Appendix S2 - Downscaling of species distribution models vol. II: spatial autocorrelation, informative priors and prediction uncertainty

Petr Keil[1,2,*], Adam M. Wilson[1], Walter Jetz[1]

[1] *Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06520, USA*

[2] *Center for Theoretical Study, Charles University, Jilska 1, 11000 Praha 1, Czech Republic*

[*] *Correspondence:* pkeil@seznam.cz

1 Simple model for spatial scaling of false absences

There is no theory that would formally link the rate of false absences (or false presences), sampling effort and spatial grain (resolution) of data. To make a preliminary first step to explore spatial scaling of false absences we performed simple simulations. We created a grid (128 × 128 grid cells) of an artificial presences and absences that followed random fractal distribution (Fig. S1): all finer-grain grid cells occupied a constant proportion (p) of an area occupied by the coarser-resolution grid cells (Fig. S1a). We are aware that real-world species distributions can deviate from fractality, but fractals represent a good approximation of species spatial aggregation at multiple scales and are more realistic than completely random or regular distributions.

To simulate the incomplete sampling coverage we randomly sampled a given proportion (s) of the finest-grain presences and up-scaled the sampled data to progressively coarser grains (Fig. S1b). We determined the *Minimum Reliable Resolution* (MRR) as the resolution at which the number of occupied grid cells of the original and the sampled data were identical (Fig. 2c). We repeated the simulation 500 times for p and s being randomly drawn from uniform distribution ($0.2 < p < 0.9$ and $0.1 < s < 0.9$). We then made the estimated MRR as a function of p and s using a simple linear model, and we plotted partial residuals of the model (Fig. S1d).

We found that with decreasing sampling effort the MRR increases and the increase is mostly linear (Fig. S1d). When we compared the lowest (0.1) and highest (0.9) sampling efforts, the difference in MRR was about 2-3 orders of magnitude on the log_2 scale. It means that, given the random sampling of 10% of the sites, we need to multiply the original grain (the area of a single grid cell) roughly 2-4 times. This can vary depending on the proportion of presences at each resolution (the p; Fig. S1d). For widespread species ($p > 0.5$) the MMR can even be finer than the mentioned 2-4 multiple of the fine-grain resolution. For relatively rare species ($p < 0.5$) the variation around the model was be substantial (Fig. S1d). In the extreme scenario considered (a rare species of p=0.2 for which we know only 10% of its real presences) we need to multiply the area of a grid cell by roughly 40, which corresponds to an 6.32-fold increase of the length of a grid cell side.

This is only a first attempt to outline the spatial scaling of false negatives in single-species distributions, and to illustrate that there can be such a thing as MRR. In reality there are problems such as spatially clumped sampling effort, false positive observations or deviations from fractality in species distributions which will influence the MMR estimates. However, these issues can be incorporated into the theory.

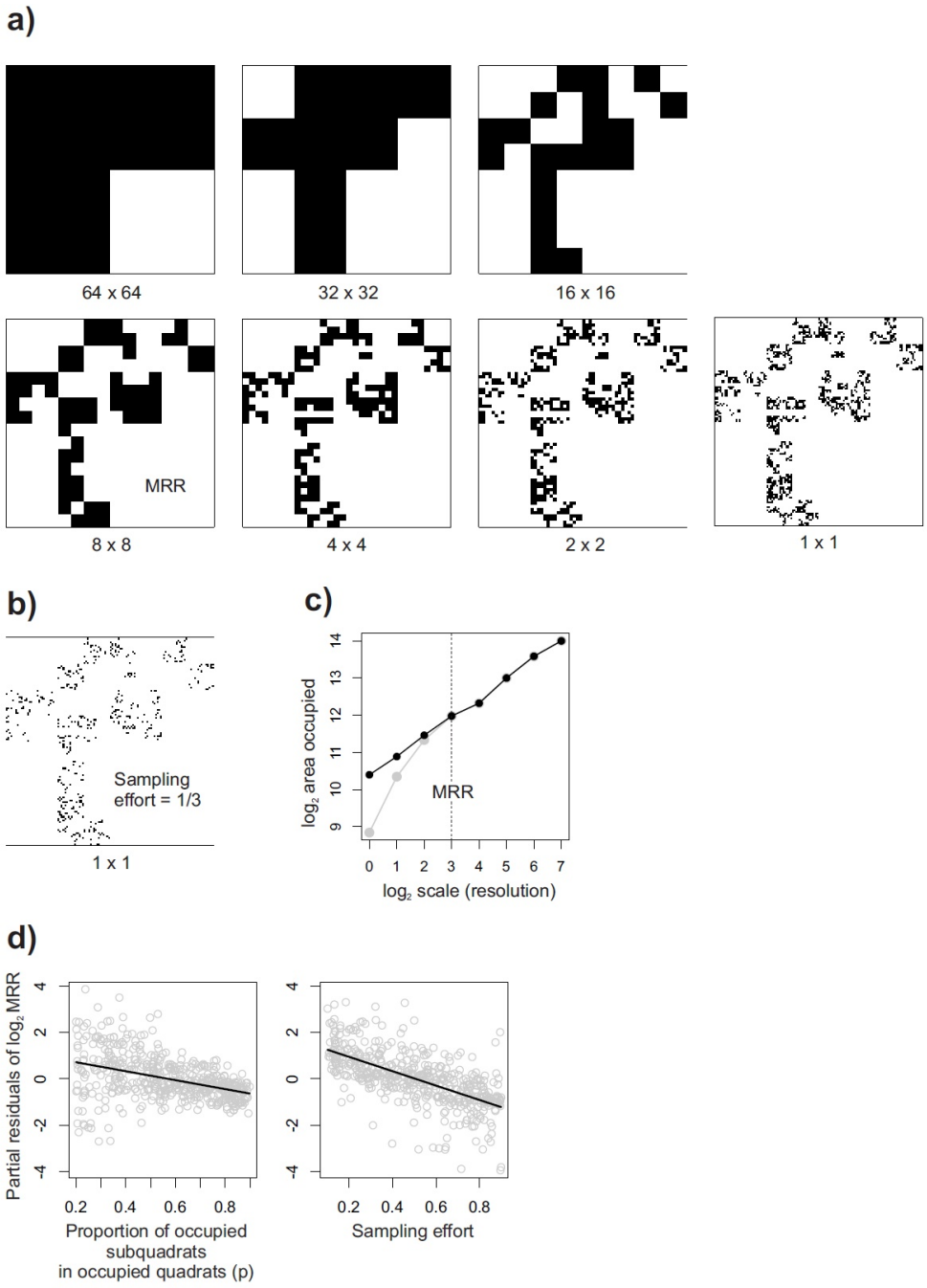

Figure S1 Simulation experiment exploring the spatial scaling of false negatives. (a) Realization of a species with random fractal distribution. For each finer-grain cell laying in an occupied coarser-grain cell we performed a Bernoulli trial with probability equal to the proportion (p). (b) To simulate incomplete observational data we randomly sampled a proportion (s, or sampling effort) of occupied grid cells at the finest resolution. (c) We estimated the MRR as the finest resolution of a grid that gives no rate of false negatives using the area-occupancy relationship. (d) We simulated the procedure (a-c) 500 times and we plotted the estimates of MRR as a function of p and s in a simple multiple regression model. The model gave R2=0.48 and formula of $log_2 MRR = 6 - 1.93 \times p - 3.1 \times s$.

2 Supplementary figures presenting details of predictions of the models

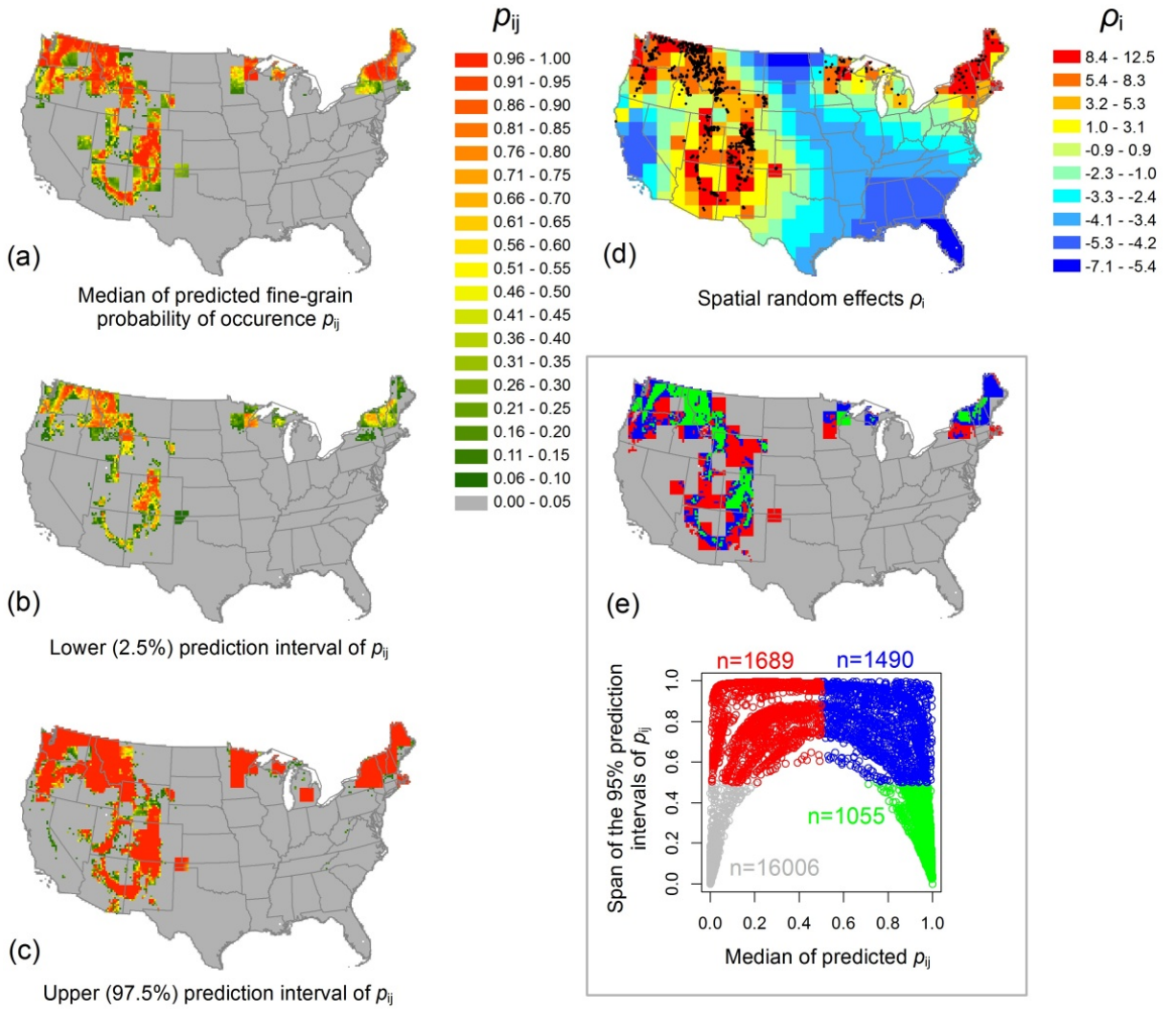

Figure S2 Predictions (at the 20 km × 20 km grain) of the **Full 2-scale model** that incorporated distributional data from two spatial scales and also the spatial random effects (but no informative priors on habitat preferences). The model was fitted using only the 751 well-surveyed grid cells of the validation dataset. Medians (a) and 95% prediction intervals (b-c) of the probability of the woodpecker occurrence. (d) Median values of the spatial random effects. (e) Fine-grain grid cells were classified into four categories, according to the predicted probability of the woodpecker's occurrence and uncertainty around this prediction.

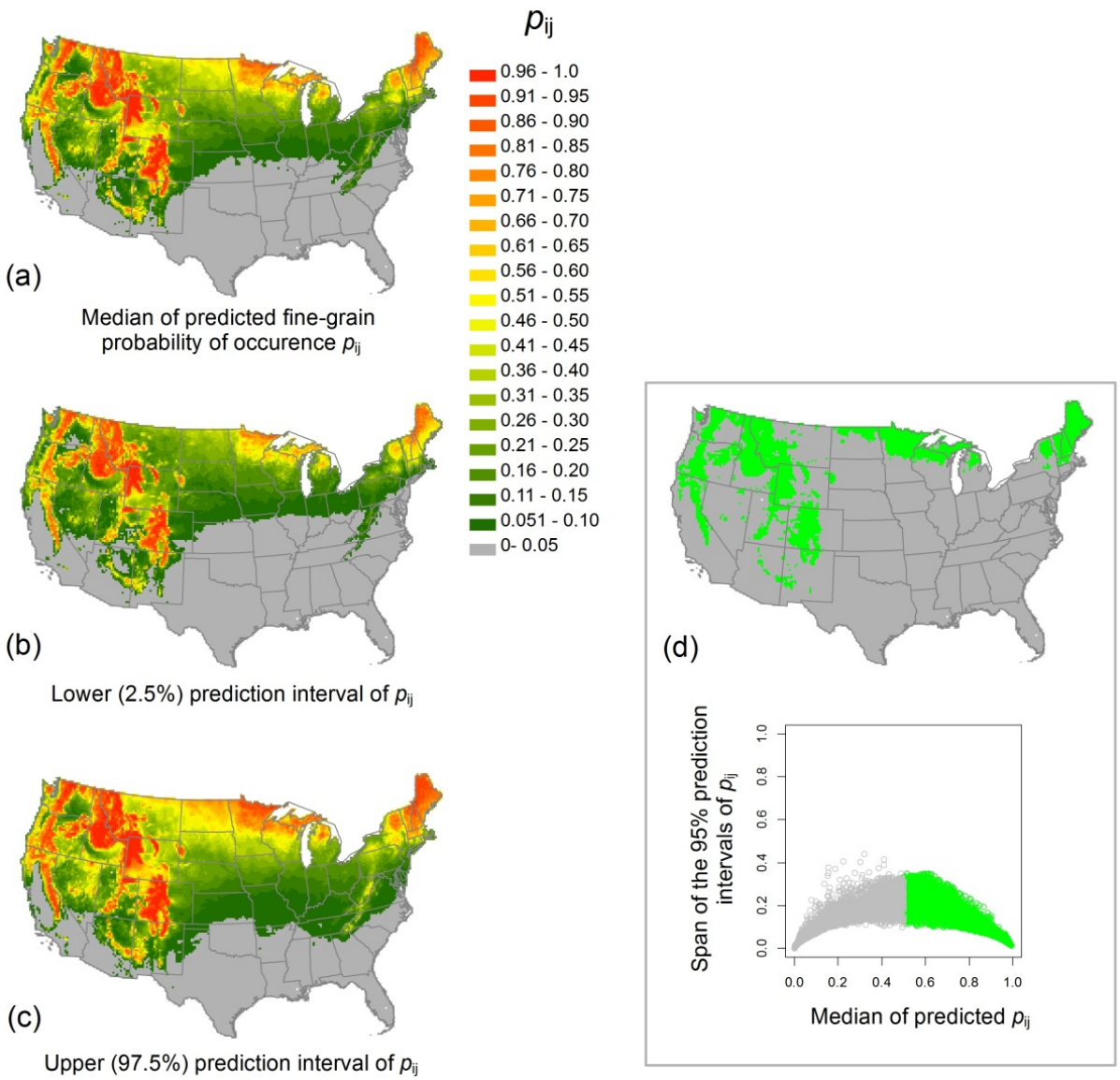

Figure S3 Predictions (at the 20 km × 20 km grain) of the **Fine-grain model** that incorporated no informative priors on habitat preferences, no spatial random effects and the distributional data only at the fine scale. The model was fitted using only the 751 well-surveyed grid cells of the validation dataset. Medians (a) and 95% prediction intervals (b-c) of the probability of the woodpecker occurrence. (d) Median values of the spatial random effects. (e) Fine-grain grid cells were classified into four categories, according to the predicted probability of the woodpecker's occurrence and uncertainty around this prediction. This model produced high probabilities of occurrence in areas where the species was never observed (e.g. in Sierra Nevada mountains).

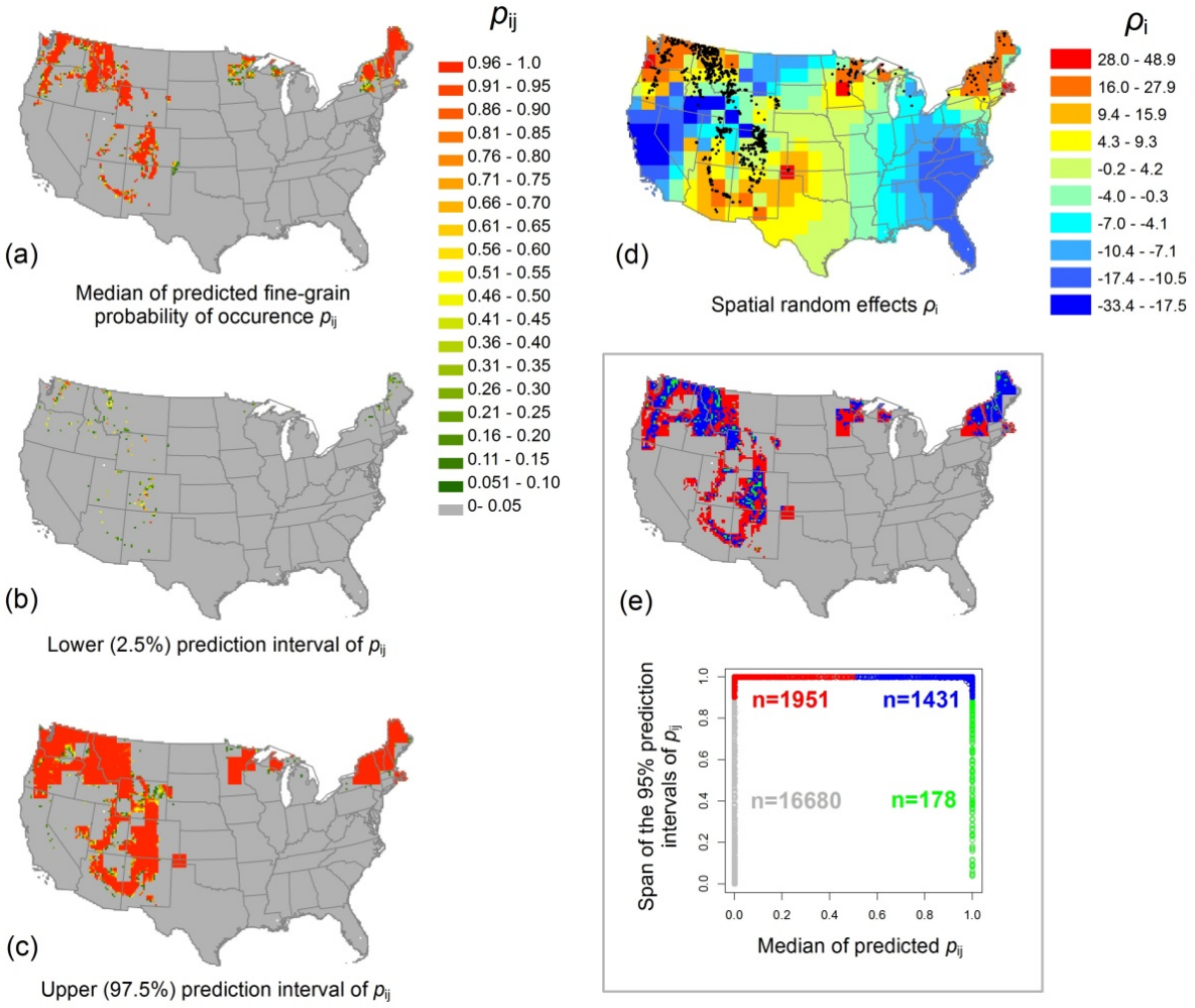

Figure S4 Predictions (at the 20 km × 20 km grain) of the **Downscaling model 2** that incorporated no informative priors on habitat preferences but used spatial random effects. Medians (a) and 95% prediction intervals (b-c) of the probability of the woodpecker occurrence. For example, a 2.5 quantile (b) of 0.51 (yellow) indicates that there is only a 2.5% probability that the true probability of presence is less than 0.51. Conversely, a 97.5% quantile (c) of 0.05 (grey) indicates that there is a 97.5% probability that the true probability of presence is less than 0.05 (virtually certain to be absent). (d) Median values of the spatial random effects. (e) Fine-grain grid cells were classified into four categories, according to the predicted probability of the woodpecker's occurrence and uncertainty around this prediction.

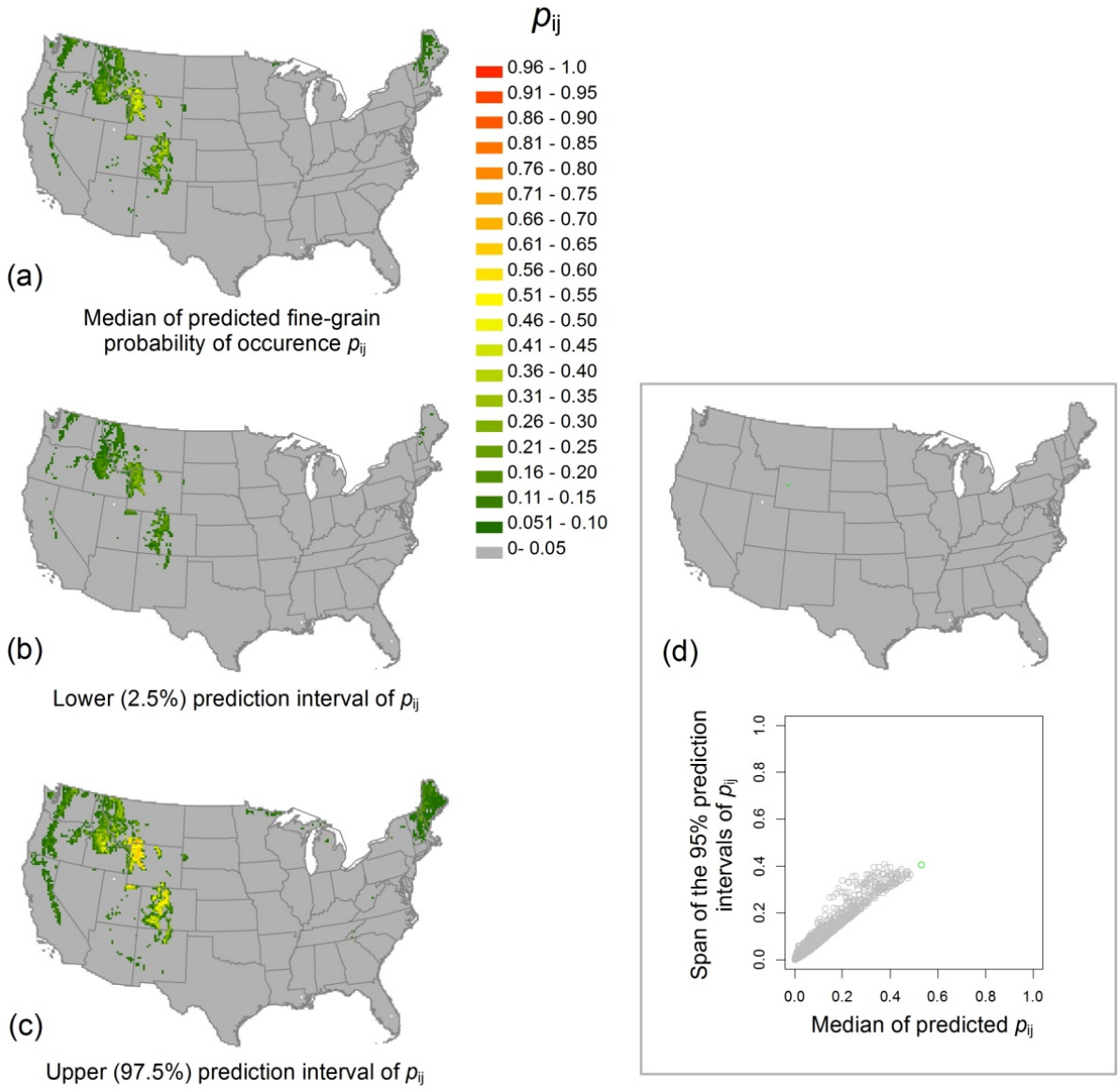

Figure S5 Predictions (at the 20 km × 20 km grain) of the **Downscaling model 3** that incorporated no informative priors on habitat preferences and no spatial random effects. Medians (a) and 95% prediction intervals (b-c) of the probability of the woodpecker occurrence. (d) Median values of the spatial random effects. (e) Fine-grain grid cells were classified into four categories, according to the predicted probability of the woodpecker's occurrence and uncertainty around this prediction. The model led to very low uncertainty in predictions, but the predictions were unrealistically low.