Appendix S3: Derivation of D: 'Deviation from the best model'

Model predictions from Simulation 1 and Simulation 2 were evaluated using a novel metric derived from first principles that we named 'deviation from the best model' (*D*). This metric compares the probability of obtaining the true (raw) abundance (i.e. before sampling occurs) in each cell based on the model prediction, with the probability of obtaining a prediction equal to the true (raw) abundance i.e. predictions produced by a perfect model (Eq. 1). For each grid cell *i*, the probability of obtaining the true (raw) abundance (pA_i) was estimated from a Poisson probability distribution with a mean equal to the predicted mean abundance ($\bar{A}i$) for that cell. The summed natural logs of these probabilities across the study area represents the overall probability of obtaining the true (raw) abundances under the model predictions. This is then expressed as a ratio against the summed natural log probabilities for each cell ($q\bar{A}_i$) that would be obtained for a perfect model where the true (raw) abundance is equal to the predicted mean abundance.

Eq. 1

$$D = \frac{\sum(\ln(pA_i|\bar{A}_i))}{\sum(\ln(q\bar{A}_i|\bar{A}_i))}$$

The rationale behind creating a new evaluation metric is that our generation of occurrence points was based on a Poisson process and was deliberately zero-inflated, so there are an extremely large proportion of 0's and 1's, and the highest 'abundance count' is only six. Therefore, a metric based on the probability of obtaining the raw data from the model, rather than a direct assessment of the actual values, would provide a more appropriate measure of model performance and fit that was not as weighted by the large proportion of zeroes in the data. Using traditional binary presence-absence classification metrics would results in penalties against large predictions of abundance in comparison to the true raw abundance: the observed values of 0 and 1 which are most common in our dataset would score more highly using binary classification methods. The magnitude of the difference between prediction and true (raw) abundance observations will scale with the mean, leading to greater "inaccuracy" in cells with large numbers of individuals, regardless of the model used. Therefore, a metric that considers the actual abundance value rather than just presence-absence is likely to provide a more accurate assessment of model predictive power.