**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 between 2016 and 2019 (Appendix 1). We read each study to recorded: (1) if ENMeval was actually used for analysis or cited for concepts, (2) which features of ENMeval were used, (3) which evaluation metric(s) were used and which were reported, (4) whether the selected settings (i.e., regularization multiplier and feature class parameters) were reported, (5) whether variation in model performance across settings was reported, and (6) which method was used to partition data into testing and training datasets. We also checked supplemental materials for articles when a reference was made to these materials in the relevant sections of the main text.

**Results**

Out of the 182 studies we reviewed, 141 (78%) used ENMeval for part of the analysis. Of these, 93% used ENMeval to evaluate model performance or select among competing models (i.e., tune for ‘optimal’ settings). The most commonly used statistic for selecting model settings was AICc (70% of studies), followed by test AUC (21%), and then omission rates (11%). About 62% of studies reported the settings selected as ‘optimal’, but only 26% reported variation of model performance across the settings used. The most common evaluation metric reported was AUC of test locations (60% of studies). Only about a quarter of the studies (26%) reported omission rates; 11% reported AUC of the training data, 11% reported TSS, 7% reported the difference between test and train and 4% reported the Boyce index. Only 47% of the studies reported more than a single evaluation statistic, and only 18% reported more than statistics. About 35% of the studies used a spatial (as opposed to random) method of partitioning data.

**Discussion points**

* Most 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 model selection criteria. 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.
* 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 outside 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.

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

