**Literature Review for ENMeval**

**Methods**

We searched Web of Science on June 5, 2019 for articles that had cited the original publication describing ENMeval (Muscarella *et al*. 2014). In total, we retrieved 182 studies published from 2016 to 2019 (Appendix 1). We read each study and recorded: (1) if ENMeval was actually used for analysis or only cited for concepts, (2) which features of ENMeval were used, (3) which method was used to partition data into testing and training datasets, (4) which evaluation metric(s) were used and whether or not their values were reported, (5) whether the selected Maxent settings (i.e., regularization multiplier and feature class combinations) were reported, and (6) whether variation in model performance across settings was reported. We also checked supplemental materials for articles when they had references in the relevant sections of the main text.

**Results**

Of the 182 studies we reviewed, 141 (78%) used ENMeval for part of their analysis and we consider only these studies in the results described below. Of these studies, 92% (131/141) used it to evaluate model performance or select among candidate models (i.e., tune to identify ‘optimal’ settings). The most commonly used metric for selecting ‘optimal’ settings was AICc (70% of studies), followed by testing AUC (21%), and then omission rates (11%). About 62% of studies that used ENMeval for evaluation statistics (81/131) reported the model settings selected as ‘optimal’ but only 26% reported variation of model performance across the settings used.

Among the 141 studies that used ENMeval as part of the analysis, 80% (113/141) reported at least one evaluation statistic and only 18% reported more than one evaluation statistic. The most commonly reported evaluation metrics were calculated on the test data: 60% of studies reported testing AUC, 26% reported omission rates (MTP and 10th percentile), 11% reported TSS, and 7% reported the difference between AUC of testing and training data. Few studies reported evaluation metrics calculated on the training data only: 11% reported training AUC and 4% reported the continuous Boyce index. It is important to note that only AUC and omission rate evaluation statistics were returned with ENMeval version 0.3.0 and previous versions, so any other metrics reported were calculated outside the software.

Most studies that used ENMeval for analysis (80%, or 112/141) reported the method used to partition testing and training data. Of these, 44% (49/112) used random data partitioning, 43% (48/112) used one of the spatial methods provided by ENMeval (i.e., block and checkerboard methods were used by 28% and 16% of studies, respectively), and 14% (16/112) used jackknife (leave-one-out) partitioning.

**Discussion points**

* The majority of studies that have cited ENMeval used the package to evaluate model performance and select ‘optimized’ model settings.
* The majority of studies have used AICc as the sole model selection criterion. There are issues with this (*haven’t there been recent critiques of this criterion?*).
* Too few studies report ‘optimal’ model settings actually used, which hinders research reproducibility and extension (like if somebody wanted to remake a model and project it to another time/place).
* Even fewer studies report variation of model performance across model settings, which prevents us from analyzing the overall effect of tuning model settings.
* Most studies report only one (or no) evaluation statistics. This presents challenges for evaluation of model performance and comparison across studies.
* Studies have used the spatial data partitioning methods of ENMeval but these should be used more. BlockCV is an alternative package for this.
* XX % of all the studied we surveyed did all the things that are suggested for SDM analyses (wish we could cite the “10 important things” paper, but could also cite the recent SDM standards paper)

**Figure 1.** Number of reviewed studies using a given evaluation metric.

