Running Head: Probabilistic ecological forecasting

Evaluating Probabilistic Ecological Forecasts

Juniper L. Simonis[1,2†], Ethan P. White[1], S. K. Morgan Ernest[1]

[1]Wildlife Ecology and Conservation, University of Florida

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

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

Abstract

Probabilistic near-term forecasting facilitates 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, many ecologists are unfamiliar with the widely used tools for evaluating probabilistic forecasts developed in other fields. We address this gap by reviewing the literature on probabilistic forecast evaluation from diverse fields including climatology, economics, and epidemiology. We present established practices for selecting evaluation data (end-sample hold out), graphical forecast evaluation (times-series plots with uncertainty, Probability Integral Transform plots), quantitative evaluation using scoring rules (log, quadratic, spherical, and ranked probability scores), and comparing scores across models (skill score, Diebold-Mariano test). We cover common approaches, highlight mathematical concepts to follow, and note decision points to allow application of general principles to specific forecasting endeavors. We illustrate these approaches with an application to a long-term rodent population time series currently used for ecological forecasting and discuss how ecology can continue to learn from and drive the cross-disciplinary field of forecasting science.

Keywords*:* continuous analysis, desert pocket mouse, ecological forecasting, end-sample holdout, forecast skill, hierarchical Bayes, prequential, score rule, time series, validation.

Introduction

Forecasting—predicting the future state of a system—is rapidly becoming an important focus of ecology (Clark et al. 2001, Pennekamp et al. 2017). Understanding the accuracy and precision of ecological forecasts is essential to improving models and using their results for decision making. Ecological forecasting has typically focused on evaluating forecasts based on point estimates – the expected value or average prediction for a state at some point in the future.

However, the uncertainty of forecasts is also essential for decision making and understanding how well models capture sources of variation (Dietz 2017). Probabilistic forecasts produce distributions of future state values allowing the predicted uncertainty to be incorporated into forecast evaluation (Dawid 1984, Dietze et al. 2018).

Properly evaluating probabilistic forecasts requires a unique set of tools and approaches. These methods have been well developed in disciplines with long forecasting histories including climatology, economics, and epidemiology, but are not familiar to many ecologists. We address this gap by reviewing the literature on probabilistic forecast evaluation from disciplines with established cultures, principles, and tools to help guide ecologists in the selection of best practices for assessing probabilistic forecast performance (Winkler 1977, Dawid 1984, Gneiting and Raftery 2007). After introducing notation and terminology, we present established practices for: 1) selecting data to hold out for evaluation; 2) graphical assessment of performance; 3) quantitative scoring of forecasts; and 4) comparing performance across models. We use these methods to analyze forecasts of a long-term study of desert rodent populations and provide simplified example code for readers to apply to their own systems.

Notation and Terminology

We base our coverage of forecast evaluation on the following notation and terminology for data, models, and approaches (Fig. 1). Consider a time series of samples n in $1 \ldots N$ of variable y (y_n in $y_{1:N}$), collected through time (t_n in $t_{1:N}$). y can be discrete or continuous and samples can be taken at fixed or variable intervals. The observed $y_{1:N}$ is but one realization drawn from the unknowable generating distribution $G_{1:N}$, where G_n is the distribution of possible states at t_n (Fig. 1a). The last datum is the *forecast origin* o (Tashman 2000). We use models m in $1 \ldots M$ to gain inference about $G_{1:N}$ and make forecasts p in $1 \ldots P$ of y after y_o, where the time between o and p is the *lead time* or *forecast horizon* ($t_{o \rightarrow (o+p)}$; Fig. 1a) and models predict samples $o+1$ to $o+P$ ($y_{(o+1):(o+P)}$) with a total forecast horizon of $t_{o \rightarrow (o+P)}$. Thus, each model m needs to fit $y_{1:o}$ then predict $y_{(o+1):(N+P)}$ with its distribution $H^{m}_{(o+1):(o+P)}$ across the horizon (Fig. 1a; **Appendix S1**). For fitting and predicting, we use data in hand to validate our models, iterating the evaluation over time via a probabilistic and sequential (*prequential* sensu Dawid 1984) approach to testing existing data, compared to validating models only after future data are collected (Makridakis et al. 1993). Prequential methods are well-defined, preferred in established fields (Dawid 1984), and implementable in ecological forecasting (Dietze et al. 2018).

Holding Out Data For Forecast Validation

The validation procedure defines how the data are split into those used to fit the model (*training data*) and those reserved to evaluate its predictions (*test data*). Because the goal of forecasting is predicting the next data in a time series (Dawid 1984), the dominant paradigm in forecasting validation is *end-sample holdout*, where the last k observations are used for testing (Fig. 1b; Fildes and Makridakis 1995, Tashman 2000). Simulation and empirical evaluations show that end-sample holdout methods produce realistic distributions for future data (Tashman 2000) and training and testing errors are also typically very weakly correlated (Makridakis 1986, Makridakis and Winkler 1989). Other approaches or modifications including cross-validation (Bergmeir et al. 2018) and end-sample holdout with buffers (Cerqueira et al. 2020) are also used in forecasting. Cross-validation, which selects test data from across the entire data set, can increase the number of evaluations, avoiding issues with few evaluations being an unstable estimate of model skill (Tashman 2000). Adding buffers to end-sample holdout has recently been proposed to address the influence of autocorrelation (Cerqueira et al. 2020). However, the main purpose of forecasting is to predict data starting at the time step following the last observation

(Fig. 1a), making standard end-sample holdout the closest validation approach to the forecasting task. Indeed, the best models validated using end-sample holdout tend to outperform those validated via cross-validation when tested on novel future data (Fildes and Makridakis 1995).

Using end-sample holdout, we define a break in our time series ($y_{1:N}$) using forecast origin t_o, resulting in a training set of o values ($y_{1:o}$) and a test set of $N-o=P$ values ($y_{(o+1):N}$). This break focuses validation on quantifying how well a model's forecast distribution $H_{(o+1):N}$ matches the observations in the test set $y_{(o+1):N}$, where matching is defined by a score (see **Scoring Functions**; Dawid 1984). To cover the range of expected values, the number of samples allocated to the test set (via the location of o) should cover at least the longest forecast horizon required by the main application (Tashman 2000). That is, if the model makes 12-month-ahead forecasts, the holdout data set should cover at least one year of observations.

One end-sample holdout results in a single forecast evaluation for each model, which can be insufficient for describing skill. This is especially true if the data display cyclic or seasonal dynamics, in which case performance of each forecast will vary as a function of its origin (Pack 1990). Therefore, we recommend using *rolling forecast origin* validation, where multiple forecasts are made with the origin moved forward in the series (Fig. 1b; Armstrong 1985). Rolling origins generate robust estimates of skill and facilitate analyses of skill as a function of factors like lead time (Makridakis and Winkler 1989). Larger holdouts allow for more forecasts of the target horizon, but may not be an option for shorter time series (Tashman 2000).

A critical decision for rolling origin evaluations is whether each step forward should include just an update to the data or if the model should also be re-fit (Tashman 2000). Although it is generally preferable to update the model at each step in the evaluation, re-optimization can be computationally intensive and requires technical knowledge not broadly available in ecology

(Tashman 2000). In prequential methods, however, iterative forecasts replace done-all-at-once evaluations, easing computational burdens (Dawid 1984, Dietze et al. 2018). This is aided via *continuous analysis* systems that re-run models when data are updated (White et al. 2019)—in essence, an automated system of rolling origin, fixed horizon, recalibrating end-sample holdout validations, to which each new (fixed origin end-sample holdout) validation is added (Fig. 1b).

Graphical Evaluation

Graphical evaluation provides key insight into model appropriateness over the training and test sets (Dietze 2017). In forecasting, where data are explicitly temporal, it is helpful to plot the time series of predictions and observed values with training data to show past dynamics (Fig. 1a). Ecological models often have multiple levels of uncertainty and non-linearities (Hooten and Hobbs 2015) not well summarized by quantiles, necessitating the plotting of distributions or representative draws (Dietze 2017). A plot of model residuals over time can highlight persisting temporal autocorrelations. In addition, a plot of predicted-vs.-observed values will ideally follow a 1:1 line with deviation appropriate to model uncertainty (Appendix S1: Fig. S1).

The *Probability Integral Transform* (PIT) is a diagnostic plot with a solid statistical basis and a long history in forecasting. It comprises the values of the predictive cumulative distribution functions (CDFs) evaluated at the observed values (Appendix S1: Table S1; Dawid 1984). If observed values match predictive distributions and the predictive distributions are continuous, the PIT has a standard uniform distribution (Dawid 1984), which can be checked informally using plots (Appendix S1: Fig. S1). The uniformity of the PIT is necessary but not sufficient for a forecast to match the generating distribution (Hamill 2001). PIT histograms and CDFs allow comparison to a uniform and deviations have specific meanings: skew indicates biased central tendency, U-shapes underdispersion, and hump-shapes overdispersion (Appendix S1: Fig. S1;

Gneiting et al. 2007). The PIT has been extended to discrete distributions via approximations that add noise (Smith 1985) or use a conditional CDF (Czado et al. 2009; Appendix S1: Table S1).

Scoring Rules

Scoring rules are quantitative measures of the fit of the forecast to the test data (Brier 1950; **Appendix S1**). The score (s) of how point observation y_n matches model m's forecast distribution (H_n^m) is measured using rule r's function S^r: $s_n^{rm} = S^r\left(H_n^m, y_n\right)$. A model's average score across multiple observations is $\bar{s}_{(o+1):N}^{rm}$ (Table 1). Here, we use a positive orientation: higher score is better. Although scores are typically framed in terms of distributions, they are defined for point forecasts and many simplify to classical point-based metrics. Key attributes of rules are encompassed in the concept of (*strict*) *propriety* (Dawid 1998; **Appendix S1**). A proper function is convex and optimizes at the true distribution; a strictly proper function is *strictly* convex and optimizes *only* at the true distribution (Good 1952, Winkler and Murphy 1968). Proper rules encourage forecasts to maximize reward and strictly proper rules ensure unique solutions (de Finetti 1962). Several strictly proper rules can handle discrete as well as continuous distributions (Table 1; Gneiting and Raftery 2007). Each rule has strengths and weaknesses and forecasters often use multiple rules to leverage their attributes (Ray and Reich 2018).

The *Log Score* is the logarithm of the predictive probability evaluated at the observed value (Table 1; Good 1952). The log score is the only proper rule that depends solely on the probability distribution at the 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). Although simple and popular, the log score can be *insensitive* to how far the true distribution is from the prediction and *hypersensitive* to small differences in probabilities (Selten 1998, Gneiting and

Raftery 2007), so caution should be used when employing it if rare values are observable.

The *Quadratic (Brier) Score* is the average squared error of the probability forecasts where the observations are either matched or not (Table 1; Brier 1950). It extends the mean squared error from point to distributional forecasts (Winkler 1996) and can be generalized to the *Power Score* (Table 1; Selten 1998). Weaknesses of the Brier score include that it is not local (it depends on events that did not happen), can result in counter-intuitive values for rare and very common events because it uses absolute differences, and can require many samples to account for inflation of score and skill score variance by autocorrelation (Wilks 20108).

The *Spherical Score* is strictly proper and symmetric, so named because it standardizes the probability to a point on the unit sphere via division by its Euclidean norm (Table 1; Roby 1965). In contrast to the log score, the spherical score is hypersensitive near medial probabilities, and thus incentivizes matching the central tendency of the predicted distribution (Selten 1998). As such, the spherical and log scores produce complementary information regarding model performance. Similar to the quadratic score, the spherical score can be generalized to the pseudospherical (Table 1; Gneiting and Raftery 2007).

The *Ranked Probability Score* (RPS) defines a squared function that compares CDFs of a forecast and observation over a discrete number of categories (Table 1; Epstein 1969). The RPS generalizes the binary quadratic score to more than two categories (Czado et al. 2009) and is expanded to continuous variables as the *Continuous RPS* (CRPS; Matheson and Winkler 1976), the integral of quadratic scores for binary forecasts at all real-valued thresholds (Table 1). Favorably, the RPS considers the shape and tendency of forecast distributions, is sensitive to distance (rewards distributions closer to the observation), uses the CDF (more stable than the PDF/PMF; Hersbach 2000), and generalizes mean absolute error (facilitating comparison of point and probabilistic forecasts; Gneiting and Raftery 2007). Concerns with the RPS include its sensitivity to unusually large predicted or observed values (Candille and Talagrand 2005) and computation, the latter of which recent work alleviates (**Appendix S1**).

Comparing Model Scores

Once models have been scored on the same data with the same function, they can be quantitatively and statistically compared to each other as their scores form empirical distributions (Makridakis and Winkler 1989, Gneiting and Raftery 2007). The *skill score* ($\dot{\bar{s}}$) standardizes skill values for comparisons. The skill score of model m is $\dot{\bar{s}}_n^m = \frac{\bar{s}_n^m - \bar{s}_n^{ref}}{\bar{s}_n^{opt} - \bar{s}_n^{ref}}$, where $\bar{s}_n^{ref}$ is the score of a reference model (e.g., the marginal distribution of the predictand such as a smooth of historical values; Gneiting and Raftery 2007) and $\bar{s}_n^{opt}$ is the score of an ideal forecast (maximal value; Murphy 1973). Skill scores are equal to 0 for the reference forecast and 1 for an optimal forecast; a positive score means the forecast was better than the reference, a negative score means it was worse. Although skill scores provide standardized comparisons, they are generally not proper (see above) even if the underlying scoring function is proper (Murphy 1973).

Frequentist tests of forecasts are robust as long as correlations among scores are modeled (Makridakis and Winkler 1989). The *Diebold-Mariano (D-M) Test* is the main method for such comparisons and evaluates the significance of differences between forecast skill using z-tests that account for correlated errors (Diebold and Mariano 1995; **Appendix S1**). The test is based on the difference between scores for any two forecasts, which has an expected value of 0 under a null hypothesis of no difference. The formal test statistic is then the standardized mean difference, which has an expected standard normal distribution under the null (Diebold and Mariano 1995). Serial autocorrelation in scores is addressable using standard robust equations (**Appendix S1**).

While scores are typically aggregated across test data for quantitative comparisons, graphing sample-level scores and comparing across models can also provide useful insight (Gneiting et al. 2007). For example, plotting scores as a function of covariates can identify model differences associated with external forces. Similarly, plots of scores as a function of lead time allow comparison of how skill decays over the forecast horizon (Petchey et al. 2015). Graphical comparisons are bolstered through a cache of evaluations built via the prequential approach (Dawid 1984, Dietz et al. 2018, White et al. 2019), as apparent patterns may be artefactual.

Example: Pocket Mouse Population Counts

To demonstrate prequential ecological forecasting, we use a subset of data collected at a long-term study in the Chihuahuan Desert (AZ, USA; Brown 1998) that is actively used for ecological forecasting (White et al. 2019). Small mammals have been trapped on 24 plots with 49 traps per plot every four weeks since 1977 (512 trappings over the course of the study; **Appendix S2**). On some plots, the dominant genus *Dipodomys* (kangaroo rats) has been excluded. Here, we model counts of the desert pocket mouse (*Chaetodipus penicillatus*) in one kangaroo-rat exclosure plot (Fig. 2a). We forecast 12 counts (following White et al. 2019) from a true origin of sample 500 as if it were the final sample and compare them to the true observations from samples 501-512.

We fit three Bayesian time series models (**Appendix S2**) with the same right-truncated Poisson observation model with log-scale mean density ($\lambda = e^{x_n}$) and maximum of 49 (the number of traps; double captures are rare: ~0.01%) and one of three process models: random walk (RW), first-order autoregressive (AR(1)), and seasonal first-order AR (sAR(1); given the species' seasonal variation; Fig. 2a). We validated the models across a training period from sample 200 to 500 using rolling origin end-sample evaluation (Figs. 1, 2) beginning with a test origin of 300 and increasing in steps of 1 to a final test origin of 499, with test data being the subsequent 12 samples (up to and including 500). For the true origin (500), the test data were 501-512: a single realization of observations (Fig. 2a,b). We fit the models using Markov Chain Monte Carlo via JAGS (Plummer 2003) in R (R Core Team 2020) (**Data S1**) and used the log (for comparison to likelihood) and rank probability (to incorporate full predictive distributions) scores for evaluations (Table 1). We graphically assessed the fit of the rolling and true origin predictions using non-random discrete PIT histograms (Table 1, Czado et al. 2009). A simplified application of these methods is detailed in **Appendix S3** for translation to other systems.

Across the rolling-origin validation test sets, the random walk and sAR(1) were both well calibrated, albeit with a slight excess of variance, as evidenced by their slightly peaked PIT histograms (Fig. 2c). Comparatively, the AR(1)'s PIT histogram showed modality at the upper range, indicating negative bias (Fig. 2c; c.f. Appendix S1: Fig. S1). The sAR(1) was the best model with respect to both scoring functions across the rolling-origins (Fig. 2d). Yet for the final test, the AR(1) performed best (Fig. 2f) because its negative bias better matched the realized data (Fig. 2b). This provides an important lesson: the best long-term model (sAR(1)) was not best for the short term. Rather, the biased AR(1) was best in this specific evaluation of this case study.

Discussion

Taking a probabilistic approach to forecasting and evaluation is important for developing models that produce both accurate point predictions and useful estimates of uncertainty (Dietze 2017). In developing approaches to evaluate probabilistic forecasts ecologists can learn from fields with more established histories of forecasting. However, knowledge and skill transfer among disciplines is not one-way in the application of probabilistic forecasting to ecology (Pennekamp et al. 2017) and there are many active areas of research in forecasting science where ecologists can make important contributions (Dietze 2017). For example, ecological data often violate assumptions of forecast evaluation approaches due to non-normality, multiple levels of hierarchical variation, uncertainty in observations, feedbacks, non-linearities, and autocorrelation (Hooten and Hobbs 2015). Thus, while standard practices developed in other disciplines provide a foundation for quantitatively evaluating probabilistic ecological forecasts, ecology can help generalize these methods, develop new tools, and further the theory of probabilistic forecasting.

As ecology continues to develop its forecasting culture we envision a key next step being the incorporation of these probabilistic evaluation methods into iterative forecasting processes. Iterative forecasts involve a series of steps including selecting models, identifying a validation approach, fitting the models to available training data, generating predictions with uncertainties for test data, and then evaluating model predictions for the test data, with each iteration involving multiple components (Dietze et al. 2018). The fitting, predicting, and evaluating steps can be automated (White et al. 2019), and should use the probabilistic methods described in this review when possible, as opposed to point-prediction methods approaches. However, much of the true potential of prequential forecasting also involves direct researcher engagement with selecting and evaluating new models and continually improving methods for evaluation (Dietze et al. 2018). We hope that the graphical methods and evaluation approaches described here and further demonstrated in **Appendix S3** can help provide a route forward for these efforts. Finally, as new models are explored and incorporated into ecological forecasts, developing ensembles of forecasts will become increasingly important. There is much to learn from fields with more established forecasting cultures about best practices for ensembling probabilistic forecasts (e.g., Gneiting et al. 2007, Ray and Reich 2018).

Acknowledgements

This work was funded by the National Science Foundation (DEB-1622425 to Ernest and

DEB-1929730 to Ernest and White) and the Gordon and Betty Moore Foundation (GBMF4563 to White) and informed by discussions at the 2019 Ecological Forecasting Initiative Conference.

Literature Cited

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

Benedetti, R. 2010. Scoring rules for forecast verification. *Monthly Weather Review* **138**:203-211.

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.

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

Brown, J. H. 1998. The desert granivory experiments at portal. In *Experimental Ecology*, W. L. Resetarits, Jr. and J. Bernardo (Eds.). Oxford University Press, Oxford, UK. pp 71-95.

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.

Cerqueira, V., L. Torgo, and I. Mozetič. 2020. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning* **109**:1997-2028.

Clark, J. S., S. R. Carpenter, M. Barber, S. Collins, A. Dobson, et al. 2001. Ecological forecasts: an emerging imperative. *Science* **293:**657–660.

Czado, C., T. Gneiting, and L. Held. 2009. Predictive model assessment for count data. *Biometrics* **65**:1254-1261.

Dawid, A. P. 1984. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)* **147**:278-292.

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*, I. J. Good (Ed.). Basic Books, New York. pp 357-363.

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

Dietze, M. 2017. Ecological Forecasting. Princeton University Press, Princeton, N. J., USA.

Dietze, M. C., A. Fox, L. M. Beck-Johnson, J. L. Betancourt, M B. Hooten, et al. 2018. Iterative near-term ecological forecasting: needs, opportunities, and challenges. *Proceedings of the Natural Academy of Sciences* **115**:1424-1432.

Epstein, E. S. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* **8**:985-987.

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

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.

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

Good, I. J. 1952. Rational decisions. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **14**:107-114.

Hamill, T. M. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly*

Weather Review **129**:550-560.

Hersbach, H. 2000. Decomposition of the Continuous Ranked Probability Score for ensemble prediction systems. *Weather and Forecasting* **15**:559-570.

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

Makridakis, S. 1986. The art and science of forecasting; an assessment and future directions. *International Forecasting* **2**:15-39.

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.

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, et al. 1993. The M2 competition: a real life judgmentally-based forecasting study. *International Journal of Forecasting* **9**:5-29.

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

Murphy, A. H. 1973. Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology* **12**:215-223.

Pack, D. J. 1990. In defense of ARIMA modeling. *International Journal of Forecasting* **6**:211-218.

Pennekamp, F., M. W. Adamson, O. L. Petchey, J. C. Poggiale, M. Aguiar, et al. 2017. The practice of prediction: What can ecologists learn from applied, ecology-related fields? *Ecological Complexity* **32**:156-167.

Petchey, O. L., M. Pontarp, T. M. Massie, S. Kefi, A. Ozgul, et al. 2015. The ecological forecast horizon, and examples of its uses and determinants. *Ecology Letters* **18**:597-611.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing,* K. Hornik, F. Leisch, and A. Zeileis, eds. ISSN 1609-395X.

R Core Team. 2020. R: A language and environment for statistical computing. v4.0.3. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

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

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

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

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

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.

Tashman, T. J. 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* **16**:437-450.

White, E. P., G. M. Yenni, S. Taylor, E. Christensen, E. Bledsoe, et al. 2019. Developing an automated iterative near-term forecasting system for an ecological study. to Methods in Ecology and Evolution **10**:332-344.

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

Winkler R. L., A. H. Murphy. 1968. “Good” probability assessors. *Journal of Applied*

Meteorology **7**:751-758.

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. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**:1-60.

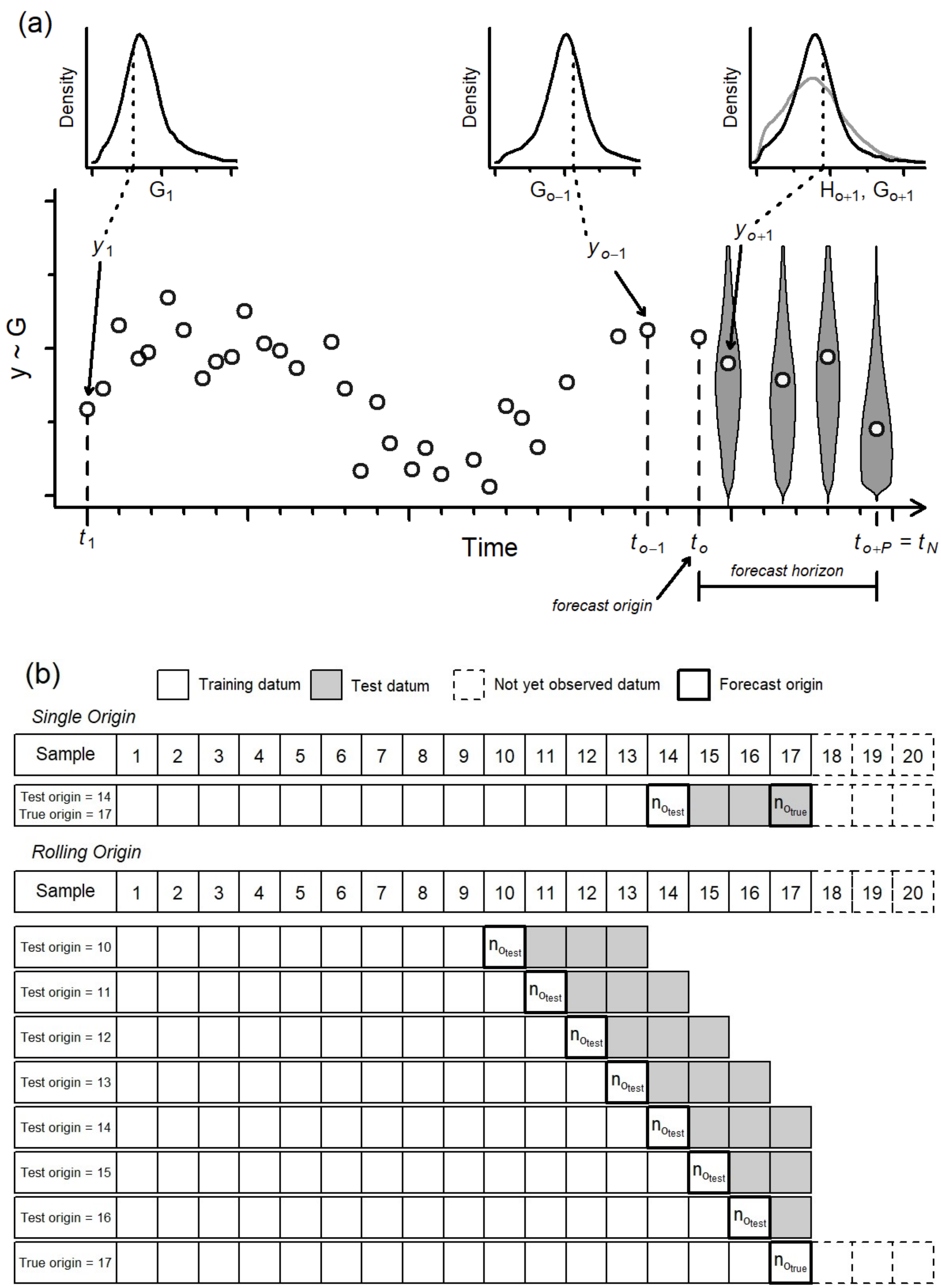
Table 1. Commonly used scoring rules, all defined as positively oriented.

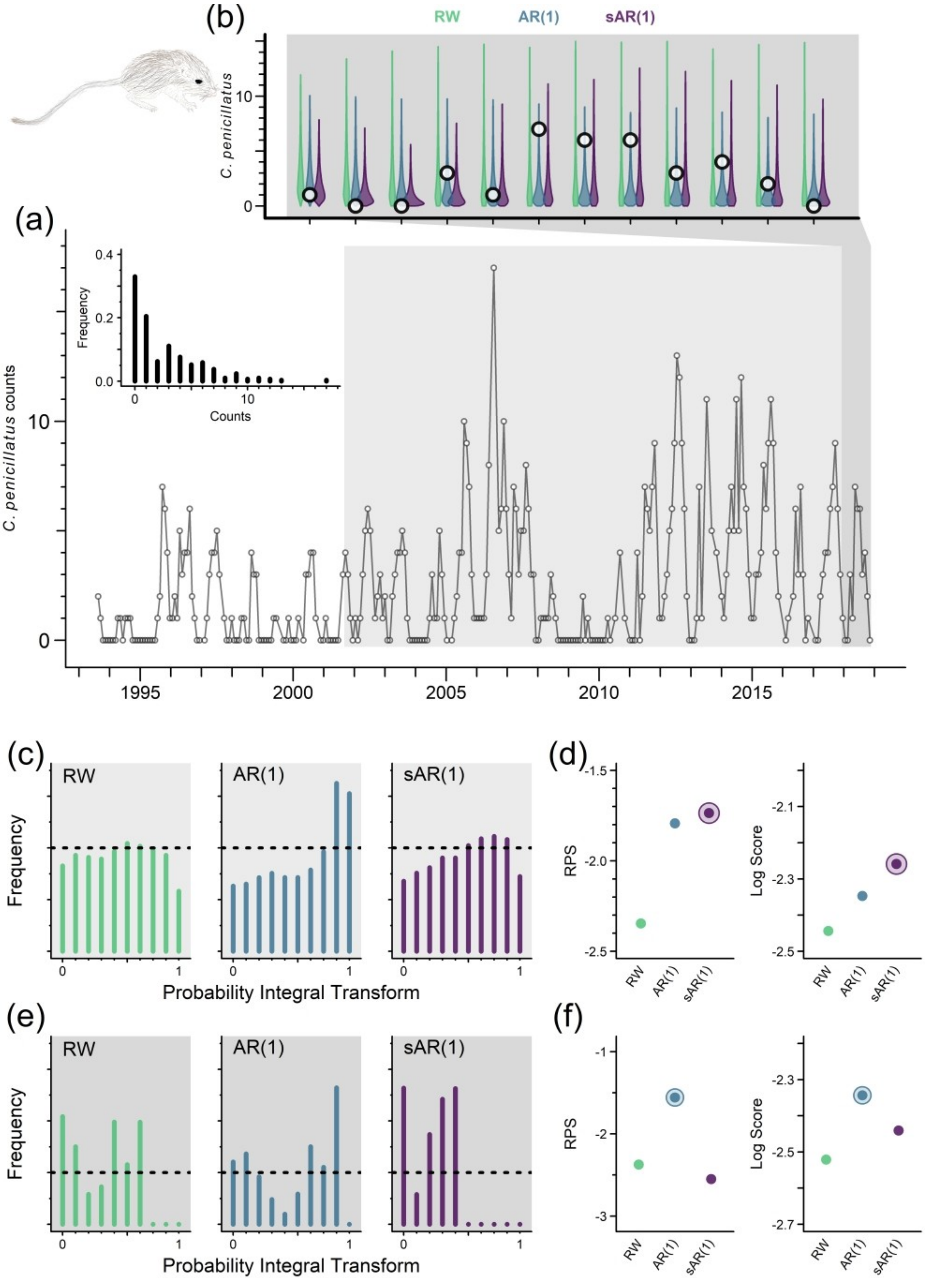
Name	Formula
Log	$\log\left(f_{H_n}(y_n)\right)$
Quadratic (Brier)	$2f_{H_n}(y_n)-\left(\left\Vert f_{H_n}(y_n)\right\Vert_2\right)^2$
Power	$\alpha\left(f_{H_n}(y_n)\right)^{\alpha-1}-(\alpha-1)\left(\left\Vert f_{H_n}(y_n)\right\Vert_\alpha\right)^\alpha$
Spherical	$\dfrac{f_{H_n}(y_n)}{\left\Vert f_{H_n}(y_n)\right\Vert_2}$
Pseudo-spherical	$\dfrac{\left(f_{H_n}(y_n)\right)^{\alpha-1}}{\left(\left\Vert f_{H_n}(y_n)\right\Vert_\alpha\right)^{\alpha-1}}$
Ranked Probability	$-\sum_{k=-\infty}^{\infty}\left(F_{H_n}(k)-1(y_n\le k)\right)^2$

n: sample, H_n: predictive distribution, y_n: observed value (i.e., a single data point), F: cumulative distribution function, f: probability density or mass function, $\Vert x\Vert_p$: p-norm of x ($\Vert x\Vert_p=\left(\sum|x|^p\right)^{\frac{1}{p}}$), α: generalization parameter, 1: the characteristic function ($1(y_n\le k)=\begin{cases}1,\wedge y_n\le k\\0,\wedge y_n>k\end{cases}$). For continuous variables, summations are replaced with integrals.

Figure 1. (a) Time series of N samples of variable y broken into a training set $y_{1:o}$ used to fit the model that will forecast the test set $y_{(o+1):(o+P)}$. At each time step t_n, the observation y_n is one realization from the underlying generating distribution G_n, shown with the insets. Probabilistic forecasts H_n are made for each time step forward from the forecast origin o at time t_o through the forecast horizon to the final sample at time $t_{o+p}=t_N$. The comparison between the forecast (grey) and generating (black) distributions for the first forecast at $o+1$ is shown in the rightmost subset. (b) Fixed and rolling origin end-sample evaluation on a mock data set of 17 observed samples and a forecast horizon of three samples. Open squares are training data, filled squares are test data, and dashed-line squares are not-yet-observed data. Origins for model test ($n_{o_{test}}$, estimates of the test data) and true ($n_{o_{true}}$, estimates of not-yet-observed data) forecasts are noted by the bold squares. As additional data are collected, the number of model tests (grey squares) grows in the rolling evaluation, whereas the fixed evaluation always has the same number of tests (three). In combination with probabilistic forecasting (a) the rolling origin approach forms the basis of the prequential approach.

Figure 2. (a) Time series and histogram of *C. penicillatus* counts in plot 19 since 1993-08-17 (sample 200). The rolling origin end-sample period (300 to 500) is denoted with the lighter grey rectangle and the final true test period (501 to 512) is the darker grey rectangle. (b) Predictive distributions for the three models (violins, delineated by color as shown by name) and observed data for the final true test period. (c,e) Probability Integral Transform histograms and (d,f) ranked probability and log scores for the models (RW: Random Walk, AR(1): first-order AutoRegressive, sAR(1): seasonal AR(1)) evaluated for the test period up to sample 500 (c,d) and for the final test with forecast origin of sample 500 (e,f). Dashed lines in (c,e) show uniform distributions and circled scores in (d, f) are best. (Sketch based on https://flic.kr/p/dhSSgy.)


(a)
Density
G1
Go−1
Ho+1, Go+1
y1
yo−1
yo+1
y ~ G
Time
t1
to−1
to
to+P = tN
forecast horizon
forecast origin
(b)
Training datum
Test datum
Not yet observed datum
Forecast origin
Single Origin
Sample
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Test origin = 14
True origin = 17
no test
no true
Rolling Origin
Sample
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Test origin = 10
Test origin = 11
Test origin = 12
Test origin = 13
Test origin = 14
Test origin = 15
Test origin = 16
True origin = 17


(b)
RW
AR(1)
sAR(1)
C. penicillatus
10
0
(a)
C. penicillatus counts
10
0
1995
2000
2005
2010
2015
Frequency
0.4
0.2
0.0
Counts
0
10
(c)
RW
AR(1)
sAR(1)
Frequency
0
1
Probability Integral Transform
(d)
RPS
-1.5
-2.0
-2.5
Log Score
-2.1
-2.3
-2.5
RW
AR(1)
sAR(1)
(e)
RW
AR(1)
sAR(1)
Frequency
0
1
Probability Integral Transform
(f)
RPS
-1
-2
-3
Log Score
-2.3
-2.5
-2.7
RW
AR(1)
sAR(1)