**Abstract**

Species distribution models (SDMs) rely on the availability of both presence and absence data to accurately predict species’ environmental niches. In the absence of reliable true absences, pseudo-absences are commonly used. Barbet-Massin et al. (2012) provided foundational guidelines for pseudo-absence generation, showing that model performance depends critically on how, where, and how many pseudo-absences are selected. Their results recommend using a large number of pseudo-absences (e.g., 10,000) with equal weighting relative to presences for regression techniques, and a 1:1 presence–pseudo-absence ratio for classification methods, alongside random selection strategies for pseudo-absences.

While these recommendations significantly advanced SDM methodology, they were based on specific modeling approaches and virtual species configurations. A key knowledge gap remains regarding the generalizability of these findings across broader pseudo-absence strategies. Notably, while Barbet-Massin et al. recommend a large number of pseudo-absences for regression techniques and a 1:1 ratio for classification models, the paper does not explicitly address how different presence–absence ratios interact with varying numbers of initial presence points. To address this, we test multiple initial presence sample sizes (20, 50, 100, 500, 1000) and apply varying pseudo-absence ratios (x1, x3, x5, x10) to systematically evaluate their impact across different modeling approaches.

Although Barbet-Massin et al. emphasize that bias in presence data can degrade model accuracy, we do not explicitly test for sampling bias in presences and instead assume our presence data is unbiased. Nonetheless, we aim to assess how different spatial thinning methods applied to pseudo-absences influence model outcomes. To do this, we first generate pseudo-absences randomly within a fixed buffer of 200 km around the presence points to simulate climatic limits of the virtual species. We then apply two thinning strategies—checkerboard and spThin() with a fixed minimum distance of a diagonal of one raster cell or 30 km accordingly—to both the pseudo-absences and the presences. This allows us to explore the potential effects of spatial filtering on model robustness and predictive performance.

We test a range of modeling algorithms (GAMs, GLMs, Random Forests, Maxent), run multiple times and averaged, then validated using several performance metrics: AUC, TSS, Kappa, Sensitivity, Specificity, Pearson Correlation Coefficient (PCC), Deviance explained (D2), and threshold-dependent statistics. Additionally, we calculate and compare the mean niche optima derived from the models against the initial environmental parameters used to define the virtual species.