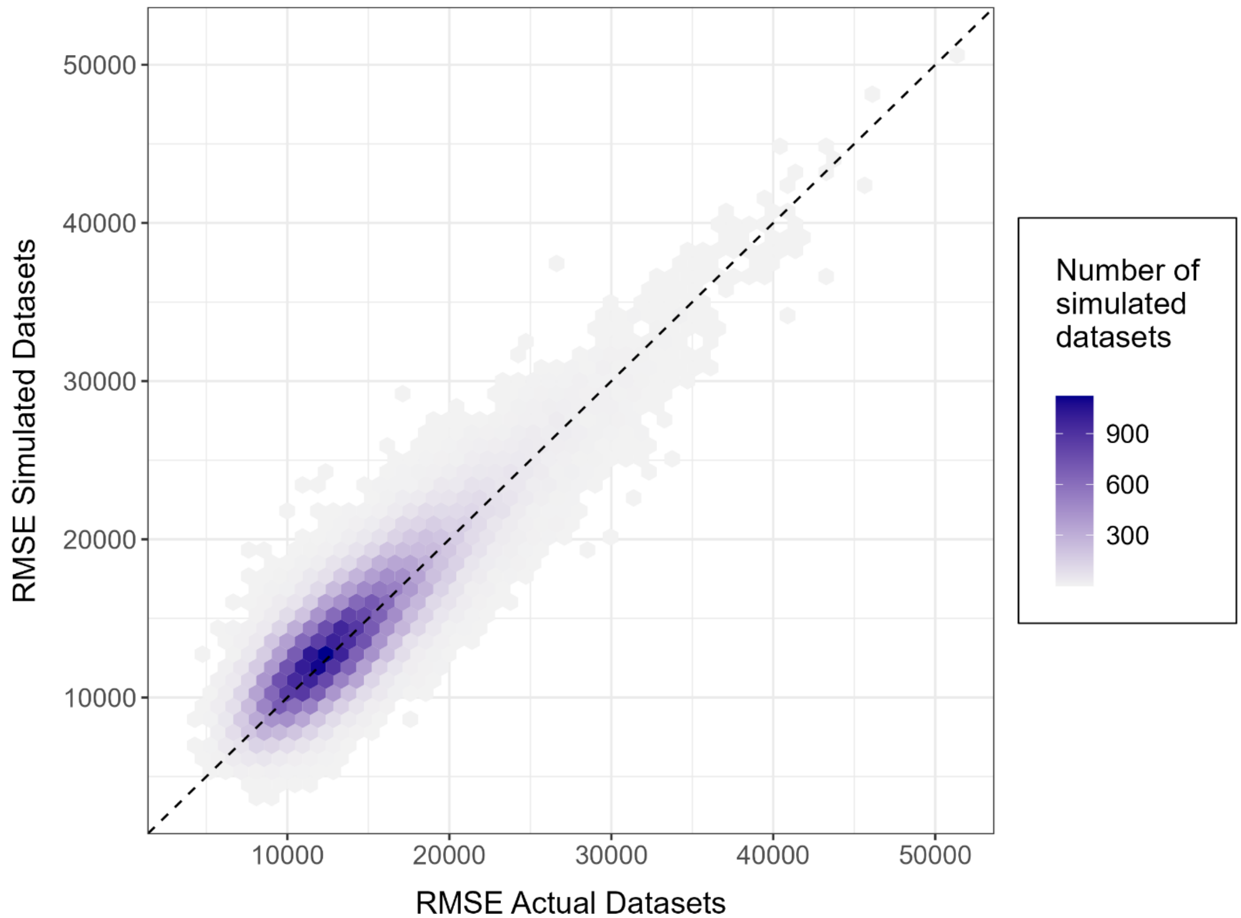

Figure S2.2. Posterior predictive check for puffin model, comparing discrepancy measures (root mean squared error; RMSE) between the observed data and fitted model (x axis) to discrepancy measures between simulated data and fitted model (y axis). Correspondence along the 1:1 line indicates that simulated datasets have similar discrepancy to fitted model as the observed data, suggesting adequate model fit.