Supplemental materials for "Integrating auxiliary data in optimal spatial design for species distribution modeling"

SM.1 – Second-order Approximation

The negative log-likelihood corresponding to (6) for one term (hence dropping the subscript i) is (dropping constants that do not depend on θ)

$$-\log[p(Y|\theta,\pi,\mathcal{D})] = \begin{cases} 0 & \text{if } n=0 \\ -\log[G(\theta)] & \text{if } n>0, Y>0 \\ -\log[G(\theta)q+1], & \text{if } n>0, Y=0 \end{cases} \tag{1}$$

for $q=(1-\pi)^n-1$. Therefore, for terms with $n=0$, corresponding elements of $\mathbf{M}(\mathbf{Y},\pi,\mathcal{D},\tilde{\boldsymbol{\theta}})$ and $\mathbf{V}(\mathbf{Y},\pi,\mathcal{D},\tilde{\boldsymbol{\theta}})$ are zero. For terms with $n>0$ and $Y>0$, the corresponding elements of $\mathbf{M}(\mathbf{Y},\pi,\mathcal{D},\tilde{\boldsymbol{\theta}})$ are

$$\frac{G'(\tilde{\theta})^2 - G''(\tilde{\theta})G(\tilde{\theta})}{G(\tilde{\theta})^2}\tilde{\theta} + \frac{G'(\tilde{\theta})}{G(\tilde{\theta})}$$

and the corresponding diagonal elements of $\mathbf{V}(\mathbf{Y},\pi,\mathcal{D},\tilde{\boldsymbol{\theta}})$ are

$$\frac{G'(\tilde{\theta})^2 - G''(\tilde{\theta})G(\tilde{\theta})}{G(\tilde{\theta})^2},$$

where G' and G'' are the first and second derivatives of G, respectively. For terms with $n > 0$ and $Y = 0$, the corresponding elements of $\mathbf{M}(\mathbf{Y}, \pi, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ are

$$\frac{q^2 G'(\tilde{\theta})^2 - qG''(\tilde{\theta})[G(\tilde{\theta})q + 1]}{[G(\tilde{\theta})q + 1]^2}\tilde{\theta} + \frac{qG'(\tilde{\theta})}{G(\tilde{\theta})q + 1}$$

and the corresponding diagonal elements of $\mathbf{V}(\mathbf{Y}, \pi, \mathcal{D}, \tilde{\boldsymbol{\theta}})$ are

$$\frac{q^2 G'(\tilde{\theta})^2 - qG''(\tilde{\theta})[G(\tilde{\theta})q + 1]}{[G(\tilde{\theta})q + 1]^2}.$$

For the probit link, $G(\theta) = \Phi(\theta)$, $G'(\theta) = \phi(\theta)$ and $G''(\theta) = -\theta\phi(\theta)$ where Φ and ϕ are the standard normal distribution and density functions, respectively.

SM.2 – Evaluation of the posterior approximation

To evaluate the quality of Section 2.3.1's approximation to the posterior of $\boldsymbol{\theta}$, we compare the posterior mean and variance of $\boldsymbol{\theta}$ obtained from the usual MCMC approximation versus the second-order approximation in three cases. The three datasets are generated as in Section 3.1 with "Donut" covariate structure and parameters fixed at $\boldsymbol{\beta} = (0, 0.5)^T$, $n_0 = 5$, and $\text{Cov}(\theta_i, \theta_j) = \sigma^2 \exp(-d_{ij}/3)$. The data are collected at $M = 100$ randomly selected locations and we compare the fidelity of the approximation for spatial standard deviations $\sigma \in \{1, 3\}$ and detection probabilities $\pi = \{0.3, 0.7\}$. Both the MCMC and Gaussian approximations assume the parameters to be fixed and known at their true values except for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ which have the same priors as Section 3.1.

Figure S1 plots the posterior mean and variance of each θ_i for the three simulated datasets. The

approximations are similar in the first and third cases with small spatial standard deviation except for sites with small θ_i and thus occupancy probability near zero. In the second case with large σ, the second-order approximation to the posterior mean is shrunk to zero and the posterior variance is underestimated. For data generated with large σ the occupancy probabilities are often close to zero or one and the normal approximation is inaccurate. However, even in the cases where the approximation is poor on the absolute scale, the ordering of sites in terms of occupancy probability remains reasonably accurate. In terms of computation, the MCMC approximation takes around 9 minutes while the second-order approximation takes less than 0.1 seconds.

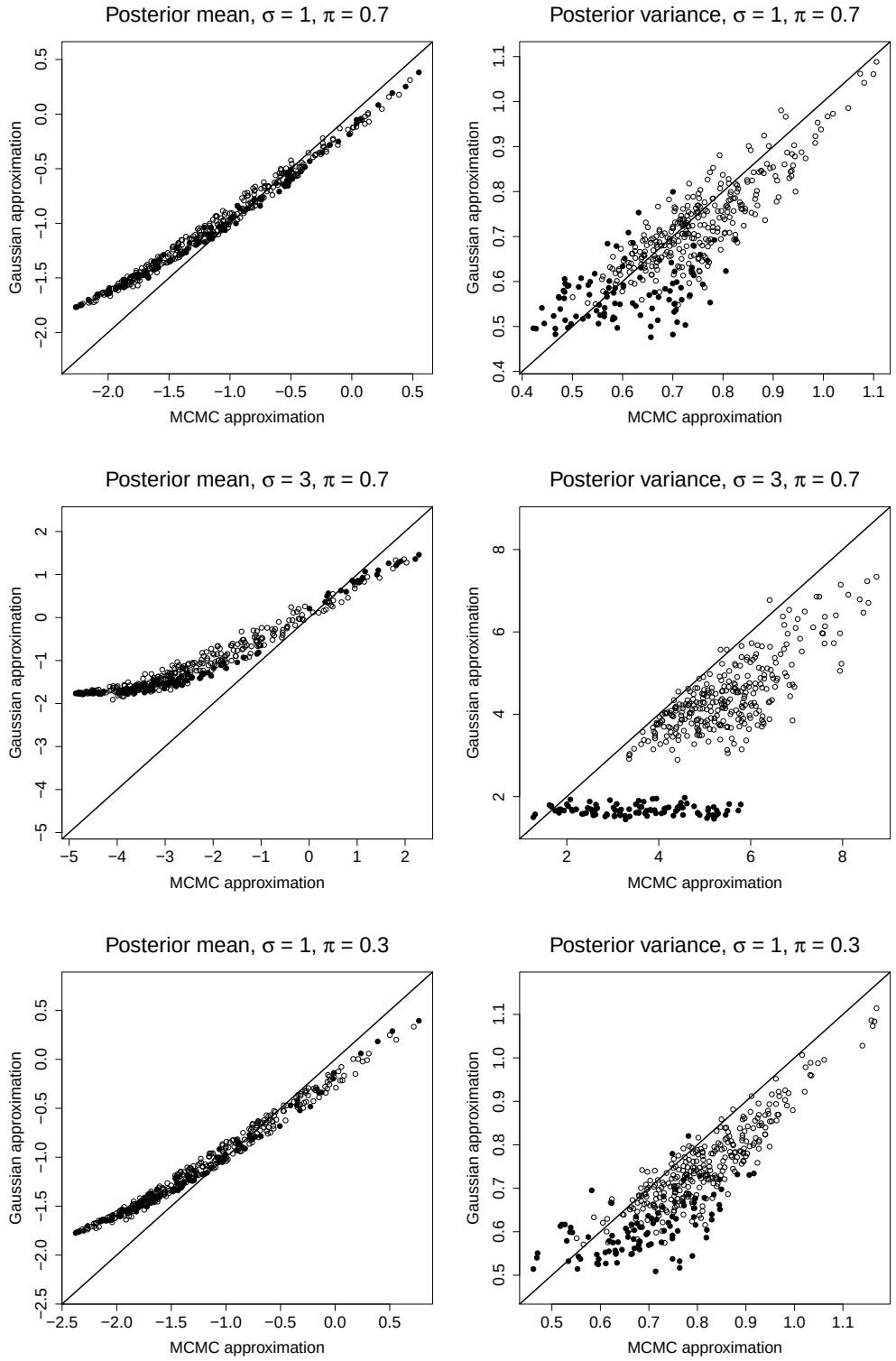
SM.3 – Priors and MCMC details

The spatial occupancy model described in Section 2.2 with probit link $G(\theta) = \Phi(\theta)$ and exponential spatial correlation can be written

$$\begin{aligned} Y_i|Z_i &\overset{indep}{\sim} \text{Binomial}(n_i, \pi Z_i) \\ Z_i &= I(\tilde{Z}_i > 0) \\ \tilde{Z}_i &\overset{indep}{\sim} \text{Normal}(\theta_i, 1) \\ \boldsymbol{\theta} &\sim \text{Normal}[\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\rho)] \end{aligned}$$

where σ^2 is the spatial variance and the (i, j) element of spatial correlation matrix $\mathbf{C}(\rho)$ is $\text{Cor}(\theta_i, \theta_j) = \exp(-d_{ij}/\rho)$. The priors are $\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\gamma}, \Lambda)$, $\pi \sim \text{Beta}(a, b)$, $\sigma^2 \sim \text{InvGamma}(c, d)$, and $\log(\rho) \sim \text{Normal}(m, s^2)$. We use uninformative priors by selecting hyperparameters $a = b = 1$, $c = d = 0.1$, $m = 0$, $s = 1$, $\boldsymbol{\gamma} = 0$ and $\Lambda = 10^2 I_p$.

Figure S1: **Approximate posterior mean and variance of θ_i for three datasets using MCMC versus second-order approximations.** Each point is the estimated posterior mean (left) or variance (right) of θ_i from the two approximation, and is shaded for sites with $n_i > 0$ and empty for sites with $n_i = 0$. The datasets were generated with different values of the spatial standard deviation (σ) and detection probability (π).



MCMC proceeds by setting initial values for all parameters and updating them in sequence from their full conditional posterior distributions. The occupancy indicators at location i $(Z_i, \tilde{Z}_i)$ are drawn simultaneously from their full conditional distribution $Z_i, \tilde{Z}_i|$rest as

$$Z_i|\text{rest} \sim \text{Bernoulli}\left[g(n_i, \theta_i)\right] \quad \text{and} \quad \tilde{Z}_i|Z_i, \text{rest} \sim \begin{cases} \text{TN}_{(-\infty,0)}(\theta_i, 1) & Z_i = 0 \\ \text{TN}_{(0,\infty)}(\theta_i, 1) & Z_i = 1 \end{cases}$$

where $g(n, \theta) = \frac{(1-\pi)^n \Phi(\theta)}{(1-\pi)^n \Phi(\theta) + 1 - \Phi(\theta)}$ and TN is the truncated normal distribution. The full conditional distributions for all $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, σ^2, and π are

$$\begin{aligned}
\boldsymbol{\theta}|\text{rest} &\sim \text{Normal}\left[(\Omega + I_N)^{-1}(\Omega\mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{Z}}), (\Omega + I_N)^{-1}\right] \\
\boldsymbol{\beta}|\text{rest} &\sim \text{Normal}\left[(\mathbf{X}^T\Omega\mathbf{X} + \Lambda^{-1})^{-1}(\mathbf{X}^T\Omega\boldsymbol{\theta} + \Lambda^{-1}\boldsymbol{\gamma}), (\mathbf{X}^T\Omega\mathbf{X} + \Lambda^{-1})^{-1}\right] \\
\pi|\text{rest} &\sim \text{Beta}\left[a + \sum_{i=1}^{N} Z_i Y_i, b + \sum_{i=1}^{N} Z_i(n_i - Y_i)\right] \\
\sigma^2|\text{rest} &\sim \text{InvGamma}\left[c + N/2, d + (\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{C}(\rho)^{-1}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})\right]
\end{aligned}$$

where $\Omega = \sigma^{-2}\mathbf{C}(\rho)^{-1}$ and $\tilde{\mathbf{Z}} = (\tilde{Z}_1, ..., \tilde{Z}_N)^T$. The spatial range parameter is transformed to $\rho^* = \log(\rho)$. The full conditional for ρ^* is proportional to

$$|\mathbf{C}(\rho)|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{C}(\rho)^{-1}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})\right] \phi\left(\frac{\rho^* - m}{x}\right)$$

where $\mathbf{C}(\rho) = \mathbf{C}[\exp(\rho^*)]$ and ϕ is the standard normal density function. The log range is updated using Metropolis sampling with random-walk normal candidate distribution tuned to have acceptance probability around 0.4. In all analyses we generate 10,000 MCMC samples and discard the

first 2,000 as burn-in. Convergence is monitored by visual inspection of the chains.

SM.4 – Sensitivity to starting values

For each of the eight combinations of β, π, and ρ we ran the exchange algorithm 10 times with different randomly-selected starting values for the m sampling locations; Figure 4 shows the best of the 10 solutions for each of the eight parameter settings. Figure S2 plots nine of the ten solutions for the scenario with high-quality auxiliary data, $\beta = 5$, high detection, $\pi = 0.7$, and strong spatial dependence in the true occupancy, $\rho = 2.0$. The 10 solutions are of course not identical, but all 10 place most of the sampling locations on the periphery of the distribution map estimated by eBird, and a few sampling locations on the edge of the domain where the eBird abundance estimate is low.

Figure S2: **Optimal spatial design for the brown-headed nuthatch for 9 different starting values.** The recommended sampling locations are white dots and the background color is the eBird abundance estimate. The designs are all for the case with true values $\beta = 5$, $\pi = 0.7$ and $\rho = 2.0$, but vary by initial configuration of sampling locations. The approximate value $\mathcal{V}(\mathcal{D})$ (times 100) is given below the map.