Received: 21 March 2022 | Revised: 1 September 2022 | Accepted: 14 September 2022

DOI: 10.1002/ecm.1557

CONCEPTS & SYNTHESIS

ECOLOGICAL MONOGRAPHS
ECOLOGICAL SOCIETY OF AMERICA

Cross validation for model selection: A review with examples from ecology

Luke A. Yates | **Zach Aandahl** | **Shane A. Richards** | **Barry W. Brook**

School of Natural Sciences, University of Tasmania, Hobart, Tasmania, Australia

Correspondence
Luke A. Yates
Email: luke.yates@utas.edu.au

Funding information
Australian Research Council, Grant/Award Number: FL160100101

Handling Editor: Timothy E. Essington

Abstract

Specifying, assessing, and selecting among candidate statistical models is fundamental to ecological research. Commonly used approaches to model selection are based on predictive scores and include information criteria such as Akaike's information criterion, and cross validation. Based on data splitting, cross validation is particularly versatile because it can be used even when it is not possible to derive a likelihood (e.g., many forms of machine learning) or count parameters precisely (e.g., mixed-effects models). However, much of the literature on cross validation is technical and spread across statistical journals, making it difficult for ecological analysts to assess and choose among the wide range of options. Here we provide a comprehensive, accessible review that explains important—but often overlooked—technical aspects of cross validation for model selection, such as: bias correction, estimation uncertainty, choice of scores, and selection rules to mitigate overfitting. We synthesize the relevant statistical advances to make recommendations for the choice of cross-validation technique and we present two ecological case studies to illustrate their application. In most instances, we recommend using exact or approximate leave-one-out cross validation to minimize bias, or otherwise k-fold with bias correction if $k < 10$. To mitigate overfitting when using cross validation, we recommend calibrated selection via our recently introduced modified one-standard-error rule. We advocate for the use of predictive scores in model selection across a range of typical modeling goals, such as exploration, hypothesis testing, and prediction, provided that models are specified in accordance with the stated goal. We also emphasize, as others have done, that inference on parameter estimates is biased if preceded by model selection and instead requires a carefully specified single model or further technical adjustments.

KEYWORDS
cross validation, information theory, model selection, overfitting, parsimony, post-selection inference

INTRODUCTION

The expression of scientific hypotheses as statistical models is a fundamental component of ecological research. In addition to expert domain knowledge, modern statistical modeling requires substantial technical considerations including the formulation of mathematical descriptions of hypothesized relationships between variables and the structure of stochastic processes (Fox et al., 2015). Statistical modeling also involves consideration of computational aspects involved with the fitting of models to data. When existing theory and empirical evidence are insufficient to uniquely inform model choice, alternative models can be formulated and the available data used to assess their relative merits (Claeskens & Hjort, 2008; Hooten & Hobbs, 2015). The use of data in this way—to assess and ultimately select among alternative models—is called model selection. As an adjunct to statistical modeling, model selection has become integral to ecological research; indeed, as Tredennick et al. (2021) assert: "confusion about how to do model selection is confusion about how to do ecology".

Cross validation is a technique based on data splitting to make predictive assessments of statistical models. Although the specific goal of a statistical analysis—such as hypothesis testing or prediction—can constrain the set of models under consideration, predictive assessment is a broadly applicable and objective basis for both model comparison and selection across a range of modeling goals (Shmueli & Koppius, 2011). As a technique for predictive assessment, cross validation is extremely flexible due to the breadth of predictive measures (scores) with which it can be used (Gneiting & Raftery, 2007), the availability of data-splitting strategies that can be employed to account for the structure of the data and/or manage computational costs and estimation bias (Arlot, 2008), and its broad applicability to both optimisation and Bayesian frameworks. Recent methodological innovations, often involving approximation methods (Vehtari et al., 2017), have also improved the computational efficiency of cross-validation algorithms. Further, when the predictive measure is log likelihood, cross-validation estimates the relative expected Kullback–Leibler divergence which is the objective of commonly adopted information-theoretic model selection methods such as Akaike's information criterion (AIC) approximation (Akaike, 1973).

An important and long-standing concern in model selection is the issue of overfitting—the inclusion of spurious variables in a selected model. For hypothesis testing, overfitting misleads future research and is considered a substantial driver of the current replicability crisis in the sciences (Benjamini, 2020). For predictive goals, overfitting degrades the generalization of predictive performance to new data. Although it is well known that information-theoretic approaches, and predictive model selection more generally, suffer a tendency to overfit, it is less known that this proclivity is predominantly due to failure to correctly account for score-estimation uncertainty (Piironen & Vehtari, 2017a). For model selection based on cross validation, an effective and easily applied mitigation strategy is the use of calibrated selection rules which identity the simplest model with comparable predictive performance to the best-scoring model (Yates et al., 2021).

In this paper we seek to provide an accessible yet comprehensive review on using and understanding cross validation for model selection, with a focus on ecological problems. Our own efforts to synthesis a coherent picture of the current state of the model-selection literature—which mostly appears as technical results in statistical journals—motivated us to create a useful reference for practitioners who are not necessarily biostatisticians. We include and explain technical aspects that are important but often overlooked, such as bias correction, estimation uncertainty, choice of predictive score, and calibrated selection rules. In addition to the technical and conceptual review we give clear recommendations on cross-validation strategies including adjustments to manage computational demands while accounting for bias and mitigating the risk of overfitting. Prior to an exposition of model scores and cross-validation techniques, the first section of this paper provides an overview of the different goals of statistical modeling, the corresponding purposes of model selection, and the merits of using predictive assessment. In particular we emphasize, as others have done (Breiman & Spector, 1992; Claeskens & Hjort, 2008), the incompatibility of model selection with inference on model parameters. Several boxes and tables are included in the paper to provide an easily referenced summary of the topics reviewed or an additional level of detail on specific subjects. We also provide two ecological case studies which are used to illustrate many of the methods recommended in the paper.

PREDICTIVE ASSESSMENT, MODELING GOALS, AND THE PURPOSE OF MODEL SELECTION

Cross validation works by splitting the available data into a pair of training and test sets where the model is fit to the training data and subsequently assessed on the basis of its predictions to the test data (Hastie et al., 2009). By repeating this process for many different splits of the data, the average predictive performance of one or more models is estimated. Cross-validated predictive performance is commonly used to estimate or tune auxiliary

parameters, often called hyperparameters, of a model (e.g., the degree of smoothness of a spline), or to make comparisons between a discrete set of models for the purpose of selection (Arlot & Celisse, 2010).

The need for model selection arises when there is uncertainty about which model, from a given set of candidates, is best suited to achieve to a specified modeling goal (Claeskens & Hjort, 2008). Common modeling goals include: (1) exploration (or data mining) to generate new hypotheses; (2) hypothesis testing for a small set of multiple-working hypotheses; (3) prediction to new data; and (4) causal inference based on estimated model parameters (Tredennick et al., 2021). Although it would be most convenient for scientific progress if all four of these goals could be achieved simultaneously using a single data set, they are for the most part mutually exclusive.

The incompatibility of these four goals can be seen by considering how models are generated in each case. Models for exploration are generally informed by existing theory, but only weakly, so that the full set of candidates will necessarily include new combinations of variables or model structures to create the possibility of discovery. In contrast, models for hypothesis testing are strongly based on both theoretical understanding and empirical evidence from previous studies (Burnham & Anderson, 2002); thus, the models are defined a priori, and model selection for hypothesis testing is confirmatory. Models for prediction are often generated from algorithms with complex and obscure internal structures and might include variables with theoretical support ranging from strong to virtually none at all. Finally, for inference on parameter estimates there should be just one model which is generated from strong theoretical understanding and substantial evidence from previous studies. Thus, the goal of causal inference is incompatible with model selection (but see Box 1 for recent methodological innovations).

Despite their differing objectives, model selection for the goals of exploration, hypothesis testing, and prediction can all be performed on the basis of predictive assessment. This may seem contradictory, especially for hypothesis testing, since it is known that the model closest to "truth" is not necessarily the best predictive model (Arif & Aaron MacNeil, 2022; Shmueli, 2010) (see Box 3 for further discussion). However, the specified modeling goal strongly constrains which models are included in the candidate set, and for this reason, the interpretation of the selected model will differ even if the same method is used to compare models across the different goals.

For example, one does not expect any one of the models within a small set of multiple-working hypotheses to be the best possible predictive model (compared, for instance, to an artificial neural network trained with all available predictors). However, given that all the candidate models have strong theoretical support, each corresponding to an alternative causal hypothesis, their predictive performance when confronted with new real-world data provides both a reality check and an objective basis for assessing their relative merits (Shmueli & Koppius, 2011). Indeed, it is predictive performance that underpins information-theoretic approaches to model selection (e.g., AIC) where models are scored based on their predictive log likelihood which is an estimate of relative expected Kullback–Leibler divergence (Akaike, 1973): a relative measure of the information lost by approximating (the usually unknown) true model with each candidate.

The merits of using predictive assessment as the basis for model selection for exploratory and predictive goals are more obvious. For exploratory analysis, the ultimate goal of generating new scientific hypotheses to perform future hypothesis tests or make inferences using new data demands that models should have a causal interpretation and that the inclusion of spurious variables is to be avoided. Although modern "interpretable" techniques such as partial dependence plots (Friedman, 2001) and Shapely additive explanations (Lundberg & Lee, 2017) can provide insight into correlative associations for almost any class of statistical model, a causal interpretation requires that the model specification is consistent with known or hypothesized casual mechanisms; for example, specified within a structural-causal framework (Pearl, 2010). For all modeling goals, predictive assessment can be used to implement strategies that seek to mitigate the risk of overfitting including statistical techniques such as regularization (e.g., penalized regression) or calibrated selection rules (Yates et al., 2021).

How does predictive assessment compare with other approaches to model selection? For explanatory goals, goodness-of-fit measures such as the proportion of variation explained would seem a natural choice for model assessment. However this is problematic for model selection because it leads to overfitting due to a preference for complex models where the increased flexibility fits noise in the data. Null-hypothesis significance testing has a long tradition in the sciences, but has sustained heavy criticism due to concerns including the a priori bias toward the null model being true, correct adjustment for multiple testing, the requirement for models to be nested, and the arbitrary nature of *p*-value thresholds (Johnson, 1999; Wasserstein & Lazar, 2016). The predictive approach has the advantage that all models are placed on an equal footing and models do not need to be nested to make comparisons (Burnham & Anderson, 2002). Still, the use of predictive model selection is not without its challenges. Predictive scores are random variables, such that the bias and variance of estimated scores impact selection decisions (Piironen & Vehtari, 2017a). When two or more models

BOX 1 Valid post-selection inference

The problem of inference after selection

Using a single data set to make selection decisions, and then subsequently making inferences based on computed statistics is rife with issues (Kabaila, 2009; Shen et al., 2004). Failure to correctly account for the impact of the selection procedure on the bias and uncertainty of effect estimates is considered to be a leading cause of the current replicability crisis in the sciences (Benjamini, 2020). In the context of model selection, using data to select a preferred model and then acting as though the model was decided upon a priori, by using the same data to estimate parameters (without further adjustments), leads to biased estimates and optimistic confidence intervals (Hjort & Claeskens, 2003). These inferential issues apply equally to selection procedures based on predictive assessment such as information criteria and cross validation, regularization such as penalized regression and shrinkage priors, as well as null-hypothesis testing.

Some technical solutions

A suite of methodological papers that offer techniques for making valid post-selection inference have emerged in the statistical literature in the last two decades (Zhang et al., 2022). These approaches are predominantly concerned with linear models and the estimation of valid confidence intervals, with specialized methods for different selection algorithms, including the LASSO (Lee et al., 2016), AIC (Charkhi & Claeskens, 2018), and forward stepwise regression (Tibshirani et al., 2016). There are more general approaches that are not conditional on a specific algorithm or the selected model (Bachoc et al., 2019; Berk et al., 2013); however, their interval estimates are more conservative than those of specialized methods. Recent advances extend valid post-selection inference to generalized linear models (Garcia-angulo & Claeskens, 2023).

While these techniques allow for valid inferences to be made in various specialized settings, they are highly technical and remain embedded in the statistical literature rather than emerging in the form of didactic publications for applied analysts (but see the R package `selectiveInference` (Tibshirani et al., 2016)). Given the restrictions on the modeling settings in which these methods apply, post-selection inference currently has limited utility for applications in ecology.

A practical way forward

If, at the outset of a statistical analysis, there is model-selection uncertainty, we need to ask whether we are really ready to make inferences on parameter estimates. Using our data to perform model comparisons and publishing the full set of results, possibly including a preferred selection, is itself a valuable research outcome. If we do decide to make inferences, Kabaila et al. (2016) have shown that the confidence intervals for the full model (with all potential predictors included) are comparable to those obtained using certain post-selection methods. This result supports the intuitive notion that the large estimation uncertainty for parameters in the full model, relative to those of a simpler submodel, is a good estimate of our overall inferential uncertainty (Harrell, 2015). Finally, one should not use model-averaged parameters, as advocated by Burnham and Anderson (2002), as a means to account for selection uncertainty when making inferences; unless all models are linear and there is no multicollinearity among the predictors, this approach is invalid (Banner & Higgs, 2017; Cade, 2015; Claeskens & Hjort, 2008).

perform comparably, assessing the "significance" of the difference between scores for the purpose of selection requires thresholds to be set, much like null-hypothesis significance testing.

PREDICTIVE MODEL SELECTION: A BRIEF OVERVIEW

Model selection uses the available data to compare and select among a set of candidate models. Models differ in their specification, such as complexity (e.g., the number of variables included and/or regularization methods), the functional relationship between variables (e.g., variable interactions or alternative non-linear functional dependencies), structure (e.g., alternative grouping, autocorrelative, or hierarchical structures), and the choice of the probability distributions. There are many technical terms commonly used with model selection; Table 1 contains a glossary of technical terms used in this paper.

For a given set of candidate models, predictive model selection is based on estimates of the predictive

TABLE 1 Glossary of terms.

Term	Definition
Calibration	Quantification of the significance of expected score differences (e.g., to determine when the performance of two models is comparable)
Confusion matrix	A matrix summarizing the predictive performance of a binary classifier
Cross validation	The use of data splitting to estimate predictive performance
Data-generating model	The true but usually unknown generating mechanism for the available data
Divergence	The difference between the expected losses of the true and an approximating model
Full model	A model that includes all available predictors
Hyperparameter	A parameter that governs model fitting or a parameter of a prior distribution
Information criterion	A within-sample estimator of Kullback–Leibler divergence
Information-theoretic	Based on the principles of information theory
Kullback–Leibler divergence	Divergence based on log-density (information) loss
LASSO	a type of penalized regression that can lead to rejection of weak predictors
Loss function	A function that numerically quantifies the predictive performance of a model
MCC	a confusion-matrix metric
Metric	(Of a confusion matrix) a statistic based on the entries of a confusion matrix
Objective function	A function that is optimized when fitting a model (e.g., log-likelihood)
One-standard error rule	A selection rule calibrated by score-estimation uncertainty
Overfitting	The inclusion of spurious predictors in a model, leading to imprecision in predictions
Penalisation	Regularization via addition of a complexity penalty to the objective function
Regularization	A statistical method to control/constrain the effective complexity of a model
Score	Out-of-sample loss, averaged over test and training data
Tuning	Using test data to determine the value of a hyperparameter (e.g., using cross validation)
TSS	a confusion-matrix metric
Underfitting	The failure to include important predictors in a selected model, leading to bias

Abbreviations: LASSO, least absolute shrinkage and selection operator; MCC, Matthew's correlation coefficient; TSS, true skill statistic.

performance (score) of each model, where performance is quantified by a chosen loss function and estimated using cross validation, or in some cases, information criteria. Although selection is based on the estimated model scores, choosing the model with the best score can lead to overfitting due to score-estimation uncertainty; thus, selection rules which account for estimation uncertainty can be applied to identify the simplest model with comparable predictive performance. The predictive model-selection process is summarized in Figure 1.

MODEL SCORES

Fundamental to predictive model selection is the a priori selection of a suitable predictive measure called a score which forms the basis for model assessments and comparisons. Starting with the notion of a loss function, this section reviews the definitions and theoretical properties of some commonly used scores for both regression and classification problems. We make recommendations for score selection according to the modeling context.

Loss functions, discrepancies, and scores

A loss function L is used to quantify predictive performance. In the regression setting, two commonly used functions are:

$$L(y,\widehat{y}) = \|y-\widehat{y}\|^2 \qquad \text{(squared error)}$$
$$L(y,\widehat{y}) = \log p(y|\widehat{y}) \qquad \text{(log likelihood)},$$

where y and $\widehat{y}$ denote vectors of observed and fitted responses, respectively, $\|\ \|^2$ is the L_2-norm (the "sum of squares") and p is the model likelihood, if available.

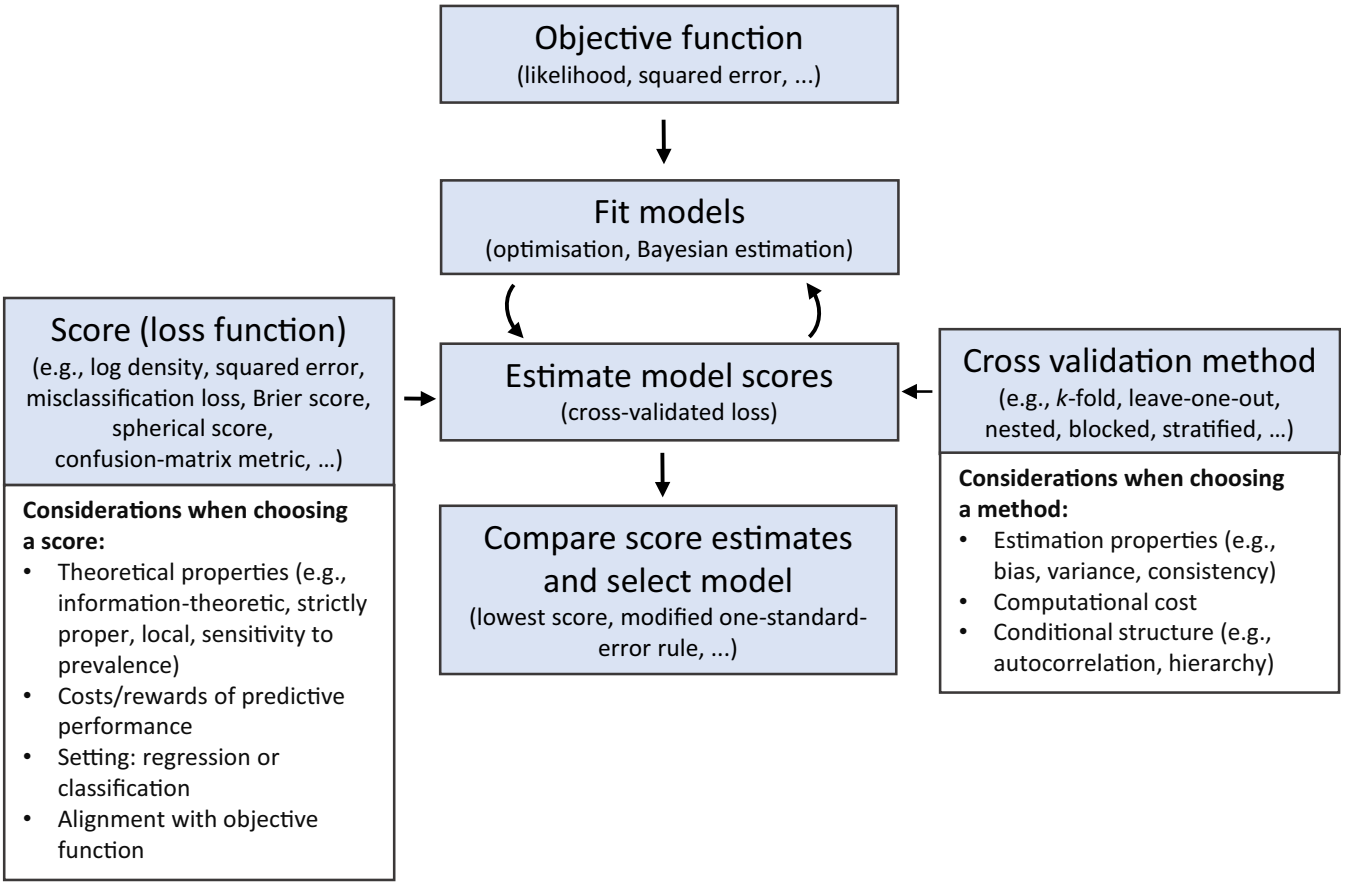

FIGURE 1 Overview of the model fitting and model selection process. Given data and a set of candidate models, the process proceeds from fitting and score estimation, to comparison and selection. Models are fit with respect to a chosen objective function, and score estimation is based on the selection of both a loss function and an estimation method—application of the latter usually requires multiple fits of each model. Statistical summaries of the estimated scores (e.g., means, standard deviations, and correlations) are used in conjunction with a chosen selection rule to compare and possibly select a preferred model.

Associated with a given loss function is a divergence and a score, defined as follows:

$$\textbf{Loss} \qquad L_m = L(y, \hat{y}_m) \tag{1}$$

$$\textbf{Divergence} \qquad D_m = E_y[L(y, y_{m^*})] - E_y[L(y, \hat{y}_m)] \tag{2}$$

$$\textbf{Score} \qquad S_m = E_x E_y\left[L\left(y, \hat{y}_{m(x)}\right)\right], \tag{3}$$

where $m = 1, ..., M$ indexes candidate models, m^* is the true (but usually unknown) data-generating model, and all expectations are taken over the distribution of the (multi-dimensional) data. The score (3) is the expected divergence (up to an additive constant) or the double-expected loss, averaging over the randomness in both the training data x and the representative test data y. The term "score" does not have a common definition across the statistical literature. In the context of predictive assessment, the loss function is often called the scoring rule (Gneiting & Raftery, 2007), while estimates of the double-expected loss are commonly called scores, such as AIC scores (Hastie et al., 2009) or cross-validation scores (James et al., 2013; Tredennick et al., 2021). Here we use the term to denote both the abstract and the estimated double expected loss. When L is log likelihood, the corresponding divergence is the Kullback–Leibler divergence, and the score (relative expected Kullback–Leibler divergence) is the usual quantity estimated in information-theoretic model selection (e.g., AIC or cross validation).

The first term in the divergence is an unknown, but fixed constant common to all models of the same data, and can therefore be ignored when assessing the relative divergence of candidate models. Despite this simplification, it is generally not possible to estimate D_m, since the data-generating model is unknown. In practice, given a finite data set, the double-expected loss (i.e., the score) is much more amenable to statistical analysis (Hastie et al., 2009). For this reason it is the score, not the divergence, that forms the basis of model selection.

Commonly-used scores and their properties

What options do we have for loss functions and their associated scores? Moreover, how do we make an appropriate choice? The choice of score must reflect the modeling problem at hand, taking account of the cost and benefits of differing predictive performance as well as the objective functions used to fit candidate models (Gneiting, 2011; Vehtari et al., 2017). For likelihood-based estimation, including both optimisation and posterior densities, log likelihood is a common choice as it is information-theoretic and it coincides with the objective function of model fitting (i.e., [maximum] log likelihood). Squared error coincides with the objective of least-squares regression and it is equivalent to log likelihood for homoscedastic Gaussian models. For classification problems, common choices are log loss, the Brier score, the spherical score, and misclassification (zero-one) loss (Gneiting & Raftery, 2007), in addition to metrics derived from the entries of the confusion matrix, such as κ, F_1, Matthew's correlation coefficient (MCC), true skill statistic (TSS), and many more (Allouche et al., 2006; Chicco et al., 2021) (see Table 2 for a summary of some common scores and Box 2 for definitions and an overview of confusion matrices and associated metrics).

In addition to alignment with the objective function, other considerations when choosing a score are the theoretical properties it possesses; the most important of these being propriety and locality (Bernardo, 1979). A score is proper if it is optimized by the true data-generating distribution, and it is strictly proper if it is uniquely optimized by this distribution; thus proper scores reward predictions which are closer to the "truth". A score is local if it depends only on the predicted density at the observed response (i.e., $p(y\,|\,\widehat{y})$).

The class of scores associated with the Bregman divergences (Zhang, 2008), which includes log likelihood, squared error, and misclassification loss are all proper scores; however, the latter is neither strictly proper nor local (Gneiting & Raftery, 2007). The non-locality, or distance insensitivity, of misclassification loss is easily demonstrated for a binary response; for example, the predicted probabilities $p = 0.6$ and $p = 0.9$ attain the same score (i.e., 0, no loss) for the response $y = 1$, using the classification threshold $c = 0.5$, despite the latter being much closer in probability. Notably, the mean absolute error $|y - \widehat{y}|$ is not a proper score (see Gneiting and Raftery (2007) for an example).

Recommendations

For likelihood-based regression models, log likelihood is the recommended loss function because it is strictly proper, information-theoretic, and accommodates a broad class of modeling structures. For classification problems, log loss is recommended when the properties of strict propriety and locality are deemed important, otherwise MCC or TSS are examples of general-purpose metrics to be used when all of the entries in the confusion matrix entries are needed to characterize the costs and rewards of predictive performance (see Box 2). These two metrics are identified because they are not sensitive to class prevalence; however, there are a plethora of options available, including simple misclassification loss, and we recommend further investigation for specific applications (Luque et al., 2019).

TABLE 2 Summary of common scores and associated loss functions.

Setting	Name of score	Loss function[a]	Propriety	Locality
Regression	log likelihood	$\log p(y\,\|\,\widehat{y})$	Strictly proper	Local
	(Mean) squared error	$\|\|y-\widehat{y}\|\|^2$	Strictly proper	Not local[b]
	(Mean) absolute error	$\|\|y-\widehat{y}\|\|^1$	Not proper	Not local
Classification[c]	log loss	$\log p_j$	Strictly proper	Local
	Brier (quadratic)	$\sum_{k\in\Omega}(I(j=k)-p_k)^2$	Strictly proper	Not local[d]
	Spherical	$p_j/\sum_{k\in\Omega}(p_k)^2$	Strictly proper	Not local[d]
	Misclassification loss[e]	$I(y\neq\widehat{y}_c)$	Proper	Not local
	Confusion-matrix metric[e,f]	$f(M_c)$	Not proper	Not local

[a]Except for $f(M_c)$, loss functions for classification are defined for a single datum.
[b]Local for homoskedastic Guassian errors.
[c]$p_k = p(k|\widehat{y})$ denotes the predictive probability of class $k\in\Omega$, where Ω indexes all possible classes. $p_j = p(y\,|\,\widehat{y})$ is the predictive probability of the observed class $y = j$. $I(x)$ is the indicator function, returning 1 if x is true and 0 otherwise.
[d]Local for $|\Omega| = 2$, that is, binary classification.
[e]The subscript c denotes the (tuneable) threshold for the binary case $\Omega = \{0, 1\}$, such that $\widehat{y}_c = I(p_1 > c)$.
[f]A summary of common metrics based on the confusion matrix $M = M_c$ is provided in Box 2.

BOX 2 Confusion matrices and metrics

Confusion matrix

A confusion matrix summarizes the predictive performance of a binary (two-class) classifier. Labeling one class positive and the other negative, the matrix entries are the counts of the four prediction outcomes: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

	True class		
	Positive	Negative	
Positive	TP	FP	Predicted class
Negative	FN	TN	

Tunable threshold

For models that predict the (positive) class probability p, rather than a dichotomous outcome, a threshold c is applied such that the predicted class is positive if $p > c$, else negative. The threshold can be treated as a hyperparameter of the model and tuned to maximize a selected metric using cross validation—nested cross validation should be used when model selection is preceded by hyperparameter tuning (see *Nested cross validation*).

Metrics

To use the confusion matrix for model comparison or parameter tuning it must be summarized as a scalar statistic or metric. Although confusion-matrix metrics are not usually strictly proper scores, they are a flexible class of scores which can be tailored to application-specific needs, such as accounting for class imbalance in the data and asymmetric cost weighting of the prediction outcomes (e.g., when a FN is more costly than a FP). The literature contains a plethora of existing metrics to choose from and we present here the definitions of some that are commonly used:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \qquad \kappa = \frac{2\times(\text{TP}\times\text{TN}-\text{FP}\times\text{FN})}{(\text{TP}+\text{FP})(\text{FP}+\text{TN})+(\text{TP}+\text{FN})(\text{FN}+\text{TN})}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad \text{TSS} = \text{sensitivity}+\text{specificity}-1$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \qquad \text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$$

$$F_1 = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}.$$

The true skill statistic (TSS, Allouche et al., 2006) and Matthew's correlation coefficient (MCC, Matthews, 1975) are particularly useful because they are not sensitive to class imbalance. Recent studies suggest that MCC is more truthful and informative than κ, F_1, accuracy, and even the strictly proper Brier score (Chicco et al., 2021).

Estimation using cross validation

Confusion-matrix metrics can be estimated using cross validation by populating the matrix with the aggregate test outcomes of a single k-fold iteration (see *Cross-validation techniques*). Repeated k-fold cross validation can be used to estimate the sampling variability of the metric where k must be less than n since leave-one-out has only one unique split. We use repeated 10-fold cross validation in the scat classification example to estimate MCC.

CROSS-VALIDATION TECHNIQUES

As summarized earlier, cross validation is the application of data splitting to estimate the predictive performance of one or more candidate models. Cross validation works by fitting each model to a subset of the available data (the training set) and then comparing the models' predictive capacities (loss) on the remaining portion of the data (the test set). To improve the estimate, the splitting procedure is usually iterated by systematically selecting different subsets of data and summarizing the overall predictive performance across iterations (Arlot & Celisse, 2010). Despite its conceptual simplicity, cross validation is a theoretically rigorous method to estimate the score (3) associated with a given loss function (1) (Zhang, 2008). This section reviews many common variants of cross validation, their associated options, and their relative merits in terms of statistical properties and computational costs. Recommendations are given according the modeling context and available computational resources.

Data-splitting schemes

There are many variants of cross validation, each characterized by different data-splitting schemes. A commonly used scheme is k-fold, where the available data is split into k (approximately) equal-sized subsets (i.e., "folds") which generates k distinct pairs of training/test sets, obtained by removing one fold at a time (the test set) from the full data set. This scheme can be applied once, for a single initial split (ordinary k-fold), or it can be repeated many times for different splits (repeated k-fold). The score estimate for a single k-fold cross validation is

$$\widehat{S}_k = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, \widehat{y}_i^{-j[i]}\right), \tag{4}$$

where $i = 1, \ldots, n$ indexes the data points, $j = 1, \ldots, k$ indexes the folds, and the superscript $-j[i]$ indicates that the training set for the fitted value excluded the fold containing the ith data point.

An alternative to k-fold is leave-d-out, which involves the repeated removal of d (randomly selected) test points. For a sufficiently large number of iterations, the mean leave-d-out estimate approaches the repeated k-fold estimate for $d \approx n/k$; however, k-fold is preferable as it guarantees of balanced draw of samples and has lower variance (see Box 3) (Burman, 1989). An important limiting case of both k-fold and leave-d-out is leave-one-out, which is equivalent to k-fold for $k = n$.

Ordinary, stratified, and blocked cross validation

For a given splitting structure, the assignment of data points to test and training sets can be uniformly random, which is typical, or it can depend on the values of one or more categorical variables (stratified cross validation), or the assignment can be determined by the spatial or temporal "distance" between training and test points (blocked cross validation).

Stratified cross validation is generally applied to latent-variable (or random-effects) models to balance group membership (i.e., the proportion of data within each group level) across training sets or to leave-one-group-out of the training set entirely. These two alternatives correspond to conditional-likelihood (i.e., prediction to existing group levels where latent effects are treated as [regularized] model parameters) and marginal-likelihood (i.e., prediction to new group levels where latent effects are "integrated out") foci, respectively (Fang, 2011; Merkle et al., 2019). We explore these two foci further in the Pinfish example (see *Examples*), where we apply both leave-one-out and leave-one-group-out to a set of non-linear hierarchical models.

Blocked cross validation omits training points within a certain "distance" of the test data to account for spatial or temporal autocorrelation. It is important to use blocked cross validation when spatial and/or temporal correlation is visible in the model residuals (e.g., in a spatial correlogram) (Fletcher & Fortin, 2018; Roberts et al., 2017). These structured data-assignment schemes ensure that the test data are conditionally—with respect to the model—independent (or at least conditionally exchangeable), a pre-requisite for cross-validated score estimates to be unbiased (Gelman et al., 2013; Milà et al., 2022).

Bias-corrected cross validation

Cross-validated score estimates are generally biased upward (i.e., the expected loss is overestimated) due to the training set being necessarily smaller than the full data set, but the bias is easily corrected (Arlot & Celisse, 2010). The bias reduces as the size of training set increases, such that $k = n$ (i.e., leave-one-out) has the minimum bias of all k values in k-fold cross validation. Indeed, the bias of leave-one-out is usually negligible and for $k \geq 10$, the bias is often small enough that correction is not needed (Hastie et al., 2009). It is advisable, when possible, to check this assertion by computing and comparing the bias correction for the least- and most- complex models.

A bias-correction term can be estimated, without any additional model fits, using the method of Burman

BOX 3 Bias-variance trade-off

Choosing the best predictive model: In model selection, the problem of finding the "best" predictive model is often viewed as a bias-variance trade-off: to find the "sweet spot" between underfitting (i.e., simple models with high bias and low variance) and overfitting (i.e., complex models with low bias and high variance) (Hastie et al., 2009). For squared-error loss, this trade-off is made explicit by the decomposition:

$$\begin{aligned}\mathrm{E}\left[\left(\widehat{y}_m - y_{m^*}\right)^2\right] &= \left(\mathrm{E}[y_{m^*}] - \mathrm{E}[\widehat{y}_m]\right)^2 + \mathrm{Var}[y_{m^*} - \widehat{y}_m] + \mathrm{Var}[\varepsilon] \\ &= \text{bias}^2 + \text{predictive variance} + \text{irreducible error}.\end{aligned}$$

This decomposition reveals why, for squared-error loss, the model closest to truth (i.e., with minimal bias) is not necessarily the best predictive model (Shmueli, 2010). Using a strictly proper score, however, such as log likelihood (see *Commonly used scores and their properties*), does at least guarantee that the true data-generating model uniquely attains the optimal true score. However, in real-world applications all models are inevitably misspecified and scores are estimated using finite data. Thus, when the goal is hypothesis testing, care must be taken to specify models based on causal hypotheses (lest correlated, but non-causal variables be selected), and to account for score-estimation uncertainty when making selection decisions (see *Mitigating overfitting using calibrated selection rules*).

Choosing k in k-fold cross validation: It is sometimes claimed that there is a bias-variance trade-off when selecting the value of k in k-fold cross validation (James et al., 2013), however the statistical literature tells a more nuanced story. Although the bias of k-fold cross validation as a score estimator is always reduced by increasing k, it is difficult to make universal statements about a bias-variance trade-off because the effect of k on the variance depends on the estimation setting (e.g., objective function or score choice) as well as the stability of the training algorithm (e.g., model sensitivity) (Arlot & Celisse, 2010). For example, $k = n$ is known to have the lowest bias and the lowest variance of all k values in linear regression (Burman, 1989)–this analytic (and asymptotic) result is conjectured to be true in other regression settings (Arlot & Celisse, 2010). In contrast, on the basis of simulation studies, it is often recommended to use a much smaller $k = 5$ or $k = 10$, especially for classification problems (Cawley & Talbot, 2010; Hastie et al., 2009).

Yet even when conclusive statements can be made about the relationship between the variance and the choice of k, the variance of the score is not necessarily the correct quantity to examine (Breiman & Spector, 1992). Arlot and Lerasle (2016) suggest that the variance of the score differences $\mathrm{Var}(S_{k,m} - S_{k,m'})$ is a more important quantity (to minimize) since model selection ultimately concerns model comparisons not score estimation per se. For least-squares density estimation, Arlot and Lerasle (2016) show that $\mathrm{Var}(S_{k,m} - S_{k,m'})$ reduces with k, however close-to-optimal values are attained for $k \geq 10$, affirming the existing advice (for this setting at least) to choose $k = 10$ as a *minimum* value when computational cost prohibits the use of leave-one-out.

(1989). For k-fold cross validation, the pointwise bias correction ν_i is the difference between the loss value for the model trained on the full data set and the average loss of the predictions from the k training folds:

$$\nu_i = L(y_i, \widehat{y}_i) - \frac{1}{k}\sum_{j=1}^{k} L\left(y_i, \widehat{y}_i^{-j}\right). \tag{5}$$

The bias-corrected score estimate is $\widehat{S}_{k^*} = \widehat{S}_k + \nu$, where $\nu = \frac{1}{n}\sum \nu_i$. Using (4) and (5), $\widehat{S}_{k^*}$ can be expressed as a pointwise sum which facilitates bias-corrected estimation of the score variance for use with calibrated model selection (see *Model selection* for details). For log-density loss, $\widehat{S}_{k^*}$ can be used to estimate the bias-corrected effective number of parameters (see *Examples*). For scores which are not evaluated pointwise (e.g., those based on the confusion matrix), it is generally not possible to estimate a bias-correction term using (5). In these cases, bias can be managed by selection of k, for example, $k > 10$.

Bias correction is important in cross validation because score-estimation bias increases with model complexity and more complex models can be over-penalized.

Failure to correct for this complexity-dependant bias results in a complexity penalty which undermines the interpretation of the selected loss function and the associated divergence and score. For example, model comparisons based on predictive log likelihood are a poor approximation to information-theoretic comparisons if the expected loss estimates have a complexity-dependent bias. Although the inclusion of complexity-dependent bias is sometimes used as a strategy to mitigate overfitting (Cawley & Talbot, 2010), we recommend using minimally biased estimates, to retain score interpretation, instead applying calibrated selection rules to account for the predominant cause of overfitting: score-estimation uncertainty (see *Model Selection* for further discussion).

Nested cross validation

In addition to discrete model selection, cross validation is commonly used to tune continuous model hyperparameters such as regularization parameters (see *Regularization techniques*) (Hastie et al., 2009). For this use, the cross-validated score estimate is computed across a range of fixed values of the hyperparameter to determine the value which optimizes the estimate. Calibrated selection rules can also be used to select the hyperparameter value (see *Mitigating overfitting using calibrated selection rules*). When using cross validation for both model selection and tuning, hyperparameters must be tuned separately for each training set using an additional inner layer of cross validation; this is called nested cross validation (Cawley & Talbot, 2010). Although it can be computationally expensive, nested cross validation is necessary to mitigate overfitting and place all models on an equal footing. We provide an illustration of nested cross validation in the scat classification example.

Approximate cross validation

Approximate cross-validation techniques provide an alternative to data splitting (with repeats) and multiple model fits, instead typically requiring just a single fit using all of the available data. Most approximate cross-validation methods express the leave-one-out score estimate as a weighted sum of the pointwise loss values for an initial fit, where the weights are a function of the estimated leverage h_i of each data point. For linear regression with squared error loss, the leverage can be computed analytically from the design matrix X (for which the element X_{ij} is the value of the jth predictor [or the indicator value of a corresponding factor level] for the ith response), permitting exact computation of the leave-one-out estimate for mean squared error (Davison & Hinkley, 1997):

$$\widehat{S}_{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{1 - H_{ii}}\right)^2, \tag{6}$$

where H_{ii} are the diagonal elements of the hat matrix $H = X(X^TX)^{-1}X^T$ (calculated in R using the function `hatvalues`). An useful variant of (6), called generalized cross validation (Craven & Wahba, 1978; Wood, 2017), is obtained by replacing H_{ii} with its average value $\frac{1}{n}\mathrm{trace}(H)$, which is generally faster to compute than the individual values. The exact method (6) has been extended as an approximate method to the loss functions associated the entire class of Bregman divergences (e.g., log likelihood; see *Commonly-used scores and their properties*) Zhang (2008); however, we are unaware of any software implementation of these approximate formulas.

In a Bayesian setting, the leave-one-out estimate of the predictive log likelihood can be approximated using smoothing techniques (Vehtari et al., 2017). In many instances, this estimate can be computed from a single set of Markov Chain Monte Carlo samples, however additional sets of samples may be needed if there are points of high leverage. The advantages of using this technique include: flexibility of application to non-linear, multi-response and hierarchical models, low computational cost, and ease of implementation using the R package `loo` (Vehtari et al., 2017).

Choosing a cross-validation technique

When choosing a cross-validation technique, there is an inherent trade-off between the statistical properties, for example, the bias and variance, and the computational cost. Here we organize recommendations and considerations for choosing cross-validation according the dependency structure of the data. An overview of the cross-validation techniques presented in this paper is provided in Table 3.

Data are conditionally independent

In terms of statistical properties, leave-one-out cross validation is generally the gold standard (see Box 3), and in many instances it is practical to compute leave-one-out exactly by fitting the model n times—computation time can be greatly reduced by using a parallel implementation on a modern multicore processor.

TABLE 3 Summary of statistical results and recommendations for common cross-validation techniques. Further details and references are provided in the main text.

Data-splitting scheme	Estimation method	Recommendations and statistical results
k-fold	Biased	For log-density regression, analytic results suggest $k \geq 10$ is required to ensure that the bias and variance are close to optimal. Similar advice, based on simulation studies, is given for classification problems (see Box 3).
	Bias-corrected	Bias-corrected estimates are recommended for $k < 10$ and blocked cross validation, when many test data are omitted. Calculation of the corrections does not increase computational cost.
	Repeated	Useful for estimating sampling variability when score estimates do not factor into a pointwise sum (e.g., confusion-matrix metrics). *k*-fold cross validation repeated *R* times has higher bias and variance than a single iteration of $(R \times k)$-fold cross validation.
	Stratified	Used for grouped data to balance group membership across folds in *k*-fold cross validation. Not applicable or meaningful to leave-one-out.
Leave-one-out	Exact	The preferred method when computationally achievable (and is asymptotically equivalent to AIC). Bias is neglible. For linear regression, a computational shortcut permits an analytic solution with a single model fit.
	Approximate	An excellent alternative to exact leave-one-out, especially for slow-fitting models. Various methods exist, some of which provide diagnostics to assess validity of the approximated estimate.
Leave-*d*-out	$d \simeq \frac{n}{K}$	Equivalent to *k*-fold for a large number of repetitions, but otherwise, unlike *k*-fold, it does not guarantee a balanced draw of test samples. Computationally inefficient.
	$d > \frac{n}{2}$	Used for asymptotic selection of the true data-generating model (consistent selection); a suggested value is $d_c = n - n/(\log(n) - 1)$ (Shao, 1993). Consistent selectors such as leave-d_c-out and BIC are not appropriate in ecology as the true model is almost certainly not in the candidate set
Blocked		Recommended when model residuals are spatially or temporally autocorrelated. Block size depends on the strength of correlation as a function of distance or time. Bias correction is sometimes required for large blocks.
Leave-one-group-out		Used for grouped data to assess model performance by predictive performance to new group levels. More than one group can be left out to reduce the no. model fits.

Abbreviation: AIC, Akaike's information criterion.

If leave-one-out is too slow, *k*-fold or approximate leave-one-out are necessary. When an approximate leave-one-out method exists for the types of models being fitted and the chosen score, then this option is usually the fastest and requires no further bias correction. Otherwise, *k*-fold can be used. In terms of minimizing bias and variance, it is better use of computational resources to perform a single estimate, setting *k* as large as possible, rather than averaging repeated estimates using smaller *k* values (Burman, 1989). For $k < 10$, bias correction is usually required.

Data have structural dependencies

When the model residuals are not independent due to temporal, spatial, or other grouping structure in the data, an appropriately structured data-splitting scheme should be selected; for example, blocked or leave-one-group-out. Structured residuals can arise because the model has failed to adequately capture non-independence of data, or because predictions are being extrapolated to new regions or group levels.

Unlike ordinary leave-one-out, the blocked version will need bias correction if the number of case deletions constituting the left-out block is large (Burman et al., 1994); for example, if deletions exceed 10% of the data. If blocked leave-one-out or leave-one-group-out are too slow to compute, alternative options include blocked *k*-fold, or leaving out more than one group out at a time; detailed recommendations for the use of blocked cross validation in ecology can found in the review by Roberts et al. (2017).

MODEL SELECTION

The selection of the best-performing model from a set of candidates is often framed as a bias-variance trade-off with respect to model complexity (see Box 3

for elaboration), where the optimal trade-off is achieved asymptotically by selecting the model with the best-estimated predictive score (Hastie et al., 2009). However, simply selecting the model with the best score is sub-optimal due to overfitting resulting from score-estimation uncertainty (Yates et al., 2021). Overfitting is a serious concern for all modeling goals since the inclusion of spurious effects leads to false discovery in exploration, false confirmation in hypothesis testing, and degrades the performance of predictive models. The first part of this section reviews techniques to mitigate overfitting by taking account of score-estimation uncertainty. The second part reviews the notion of regularization, wherein cross validation is often used to tune model complexity in a continuous manner, via shrinkage, rather than choosing among discrete alternative models.

Mitigating overfitting using calibrated selection rules

Overfitting due to estimation uncertainty in predictive scores occurs because of the asymmetry in the predictive cost of excluding an important variable versus including a spurious one (Tredennick et al., 2021). Including a spurious variable will degrade predictive performance, however the resulting score difference is usually small relative to the estimation uncertainty of the score differences; thus, an overfitted model is often selected due to random variation alone (Yates et al., 2021). On the other hand, excluding an important variable usually results in a score difference that is large relative to estimation uncertainty, which makes underfitting much less probable than overfitting.

An effective strategy to mitigate overfitting due to estimation uncertainty is to select the least-complex model whose performance is deemed "comparable" to best scoring model (Piironen, Paasiniemi, & Vehtari, 2020). To quantify the notion of comparable, the set of cross-validated loss values (e.g., the pointwise losses using leave-one-out) can be used to estimate the sampling variability of the score estimates which, in turn, can be used to visualize and suggest nominal performance thresholds for the possible selection of a simpler model.

The original one-standard-error rule, proposed by Breiman et al. (1984), is an example of this approach, where the performance threshold is simply the standard error of the best score, σ_{best}; that is, a simpler model is selected when its score estimate is within σ_{best} of the best score. However, this threshold can be problematic because it fails to account for the covariance of the scores (e.g., all models will predict poorly to a given outlier). This typically results in overestimating the relative variation between models, which can lead to underfitting.

Our recent modification of the one-standard-error rule addresses this issue by using the pairwise correlation coefficient of the best score with each alternative score (denoted $\rho_{\text{best},m}$) to define the following correlation-adjusted standard errors, indexed by model m (Yates et al., 2021):

$$\sigma_m^{\text{adj}} \equiv \sigma_{\text{best}} \sqrt{1 - \rho_{\text{best},m}}. \tag{7}$$

To visualize and apply the modified one-standard-error (m-OSE) rule, σ_m^{adj} is added as an error bar to a plot of the score estimates of each model, where $\sigma_{\text{best}}^{\text{adj}} = 0$ since $\rho_{\text{best,best}} = 1$. The model selected by the rule is the least-complex model whose adjusted error interval includes the estimate of the best-scoring model (e.g., see Figures 2, 5a,b). The variance terms can be estimated using a normal approximation of the distribution of the loss values arising from cross-validation folds, pointwise log-likelihood values, or non-parametric bootstrap samples. The definition (7) is derived so that $\sigma_m^{\text{adj}} = \sigma_{\text{best}}$ when model scores are independent (i.e., the original one-standard-error rule), and $\sigma_m^{\text{adj}} = 0$ when model scores are maximally correlated (i.e., the best-scoring model is selected). In some instances, the model with the best score coincides with the model selected by the modified rule, providing assurance in such cases that the best-scoring model is not overfit. We illustrate the use of the modified one-standard-error rule in both of the examples herein.

Another measure of estimation uncertainty that appears in the recent model-selection literature is the standard error of the difference between the best score and each alternative model (Piironen, Paasiniemi, & Vehtari, 2020):

$$\begin{aligned}\sigma_m^{\text{diff}} &= \sqrt{\text{Var}(\Delta S_m)} \\ &= \sqrt{\sigma_m^2 + \sigma_{\text{best}}^2 - 2\rho_{\text{best},m}\sigma_m\sigma_{\text{best}}},\end{aligned} \tag{8}$$

where $\Delta S_m = S_{\text{best}} - S_m$. Estimates of ΔS_m sample the null-hypothesis distribution that the mth model does not improve upon the best-scoring model; thus, σ_m^{diff} can be used to calculate probabilities related to pairwise model comparisons.

The correlation-adjusted σ_m^{adj} is closely related to σ_m^{diff}. Indeed, the former is obtained from a generalization of the latter, subject to the aforementioned selection conditions for $\rho = 0$ and 1. When used as a performance threshold in a modified one-standard-error rule, the two

thresholds will often select the same model, since both account for correlation in the score estimates (see Yates et al. (2021) for further discussion). When the data set is small ($n < 100$) or performance estimates of the best models are similar (i.e., those with $\Delta\widehat{S} < 4$), Sivula et al. (2020) have shown that both the bootstrap and the normal approximation provide unreliable estimates of the standard error for log-likelihood loss. However, in the context of mitigating overfitting, these estimation issues are most likely benign, because the simpler model is favored as a consequence.

For hypothesis generation and predictive goals, estimates of the correlation of model scores can also be used to assess the merits of model enlargement (Garthwaite & Mubwandarikwa, 2010; Gelman et al., 2013). For example, when a subset of the best-scoring models have similar score estimates but low correlation, then improved predictive performance may be obtained using model averaging of the predictions (not parameters) (Dormann et al., 2018). It is also possible that each of these best-scoring models captures independent aspects of the true data-generating process, which could inform future theoretical developments and subsequent hypothesis-driven studies.

Estimated performance thresholds such as σ_m^{adj} are a useful guide for selecting a preferred model for the purpose of mitigating overfitting, but we caution against their indiscriminate use without appropriate model checking. Laudable though the goal of using data-driven processes to make objective selection decisions is, we reiterate the perspective expressed by Vehtari et al. (2019) "that we need to abandon the idea that there is a device that will produce a single-number decision rule". We suggest that model-selection analyses be published with a summary of the score estimates for all candidate models, including their variation and covariation as described in this section; this gives readers the opportunity to directly interpret performance comparisons and selection decisions.

Regularization techniques

Regularization includes a large suite of statistical methods for controlling/constraining the complexity of a model. Indeed, the use of calibrated selection rules, and variable-selection approaches more generally, can be viewed as a type of discrete regularization, where a subset of parameters in the global model are set to zero, thereby reducing model complexity. This discrete approach can be contrasted with continuous regularization techniques that use penalized regression such as the least absolute shrinkage and selection operator (LASSO, Zou & Hastie, 2005), non-uniform or sparsifying priors for Bayesian approaches (Piironen & Vehtari, 2017b; van Erp et al., 2019), or the inclusion of hierarchical structures. All of these techniques constrain parameter estimates, shrinking them toward zero (or a distributional mean), but without necessarily removing them from the (global) model. For a detailed summary of regularization methods and their interpretation, see Hooten and Hobbs (2015).

For continuous approaches, the effective reduction in complexity is governed by one or more regularization parameters, which can be estimated from the data—Bayesian priors are a natural exception, unless adopting the empirical Bayes approach. For hierarchical approaches, the associated regularization parameters are implicitly estimated given that they are a function of the estimated scale parameters. In penalized regression, cross validation is typically used to tune the regularization parameters for optimal predictive performance with respect to a chosen score (see *Nested cross validation*). For example, in elastic net regression, the regularized parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_p)$ are those minimizing the penalized function (Zou & Hastie, 2005):

$$f\left(\mathbf{x}; \widehat{\boldsymbol{\theta}}\right) + \lambda\left(\alpha\|\widehat{\boldsymbol{\theta}}\|^1 + (1-\alpha)\|\widehat{\boldsymbol{\theta}}\|^2\right), \tag{9}$$

where f is the objective function (usually mean squared error or negative log likelihood), and $\|\boldsymbol{\theta}\|^1 = \sum_{j=1}^{p} |\theta_j|$ and $\|\boldsymbol{\theta}\|^2 = \sum_{j=1}^{p} \theta_j^2$ are the L_1- and L_2-norm of the vector of model parameters, respectively. The regularization parameter λ determines the strength of the penalty, enforcing a trade-off between the size of the model's parameter estimates (the shrinkage or effective complexity) and the minimized value of the unconstrained objective function f. The hyperparameter α ($0 \le \alpha \le 1$) indexes a family of regularized models for which the extreme values $\alpha = 1$ and $\alpha = 0$ correspond to LASSO and ridge regression, respectively. Both α and λ can be tuned using cross validation (e.g., using a two-dimensional grid of candidate (α, λ) values). A further useful property of elastic net regression (in addition to the regularization of model complexity) is its remarkable tolerance to the inclusion of correlated variables via the parameter-grouping effect of the penalization strategy, which alleviates the deleterious effects of multicollinearity on estimation (Dormann et al., 2013; Hastie et al., 2015). We illustrate the use of LASSO, ridge, and elastic net regularization in the first example.

Some regularization techniques such as the LASSO and sparsifying priors can be viewed as both continuous—in the sense of providing constrained parameter estimation—and discrete—in that a subset of the model parameters can be shrunk all the way to zero, or at least very close to zero,

leading ultimately to their removal. A recently developed method that makes use of both continuous regularization and discrete selection is Bayesian projective inference (Piironen, Paasiniemi, & Vehtari, 2020). The method uses a global reference model (usually the full model) to simulate new data for the purposes of both model selection and to project the uncertainty of the global posterior on to the subspace of a selected submodel. Advantages of using a reference model in lieu of fitting submodels directly to the original data include a reduction in the variance of model selection, improved calibration of post-selection predictive uncertainties, and resilience to overfitting in variable-selection problems where the number of predictors exceeds the number of data points. We illustrate the use of Bayesian projective inference in the first example.

EXAMPLES

In this section we use two biological case studies to illustrate many of the cross-validation-based model-selection approaches described in this paper. To accompany these analyses, we provide an online code repository for reproducing the complete workflow, including: data preparation, model fitting, model selection, and plot creation (see *Data availability statement*). The code provides a template which other users can readily adapt and customize to use cross validation in their statistical learning problems.

Scat classification

Data on animal feces in coastal California is recorded in the scat data set available via the R package **`caret`** (Kuhn, 2022). The data consist of DNA-verified species designations as well as fields related to the time and place of the collection and the morphology of the scat itself. The aim of the analysis is to predict the biological family (felid or canid) for each scat observation, based on eight morphological characteristics, scat location, and carbon-to-nitrogen ratio; see Yates et al. (2022) and the original data publication (Reid, 2015) for further details concerning the data.

Part 1: Logistic regression with MCC score

We begin in a parametric setting, using logistic regression to model the binary class probabilities. We fit generalized linear models using maximum likelihood, but compare model performance to select variables using estimates of

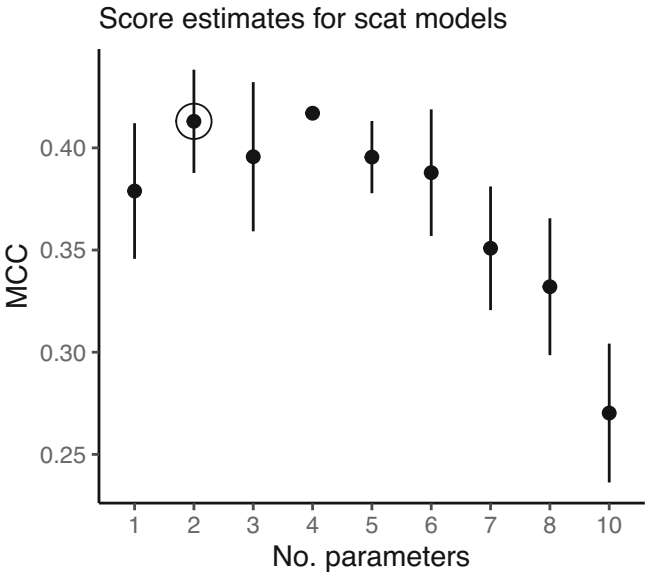

FIGURE 2 Comparison of logistic models using repeated 10-fold cross validation to estimate Matthew's correlation coefficient (MCC). The dots and bars depict the mean MCC estimate and the modified standard error (7), respectively. After applying the modified one-standard-error rule, the selected model comprises two predictors: Carbon–nitrogen ratio and the number of scat pieces.

MCC, as described in Box 2. Setting the probability threshold equal to the prevalence (felid: 0.514) (Liu et al., 2005), we use 10-fold cross validation, repeated 50 times, to generate a MCC estimate for each model for each repetition. Figure 2 shows the mean MCC estimate and modified standard error (7) of the highest scoring model for each level of model complexity using all combinations of the 10 predictors (1024 models in total). To mitigate overfitting, we applied the modified one-standard-error rule which suggests selection of the two-parameter model comprising the predictors carbon–nitrogen ratio and the number of scat pieces; the one-predictor model comprising only carbon–nitrogen ratio is close in score and should also be considered. The mean model scores and estimation uncertainty for the top 10% of all the models are shown in Appendix S1: Figure S1. The model fitting took 220 s using 40 Intel i7 cores.

As an alternative to selection among a discrete set of models, we illustrate the use of penalized regression (9), applying LASSO ($\alpha = 1$) and ridge ($\alpha = 0$) regularization to the global logistic model. In each case, we tune the regularization parameter λ using cross-validated MCC estimates with the same set of k-fold data splits used for the discrete approach. The cross-validation-selected LASSO model comprised just one (regularized) parameter associated with the predictor carbon–nitrogen ratio (Figure 3a,b). Despite its simplicity, the lasso model had a higher cross-validated MCC estimate than the

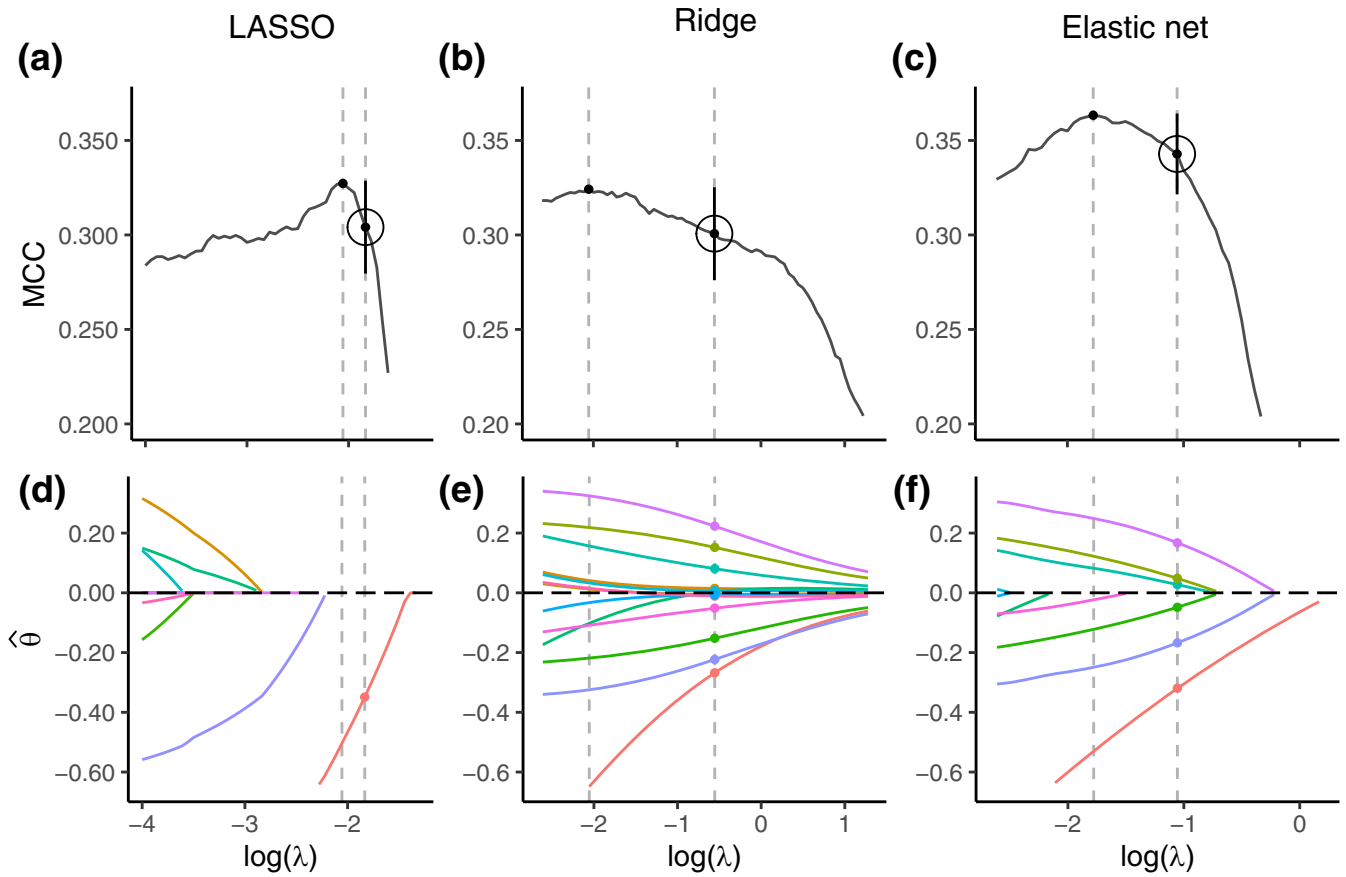

FIGURE 3 Penalized regression for scat classification models: least absolute shrinkage and selection operator (LASSO), ridge, elastic net regression. Plots (a), (b), and (c) show the cross-validated Matthew's correlation coefficient (MCC) estimates as a function of the logged regularization parameter λ; the circled point is the largest λ value (i.e., the most regularized model) within one standard error (7) of the value that maximizes MCC. Plots (d), (e), and (f) show the corresponding trajectories of the regularized estimates of the model parameters $\widehat{\theta}$, each denoted by a unique color (labels ommited for clarity). The colored points denote the final parameter estimates for the selected λ-values: For all cases, significant shrinkage is evident; for the LASSO, just one non-zero predictor is retained (carbon–nitrogen ratio); for elastic net, six non-zero predictors are retained).

ridge model which kept all 10 predictors (by construction), strongly regularizing all of the associated parameters (Figure 3b,e). We also applied elastic-net regularization to the scat classification problem for which the optimal value of the tuned hyperparameter was $\alpha = 0.175$ (see Appendix S1: Figure S2) which selected six predictors (Figure 3c,f). We used the R package **`glmnet`** to fit the penalized regression models (Friedman et al., 2010).

Part 2: Regularizing priors and Bayesian projective inference

In a Bayesian setting, ridge- and LASSO-type regression can be implemented using Gaussian and Laplacian (two-sided exponential) priors, respectively (Hastie et al., 2015). An alternative choice is the regularized horseshoe prior (Piironen & Vehtari, 2017a) which provides support for a proper subset of parameters to be far from zero. Using the **`brms`** package, we fit the global logistic scat model using: (1) Laplacian priors (i.e., LASSO), (2) regularized horseshoe priors, and (3) weakly-informative Gaussian priors (i.e., weak ridge-type regression). Using the **`loo`** package to estimate approximate leave-one-out cross-validation scores, the horseshoe variant was the best predictive model, followed by the LASSO ($\Delta\widehat{S} = 2.1$, $\sigma^{\text{diff}} = 0.9$), and the weakly-informative priors ($\Delta\widehat{S} = 7.9$, $\sigma^{\text{diff}} = 3.4$). Appendix S1: Figure S3 shows the posterior distributions for all 11 parameters for each of the three models; the strongly regularizing effect of the horseshoe prior is clearly visible, leaving only one variable (carbon–nitrogen ratio) with support far from zero.

To assess the merits of using a simpler submodel, we apply Bayesian projective inference (see *Regularization techniques* for details). Using the **`projpred`** package (Piironen, Paasiniemi, Catalina, et al., 2020) and taking as a reference model the fitted regularized horseshoe

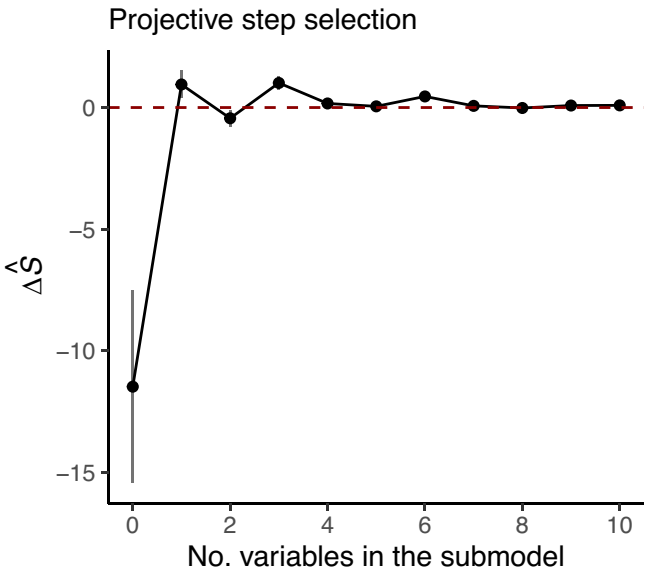

FIGURE 4 Bayesian projective inference applied to Bernoulli scat-classification models. The dots and bars depict the mean and standard error of the difference of the approximate leave-one-out log likelihood estimates $\Delta\widehat{S}$, with respect to the reference model (dashed line), for each increment of the projective step-selection process. The simplest submodel, with performance comparable to the reference model, contains only one-predictor: Carbon–nitrogen ratio.

model, we use approximate leave-one-out-based forward step selection to determine the optimal size of the alternative submodel and which variables to include. The selection results shown in Figure 4, suggest that the inclusion of just one variable (carbon–nitrogen ratio) has the best estimated predictive performance. As a further step (see Appendix S1: Figure S4), post-selection posterior densities can be generated by projection of the reference posterior onto the selected submodel.

Part 3: Nested cross validation to compare logistic and random forests models

Machine-learning methods are increasingly used for classification problems, but their non-parametric (or "black-box") nature makes comparison with parametric approaches difficult. Here we illustrate the use of cross validation to compare the classification skill of a discretely selected parametric modeling approach (as per part 1) to a tuned machine-learning random forest algorithm (an ensemble of decision trees) (Breiman, 2001). We use nested cross validation to completely separate model training, which includes variable selection and hyperparameter tuning, from the estimation of predictive performance. For the outer folds, we use 10-fold cross validation, repeated 50 times, and for the inner folds ordinary 10-fold cross validation. Thus, for each of the 500 outer training sets, 10 inner folds are used to select variables for the (discrete) logistic models, to tune the tree depth for the random forest models, and to tune the probability threshold for all models (see Box 2: *Tunable threshold*). Each repetition of the outer 10-fold cross validation generates a single MCC estimate for each selection approach. To reduce computation time, we set the number of trees in the random forest to the fixed value 500 and we included only the top 10% of the 1024 discrete logistic models based on the MCC scores estimated in part 1. The analysis took 280 s using 40 cores.

The tuned random-forest model performs better than the discretely-selected logistic model: the mean MCC score estimates are 0.25 and 0.375, respectively, and the modified standard error of their difference is 0.06. Here, to simplify the exposition, we used the linear logistic models introduced in Part 1, but a fairer comparison of these learning algorithms would permit the inclusion of non-linear and interaction terms in the logistic models. We further simplified the presentation by tuning only a single hyperparameter, however complex learning algorithms typically contain large numbers of hyperparameters for which tuning can be computationally expensive; for an overview of existing and emerging techniques for hyperparameter tuning or optimisation, see Feurer and Hutter (2019).

PINFISH GROWTH

Length, age and sex data for pinfish (*Lagodon rhomboides*) from Tampa Bay, are recorded in the pinfish data set available via the R package **`fishmethods`** (G. A. Nelson, 2022). The aim of the analysis is to determine which allometric growth function (see below) best describes the relationship between the length and age measurements of pinfish, while accounting for the effect of sex and haul on the model parameters (the data comprise measurements from 45 separate fishery hauls, see (G. Nelson, 2002) for details). We take as candidate growth models the following commonly used non-linear functions (Tjørve & Tjørve, 2010):

$$\begin{aligned} &\text{Gompertz (G)} && L_{\mathrm{G}}(a) = L_0 e^{-e^{(-K(a-t_0))}} \\ &\text{logistic (log)} && L_{\log}(a) = L_0/\left(1+e^{-K(a-t_0)}\right) \\ &\text{von Bertalanffy (vB)} && L_{\mathrm{vB}}(a) = L_0\left(1-e^{-K(a-t_0)}\right), \end{aligned} \quad (10)$$

where a is the age and L is the (modeled) length. For each function, the parameters L_0, K, and t_0 denote the length asymptote, the growth rate, and the initial length, respectively. From inspection of Figure 5c, there is an obvious effect of haul on L_0; thus, we model L_0 hierarchically, using haul-level intercepts (i.e., Gaussian distributed random intercepts). For each growth function

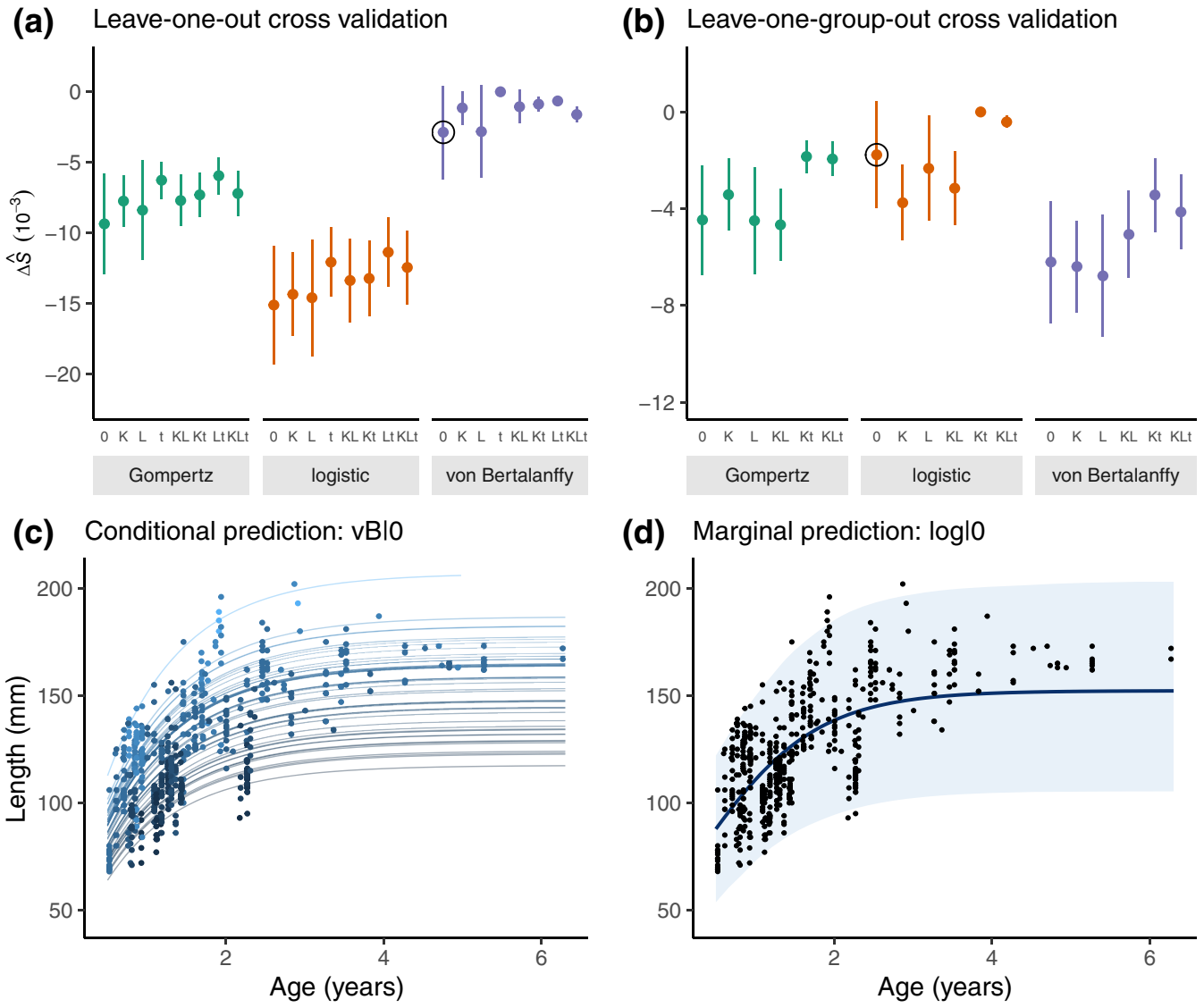

FIGURE 5 Model comparison and predicted curves for pinfish growth models. The plots on the left, (a) and (c), and right, (b) and (d), relate to a conditional and marginal focus, respectively. Applying the modified one-standard-error rule to each focus, the fitted growth curves of the selected models, circled in (a) and (b), are plotted alongside the data in (c) and (d). The data and curves in (c) are colored by haul according to the modeled haul-level length asymptote $L_{0,\text{haul}}$. The dots and bars in (a) and (c) are the mean and standard error (7) of the score differences, respectively. The model variants **x**|**t** and **x**|**Lt** are omitted from (d) for clarity due to their mean $\Delta\widehat{S}$ estimates being significantly more negative the others. The envelope in (d) is the 95% credible interval.

x = G, log, vB, the most complex model **x**|**LKt** includes the two-level predictor sex as a fixed effect on L_0, K, and t_0, specified as follows:

$$\mathbf{x}\,|\,\mathbf{LKt} \qquad \begin{aligned} y_i &\sim N(\mu_i, \sigma) \\ \mu_i &= L_x(a_i; L_{0,i}, K_i, t_{0,i}) \\ L_{0,i} &= \beta_{\text{L},0} + \beta_{\text{L},\text{sex}_1[\text{i}]} + \text{b}_{\text{haul[i]}} \\ b &\sim N(0, \tau) \\ K_i &= \beta_{\text{K},0} + \beta_{\text{K},\text{sex}_1[\text{i}]} \\ t_{0,i} &= \beta_{\text{t},0} + \beta_{\text{t},\text{sex}_1[\text{i}]}, \end{aligned} \tag{11}$$

from which simpler submodels are obtained by setting a subset of the terms $\beta_{\cdot,\text{sex}_1}$ to zero. For example, **vB**|**K** is the von Bertalanffy growth model including sex as a fixed effect on K only (i.e., $\beta_{L,\text{sex}_1} = \beta_{t,\text{sex}_1} = 0$). All models include the haul-level L_0 intercepts $b_{\text{haul}[i]}$. We fit all 24 models in a Bayesian framework using R package **brms** (Bürkner, 2017). Model priors were normal or student-t distributions with standard deviations in the range of 0.3 to 10—sufficiently narrow to aid chain convergence while large enough to exert minimal influence on the posterior distributions (see Yates et al., 2022 for further details).

We illustrate two approaches to model selection using cross validation:

1. Conditional focus using approximate approximate leave-one-out cross validation; and
2. Marginal focus using leave-one-group-out cross validation.

The conditional focus concerns model predictions to existing hauls, conditional on the haul-level L_0 estimates which are treated as (regularized) model parameters. In this scenario, individual fish measurements constitute conditionally independent test samples, thereby permitting the use of leave-one-out cross validation to estimate and compare model performance. A possible interpretation of this focus is that, after accounting for unmeasured effects of haul on the data-generating process (e.g., changes in the measurement process due to differing apparatus or human expertise), estimates of conditional model performance quantify the capacity of candidate growth functions to predict within-haul growth trends and variation.

The marginal focus concerns model predictions to new hauls, where the estimate of between-haul variation, τ in (11), is used in combination with the residual variation σ to explain the total variability of the modeled response around the mean population-level growth curve. The use of leave-one-group-out cross validation in this case provides a direct means to estimate the predictive merits of the marginalized model, without having to actually integrate (i.e., marginalize) the likelihood over the hierarchical distribution of haul-level intercepts (this can be slow and difficult for non-linear models). A possible interpretation of the marginal focus is that having estimated the haul-level variability of the length asymptote, leave-one-group-out cross validation quantifies the capacity of candidate growth functions to model the population-level mean growth curve, predicting to new fish measurements in new hauls.

The model-selection results for both types of focus are shown in Figure 5a,b, where the modified one-standard-error rule has been applied using the standard error (7). The best-performing growth model differed between the two types of focus, highlighting the importance of selecting a data-splitting scheme according to the predictive focus. However, in both types of focus, the application of the modified one-standard-error rule suggested selection of the least complex variant $\mathbf{x}|\mathbf{0}$ which excluded sex as a fixed effect on any of the model parameters. This does not mean that sex does not have an effect on fish growth, only that, given the available data and the specified set of candidate models, accounting for sex does not change which growth model is selected. Figure 5c,d show the fitted curves of the selected model for each focus.

In many instances, researchers will have an a priori reason to model a group-level effect as either fixed or random. For example, random effects may be chosen to "borrow strength" across group levels, to account for unmeasured group-level effects that vary around a population-level mean, to partition variance via marginalization, or to provide regularization to improve model prediction and/or convergence of the parameter-estimation algorithm (Hobbs & Hooten, 2015). It is possible, however, to use the data (via cross validation) to investigate whether a random effect or a fixed effect is the preferred choice, or at least to compare the two sets of estimates (e.g., to determine the amount of shrinkage). To illustrate, we apply approximate leave-one-out cross validation to the fixed and random variants of the conditional model $\mathbf{vB}|\mathbf{0}$, obtaining an estimated score difference of 2.17 ($\sigma_{\text{diff}} = 1.85$) in favor of the random model. To compare the complexity of each model, we compute the effective number of parameters (Gelman et al., 2013):

$$p_{\text{CV}} = \ell_{\text{WS}} - \ell_{\text{CV}}, \tag{12}$$

where $\ell_{\text{WS}} = \sum_{i=1}^{n} \log p(y_i|y)$ and $\ell_{\text{CV}} = \sum_{i=1}^{n} \log p\left(y_i|y^{-k[i]}\right)$ are the within-sample (i.e., trained on the full data set) and cross-validated log likelihood of the data, respectively. Using the approximate leave-one-out cross-validation estimates, the effective number of parameters for the random model is 42.6 compared to 45.7 for the fixed model—the latter is less than the nominal count of 48 fixed parameters due to the inherent regularization of the Bayesian estimation process. The weak preference for the random model, and the reduction of just 3.1 (SE = 0.6) effective parameters, is reflected in the mild shrinkage of the haul-level L_0 estimates for the random model compared to the fixed (Appendix S1: Figure S5).

For clarity of exposition, we have omitted more complex models, such as those including haul-level intercepts for K_0 and t_0. The inclusion of more than one group effect will generally lead to a multivariate hierarchical structure with corresponding correlation terms. However, assuming there are sufficient data to estimate the model parameters, model selection using cross validation can be implemented in these more complex cases in the same way as we have demonstrated above.

DISCUSSION

We have presented a comprehensive review on understanding and applying cross validation for model

selection, with a focus on how ecologists would use this approach. This includes an overview of commonly used techniques, theoretical aspects and recent developments, as well as practical guidance for implementation. Although it is difficult to provide universal advice for using cross validation, since the choice of data-splitting scheme and score depends on the modeling context, in most instances we recommend leave-one-out or approximate leave-one-out to minimize bias. Otherwise we recommend k-fold with k set as large as practicably possible. If $k < 10$ bias correction should be used when available. Further, the use of blocked, leave-one-group-out, or stratified data splitting should be investigated when the data are autocorrelated or hierarchically structured. To mitigate overfitting, we recommend calibrated selection via the modified one-standard-error rule (Yates et al., 2021) which accounts for the predominant cause of overfitting: score-estimation uncertainty.

There are many software packages for the R programming language that aid the implementation of cross validation. The packages **caret** (Kuhn, 2022) and **tidymodels** (Kuhn & Wickham, 2020) are particularly versatile, incorporating a broad class of model types (via commonly used auxiliary packages) ranging from generalized linear models and penalized regression to (tuned) machine-learning methods and certain classes of Bayesian models. The **tidymodels** package is a composite of several helper packages including **rsample** for data splitting (e.g., nested k-fold) and **yardstick** for estimating both pre- and user-defined model scores; these packages can be used independently of the integrated workflow of the parent package. Bias-corrected k-fold cross validation can be performed using the packages **bestglm** (McLeod et al., 2020) or **boot** (Canty & Ripley, 2021), although support is limited to generalized linear models; the latter permits the use of custom loss functions. For fully Bayesian approaches, the package **loo** (Vehtari et al., 2017) implements an approximate leave-one-out method, requiring as an input the log-likelihood evaluated at a set of posterior simulations of the parameter estimates. All of the aforementioned packages support parallel processing. When existing software is unsuitable, custom code can be developed. Generally speaking, if the set of candidate models is able to be fit to a training subset of the data and subsequently predicted to a test set, then cross-validation implementation is usually possible, requiring only data-splitting, repeated model-fitting, and subsequent aggregation of the score estimates. Existing software can help with each of these steps.

The ever-increasing uptake of Bayesian methodology in ecology in recent decades has been facilitated by the steady publication of both pedagogical texts (e.g., Hoeting et al. (1999); Ellison (2004); Gelman et al. (2013); Hobbs and Hooten (2015)) and integrated software (e.g., BUGS, JAGS, PyMC, Stan). Recently, there has been a surge in the development of Bayesian model-selection techniques (e.g., Vehtari et al. (2017) and Bürkner et al. (2020, 2021)), wherein information-theoretic methods based on cross validation have played a central role. The parallel development of Hamiltonian Monte Carlo (HMC, Neal, 2011) sampling methods has brought significant efficiency gains which are readily accessible to ecologists via the **rstan** (Stan Development Team, 2020) package or its user-friendly front-end packages **rstanarm** (Goodrich et al., 2020) and **brms** (Bürkner, 2017). The combination of Hamiltonian Monte Carlo with approximate leave-one-out cross validation permits highly efficient model fitting and score estimation for a broad range of model-selection contexts, including historically challenging model classes such as non-factorized data models (e.g., observation-level latent-variable models) (Bürkner et al., 2021)—equivalent methods do not exist outside of the Bayesian setting. None of these implementations require informative priors, although they could certainly be used where appropriate. Thus, there is a compelling case for the use of Bayesian methods for model specification and model selection in ecology (at least for likelihood-based inference) where the increasing use of complex hierarchical model structures is a natural fit with sampling-based estimation which has robust estimation properties and a high tolerance for model complexity.

In most model-selection problems, cross validation can be used to estimate scores and suggest a preferred model, but is the selected model good enough and why cannot we make valid inference after selection? In terms of model checking, it is important to note that cross validation is not a replacement for data simulation (e.g., posterior predictive checks or the parametric bootstrap) or within-sample measures such as the proportion of deviance explained or graphical checks of residual distributions. These techniques are used to assess the adequacy of the selected model to generate plausible data where expert knowledge is often required to determine what constitutes adequate in a given modeling context (Gabry et al., 2019). Given the selected model has passed model checking, the issue of valid inference remains. Although it is common practice, it is highly problematic to make inferences using the parameter estimates of the selected model fit to the full data set without accounting in some way for model-selection uncertainty—problems include inflated effect sizes (i.e., selection-induced bias) and underestimated uncertainty intervals (Hjort & Claeskens, 2003). Some recently developed (and highly

technical) methods exist to make valid post-selection inferences, although much work is needed to make these relevant and accessible to ecological analysts (see Box 1 for further discussion).

Cross validation is clearly a practical and versatile technique for model selection in ecology: a research field in which statistical modeling has been increasingly dominated over the last two decades by the use of information criteria (especially AIC). Yet given the broad applicability of cross validation, and the now near-ubiquitous availability of multi-core parallel computing, one might ask whether the prominent role of information criteria is coming to an end. Information Criteria are generally easier and faster to compute than conventional cross-validation estimates (except leave-one-out in linear regression), but the need for bias correction imposes limitations on the data and models, leading to an ever-growing set of information-criterion variants to accommodate specialized settings; for example, AIC_c (corrected) (marginal) mAIC, (quasi) QAIC, (focussed) FIC, and the widely applicable information criterion (WAIC). High-performance parallel computation and the development of accurate approximate cross-validation methods now makes cross validation comparable with information criteria for speed, obviating the need for such a large number of specialized variants, while at the same time increasing the reach of model-selection techniques into complex modeling contexts where traditional information criteria remain inapplicable (e.g., deep-learning approaches such as artificial neural networks). In a Bayesian setting, Vehtari et al. (2017) advocate for the use of (approximate) leave-one-out over the widely applicable information criterion (WAIC) due to improved stability, although both are considered superior to the commonly used AIC and DIC. Given the demonstrable utility and generality of cross validation for comparing a diverse range of statistical models, fitted to simple or highly complex data sets, we foresee its increasingly widespread adoption in the domain of ecology.

ACKNOWLEDGMENTS

We thank Carsten Dormann, one anonymous reviewer, and the editor for their detailed and constructive feedback which greatly improved the manuscript. This work was funded by the Australian Research Council grant FL160100101. We are grateful to Leon Barmuta for helpful discussions and feedback. Open access publishing facilitated by University of Tasmania, as part of the Wiley - University of Tasmania agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sets utilized for this research are "scat data" available from Kuhn (2022) https://cran.r-project.org/package=caret and "pinfish data" available from Nelson (2022) https://cran.r-project.org/package=fishmethods. Novel code (Yates et al., 2022) is available at https://doi.org/10.5281/zenodo.7094841.

ORCID

Luke A. Yates https://orcid.org/0000-0002-1685-3169
Zach Aandahl https://orcid.org/0000-0002-9412-8288
Shane A. Richards https://orcid.org/0000-0002-9638-5827
Barry W. Brook https://orcid.org/0000-0002-2491-1517

REFERENCES

Akaike, H. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, edited by B. N. Petrov and F. Csaki, 267–81. Budapest: Academiai Kiado.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. "Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS)." *Journal of Applied Ecology* 43(6): 1223–32. https://doi.org/10.1111/j.1365-2664.2006.01214.x.

Arif, S., and M. Aaron MacNeil. 2022. "Predictive Models aren't for Causal Inference." *Ecology Letters* 25(8): 1741–5. https://doi.org/10.1111/ele.14033.

Arlot, S. 2008. "V-Fold Cross-Validation Improved: V-Fold Penalization."

Arlot, S., and A. Celisse. 2010. "A Survey of Cross-Validation Procedures for Model Selection." *Statistics Surveys* 4: 40–79. https://doi.org/10.1214/09-SS054.

Arlot, S., and M. Lerasle. 2016. "Choice of V for V-Fold Cross-Validation in Least-Squares Density Estimation." *Journal of Machine Learning Research* 17: 1–50.

Bachoc, F., H. Leeb, and B. M. Pötscher. 2019. "Valid Confidence Intervals for Post-Model-Selection Predictors." *Annals of Statistics* 47(3): 1475–504. https://doi.org/10.1214/18-AOS1721.

Banner, K. M., and M. D. Higgs. 2017. "Considerations for Assessing Model Averaging of Regression Coefficients." *Ecological Applications* 27(1): 78–93. https://doi.org/10.1002/eap.1419.

Benjamini, Y. 2020. "Selective Inference: The Silent Killer of Replicability." *Harvard Data Science Review* 2(4). https://doi.org/10.1162/99608f92.fc62b261.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao. 2013. "Valid Post-Selection Inference." *Annals of Statistics* 41(2): 802–37. https://doi.org/10.1214/12-AOS1077.

Bernardo, J. M. 1979. "Expected Information as Expected Utility." *The Annals of Statistics* 7(3): 689–90. https://doi.org/10.1214/aos/1176344689.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45(1): 5–32. https://doi.org/10.1023/a:1010933404324.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees* 1–358. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software.

Breiman, L., and P. Spector. 1992. "Submodel Selection and Evaluation in Regression. The X-Random Case." *International Statistical Review / Revue Internationale de Statistique* 60(3): 291–312. https://doi.org/10.2307/1403680.

Bürkner, P.-C. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80(1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Bürkner, P.-C., J. Gabry, and A. Vehtari. 2020. "Approximate Leave-Futureout Cross-Validation for Bayesian Time Series Models." *Journal of Statistical Computation and Simulation* 90(14): 2499–523. https://doi.org/10.1080/00949655.2020.1783262.

Bürkner, P.-C., J. Gabry, and A. Vehtari. 2021. "Efficient Leave-One-out Cross-Validation for Bayesian Non-factorized Normal and Student-t Models." *Computational Statistics* 36(2): 1243–61. https://doi.org/10.1007/s00180-020-01045-4.

Burman, P. 1989. "A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods." *Biometrika* 76(3): 503–14. https://doi.org/10.1093/biomet/76.3.503.

Burman, P., E. Chow, and D. Nolan. 1994. "A Cross-Validatory Method for Dependent Data." *Biometrika* 81(2): 351–8. https://doi.org/10.1093/biomet/81.2.351.

Burnham, K. P., and D. R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. 515. New York: SpringerVerlag.

Cade, B. S. 2015. "Model averaging and muddled multimodel inferences." *Ecology* 96(9): 2370–82. https://doi.org/10.1890/14-1639.1.

Canty, A., and B. D. Ripley. 2021. "boot: Bootstrap R (S-Plus) Functions." R package version 1.3-28. https://cran.r-project.org/package=boot.

Cawley, G. C., and N. L. C. Talbot. 2010. "On over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation." *Journal of Machine Learning Research* 11: 2079–107.

Charkhi, A., and G. Claeskens. 2018. "Asymptotic Post-Selection Inference for the Akaike Information Criterion." *Biometrika* 105(3): 645–64. https://doi.org/10.1093/biomet/asy018.

Chicco, D., M. J. Warrens, and G. Jurman. 2021. "The Matthews Correlation Coefficient (MCC) Is more Informative than Cohen's Kappa and Brier Score in Binary Classification Assessment." *IEEE Access* 9: 78368–81. https://doi.org/10.1109/ACCESS.2021.3084050.

Claeskens, G., and N. L. Hjort. 2008. "Model Selection and Model Averaging." In *Cambridge Series in Statistical and Probabilistic Mathematics*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511790485.

Craven, P., and G. Wahba. 1978. "Smoothing Noisy Data with Spline Functions." *Numerische Mathematik* 31(4): 377–403. https://doi.org/10.1007/bf01404567.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and their Application*. New York: Cambridge University Press. https://doi.org/10.1017/cbo9780511802843.

Dormann, C. F., J. M. Calabrese, G. Guillera-Arroita, E. Matechou, V. Bahn, K. Bartoń, C. M. Beale, et al. 2018. "Model Averaging in Ecology: A Review of Bayesian, Information-Theoretic, and Tactical Approaches for Predictive Inference." *Ecological Monographs* 88(4): 485–504. https://doi.org/10.1002/ecm.1309.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. García Marquéz, et al. 2013. "Collinearity: A Review of Methods to Deal with it and a Simulation Study Evaluating their Performance." *Ecography* 36(1): 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Ellison, A. M. 2004. "Bayesian Inference in Ecology." *Ecology Letters* 7(6): 509–20. https://doi.org/10.1111/j.1461-0248.2004.00603.x.

Fang, Y. 2011. "Asymptotic Equivalence between Cross-Validations and Akaike Information Criteria in Mixed-Effects Models." *Journal of Data Science* 9: 15–21.

Feurer, M., and F. Hutter. 2019. "Hyperparameter Optimization." In *Automated Machine Learning: Methods, Systems, Challenges*, edited by F. Hutter, L. Kotthoff, and J. Vanschoren, 3–33. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-51.

Fletcher, R., and M.-J. Fortin. 2018. *Spatial Ecology and Conservation Modeling*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-01989-1.

Fox, G. A., S. Negrete-Yankelevich, and V. J. Sosa. 2015. *Ecological Statistics: Contemporary Theory and Application*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199672547.001.0001.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29(5): 1189–232. https://doi.org/10.1214/aos/1013203451.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1): 1–22.

Gabry, J., D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. 2019. "Visualization in Bayesian Workflow." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182(2): 389–402. https://doi.org/10.1111/rssa.12378.

Garcia-angulo, A. C., and G. Claeskens. 2023. "Optimal Finite Sample Post-Selection Confidence Distributions in Generalized Linear Models." *Journal of Statistical Planning and Inference* 222: 66–77. https://doi.org/10.1016/J.JSPI.2022.06.001.

Garthwaite, P. H., and E. Mubwandarikwa. 2010. "Selection of Weights for Weighted Model Averaging." *Australian and New Zealand Journal of Statistics* 52(4): 363–82. https://doi.org/10.1111/j.1467-842X.2010.00589.x.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian Data Analysis*, [in English]. Hardcover. 675. Boca Raton: Chapman/Hall/CRC.

Gneiting, T. 2011. "Making and Evaluating Point Forecasts." *Journal of the American Statistical Association* 106(494): 746–62. https://doi.org/10.1198/jasa.2011.r10138.

Gneiting, T., and A. E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102(477): 359–78. https://doi.org/10.1198/016214506000001437.

Goodrich, B., J. Gabry, I. Ali, and S. Brilleman. 2020. "rstanarm: Bayesian applied regression modeling via Stan." R package version 2.19.3. https://mc-stan.org/rstanarm/.

Harrell, F. E. 2015. *Regression Modeling Strategies*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd ed. 282. New York: Springer. https://doi.org/10.1007/978-0-387-84858-7.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical Learning with Sparsity*. Boca Raton: Chapman/Hall/CRC. https://doi.org/10.1201/b18401.

Hjort, N. L., and G. Claeskens. 2003. "Frequentist Model Average Estimators." *Journal of the American Statistical Association* 98(464): 879–99. https://doi.org/10.1198/016214503000000828.

Hobbs, N. T., and M. B. Hooten. 2015. *Bayesian Models: A Statistical Primer for Ecologists*. Princeton: Princeton University Press.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. "Bayesian model averaging: A tutorial." *Statistical Science* 14(4): 382–401. https://doi.org/10.1214/ss/1009212519.

Hooten, M. B., and N. T. Hobbs. 2015. "A Guide to Bayesian Model Selection for Ecologists." *Ecological Monographs* 85(1): 3–28. https://doi.org/10.1890/14-0661.1.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning*, Vol 103. Springer Texts in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7.

Johnson, D. H. 1999. "The Insignificance of Statistical Significance Testing." *The Journal of Wildlife Management* 63(3): 763–72.

Kabaila, P. 2009. "The Coverage Properties of Confidence Regions after Model Selection." *International Statistical Review* 77(3): 405–14. https://doi.org/10.1111/j.1751-5823.2009.00089.x.

Kabaila, P., A. H. Welsh, and W. Abeysekera. 2016. "Model-Averaged Confidence Intervals." *Scandinavian Journal of Statistics* 43(1): 3548. https://doi.org/10.1111/sjos.12163.

Kuhn, M., and H. Wickham. 2020. "Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles." https://www.tidymodels.org.

Kuhn, M. 2022. "caret: Classification and Regression Training." R package version 6.0-90., https://cran.r-project.org/package=caret.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor. 2016. "Exact Post-Selection Inference, with Application to the Lasso." *Annals of Statistics* 44(3): 907–27. https://doi.org/10.1214/15-AOS1371.

Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. "Selecting Thresholds of Occurrence in the Prediction of Species Distributions." *Ecography* 28(3): 385–93. https://doi.org/10.1111/j.0906-7590.2005.03957.x.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems,* edited by U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus, 4768–77. California: Curran Associates Inc.

Luque, A., A. Carrasco, A. Martín, and A. d. l. Heras. 2019. "The Impact of Class Imbalance in Classification Performance Metrics Based on the Binary Confusion Matrix." *Pattern Recognition* 91: 216–31. https://doi.org/10.1016/j.patcog.2019.02.023.

Matthews, B. W. 1975. "Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme." *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405: 442–51. https://doi.org/10.1016/0005-2795(75)90109-9.

McLeod, A. I., X. Changjiang, and Y. Lai. 2020. "bestglm: Best Subset GLM and Regression Utilities." R package version 0.37.3. https://cran.r-.

Merkle, E. C., D. Furr, and S. Rabe-Hesketh. 2019. "Bayesian Comparison of Latent Variable Models: Conditional Versus Marginal Likelihoods." *Psychometrika* 84(3): 802–29. https://doi.org/10.1007/s11336-019-09679-0.

Milà, C., J. Mateu, E. Pebesma, and H. Meyer. 2022. "Nearest Neighbour Distance Matching Leave-One-out Cross-Validation for Map Validation." *Methods in Ecology and Evolution* 13(6): 1304–16. https://doi.org/10.1111/2041-210X.13851.

Neal, R. 2011. "MCMC Using Hamiltonian Dynamics." In *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. L. Jones, and X. -L. Meng. London: Chapman and Hall/CRC. https://doi.org/10.1201/b10905-6.

Nelson, G. 2002. "Age, Growth, Mortality, and Distribution of Pinfish (Lagodon Rhomboides) in Tampa Bay and Adjacent Gulf of Mexico Waters." *Fishery Bulletin* 100: 582–92.

Nelson, G. A. 2022. "fishmethods: Fishery Science Methods and Models." R package version 1.11-3. https://cran.r-project.org/package=fishmethods.

Pearl, J. 2010. "An Introduction to Causal Inference." *The International Journal of Biostatistics* 6(2): 7. https://doi.org/10.2202/1557-4679.1203.

Piironen, J., M. Paasiniemi, A. Catalina, and A. Vehtari. 2020. "projpred: Projection Predictive Feature Selection." R package version 2.0.2. https://mc-stan.org/projpred/.

Piironen, J., M. Paasiniemi, and A. Vehtari. 2020. "Projective Inference in High-Dimensional Problems: Prediction and Feature Selection." *Electronic Journal of Statistics* 14(1): 2155–97. https://doi.org/10.1214/20-EJS1711.

Piironen, J., and A. Vehtari. 2017a. "Comparison of Bayesian Predictive Methods for Model Selection." *Statistics and Computing* 27(3): 711–35. https://doi.org/10.1007/s11222-016-9649-y.

Piironen, J., and A. Vehtari. 2017b. "Sparsity Information and Regularization in the Horseshoe and Other Shrinkage Priors." *Electronic Journal of Statistics* 11(2): 5018–51. https://doi.org/10.1214/17-EJS1337SI.

Reid, R. E. B. 2015. "A Morphometric Modeling Approach to Distinguishing among Bobcat, Coyote and Gray Fox Scats." *Wildlife Biology* 21(5): 254–62. https://doi.org/10.2981/wlb.00105.

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, et al. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40(8): 913–29. https://doi.org/10.1111/ecog.02881.

Shao, J. 1993. "Linear Model Selection by Cross-Validation." *Journal of Statistical Planning and Inference* 128(1): 231–40. https://doi.org/10.1016/j.jspi.2003.10.004.

Shen, X., H. C. Huang, and J. Ye. 2004. "Inference after Model Selection." *Journal of the American Statistical Association* 99(467): 751–62. https://doi.org/10.1198/016214504000001097.

Shmueli, G. 2010. "To Explain or to Predict?" *Statistical Science* 25(3): 289–310. https://doi.org/10.1214/10-STS330.

Shmueli, G., and O. R. Koppius. 2011. "Predictive Analytics in Information Systems Research." *Management Information Systems Quarterly* 35(3): 553–72. https://doi.org/10.2307/23042796.

Sivula, T., M. Magnusson, and A. Vehtari. 2020. "Uncertainty in Bayesian Leave-One-out Cross-Validation Based Model Comparison."

Stan Development Team. 2020. "RStan: the R interface to Stan." R package version 2.21.2. https://mc-stan.org/rstan/.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani. 2016. "Exact Post-Selection Inference for Sequential Regression Procedures." *Journal of the American Statistical Association* 111(514): 600–20. https://doi.org/10.1080/01621459.2015.1108848.

Tjørve, E., and K. M. C. Tjørve. 2010. "A Unified Approach to the Richards-Model Family for Use in Growth Analyses: Why we Need Only Two Model Forms." *Journal of Theoretical Biology* 267(3): 417–25. https://doi.org/10.1016/j.jtbi.2010.09.008.

Tredennick, A. T., G. Hooker, S. P. Ellner, and P. B. Adler. 2021. "A Practical Guide to Selecting Models for Exploration, Inference, and Prediction in Ecology." *Ecology* 102(6): e03336. https://doi.org/10.1002/ecy.3336.

van Erp, S., D. L. Oberski, and J. Mulder. 2019. "Shrinkage Priors for Bayesian Penalized Regression." *Journal of Mathematical Psychology* 89: 31–50. https://doi.org/10.1016/j.jmp.2018.12.004.

Vehtari, A., A. Gelman, and J. Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC." *Statistics and Computing* 27(5): 1413–32. https://doi.org/10.1007/s11222-016-9696-4.

Vehtari, A., D. P. Simpson, Y. Yao, and A. Gelman. 2019. "Limitations of "Limitations of Bayesian Leave-One-out Cross-Validation for Model Selection"." *Computational Brain& Behavior* 2(1): 22–7. https://doi.org/10.1007/s42113-018-0020-6.

Wasserstein, R. L., and N. A. Lazar. 2016. "The ASA Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70(2): 129–33. https://doi.org/10.1080/00031305.2016.1154108.

Wood, S. N. 2017. *Generalized Additive Models*. New York: Chapman/Hall/CRC. https://doi.org/10.1201/9781315370279.

Yates, L. A., Z. Aandahl, S. A. Richards, and B. Brook. 2022. "Cross validation for model selection code. V1.0.0." https://doi.org/10.5281/zenodo.7094841.

Yates, L. A., S. A. Richards, and B. W. Brook. 2021. "Parsimonious Model Selection Using Information Theory: A Modified Selection Rule." *Ecology* 102: e03475. https://doi.org/10.1002/ecy.3475.

Zhang, C. 2008. "Prediction Error Estimation under Bregman Divergence for Non-parametric Regression and Classification." *Scandinavian Journal of Statistics* 35(3): 496–523. https://doi.org/10.1111/j.1467-9469.2008.00593.x.

Zhang, D., A. Khalili, and M. Asgharian. 2022. "Post-Model-Selection Inference in Linear Regression Models: An Integrated Review." *Statistics Surveys* 16: 86–136. https://doi.org/10.1214/22-SS135.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection Via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2): 301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Yates, Luke A., Zach Aandahl, Shane A. Richards, and Barry W. Brook. 2023. "Cross Validation for Model Selection: A Review with Examples from Ecology." *Ecological Monographs* 93(1): e1557. https://doi.org/10.1002/ecm.1557