Running Head: Evaluating probabilistic forecasts

Methods for Evaluating Probabilistic Ecological Forecasts

Juniper L. Simonis1,2†, Ethan P. White1, S. K. Morgan Ernest1

1Wildlife Ecology and Conservation, University of Florida

2DAPPER Stats, 3519 NE 15th Ave., Suite 467, Portland, OR 97212, USA

†Corresponding author; e-mail: simonis@dapperstats.com

**Abstract**

Effective near-term forecasting allows for the evaluation of model predictions against observations and is of pressing need in ecology to inform environmental decision making and effect societal change. Despite this imperative, we presently lack a set of robust, standardized, and general mathematical tools for evaluating probabilistic forecasts in ecology, impeding quantitative model comparison and ensemble model construction. We address this gap here by bringing to bear in ecology an extensive literature on probabilistic forecast evaluation from diverse applied mathematical fields (e.g. climatology, economics, epidemiology). Recognizing the variety of ecological data and projects and appreciating the variety of tools developed, rather than lobby for a specific singular metric for evaluation, we cover the breadth of options (including frequentist, information-theoretic, tactical, and Bayesian approaches), highlight mathematical concepts to follow, and focus on decision points for practitioners to allow easy application of general principles to specific forecasting endeavors. We show an example application using a long-term rodent population time series before finishing with a discussion of how ecology can continue to learn from and help drive the field of forecasting.

*Keywords:* Bayesian Model Averaging, continuous analysis, ecological forecasting, ensemble forecast, end-sample holdout, forecast skill, prequential, score rule, stacked generalization, time series, validation.

**Introduction**

Although historically present in certain subdisciplines, forecasting is rapidly becoming an important becoming a focus of ecological science in both applied and fundamental settings (Clark et al. 2001, Pennekamp et al. 2017, Dietze et al. 2018). In order to adequately represent the inherent uncertainty associated with model fitting, the natural variability of ecosystems, and the consequences for decision making, ecological forecasts must be probabilistic (Dawid 1984, Williams and Hooten 2016, Dietze et al. 2018). This uncertainty occurs at two fundamental levels: uncertainty in the values forecasted by a given model and uncertainty among models as to how well they represent the true, unobservable distribution (Hooten and Hobbs 2015, Krushke 2015). Accordingly, ecologists are beginning to produce forecasts as distributions that include within- and among-model uncertainty (Dietz 2017, Simonis et al. 2018, Harris et al. 2018).

Despite the necessity of distributional forecasts, however, the majority of metrics (for example, Root Mean Squared Error (RMSE), Mean Absolute Scaled Error (MASE), correlation coefficient) proposed for and used in ecological forecasting are based on point forecasts, typically evaluated using the expected value of the predicted distribution (Hyndman and Koehler 2006, Ward et al. 2014, Petchey et al. 2015, Dietze et al. 2018). Although these measures provide useful information about how well the central tendency of the forecasted distribution matches the observed values, they are not (necessarily) uniquely optimized by the true probability distribution (that is, they are not *strictly proper*, see **Scoring Rules** and Appendix A; Gneiting and Raftery 2007), they cannot distinguish among forecasts with the same central tendencies but different variances (Gneiting et al. 2007, Czado et al. 2009), and their use can lead to “grossly misguided inferences” in certain situations (Gneiting 2011).

Driven in particular through application in fields such as meteorology, econometrics, and epidemiology, statistical forecasters have developed a set of philosophical principles and tools for evaluating probabilistic forecasts produced by competing models (Winkler 1977, Dawid 1984, Gneiting and Raftery 2007, Ray and Reich 2018). Our goal with the present manuscript is therefore to support the use of probabilistic forecasts in ecology by bringing this set of tools to bear in an ecological setting. Embracing the variety of ecological variables that could be forecast (Dietz 2018) and drawing parallels with ecological Bayesian model checking, which also compares probabilistic predictions from one or more models to observed values (Conn et al. 2018), we recognize that there is no singular best metric or approach to evaluating all ecological forecasts. As such, we provide an overview of available methods with supporting literature, drawing on standard forecasting practices in other disciplines while also noting that forecasting statistics is an active field of research to which ecological studies can contribute important real-world applications of theoretical approaches.

**Context, Notation, and Terminology**

To begin, we define notation and a context in which we are conducting an ecological forecasting study (Fig. 1; Dietz 2018). Consider a time series of samples ( in ) of an ecological variable ( in ... or simply ), where each is an observation collected at time (in ). The variable can be discrete or continuous and the samples can be taken at fixed or variable time intervals. The observed time series is one realization drawn from the unknowable generating distribution through time ( is the distribution at time ) (Fig. 1).

Our goal is to make ( in ) forecasts of subsequent to by using ( in ) models to gain inference about . Following standard terminology (c.f., e.g., Tashman 2000), the last datum in the observed time series used to make a forecast ( when working with the full time series, but more generally) is called the *forecast origin*, and the time between the forecast origin and a specific datum being forecast () is the *lead time* or *forecast horizon* ( generally; Fig. 1). Thus, when working with the full time series, we are predicting for samples to () and the full forecast horizon is .

Each model therefore needs to be able to [1] fit and [2] predict (Fig. 1). Because we are working in a probabilistic framework, [2] means that every model needs to produce a forecast distribution across all samples within the forecast horizon (; Fig. 1). See Appendix A for the formal description of probabilistic forecasts used here. In order to properly execute both tasks (including providing skill metrics for model fits), we use the data in hand to validate our models and iterate the evaluation over time using the prequential approach (*prequential* is a portmanteau of *probabilistic* and *sequential*; Dawid 1984, Dietze et al. 2018). This is opposed to validating models after future data are collected (Makridakis et al. 1993).

**Validation Procedure**

*General methodology*

The validation procedure defines how the observed data are used to [1] fit the model (the *training data*) and [2] evaluate its predictions (the *test data*). When considering a set of multiple models, it is crucial that they be tested on the same challenge (i.e., validated using the same data set) so that they can be quantitatively and directly compared (Wolpert 1992, Hastie et al. 2011, Ray and Reich 2018). There is an array of potential validation procedures available, including in forecasting settings specifically (Stone 1974, Tashman 2000, Arlot and Celisse 2010). Based upon the ultimate task at hand being predicting the next set of data in a time series (Dawid 1984), however, the dominant paradigm in forecasting model validation is understandably based on an *end-sample holdout*, where the last observations are held out for testing (Fig. 2; Fildes and Makridakis 1995 and references therein, Tashman 2000). This is in comparison to cross-validation approaches that select test data from across the entire data set (e.g., leave- -out, -fold; Fig. 2; Arlot and Celisse 2010) and asymptotic approximations like AIC (Stone 1977).

End-sample holdout methods have been supported through simulation- and empirical-based evaluations of forecast performance, which show that they produce more realistic error distributions for future data than in-sample cross-validation (Tashman 2000). Further, within-model (training) errors are often only very weakly correlated with outside-model (test) errors (Makridakis 1986, Makridakis and Winkler 1989, Pant and Starbuck 1990, Hastie et al. 2011), indicating that models being used to forecast will perform better on novel data when they have been validated via end-sample holdout methods than cross-validation methods (Fildes and Makridakis 1995). Despite this historical and logical impetus for forecast validation using end-sample holdout procedures, recent work suggests that cross-validation may be applicable to time series (Bergmeir and Benitez 2012, Bergmeir et al. 2018), although further work is needed.

A main motivation behind cross-validation is to increase the number of evaluations, as a single evaluation of a model’s forecast is likely an unstable estimate of its overall forecast capability (Tashman 2000). And indeed, a typical end-sample holdout validation on its own provides only a single evaluation, whereas cross-validation procedures aggregate multiple evaluations (Arlot and Celisse 2010, Bergmeir et al. 2018). However, the purpose of forecasting is to predict the next set of out-of-sample data (Fig. 1), and if done in a prequential manner (Dawid 1984) as is the goal (Dietze et al. 2018) and a reality (White et al. 2019) for ecological forecasting, the number of forecast-relevant evaluations grows over time (Fig. 2).

*Algorithmic specifics*

Given that the state of the field in indicates that end-sample holdout validation is the most appropriate procedure, we leverage this approach by defining a break in our time series () between the training and test sets based on a forecast origin , resulting in a training set of values () and a test set of values (). This break focuses the validation on quantifying how well a model’s forecast distribution matches the observations in the test set , where matching is defined by a scoring function, see **Scoring Functions**) (Dawid 1984). In determining the number of samples to allocate to the test set (via the location of ), they should cover at least as much time as the longest forecast horizon required by the main application (Tashman 2000). That is, if the model is required to make a 12-month-ahead forecast, the holdout data set should cover at least 12 months of observations for testing.

As mentioned earlier, at this point our end-sample holdout has resulted in a single forecast series to be evaluated for each model. Having a single overall evaluation for each model could be insufficient, especially if the time series displays dynamics such as cyclical seasonality, in which case model performance will vary as a function of forecast origin (Pack 1990, Fildes 1992). This deficiency points to the utility of a *rolling forecast origin* validation, where multiple forecasts are made with the origin moved forward in the series (Fig. 2; Armstrong and Grohman 1972, Armstrong 1985). A rolling origin generates a more robust estimate of skill and facilitates analysis of performance as a function of (for example) lead time (Makridakis and Winkler 1989) or leveraged in constructing ensembles (Wolpert 1992, Ray and Reich 2018). If a rolling origin is used, a larger holdout set will allow for more forecasts of the target horizon, but may not be an option for shorter time series (Tashman 2000). When implementing a rolling origin forecast evaluation, a critical decision point is whether each step forward in the process should include just an update to the data or if the model should also be re-optimized (Tashman 2000). Although it is arguably preferable to update the model fit with each step forward in the evaluation, re-optimization can be computationally intensive, may not provide any marked change to the model parameters if the sample size is large and dynamics are slow (Tashman 2000, Tashman and Hoover 2001). Indeed, despite its importance, there remains substantial research to be done regarding model updating in rolling origin forecasts (Tashman 2000).

In the context of iterative forecasting within the prequential framework (Dawid 1984, Dietz et al. 2018), however, recurrent forecasts replace the done-all-at-once evaluations, thereby easing the computational burden of rolling origin validation. This is facilitated through a process known as *continuous analysis* that re-runs the models when the data have been updated (Beaulieu‐Jones and Greene 2017, White et al. 2019). In essence, continuous analysis creates a long-term system of rolling origin, fixed horizon, recalibrating end-sample holdout validations, to which each new validation (which is actually conducted as a fixed origin end-sample holdout validation) is added automatically (Fig. 2).

**Graphical Evaluation**

Upon completion of a model's fit, graphical evaluation provides key information regarding the basic appropriateness and success of the model.

*Prediction Plots*

As with any statistical model, a critical first evaluation for a forecast is plotting the predicted distribution and the observed values against each other (Dietze 2018). In the context of forecasting, where the data have an explicit time component, it is important to plot the time series of the prediction distributions and observed values with at least some of the training data to show past dynamics (Fig. 3). In addition, an informative general statistical plot is the x-y plot of predicted-observed values, which ideally follow a 1:1 relationship (Fig. 3). In both of these plots, it is important to recognize that predictive ecological models encompass multiple levels of uncertainty and often include non-linearities (Hooten and Hobbs 2015), resulting in distributions that are typically not well summarized using quantiles alone, and so either distributions or representative draws from them should be explicitly shown (Fig. 3; e.g., Simonis et al. 2018).

*Probability Integral Transform (PIT) Plot*

The Probability Integral Transform (PIT) is an informative summary with a solid statistical basis and a long history in forecasting. It simply represents the values of the predictive CDFs evaluated at the observed values through the time series (Table 1; Rosenblatt 1952, Dawid 1984, Diebold et al. 1998). If the observed values match the predictive distributions and the predictive distributions are continuous, the PIT has a standard uniform distribution (Dawid 1984), which can be checked informally using graphics (Fig. 3; Diebold et al. 1998, Gneiting et al. 2007). Although it should be noted that the uniformity of the PIT values is a necessary but not sufficient condition for a forecast to be considered *ideal* (match the generating distribution; Hamill 2001). Plotting the PIT as a histogram or CDF allows for comparison to a uniform, where deviations have particular meanings: skewed histograms indicate bias in the central tendencies, U-shaped histograms indicate undispersed predictive distributions, and hump-shaped histograms indicate overdispersed predictive distributions (Diebold et al. 1998, Gneiting et al. 2007).

The PIT was originally designed for use with continuous variables and the uniformity expectation does not hold for discrete-valued variables (Czado et al. 2009). To address this, the PIT has been extended to non-continuous distributions through multiple approaches (Table 1). The classical randomized PIT approximation adds noise using the standard uniform distribution (Smith 1985, Fruhwirth-Schnatter 1996, Lisenfeld et al. 2006). A non-randomized approximation replaces the randomized PIT with its conditional (given the observed count) CDF, which is then averaged across the set of observations to create a mean CDF that has an expectation of uniformity and can be compared to the CDF of the identity function (Table 1; Czado et al. 2009).

**Scoring Rules**

*General definition*

A scoring rule () defines a scoring function () that measures how well a point value matches a distribution (Brier 1950, McCarthy 1956, de Finetti 1962, Winkler 1967, Savage 1971; Appendix A). The score () of sample ’s observed value and the forecast distribution for that sample under model () using rule is

1

and scores are typically averaged across the forecasted values (producing ). Scoring functions can be negatively oriented (lower is better) or positively oriented (higher is better) and classical references for scores include both orientations. Here we use a positive orientation for all rules, such that the maximal score is always the best (Table 2). Although scoring rules are generally framed in terms of probabilistic distributions, they are still defined under the case of a point forecast (e.g., the expected value of the forecast distribution; Appendix A), in which case some of them simplify to classical point forecast evaluation metrics (see below).

A key set of characteristics about scoring rules are encompassed in the concepts of *propriety* and *strict propriety* (Winkler 1977, Dawid 1998, Gneiting and Raftery 2007; Appendix A). Simply, a proper scoring function is convex and achieves the best score at the true distribution, whereas a strictly proper scoring function is strictly convex and achieves it *only* at the true distribution (Brier 1950, Good, 1952, Winkler and Murphy 1968, Savage 1971). Proper scoring rules encourage honest forecasts that maximize reward and strictly proper rules ensure unique solutions (de Finetti 1962, Winkler 1977, Garthwaite et al. 2005).

*Specific scoring rules*

We now highlight specific scoring rules of potential use in evaluating probabilistic ecological forecasts (Table 2) that have functions defined for discrete (binary, categorical, count) and continuous variables (Gneiting and Raftery 2007, Czado et al. 2009). There are, in essence, an infinite number of rules, so we focus on those which are strictly proper and commonly used in forecasting. There is no automatic guideline for choosing a scoring rule, as each has its strengths and weaknesses that need to considered given the particular forecasting task at hand. Indeed, it is likely that a forecaster may wish to use multiple scores (either in combination or separately) to leverage their strengths for specific tasks (Czado et al. 2009, Ray and Reich 2018).

Logarithmic score

The logarithmic (log) scoring function is the log of the predictive probability evaluated at the observed value (Table 2; Good 1952). The log score is the sole proper scoring rule that depends only on the probability distribution at the specifically observed count (i.e., it is *local* Benedetti 2010). It is relatively simple to calculate and corresponds to a number of classic properties including Shannon entropy, Kullback-Leibler divergence, and predictive deviance (Gneiting and Raftery 2007). For these reasons, the log score is a commonly used measure of forecast skill (Gneiting and Raftery 2007). For example, the United States Center for Disease Control conducts forecast competitions for diseases such as influenza and typically evaluates models with the log score (Ray and Reich 2018, McGowan et al. 2019, Reich et al. 2019).

Despite its simplicity and popularity, the log score has its faults (Gneiting and Raftery 2007). In particular, it is *insensitive* (does not adequately respond to how far the true distribution is from the predicted one) in some instances and *hypersensitive* (reacts strongly to small differences in small probabilities and can therefore produce unacceptable values) in other conditions (Selten 1998). As such, caution should used when employing the log score, in particular if it is possible for rare (with respected to predicted density) values to be observed.

Quadratic (Brier) and Power Scores

The quadratic scoring function is the mean squared error of the probability forecasts where the observations are matched or not (Table 2; Brier 1950). It is a natural extension of the mean squared errors from point forecasts to distributional forecasts (Winkler 1996) and has both continuous and discrete calculations (Table 2), resulting in its application in a variety of forecasting scenarios (Wecker 1989, Selten 1998, Gneiting and Raftery 2007). Further, the quadratic scoring function can be generalized to a power scoring function using a single parameter allowing for a more flexible measure of the error metric (Selten 1998, Dawid 2006).

The quadratic score, while commonly used, is not without shortcomings (Jewson 2018). Namely, it is not local (it depends on events that did not happen), results in counter-intuitive rankings for very rare and very common events (because it uses an absolute difference in probability), and can require quite large sample sizes to account for variance inflation associated with autocorrelated forecasts and values (Benedetti 2010, Wilks 2010).

Spherical Score

The spherical score is a strictly proper, symmetric function so named because it transforms the probability to a point on the unit sphere by dividing the probability by its norm (Table 2; Roby 1965, Jose 2009). The score has a strong connection to the statistical notion of *surprise* (*sensu* Good 1971). Similarly to the quadratic score, the spherical score can be generalized to a pseudospherical score through a control parameter (Table 2; Good 1971, Dawid 2006, Gneiting and Raftery 2007). The spherical score is mentioned in many general coverages of scoring rules (e.g., Gneiting and Raftery 2007, Czado et al. 2009), but is not implemented frequently. Conversely to the log score (which is hypersensitive at extreme probabilities), the spherical score exhibits hypersensitivity near midpoint probabilities (Selten 1998).

Ranked Probability Score (RPS)

The Ranked Probability Score (RPS) defines a squared function that compares the CDF of a forecast and the CDF of the corresponding observation over a discrete number of categories (Table 2; Epstein 1969, Murphy 1969, 1971). In that way, the RPS is a discrete generalization of the classical binary quadratic score (Brier 1950) from two to more than two categories of outcome (Czado et al. 2009). The RPS can be expanded to continuous variables, at which point it is often referred to as the Continuous Ranked Probability Score or CRPS (Matheson and Winkler 1976, Hersbach 2000), which is the integral of the quadratic scores for the associated binary probability forecasts at all real-valued thresholds (Table 2; Dawid 2006, Gneiting and Raftery 2007). Historically, computation of the continuous version proved difficult (Krüger et al. 2019), but recent work (Gneiting and Raftery 2007) has shown that it can be empirically calculated, including an approximation using a series of draws from a distribution (e.g., samples from a posterior; Appendix A; Gneiting and Raftery 2007, Krüger et al. 2019).

The RPS has many favorable qualities: it considers the shape and tendency of the forecast distribution, is sensitive to distance (rewarding distributions for being closer to the observation), and only depends on the CDF, which is a more direct and stable function than the PDF or PMF (Hersbach 2000, Gneiting and Raftery 2007). Further, it generalizes the mean absolute error, which it reduces to if the forecast is a point measure, providing a metric to compare deterministic and probabilistic forecasts (Gneiting and Raftery 2007). As such, it has been successfully applied in many forecasting contexts (Gneiting et al. 2005, Velaquez et al. 2010, Sigrist et al. 2012). That being said, RPS is not without potential issues to be considered, as well. In particular, the function is sensitive to any predicted or observed values being unusually large when comparing competing forecasters (which may or may not be desirable; Candille and Talagrand 2005).

**Comparing Scores**

Recall that the scores for a set of models forecasting the same data measured using the same scoring function can be directly compared and form a coherent empirical distribution that can be analyzed (Makridakis and Winkler 1989, Gneiting and Raftery 2007). In that vein, many tools have been developed for evaluating forecast scores that will be useful in ecological settings.

*Graphical comparisons*

Although scores are typically aggregated across the test observations for the purposes of quantitative comparisons, graphical comparisons of sample-level score values can be used for diagnostic evaluations (Gneiting et al. 2007). For example, plotting scores as a function of covariates (e.g., temperature) could highlight if abnormal deviations are associated with external forces. Similarly, plots of scores as a function of lead time indicate how the skill of a model (or skills of models) decay over the forecast horizon (Petchey et al. 2015). Such graphical comparisons are bolstered through a cache of historical forecast evaluations as built via the prequential or iterative forecasting approach (Dawid 1984, Dietz et al. 2018, White et al. 2019), as apparent patterns in scores from a limited number of forecast evaluations may be artefactual.

*Skill scores*

The *skill score* () allows for standardized comparison among forecast scores, where the skill score of model across a set of observations using a particular scoring function is

2

where is the score of a reference model (must be selected) and is the score of an ideal forecast (set to the maximal value the scoring function can obtain) across the same set of observations using the same scoring function (Murphy 1973). In climatological forecasting, the reference model is often a simple yet reasonable reduced-complexity model, such as an estimate of the marginal predictive distribution (Gneiting and Raftery 2007). Skill scores take the value of 0 for the reference model forecast, 1 for an optimal forecast; a positive score means the model’s forecast was better than the reference, a negative score means it was worse than the reference. While skill scores provide a standardized comparison, however, they are generally not proper even if the underlying scoring function is proper, although the importance of the impropriety for inference is unresolved (Murphy 1973, Briggs and Ruppert 2005, Gneiting and Raftery 2007).

*Diebold-Mariano Test*

Frequentist comparisons of forecasts’ scores are robust as long as the non-independence of values is accounted for (Makridakis and Winkler 1989, Diebold and Mariano 1995, Hamill 1999). To that end, a variety of tests have been developed, but the Diebold-Mariano (D-M) test has become the workhorse of frequentist significance-based forecast comparison (Diebold 2015). The statistical significance (in a frequentist sense) of the difference between pairs of forecasts can be evaluated using the Diebold-Mariano test (Diebold and Mariano 1995, D’Agostino et al. 2012, Gneiting and Katzfuss 2014, Diebold 2015, Christensen et al. *unpublished*), which are based on the z-test and include methods for correlated error structures (Table 4; Diebold and Mariano 1995, Diebold 2015). The basis of the D-M test is the differential between forecast scores, for example the differential between scores measured using the same function on forecasts produced by models for observation is simply

3

which has an expected value of 0 under a null hypothesis of no difference between forecast scores (Diebold and Mariano 1995). The metric for a forecast series from to is the mean differential across the series () divided by an estimate of the standard deviation of the mean differential (; see Appendix A) times the square root of the sample size ():

4

which has an expected distribution of a standard normal (mean 0, standard deviation 1) under the null hypothesis of no difference between forecasts (Diebold and Mariano 1995, Diebold 2015). Serial correlation among the differentials can be addressed by using a formula for the standard deviation that is robust to heteroskedasticity and autocorrelation (Diebold and Mariano 1995, Diebold 2015; Appendix A). These robust calculations can require substantial test data sizes, but bootstrapping approaches have been developed to mitigate these issues (D’Agostino et al. 2012).

Although the D-M test was initially proposed as a pairwise comparison between two forecasts (Diebold and Mariano 1995), it has recently been extended to multiple comparisons among more than two forecasts using permutation-based (D’Agostino et al. 2012) and closed-form (Christensen et al. *unpublished*) calculations. These methods are promising for frequentist comparisons among multiple forecasts, but are still quite novel and will require additional theoretical and application evaluation to determine their efficacy and utility in ecological forecasting. For example, the closed-form multivariate D-M test appears to require extensive quantities of data, although finite sample corrections exist (Christensen et al. *unpublished*).

**Ensemble Building**

We turn our attention now to the task of building ensemble models by aggregating components of our base model set. Ensembles are important for providing a coherent forecast while embracing model uncertainty (Gneiting and Raftery 2005, Dietz 2018, Dietze et al. 2018). Perhaps unsurprisingly, the literature on ensemble construction is broad (Dormann et al. 2018), so we focus here on the (still extensive) tools proposed for forecasting specifically. There is no globally optimal ensemble-building method; rather, specific situations call for particular methods (Winkler 1986, Gneiting and Katzfuss 2014) and it will often be worthwhile to construct multiple ensemble models (Ray and Reich 2018). However, caution must be exercised in selecting ensemble building tools, as even commonly used methods have limitations that are often not fully appreciated (Hora 2004, Gneiting et al. 2005, Tebaldi and Knutti 2007, Yao et al. 2018). Here we discuss ensembles of quantitative predictions, although the concepts carry over to classification predictions. And given the context of probabilistic ecological forecasting, we focus on building ensembles that produce probability distributions (as CDFs or PDFs/PMFs).

*Simple averaging*

Generally, the most basic ensemble model is a naïve, equal-weighted average of all models included in the system (Table 3). Although simple, the equal-weight ensemble provides an informative benchmark against which more complex ensembles can be compared (Hora 2010, Ranjan and Gneiting 2010, Ray and Reich 2018). Interestingly, in some situations, the simple equal-weight average ensemble can even be one of the better models from the full set (McGowan et al. 2018). The next step up in complexity is to use unequal weighting, for example based on the predictive performance of the models over the training data set (Table 2), which gives more weight to the models that were better during validation, but does not properly account for model uncertainty and assumes that all models were on equal footing before the validation procedure.

*Bayesian model averaging*

Bayesian model averaging (BMA) accounts for model uncertainty and pre-existing information about the skill of models by including a prior distribution on the set of possible models and integrating over it in producing the posterior distribution of interest (Table 3; Draper 1995, Hoeting et al. 1999). As such, it provides a powerful approach for ensemble building, producing ensembles that are more skilled in a likelihood sense than all of the solo models (Madigan and Raftery 1994, Raftery et al. 2005). In the context of flexibly focused forecasting, BMA can operate on any utility to calculate model probabilities, and so it is possible to construct an ensemble based on scoring functions other than the log score.

Although BMA is a powerful and popular technique in forecasting (Dietz et al. 2018), it does exhibit some critical limitations. Importantly, despite often being compared to model combination techniques (like stacking, see *Stacked generalization*), BMA is actually more akin to a soft version of Bayesian Model Selection (BMS), operating under the assumption that the true model exists within those included (Minka *unpublished*). Formally, this means that BMA is appropriate under a model set (as defined in **Appendix A**) that is *closed* (includes the true data generating model, DGM) but not under a model space that is *complete* (could include the DGM but does not) or *open* (could not include the DGM, for example because it is too complex; Bernardo and Smith 1994, Monteith et al. 2011, Yao et al. 2018). In the open and complete cases, BMA will asymptotically select the one model that is best with respect to its score, resulting in a winner being selected rather than an ensemble being made from multiple models (Yao et al. 2018). Indeed, as sample size increases, BMA actually converges on BMS, thus the problem becomes worse with more data. This is clearly a concern in a forecasting space, where data streams grow over time under the prequential approach. Further, the marginal likelihood (and thus model posterior) is sensitive to the prior used within each model, which can result in among-model bias and related issues (Yao et al. 2018).

*Stacked generalization*

A dominant method of forecast ensemble building is known as stacked generalization (or *stacking*), which involves formalizing the ensemble as a meta-model of the component models (Table 3, Wolpert 1992, Breiman 1996, Clarke 2003, Gneiting and Katzfuss 2014). Stacking is often referred to as *model combination* or *aggregation* (Gneiting and Ranjan 2013) and is functionally a second-level model that takes the results of the main (first-level) models as given and combines them in an optimal fashion with respect to the training data (Wolpert 1992). Stacking is quite broad, notably here encompassing generalized regression models including non-homogeneous Gaussian and logistic regressions, which are commonly used in forecasting (Gneiting et al. 2005, Wilks 2006, Wilks and Hamill 2007). As with any statistical model, the mathematical and computational options for a stacking meta-model are vast and decisions should be based on the specific task at hand. In the context of probabilistic forecasting, a logical feature to base an ensemble upon is the resulting probability distribution (CDF, PDF/PMF) and it would be sensible to use the same scoring function (or set of functions) to evaluate the models for ensemble building as was used to score the component models for validation (Clarke 2003, Gneiting and Ranjan 2013, Yao et al. 2019).

As a tactical approach (Dormann et al. 2018), stacking models can be fit using a variety of methods, the most commonly used are cross-validation and maximum likelihood (Clarke 2003, Gneiting et al. 2005). The CDC, for example, uses leave-one-out cross-validation combined with a degenerate Expectation-Maximization routine to fit their implementation of model stacking (Ray and Reich 2018, Reich et al. 2019). Recent work has brought Bayesian tools to bear in stacking, with promising results (Table 3; Monteith et al. 2011, Clyde and Iversen 2013, Le and Clark 2017, Yao et al. 2018).

*Pooling equations*

Thus far, mirror the forecasting literature, we have focused on ensembles based on weighted linear combinations of base models (a.k.a. a *linear pool*; Table 3). While these linear pools tend to perform better than individual models, they are actually suboptimal themselves because they are uncalibrated with respect to the observed distribution, even if the individual models were all properly calibrated (Hora 2004, Ranjan and Gneiting 2010). As a result, weighted linear combination ensembles fail to be flexibly dispersive, even in well-defined cases that call for it (Gneiting and Ranjan 2013, Gneiting and Katzfuss 2014). To address this shortcoming, non-linear recalibration methods have been developed, most notably the spread-adjusted and beta-transformed linear pool equations (Table 3; Dawid et al. 1995, Ranjan and Gneiting 2010, Gneiting and Ranjan 2013). However, it should be noted that in settings where training data are relatively scarce (often the case in ecological studies) a standard linear pool might be the optimal method, due to parsimony concerns, because estimating non-linear pooling equations requires more degrees of freedom (Gneiting and Ranjan 2013).

***WORKING HERE***

**Worked Example: Kangaroo Rat Population Counts**

Implementation

scoringRules package (Jordan et al. 2018a, Jordan et al. 2018b) in R (R Core Team 2018)

**Discussion**

Ongoing development of the field

Kernel Scoring: Further, there is ongoing development (albeit with minimal application to date) of general (e.g., kernel-based) scoring rules that admit an exceptionally wide range of possible strictly proper scoring functions (Dawid 1998, Dawid 2006, Gneiting and Raftery 2007)

Areas where ecological forecasting applications can help push the field generally

Validation for time series with non-linear dynamics and strong autocorrelation

Kernel scoring rules

**Acknowledgements**

The motivating study—the Portal Project—has been funded nearly continuously since 1977 by the National Science Foundation, most recently by DEB-1622425 to S. K. M. Ernest. Much of this work was supported by the Gordon and Betty Moore Foundation’s Data-Driven Discovery Initiative through Grant GBMF4563 to E. P. White.

**Literature Cited**

Arlot, S. and A. Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* **4**:40-79. 10.1214/09-SS054

Armstrong, J. S., and M. C. Grohman. 1972. A comparative study of methods for long-range market forecasting. *Management Science* **19**:211-221. 10.1287/mnsc.19.2.211

Armstrong, J. S. 1985. *Long-range forecasting*. Wiley Interscience. New York, New York, USA.

Beaulieu‐Jones, B. K. and C. S. Greene. 2017. Reproducibility of computational workflows is automated using continuous analysis. *Nature Biotechnology* **35**:342-346. 10.1038/nbt.3780

Benedetti, R. 2010. Scoring rules for forecast verification. *Monthly Weather Review* **138**:203-211. 10.1175/2009MWR2945.1.

Bergmeir, C. and J. M. Benitez. 2012. On the use of cross-validation for time series predictor evaluation. *Information Sciences* **191**:192-213. 10.1016/j.ins.2011.12.028

Bergmeir, C., R. J. Hyndman, and B. Koo. 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis* **120**:70-83. 10.1016/j.csda.2017.11.003

Bernardo, J. M. and A. F. M. Smith. 1994. *Bayesian Theory*. John Wiley & Sons, New York, NY, USA.

Breiman, L. 1996. Stacked Regressions. *Machine Learning* **24**:49-64. 10.1007/BF00117832

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**:1-3.

Briggs, W. and D. Ruppert. 2005. Assessing the skill of yes/no predictions. *Biometrics* **61**:799-807. 10.1111/j.1541-0420.2005.00347.x

Candille, G. and O. Talagrand. 2005. Evaluation of probabilistic prediction systems of a scalar variable. *Quarterly Journal of the Royal Meteorological Society* **131**:2131-2150. 10.1256/qj.04.71

Clark, J. S., S. R. Carpenter, M. Barber, S. Collins, A. Dobson, J. A. Foley, D. M. Lodge, M. Pascual, R. Pielke Jr., W. Pizer, C. Pringle, W. V. Reid, K. A. Rose, O. Sala, W. H. Schlesinger, D. H. Wall, and D. Wear. 2001. Ecological forecasts: an emerging imperative. *Science* **293:**657–660. 10.1126/science.293.5530.657

Clarke, B. 2003. Comparing Bayes Model Averaging and stacking when model approximation error cannot be ignored. *Journal of Machine Learning Research* **4**:683-712. 10.1162/153244304773936090

Clyde, M. A. and E.S. Iversen. 2013. Bayesian Model Averaging in the ℳ-Open framework. In *Bayesian Theory and Applications*, P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens (Eds.). Oxford University Press, Oxford, England. pp. 483-498.

Conn, P. B., D. S. Johnson, P. J. Williams, S. R. Melin, and M. B. Hooten. 2018. A guide to Bayesian model checking for ecologists. *Ecological Monographs* **88**:526–542. 10.1002/ecm.1314

Czado, C., T. Gneiting, and L. Held. 2009. Predictive model assessment for count data. *Biometrics* **65**:1254-1261. 10.1111/j.1541-0420.2009.01191.x

D’Agostino A., K. McQuinn, and K. Whelan. 2012. Are some forecasters really better than others? *Journal of Money, Credit, and Banking* **44**:715–32. 10.1111/j.1538-4616.2012.00507.x

Dawid, A. P. 1984. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)* **147**:278-292. 10.2307/2981683

Dawid, A. P. 1998. Coherent Measures of Discrepancy, Uncertainty and Dependence, with Applications to Bayesian Predictive Experimental Design. Research Report 139, University College London, Dept. of Statistical Science.

Dawid, A. P. 2006. The Geometry of Proper Scoring Rules. Research Report 268, University College London, Dept. of Statistical Science.

Dawid, A. P., M. H. DeGroot, and J. Mortera. 1995. Coherent combination of experts’ opinions (with discussion and rejoinder). *Test* **4**:263-313. 10.1007/BF02562628

de Finetti, B. 1962. Does It make sense to speak of 'Good Probability Appraisers'?. In *The Scientist Speculates: An Anthology of Partly- Baked Ideas*, I. J. Good (Ed.). Basic Books, New York. pp 357-363.

Diebold, F. X. 2015. Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business and Economic Statistics* **33**:1-1. 10.1080/07350015.2014.983236

Diebold F. X. and R. S. Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**:253-263. 10.1080/07350015.1995.10524599

Diebold, F. X., T. A. Gunther, and A. S. Tay. 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**:863-883. 10.2307/2527342

Dietze, M. 2017. Ecological Forecasting. Princeton University Press, Princeton, New Jersey, USA.

Dietze, M. C., A. Fox, L. M. Beck-Johnson, J. L. Betancourt, M B. Hooten, C. S. Jarnevich, T. H. Keitt, M. A. Kenney, C. M. Laney, L. G. Larsen, H. W. Loescher, C. K. Lunch, B C. Pijanowski, J. T. Randerson, E. K. Read, A. T. Tredennick, R. Vargas, K. C. Weathers, and E. P. White. 2018. Iterative near-term ecological forecasting: needs, opportunities, and challenges. *Proceedings of the Natural Academy of Sciences* **115**:1424-1432. 10.1073/pnas.1710231115

Dormann, C. F., J. M. Calabrese, G. Guillera‐Arroita, E. Matechou, V. Bahn, K. Bartoń, C. M. Beale, S. Ciuti, J. Elith, K. Gerstner, J. Guelat, P. Keil, J. J. Lahoz‐Monfort, L. J. Pollock, B. Reineking, D. R. Roberts, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, S. N. Wood, R. O. Wüest, F. Hartig. 2018. Model averaging in ecology: a review of Bayesian, information‐theoretic, and tactical approaches for predictive inference. *Ecological Monographs* **88**:485-504. 10.1002/ecm.1309

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B: Methodological* **57**:45-97. 10.2307/2346087

Epstein, E. S. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* **8**:985-987. 10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2

Fildes, R. 1992. The evaluation of extrapolative forecasting methods. *International Journal of Forecasting* **8**:81-98. 10.1016/0169-2070(92)90009-X

Fildes, R., & Makridakis, S. 1995. The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review* **63**:289-308. 10.2307/1403481

Fruhwirth-Schnatter, S. 1996. Recursive residuals and model diagnostics for normal and non-normal state space models. *Environmental and Ecological Statistics* **3**:291-309. 10.1007/BF00539368

Garthwaite, P. H., J. B. Kadane, and A. O’Hagan. 2005, Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* **100**:680-700. 0.1198/016214505000000105

Gneiting, T., A. E. Raftery, A. H., Westveld III, and T. Goldman. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review* **133**:1098-1118. 10.1175/MWR2904.1

Gneiting, T., F. Balabdaoui, and A. E. Raftery. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **69**:243-268. 10.1111/j.1467-9868.2007.00587.x

Gneiting T. and A. E. Raftery. 2005. Weather forecasting with ensemble methods. *Science* **310**:248-49. 10.1126/science.1115255

Gneiting, T. and A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**:359-378. 10.1198/016214506000001437.

Gneiting, T. 2011. Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**:746-762. 10.1198/jasa.2011.r10138

Gneiting T. and R. Ranjan. 2013. Combining predictive distributions. *Electronic Journal of Statistics* **7**:1747-82. 10.1214/13-EJS823

Gneiting, T. and M. Katzfuss. 2014. Probabilistic forecasting. *Annual Review of Statistics and Its Application* **1**:125-151. 10.1146/annurev-statistics-062713-085831

Good, I. J. 1952. Rational decisions. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **14**:107-114. 10.1111/j.2517-6161.1952.tb00104.x

Good, I. J. 1971. Comment on “Measuring Information and Uncertainty,” by R. J. Buehler. In *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott, eds. Holt, Rinehart and Winston, Toronto, Canada. pp. 337–339

Hamill, T. M. 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting* **14**:155-167. 10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2

Hamill, T. M. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review* **129**:550-560. 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

Hastie T., R. Tibshirani, and J. Friedman. 2011. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, New York, New York, USA.

Hersbach, H. 2000. Decomposition of the Continuous Ranked Probability Score for ensemble prediction systems. *Weather and Forecasting* **15**:559-570. 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hoeting, J. A., D. M. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: a tutorial. *Statistical Science* **14**:382-401. 10.1214/ss/1009212519

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

Hora, S. C. 2004. Probability judgements for continuous quantities: Linear combinations and calibration. *Management Science* **50**:597-604. 10.1287/mnsc.1040.0205

Hora, S. C. 2010. An analytic method for evaluating the performance of aggregation rules for probability densities. *Operations Research* **58**:1440-1449. 10.1287/opre.1100.0789

Hyndman R. J. and A. B Koehler. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* **22**:679-688. 10.1016/j.ijforecast.2006.03.001

Jewson, S. The problem with the Brier score. 2018. *arXiv*. arXiv:physics/0401046

Jordan A., F. Krüger F, and S. Lerch. 2018a. Evaluating Probabilistic Forecasts with scoringRules. *arXiv*. arXiv:1709.04743

Jordan A., F. Krüger, and S. Lerch. 2018b. scoringRules: Scoring Rules for Parametric and Simulated Distribution Forecasts. R package version 0.9.5, URL https://CRAN.R-project.org/package=scoringRules.

Jose, V. R. 2009. A characterization for the spherical scoring rule. *Theory and Decision* **66**:263-281. 10.1007/s11238-007-9067-x

Krüger, F., S. Lerch, T. Thorarinsdottir, and T. Gneiting. 2019. Predictive inference based on Markov Chain Monte Carlo output. *arXiv*. arXiv:1608.06802

Kruschke, J. K. 2015. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. Academic Press, San Diego, California, USA.

Le, T. and B. Clarke. 2017. A Bayes interpretation of stacking for ℳ-complete and ℳ-open settings. *Bayesian Analysis* **12**:807-829. 10.1214/16-BA1023

Madigan, D. and A. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam’s Window. *Journal of the American Statistical Association* **89**:1535-1546. 10.1080/01621459.1994.10476894

Makridakis, S. 1986. The art and science of forecasting; an assessment and future directions. *International Forecasting* **2**:15-39. 10.1016/0169-2070(86)90028-2

Makridakis, S. and Winkler, R. L. 1989. Sampling distributions of post-sample forecasting errors. *Journal of the Royal Statistical Society, Series C: Applied Statistics* **38**:331-342. 10.2307/2348063

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T.  Mills, J. K. Ord, and L. F. Simmons. 1993. The M2 competition: a real life judgmentally-based forecasting study. *International Journal of Forecasting* **9**:5-29. 10.1016/0169-2070(93)90044-N

Matheson, J. E., and R. L. Winkler. 1976. Scoring rules for continuous probability distributions. *Management Science* **22**:1087-1095. doi.org/10.1287/mnsc.22.10.1087

McCarthy, J. 1956. Measures of the value of information. *Proceedings of the National Academy of Sciences* **42**:654-655. 10.1073/pnas.42.9.654

McGowan, C. J., M. Biggerstaff, M. A. Johansson, K. M. Apfeldorf, M. Ben-Nun, L. Brooks, M. Convertino, M. Erraguntla, D. C. Farrow, J. Freeze, S. Ghosh, S. Hyun, S. Kandula, J. Lega, Y. Liu, N. Michaud, H. Morita, J. Niemi, N. Ramakrishnan, E. L. Ray, N. G. Reich, P. Riley, J. Shaman, R. Tibshirani, A. Vespignani, Q. Zhang, C. Reed, and The Influenza Forecasting Working Group. 2019. Collaborative efforts to forecast seasonal influenza in the United States, 2015-2016. 2019. *Scientific Reports* **9**:683. 10.1038/s41598-018-36361-9

Monteith, K., J. L. Carroll, K. Seppi and T. Martinez. 2011. Turning Bayesian model averaging into Bayesian model combination. *Proceedings of the International Joint Conference on Neural Networks* 2657-2663. 10.1109/IJCNN.2011.6033566

Murphy, A. H. 1969. On the ‘‘ranked probability score.’’ *Journal of Applied Meteorology* **8**:988-989. 10.1175/1520-0450(1969)008<0988:OTPS>2.0.CO;2

Murphy, A. H. 1971: A note on the ranked probability score. *Journal of Applied Meteorology* **10**:155-156. 10.1175/1520-0450(1971)010<0155:ANOTRP>2.0.CO;2

Murphy, A. H. 1973. Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology* **12**:215-223. 10.1175/1520-0450(1973)012<0215:HASSFP>2.0.CO;2

Pack, D. J. 1990. In defense of ARIMA modeling. *International Journal of Forecasting* **6**:211-218. 10.1016/0169-2070(90)90006-W

Pant, P. N. and Starbuck, W. H. 1990. Innocents in the forest: forecasting and research methods. *Journal of Management* **16**:433-460. 10.1177/014920639001600209

Pennekamp, F., M. W. Adamson, O. L. Petchey, J.-C. Poggiale, M. Aguiar, B. W. Kooi, D. B. Botkin, and D. L. DeAngelis. 2017. The practice of prediction: What can ecologists learn from applied, ecology-related fields? *Ecological Complexity* **32**:156-167. 10.1016/j.ecocom.2016.12.005

Petchey, O. L., M. Pontarp, T. M. Massie, S. Kefi, A. Ozgul, M. Weilenmann, G. M. Palamara, F. Altermatt, B. Matthews, J. M. Levine, D. Z. Childs, B. J. McGill, M. E. Schaepman, B. Schmid, P. Spaak, A. P. Beckerman, F. Pennekamp and I. S. Pearse. 2015. The ecological forecast horizon, and examples of its uses and determinants. *Ecology Letters* **18**:597-611. 10.1111/ele.12443

R Core Team. 2018. R: A language and environment for statistical computing. v3.5.1. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raftery, A. E. T. Gneiting, F. Balabdaoui, and M. Polakowski. 2005. Using Bayesian Model Averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**:1155-1174. 10.1175/MWR2906.1

Ranjan R. and T. Gneiting. 2010. Combining probability forecasts. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **72**:71-91. 10.1111/j.1467-9868.2009.00726.x

Ray, E. L. and N. G. Reich. 2018. Prediction of infection disease epidemics via weighted density ensembles. *PLoS Computational Biology* **14**:e1005910. 10.1371/journal.pcbi.1005910

Reich, N. G., C. J. McGowan, T. K. Yamana, A. Tushar, R. L. Ray, D. Osthus, S. Kandula, L. C. Brooks, W. Crawford-Crudell, G. C. Gibson, E. Moore, R. Silva, M. Biggerstaff, M. A. Johansson, R. Rosenfeld, and J. Shaman. 2019. A collaborative multi-model ensemble for real-time influenza season forecasting in the U.S. *bioRxiv* 10.1101/566604

Roby, T. B. 1965. *Belief States: A Preliminary Empirical Study*. Decision Science Laboratory, United States Air Force. L.G. Hascom Field, Bedford, Massachusetts, USA.

Rosenblatt, M. 1952. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* **23**:470-472. 10.1214/aoms/1177729394

Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66**:783-801. 10.1080/01621459.1971.10482346

Selten, R. 1998. Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics* **1**:43-62. 10.1023/A:1009957816843

Sigrist, F., H. R. Kunsch, and W. A. Stahel. 2012. A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *The Annals of Applied Statistics* **6**:1452-1477. 10.1214/12-AOAS564

Simonis, J. L., R. B. Harrison, S. T. Long, D. R. Rabon Jr., W. T. Waddell, and L. J. Faust. 2018. Managed movement increases metapopulation viability of the endangered red wolf. *The Journal of Wildlife Management* **82**:573-582.10.1002/jwmg.21397

Smith, J. Q. 1985. Diagnostic checks of non-standard time series models. *Journal of Forecasting* **4**:283-291. 10.1002/for.3980040305

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **36**:111-147. 10.1111/j.2517-6161.1974.tb00994.x

Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **39**:44-47. 10.1111/j.2517-6161.1977.tb01603.x

Tashman, T. J. 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* **16**:437-450. 10.1016/S0169-2070(00)00065-0

Tashman, L. J. and J. Hoover. 2001. An evaluation of forecasting software. In *Principles of Forecasting*, J. S. Armstrong, ed. Kluwer Academic Publishers, Norwell, Massachusetts, USA. pp. 651-676.

Tebaldi, C. and R. Knutti. 2007. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **365**:2053-275. 10.1098/rsta.2007.2076

Velaquez, J. A., F. Anctil, and C. Perrin. 2010. Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. *Hydrology and Earth Systems Sciences* **12**:2303-2317. 10.5194/hess-14-2303-2010

Ward, E. J., E. E. Holmes, J. T. Thorson, and B. Collen. 2014. Complexity is costly: a meta-analysis of parametric and non-parametric methods for short-term population forecasting. *OIkos* **123**:652-661. 10.1111/j.1600-0706.2014.00916.x

Wecker, W. B. 1989. Comment: Assessing the accuracy of time series model forecasts of count observations. *Journal of Business and Economic Statistics* **7**:418-419. 10.2307/1391640

White, E. P., G. M. Yenni, S. Taylor, E. Christensen, E. Bledsoe, J. L. Simonis, and S. K. M. Ernest. 2019. Developing an automated iterative near-term forecasting system for an ecological study. to Methods in Ecology and Evolution **10**:332-344. 10.1111/2041-210X.13104

Wilks, D. S. 2006. Comparison of ensemble-MOS methods in the Lorenz ’96 setting. *Meteorological Applications* **13**:243-256. 10.1017/S1350482706002192

Wilks, D. S. 2010. Sampling distributions of the Brier score and Brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society* **136**: 2109-2118. 10.1002/qj.709.

Wilks, D. S. and T. M. Hamill. 2007. Comparison of Ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review* **135**:2379-2390. https://doi.org/10.1175/MWR3402.1

Winkler, R. L. 1967. The quantification of judgment: some methodological suggestions. *Journal of the American Statistical Association* **62**:1105-1120. 10.1080/01621459.1967.10500920

Winkler R. L., A. H. Murphy. 1968. “Good” probability assessors. *Journal of Applied Meteorology* **7**:751-758. 10.1175/1520-0450(1968)007<0751:PA>2.0.CO;2

Winkler, R. L. 1977. Rewarding expertize in probability assessment. In *Decision Making and Change in Human Affairs*, H. Jungermann and G. de Zeeuw, eds. D. Reidel, Dordrecht, Holland. pp. 127-140.

Winkler, R. L. 1986. Comment on “Combining probability distributions: A critique and an annotated bibliography”. *Statistical Science* **1**:138-140. 10.1214/ss/1177013828

Winkler, R. L. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**:1-60. 10.1007/BF02562681

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* **5**:241-259. 10.1016/S0893-6080(05)80023-1

Yao, Y., A. Vehtari, D. Simpson, and A. Gelman. 2018. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis* **13**:917-1007. doi:10.1214/17-BA1091

**Table 1**. Calculations of the Probability Integral Transform (PIT).

|  |  |
| --- | --- |
| Type | Equation |
| Continuous Original |  |
| Discrete Randomized | where ) |
| Discrete Non-randomized |  |

: sample, : predictive distribution, : observed value, : cumulative distribution function.

The original and randomized discrete individual PIT values are calculated observation-by-observation, whereas the non-randomized PIT is constructed in aggregate by integrating the CDF of the conditional randomized PIT () over the observed values (Czado et al. 2009).

**Table 2.** Commonly used scoring rules, all defined as positively oriented.

|  |  |
| --- | --- |
| Name | Formula |
| Log |  |
| Quadratic (Brier) |  |
| Power |  |
| Spherical |  |
| Pseudo-spherical |  |
| Ranked Probability |  |

: sample, : predictive distribution, : observed value, : cumulative distribution function, : probability density or mass function, : -norm of (), : generalization parameter, : the characteristic function (). For continuous variables, summations are replaced with integrals.

**Table 3.** Probabilistic forecasting ensemble calculations.

|  |  |
| --- | --- |
| Name | Equation |
| Naïve | where |
| Simple weight | where |
| BMA | where |
| Stacking | where |
| Bayesian Model Stacking | where |
| Spread-adjusted linear pool | where can be any of the above equations and can be 0 |
| Beta-transformed linear pool | where can be any of the above equations and can be 0 |

: model (in total), : sample, : predictive distribution, : CDF, : PDF or PMF, : weight of model (all , and ), : score of across the training data set, : prior distribution of the models, : additive error term (can be a complex term, as in non-homogeneous regression, Gneiting et al. 2005), : observed value, : beta function with parameters and , , and : unique median of . All methods are written as dynamic across samples; simplified versions integrate (average) across samples.

[General forecasting context figure]

**Figure 1.**

[Validation figure: procedures, prequential validation and rolling origins]

**Figure 2.**

[critical graphic evaluations: time series, XY, PIT]

**Figure 3.**

**APPENDIX A**

*Statistical Definition of a Probabilistic Forecast*

The distributions from all potential models (the distributions explicitly explored as the model set, as well as distributions for all other models that *could have been* evaluated but were not, which are part of the model space) and the generating distribution , (which may or may not be incorporated in the models) form the sample space . Then, defines a workable set of distributions on , in that it is closed under countability and complementarity ( is a “-algebra” of ) and is a general convex class of probability measures that exist on . A probabilistic forecast for our ecological variable is then any measure that exists on .

*Further Background on Scoring Rules*

A scoring function must be defined on the sample space and be able to take values on the extended real line (including negative and positive infinity), (Good 1952, de Finetti 1962). Scoring functions tend to be real-valued in their output, but can allow for infinite values for scores, as the logarithmic rule does (Good 1952). However, a scoring function must be measurable with respect to (the workable space) and *quasi-integrable* (have a defined integral for at least one of its positive or negative parts; Bauer 2001) with respect to all of (the full class of possible convex probability measures) (Winkler 1967, Savage 1971, Gneiting and Raftery 2007).

Recognizing that the actual observations are a single realization of the true process, the expected value of across the distribution of possible observations is

A1

Further, although scoring rules are generally framed in terms of probabilistic distributions, they are still defined under the case of a point forecast. For example, a scoring function can be used to measure the score for an observed value and the expected value of the forecast distribution:

A2

A key set of characteristics about scoring rules are encompassed in the concept of *propriety* (Winkler 1977, Dawid 1998, Gneiting and Raftery 2007). If a scoring rule is *proper*, then the function is convex and the maximal (best) score value is achieved by using the true generating probability distribution (Brier 1950, Good, 1952, Winkler and Murphy 1968). That is, is proper if

for all and A3

Proper scoring rules encourage honest forecasts that maximize reward (de Finetti 1962, Winkler 1977, Garthwaite et al. 2005). Further, a *strictly proper* scoring rule requires a strictly convex scoring function with a unique maximum, which must score a forecast distribution as best if, and only if, the distribution suggests the observed value as the forecast (Savage 1971, Gneiting and Raftery 2007). The score’s unique optimum is then located at the true distribution:

if and only if A4

The propriety of scoring functions holds through linear (additive and multiplicative) transformations. That is, if is a proper or strictly proper scoring rule defined for a probability distribution and observation , and is

A5

then is also proper or strictly proper, as long as and is integrable with respect to .

*Standard Deviation Estimates in the Diebold-Mariano Test*

Diebold and Mariano (1995) defined the general equation for the standard deviation estimate as

A6

where is a consistent estimator of the variance. If the forecasts’ score values are independent, a simple equation can be used for :

A7a

In the presence of autocorrelation, becomes the weighted sum of the sample covariances:

A7b

where is the lag window, is the truncation lag, and

A8

(Diebold and Mariano 1995, Diebold 2015). These approximations can require substantial test data sizes to ensure robustness and bootstrapping (permutation) approaches to the D-M test can mitigate sample size issues (D’Agostino et al. 2012).

*Empirical Calculation of Continuous Ranked Probability Score*

Historically, computation of the Continuous Ranked Probability Score proved difficult (Krüger et al. 2019). However, recent work has shown that it can be empirically calculated as

A9

where and are independent random variables with distribution Gneiting and Raftery 2007). This calculation can be approximated using a series of draws from , , such as from MCMC (Gneiting and Raftery 2007, Krüger et al. 2019):

A10

**Literature Cited**

Bauer, H. 2001. *Measure and Integration Theory*. Walter de Gruijter, Berlin, Germany.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**:1-3.

D’Agostino A., K. McQuinn, and K. Whelan. 2012. Are some forecasters really better than others? *Journal of Money, Credit, and Banking* **44**:715–32. 10.1111/j.1538-4616.2012.00507.x

Dawid, A. P. 1998. Coherent Measures of Discrepancy, Uncertainty and Dependence, with Applications to Bayesian Predictive Experimental Design. Research Report 139, University College London, Dept. of Statistical Science.

de Finetti, B. 1962. Does It make sense to speak of 'Good Probability Appraisers'?. In *The Scientist Speculates: An Anthology of Partly- Baked Ideas*, I. J. Good (Ed.). Basic Books, New York. pp 357-363.

Diebold, F. X. 2015. Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business and Economic Statistics* **33**:1-1. 10.1080/07350015.2014.983236

Diebold F. X. and R. S. Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**:253-263. 10.1080/07350015.1995.10524599

Garthwaite, P. H., J. B. Kadane, and A. O’Hagan. 2005, Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* **100**:680-700. 0.1198/016214505000000105

Gneiting, T. and A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**:359-378. 10.1198/016214506000001437.

Good, I. J. 1952. Rational decisions. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **14**:107-114. 10.1111/j.2517-6161.1952.tb00104.x

Krüger, F., S. Lerch, T. Thorarinsdottir, and T. Gneiting. 2019. Predictive inference based on Markov Chain Monte Carlo output. *arXiv*. arXiv:1608.06802

Savage, L. J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66**:783-801. 10.1080/01621459.1971.10482346

Winkler, R. L. 1967. The quantification of judgment: some methodological suggestions. *Journal of the American Statistical Association* **62**:1105-1120. 10.1080/01621459.1967.10500920

Winkler, R. L. 1977. Rewarding expertize in probability assessment. In *Decision Making and Change in Human Affairs*, H. Jungermann and G. de Zeeuw, eds. D. Reidel, Dordrecht, Holland. pp. 127-140.

Winkler R. L., A. H. Murphy. 1968. “Good” probability assessors. *Journal of Applied Meteorology* **7**:751-758. 10.1175/1520-0450(1968)007<0751:PA>2.0.CO;2