WILEY ECOLOGY

Evaluating Probabilistic Ecological Forecasts

Journal:	*Ecology*
Manuscript ID	Draft
Wiley - Manuscript type:	Statistical Reports
Date Submitted by the Author:	n/a
Complete List of Authors:	Simonis, Juniper; University of Florida, Wildlife Ecology and Conservation; DAPPER Stats White, Ethan; University of Florida, Wildlife Ecology and Conservation Ernest, S. K. Morgan; University of Florida
Substantive Area:	Statistics (general) < Statistics and Modeling < Theory < Substantive Area, Population Cycles < Population Dynamics and Life History < Population Ecology < Substantive Area, Population Ecology < Substantive Area, Statistics and Modeling < Theory < Substantive Area
Organism:	Rodents < Mammals < Vertebrates < Animals
Habitat:	Desert < Temperate Zone < Terrestrial < Habitat
Geographic Area:	Southwest US (AZ, CA, CO, NM, NV, UT) < United States < North America < Geographic Area
Key words/phrases:	continuous analysis, desert pocket mouse, ecological forecasting, end-sample holdout, forecast skill, hierarchical Bayes, prequential, score rule, time series, validation
Abstract:	Effective 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, we presently lack a set of robust, standardized, and general mathematical tools for evaluating probabilistic forecasts in ecology, impeding quantitative model comparison. We address this gap by bringing to bear an extensive literature on probabilistic forecast evaluation from diverse fields including climatology, economics, and epidemiology. Recognizing the breadth of ecological data and appreciating the variety of tools developed, rather than lobby for a specific singular metric for evaluation, we cover the range of options, highlight mathematical concepts to follow, and note decision points for practitioners to allow easy application of general principles to specific forecasting endeavors. We exemplify concepts with an application using a long-term rodent population time series and finish with a discussion of how ecology can continue to learn from, as well as help drive, forecasting science.

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

Effective 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, we presently lack a set of robust, standardized, and general mathematical tools for evaluating probabilistic forecasts in ecology, impeding quantitative model comparison. We address this gap by bringing to bear an extensive literature on probabilistic forecast evaluation from diverse fields including climatology, economics, and epidemiology. Recognizing the breadth of ecological data and appreciating the variety of tools developed, rather than lobby for a specific singular metric for evaluation, we cover the range of options, highlight mathematical concepts to follow, and note decision points for practitioners to allow easy application of general principles to specific forecasting endeavors. We exemplify concepts with an application using a long-term rodent population time series and finish with a discussion of how ecology can continue to learn from, as well as help drive, 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 is rapidly becoming an important focus of ecological science in applied and fundamental settings (Clark et al. 2001, Pennekamp et al. 2017). While the number of ecological forecasts is increasing, the ways in which the performance of these forecasts are evaluated is highly varied. 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 point estimates of states (e.g., population size), but embracing uncertainty is essential for understanding the range of possible futures (Dietz 2017). Uncertainty in ecological forecasts emerges from multiple sources of stochasticity, a lack of definitive mechanisms, and the inherent uncertainty of model fitting (Hooten and Hobbs 2015). A critical approach for capturing and communicating the variation inherent in forecasts is producing probabilistic distributions of future state values (Dawid 1984, Dietze et al. 2018), as forecasts with the same central tendency can vary substantially in the reasonableness of their fits (Fig. 1a).

Despite the necessity of making and evaluating probabilistic forecasts, most ecological studies evaluate forecast point estimates, disregarding the full predicted distribution (Ward et al. 2014, Petchey et al. 2015). Methods focused on central tendency matching of forecasts to data, however, are not uniquely optimized by the true probability distribution and cannot distinguish among forecasts (Gneiting et al. 2007, Czado et al. 2009), possibly causing "grossly misguided inferences" (Gneiting 2011). Learning from disciplines with established cultures, principles, and tools can help ecology define best practices for assessing probabilistic forecast performance (Winkler 1977, Dawid 1984, Gneiting and Raftery 2007). Here we present reputable methods for [1] measuring how well forecasted values match observations (the model's *skill*), [2] comparing models to a baseline to quantify information content, and [3] comparing models to one another.

Context, Notation, and Terminology

Consider a time series of samples n in 1...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 at time t_n (Fig. 1b). The last datum is the *forecast origin* o (Tashman 2000). We use models m in 1...M to gain inference about $G_{1:N}$ and make forecasts p in $1 \ldots P$ of y subsequent to y_o, where the time between o and p is the *lead time* or *forecast horizon* ($t_{o \to (o+p)}$; Fig. 1b) 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. 1b; **Appendix A**). For both tasks, we use the data in hand to validate our models, iterating the evaluation over time using 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).

Forecast Validation

The validation procedure defines how the observed data are split into those used to fit the model (*training data*) and those reserved to evaluate its predictions (*test data*). There is an array of validation procedures available for forecasting specifically (Stone 1977, Tashman 2000, Arlot and Celisse 2010). Based upon the ultimate task being predicting the next data in a time series (Dawid 1984), however, the dominant paradigm in forecasting validation is based on *end-sample holdout*, where the last k observations are held out for testing (Fig. 1c; Fildes and Makridakis 1995, Tashman 2000), rather than cross-validations that select test data from across the entire data set (e.g., leave- k -out; Arlot and Celisse 2010) and approximations like AIC (Stone 1977).

End-sample holdout methods are supported through simulation and empirical evaluations, where they produce more realistic distributions for future data than in-sample cross-validation (Tashman 2000). Training and testing errors are also often very weakly correlated (Makridakis 1986, Makridakis and Winkler 1989), indicating that models used to forecast will perform better on novel data when they have been validated via end-sample holdout than cross-validation (Fildes and Makridakis 1995). Despite the impetus for forecast validation using end-sample holdouts, however, cross-validation do apply to time series (Bergmeir et al. 2018). A motivation of cross-validation is to increase the number of evaluations, as a single evaluation is likely an unstable estimate of model skill (Tashman 2000) and a typical end-sample holdout provides only a single evaluation, whereas cross-validation aggregates many (Arlot and Celisse 2010, Bergmeir et al. 2018). However, the purpose of forecasting is to predict out-of-sample data (Fig. 1b), and if done prequentially (Dawid 1984), as is the goal (Dietze et al. 2018) and a reality (White et al. 2019) for ecological forecasting, the number of evaluations grows over time (Fig. 1c).

Using end-sample holdout, we define a break in our time series ($y_{1:N}$) between training and test sets based on a 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 the 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). The number of samples allocated to the test set (via the location of o) should cover at least as much time as 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 to be evaluated for each model, which is insufficient for describing skill. This is especially true if the series displays cyclic dynamics (common in ecology), in which case model performance will vary as a function of forecast 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. 1c; 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 arguably preferable to update the model with each step in the evaluation, re-optimization can be computationally intensive and may not provide notable changes to parameters (Tashman 2000). In prequential forecasting, however, recurrent forecasts replace the done-all-at-once evaluations, easing the computational burden (Dawid 1984, Dietz 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. 1c).

Graphical Evaluation

Graphical evaluation provides key insight into model appropriateness over the training and test sets. A useful first figure for a forecast is the predicted distribution and the observed values against each other (Dietze 2017). In forecasting, where the data are explicitly temporal, it is important to plot the time series of prediction distributions and observed values with some training data to show past dynamics (Fig. 1b). In addition, an informative plot is the x-y scatter of predicted-vs.-observed values, which ideally follows a 1:1 line, albeit not too closely (Fig. A1). Noting that ecological models often have multiple levels of uncertainty and non-linearities (Hooten and Hobbs 2015), their forecasted distributions are often not well summarized using quantiles (Fig. 1a). Rather, distributions or representative draws should be shown (Dietze 2017).

The *Probability Integral Transform* (PIT) is a diagnostic plot with a solid statistical basis and a long history in forecasting comprising the values of the predictive cumulative distribution functions (CDFs) evaluated at the observed values (Table A1; Rosenblatt 1952, 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 graphical plots (Fig. A1). The uniformity of the PIT is necessary but not sufficient for a forecast to match the generating distribution, however (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 (Fig. A1; Gneiting et al. 2007). The PIT has been extended to non-continuous distributions via approximations that add uniform noise (Smith 1985) or use the PIT's conditional CDF (Czado et al. 2009; Table A1).

Scoring Rules

A scoring rule r's function S^r measures how well a point matches a distribution (Brier 1950; **Appendix A**). The score s of observation y_n and model m's forecast H_n^m using rule r is $s_n^{rm} = S^r\left(H_n^m, y_n\right)$ and the model's average score is $\bar{s}_{(o+1):N}^{rm}$ (able 1). We use here 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 forecast metrics. Key attributes of rules are encompassed in the concept of (*strict*) *propriety* (Dawid 1998; **Appendix A**). 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 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 binarily 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 a more flexible *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 variance by autocorrelation (Benedetti 2010, Jewson 2018).

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 Euclidian norm (Table 1; Roby 1965). The spherical score is connected to the statistical notion of *surprise* and, similar to the quadratic, can be generalized (Table 1; Gneiting and Raftery 2007). The spherical score is discussed in texts covering scoring rules (e.g., Czado et al. 2009) but is not used frequently. In contrast to the log, the spherical score is hypersensitive near medial probabilities (Selten 1998).

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 A**).

Comparing Model Scores

Models can be quantitatively and statistically compared as long as they are scored on the same data using the same function, as their scores form an empirical distribution (Makridakis and Winkler 1989, Gneiting and Raftery 2007). Scores are typically aggregated across test data for quantitative comparisons, although graphing sample-level scores can provide useful insight (Gneiting et al. 2007). For example, plotting scores as a function of covariates could highlight if abnormal deviations are associated with external forces. Similarly, plots of scores as a function of lead time indicate 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.

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 predictive distribution; 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 values are modeled (Makridakis and Winkler 1989). The *Diebold-Mariano (D-M) Test* is the main method for frequentist comparisons and evaluates the significance of differences between forecasts using z-tests that account for correlated errors (Diebold and Mariano 1995; **Appendix A**). The D-M test is based on the differential between scores for forecasts, which has an expected value of 0 under a null hypothesis of no difference. The formal test statistic is then the standardized mean differential, which has an expected standard normal distribution under the null (Diebold and Mariano 1995). Serial autocorrelation may be addressable using robust formulae (**Appendix A**).

Example: Pocket Mouse Population Counts

To demonstrate prequential ecological forecasting, we use a subset of the data collected at a long-term study in the Chihuahuan Desert (AZ, USA; Brown 1998). Here, we focus on counts of the desert pocket mouse (*Chaetodipus penicillatus*) in one kangaroo-rat exclosure plot (Fig. 2). 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 the forecasts to actual observations for samples 501-512.

We fit three hierarchical Bayesian time series models (**Appendix B**), each with a truncated Poisson observation with a log-scale mean density ($\lambda = e^{x_n}$) and a maximum of 49 (the number of traps; double captures are rare: ~0.01%) and one of three process models: a random walk (RW), a first-order autoregressive (AR(1)), or a seasonal first-order AR (sAR(1); given the species' cycling; Fig. 2). We validated the models across a training period from sample 200 to 500 using a rolling origin end-sample evaluation (Figs. 1,2) beginning with a test origin of sample 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 sample 500). For the true origin (500), the test data were samples 501-512: a single realization of observations (Fig. 2). We fit the models using Markov Chain Monte Carlo via JAGS (Plummer 2003) accessed through R (R Core Team 2018) and used the log (for comparison to likelihood methods) and rank probability (to incorporate full predictive distributions) scores for evaluations (Table 1). We graphically assessed the fit of the predictions to both portions of the data using PIT histograms (non-randomized discrete calculation; Czado et al. 2009). See **Appendix B** for model details and **Appendix C** for code.

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

Discussion

Probabilistic forecasting has broad scientific and practical application with a rich history of mathematical and computational development driven by real world needs (Dawid 1984). Ecologists have embraced probabilistic forecasting in theory (Clark et al. 2011, Pennekamp et al. 2017) and practice (Ward et al. 2014, White et al. 2019). There persists, however, a knowledge gap with respect to tools used to evaluate probabilistic forecasts, which we hope this review has helped address. Embracing the variety of ecological variables that could be forecast, we recognize that there is no singular best metric or approach to evaluating all ecological forecasts. Thus, what we provided here should be considered an introduction to available methods drawn from standard forecasting approaches in other disciplines with a focus on current best practices.

Knowledge and skill transfer among disciplines is not one-way in the application of probabilistic forecasting to ecology (Pennekamp et al. 2017). Indeed, despite its rich history, forecasting science has many lines of inquiry with relevance to ecologists (Dietze 2017), such as the generalized kernel-based scoring rules (Dawid 1998, Gneiting and Raftery 2007). Ecological data bend or outright break assumptions of statistical methods due to non-normality, multiple levels of hierarchical variation, feedbacks, non-linearities, and autocorrelation (Hooten and Hobbs 2015). Many tools used to evaluate probabilistic forecasts make strong assumptions about model-generated distributions for which ecological data can provide important test cases. Standard practices developed in other disciplines provide a foundation for quantitatively evaluating probabilistic ecological forecasts. Simultaneously, ecology can help generalize existing methods, develop new tools, and further the theory of statistical forecasting.

Acknowledgements

This work was funded by the National Science Foundation (DEB-1622425 to Ernest) and the Gordon and Betty Moore Foundation (GBMF4563 to White) and informed by discussions at the 2019 Ecological Forecasting Initiative Conference.

Literature Cited

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

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.

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.

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

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.

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

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. 2018. R: A language and environment for statistical computing. v3.5.1. 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.

Rosenblatt, M. 1952. Remarks on a multivariate transformation. *The Annals of Mathematical*

Statistics **23**:470-472.

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.

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. *Olkos* **123**:652-661.

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.

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.

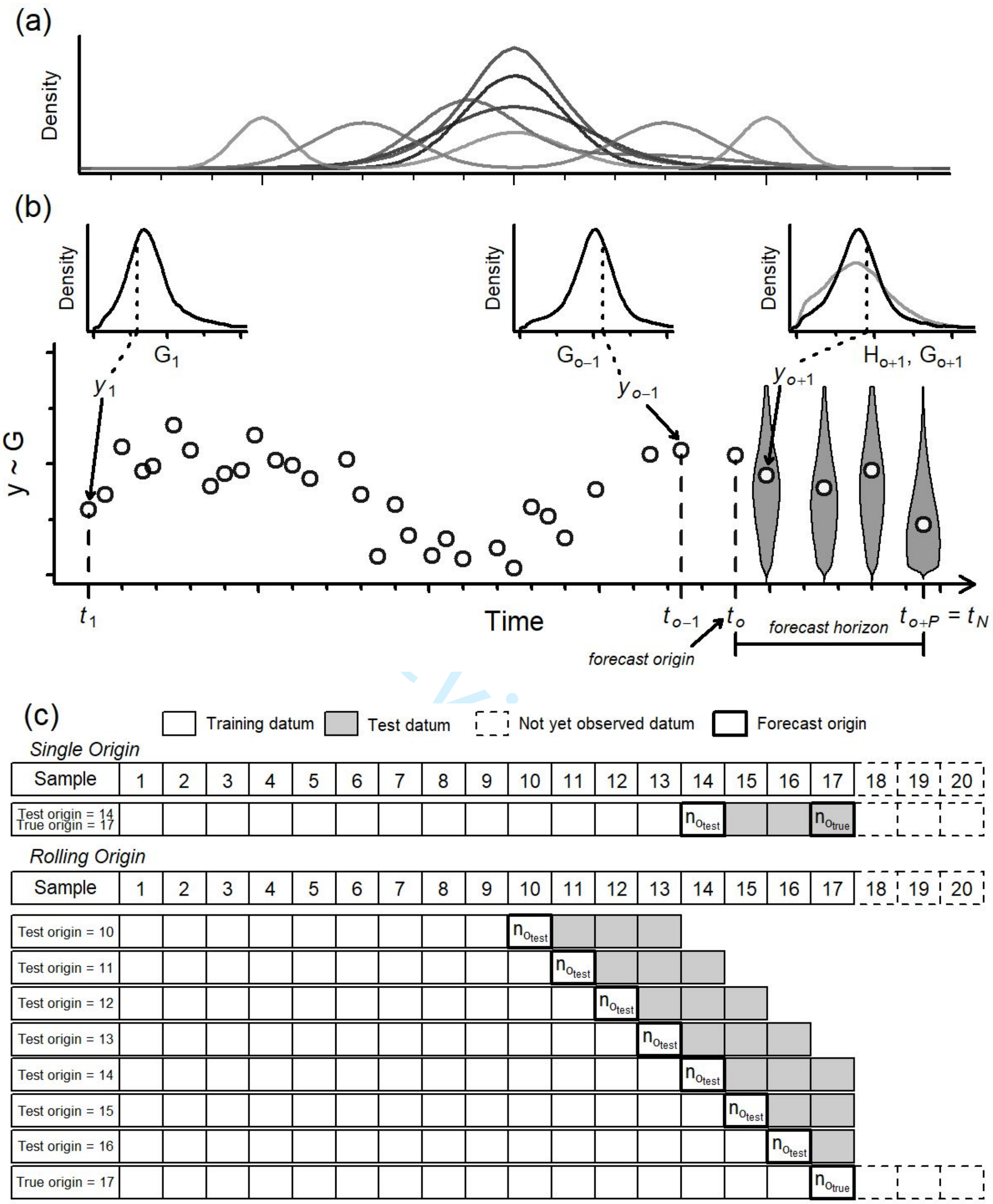
Table 1. Commonly used scoring rules, all defined as positively oriented.

Name	Formula
Log	$log(f_{H_n}(y_n))$
Quadratic (Brier)	$2f_{H_n}(y_n) - \left(\|f_{H_n}(y_n)\|_2\right)^2$
Power	$\alpha(f_{H_n}(y_n))^{\alpha-1} - (\alpha-1)\left(\|f_{H_n}(y_n)\|_\alpha\right)^\alpha$
Spherical	$\frac{f_{H_n}(y_n)}{\|f_{H_n}(y_n)\|_2}$
Pseudo-spherical	$\frac{(f_{H_n}(y_n))^{\alpha-1}}{\left(\|f_{H_n}(y_n)\|_\alpha\right)^{\alpha-1}}$
Ranked Probability	$-\sum_{k=-\infty}^{\infty}\left(F_{H_n}(k) - \mathbb{1}(y_n \leq k)\right)^2$

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

Figure 1. (a) Wildly different distributions with the same mean. (b) Time series of N samples of variable y broken into a training set $y_{1:o}$ used to 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 distributions for the first forecast at $o + 1$ is shown in the rightmost subset. (c) 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 (b) 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,c) Probability Integral Transform histograms (left three columns) and ranked probability and log scores (right two columns) for the models (RW: Random Walk, AR(1): first-order AutoRegressive, cAR(1): cyclic AR(1)) evaluated for the test period up to sample 500 (b) and for the final test with forecast origin of sample 500 (c). Dashed lines show uniform distributions and circled scores are best. (d) Predictive distributions for the three models (violins, delineated by grey tone) and observed data for the final true test period. (Sketch based on https://flic.kr/p/dhSSgy.)



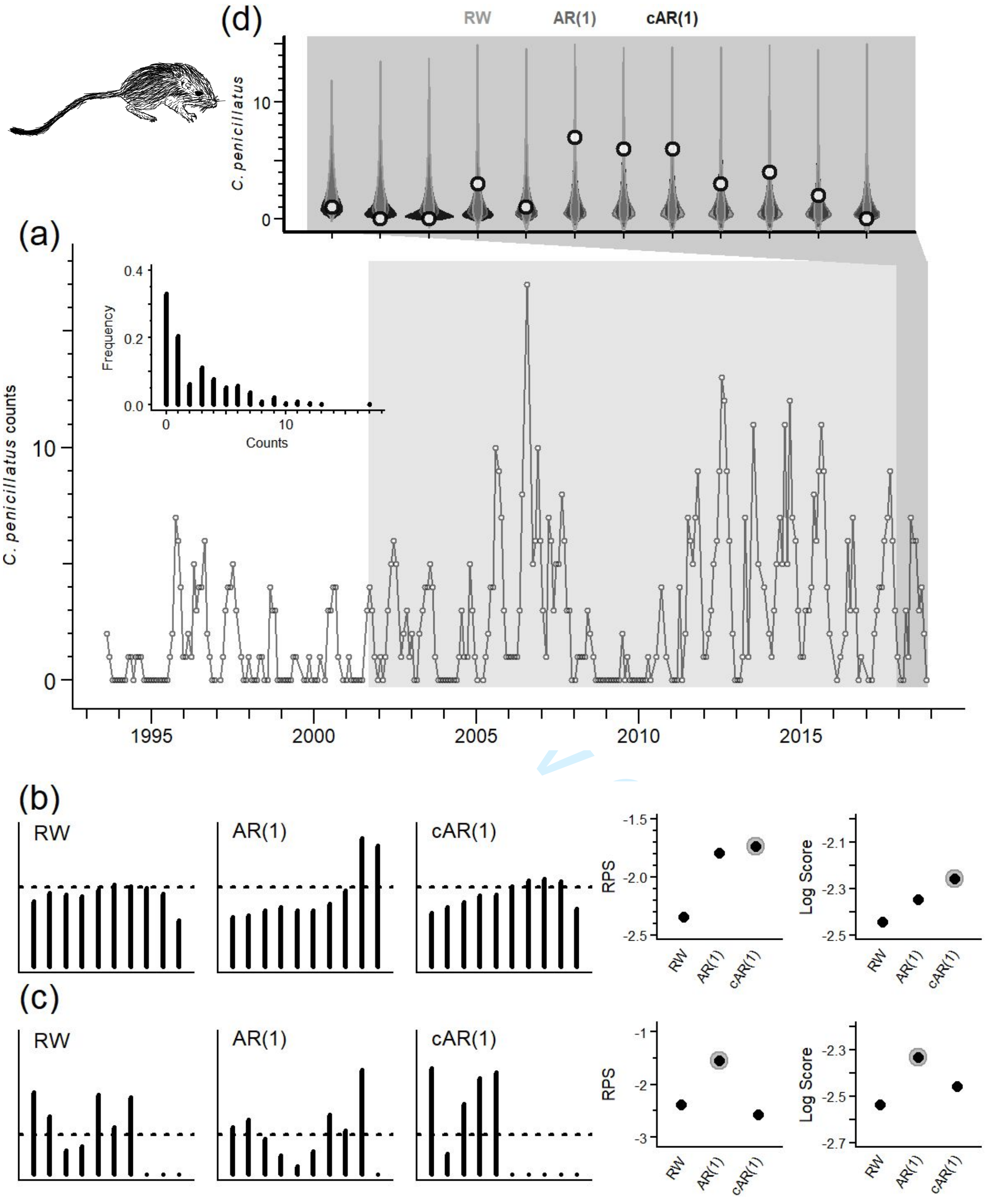
□ Training datum ▨ Test datum ┆ Not yet observed datum □ Forecast origin

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														$n_{o_{test}}$	▨	▨	$n_{o_{true}}$			

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										$n_{o_{test}}$	▨	▨	▨							
Test origin = 11											$n_{o_{test}}$	▨	▨	▨						
Test origin = 12												$n_{o_{test}}$	▨	▨	▨					
Test origin = 13													$n_{o_{test}}$	▨	▨	▨				
Test origin = 14														$n_{o_{test}}$	▨	▨	▨			
Test origin = 15															$n_{o_{test}}$	▨	▨			
Test origin = 16																$n_{o_{test}}$	▨			
True origin = 17																	$n_{o_{true}}$			


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

APPENDIX A

Additional theoretical and mathematical background for concepts in the main text.

Statistical Definition of a Probabilistic Forecast

The distributions from all potential models $\mathcal{M}$ (the M 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 $G_{1:N}$, (which may or may not be incorporated in the $\mathcal{M}$ models) form the sample space Ω. Then, $\mathcal{A}$ defines a workable set of distributions on Ω, in that it is closed under countability and complementarity ($\mathcal{A}$ is a "σ-algebra" of Ω) and $\mathcal{P}$ is a general convex class of probability measures that exist on (Ω,$\mathcal{A}$). A probabilistic forecast for our ecological variable is then any measure that exists on $\mathcal{P}$.

Scoring Functions and Rules

A scoring function must be defined on the sample space Ω and be able to take values on the extended real line (including negative and positive infinity), $\overline{\mathbb{R}} = [-\infty,\infty]$ (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 $\mathcal{A}$ (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 $\mathcal{P}$ (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 s_n^{rm} across the distribution of possible observations is

$$E\left[s_n^{rm}\right] = S^r\left(H_n^m,G_n\right) \qquad \text{A1}$$

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:

$$s_n^{rm} = S^r\left(E\left[H_n^m\right], y_n\right) \qquad \text{A2}$$

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, S^r is proper if

$$S^r(G_n, G_n) \geq S^r\left(H_n^m, G_n\right) \text{ for all } M \in \mathcal{M} \text{ and } H_n^m, G_n \in \mathcal{P} \qquad \text{A3}$$

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:

$$S^r(G_n, G_n) = S^r\left(H_n^m, G_n\right) \text{ if and only if } H_n^m = G_n \qquad \text{A4}$$

The propriety of scoring functions holds through linear (additive and multiplicative) transformations. That is, if S^1 is a proper or strictly proper scoring rule defined for a probability distribution H and observation y, and S^2 is

$$S^2(H, y) = cS^1(H, y) + q(y) \qquad \text{A5}$$

then S^2 is also proper or strictly proper, as long as $c > 0$ and q is integrable with respect to $\mathcal{P}$.

Test Statistics in the Diebold-Mariano Test

The *Diebold-Mariano Test* (D-M Test) is the primary approach for frequentist forecast comparison, which evaluates the significance of the difference between pairs of forecasts using z-tests while accounting for correlated error (Diebold and Mariano 1995, Diebold 2015). Its basis is the differential (d) between scores for models $m = 1{,}2$ on observation n:

$$d_n^{m\,=\,1,2} = s_n^{m\,=\,1} - s_n^{m\,=\,2} \tag{A6}$$

with an expected value of 0 under a null hypothesis of no difference between models. For a series, the test statistic is the mean differential across values ($\overline{d}^{m\,=\,1,2}$) divided by an estimate of its standard deviation ($\hat{\sigma}_{\overline{d}^{m\,=\,1,2}}$) times the square root of the sample size ($N - n_o$):

$$DM^{m\,=\,1,2} = \sqrt{N - n_o}\frac{\overline{d}^{m\,=\,1,2}}{\hat{\sigma}_{\overline{d}^{m\,=\,1,2}}} \tag{A7}$$

which has an expected standard normal (mean 0, standard deviation 1) distribution under the null hypothesis of no difference among models (Diebold and Mariano 1995, Diebold 2015).

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*).

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

$$\hat{\sigma}_{\overline{d}_{n_o\,+\,1:N}^{r;m\,=\,1,2}} = \sqrt{\frac{\hat{w}(0)}{N - n_o}} \tag{A8}$$

where $\hat{w}(0)$ is a consistent estimator of the variance. If the forecasts' score values are independent, a simple equation can be used for $\hat{w}(0)$:

$$\hat{w}(0) = \sum_{n = n_o + 1}^{N} \left(S^r\left(H_n^{r;m = 1}, y_n\right) - S^r\left(H_n^{r;m = 2}, y_n\right) \right)^2 \qquad \text{A9a}$$

In the presence of autocorrelation, $\hat{w}(0)$ becomes the weighted sum of the sample covariances:

$$\hat{w}(0) = \sum_{\tau = -(N - n_o - 1)}^{N - n_o - 1} l\left(\frac{\tau}{DM(N - n_o)} \right) \hat{\gamma}(\tau) \qquad \text{A9b}$$

where $l\left(\frac{\tau}{DM(N - n_o)} \right)$ is the lag window, $DM(N - n_o)$ is the truncation lag, and

$$\hat{\gamma}(\tau) = \frac{1}{N - n_o} \sum_{n = |\tau| + 1}^{N - n_o} \left(d_n^{r;m = 1,2} - \overline{d}_{n_o + 1:N}^{r;m = 1,2} \right) \left(d_{n - |\tau|}^{r;m = 1,2} - \overline{d}_{n_o + 1:N}^{r;m = 1,2} \right) \qquad \text{A10}$$

(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). Expansion of the D-M test allows for use of the robust frequentist approach to comparison (e.g., Hamill 1999) in ecological settings.

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

$$S^{rp}(H_n, y_n) = E_{H_n}|Y_n - y_n| - \frac{1}{2} E_{H_n}|Y_n - Y'_n| \qquad \text{A11}$$

where Y_n and Y'_n are independent random variables with distribution H_n Gneiting and Raftery 2007). This calculation can be approximated using a series of D draws from H_n, $Y_n^1 ... Y_n^S$, such as from MCMC (Gneiting and Raftery 2007, Krüger et al. 2019):

$$S^{rp}(H_n, y_n) = \frac{1}{D} \sum_{i = 1}^{D} \left| Y_n^i - y_n \right| - \frac{1}{2D^2} \sum_{i,j = 1}^{D} \left| Y_n^i - Y_n^j \right| \qquad \text{A12}$$

Models with Characteristic Predictive Distributions

Figure A1 shows seven models with different characteristic predictive distributions and the resulting graphical consequences. Here we give a bit more detail about the models, and **Appendix C** contains the relevant code for implementation.

The underlying generating distribution is a Poisson model with a sinusoidal factor, slope, and intercept:

$$y_n \sim Poisson\left(\lambda_n = 8 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right)\right) \tag{A13}$$

where x_n ranged from 1 to 50 and there were 35 total values. This was used for the generating distribution as well as to generate the true observations ($\dot{y}_n$). The positively and negatively biased models had simple offsets:

$$y_n \sim Poisson\left(\lambda_n = 10 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right)\right) \tag{A14}$$

$$y_n \sim Poisson\left(\lambda_n = 6 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right)\right) \tag{A15}$$

The too accurate model simply recycled the observed value as the mean of the Poisson:

$$y_n \sim Poisson(\lambda_n = \dot{y}_n) \tag{A16}$$

whereas the too precise model was based on a rounded-normal approximation to the Poisson with a reduced standard deviation compared to the standard Poisson:

$$y_n \sim Round\left(Normal\left(\mu_n = 8 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right), \sigma_n = \frac{\sqrt{\mu_n}}{1.6}\right)\right) \tag{A17}$$

The too imprecise model was a negative binomial with the mean of the standard Poisson model, but addition variance modeled via the size parameter ω:

$$y_n \sim NegBinom\left(\mu_n = 8 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right), \omega = 1\right) \tag{A18}$$

And the bimodal model was a combination of two Poisson distributions in equal proportions:

$$y_n \sim Poisson\left(\lambda_n = \begin{cases} 3 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right) at\ p = 0.5 \\ 13 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right) at\ p = 0.5 \end{cases}\right) \tag{A19}$$

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.

Christensen, J. H., F. X. Diebold, G. D. Rudebusch, and G. H. Strasser. 2008. Multivariate comparison of predictive accuracy. Unpublished working paper. http://www.econ.uconn.edu/Seminar.

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

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.

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.

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

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.

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. 1999. Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting* **14**:155-167.

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.

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

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.

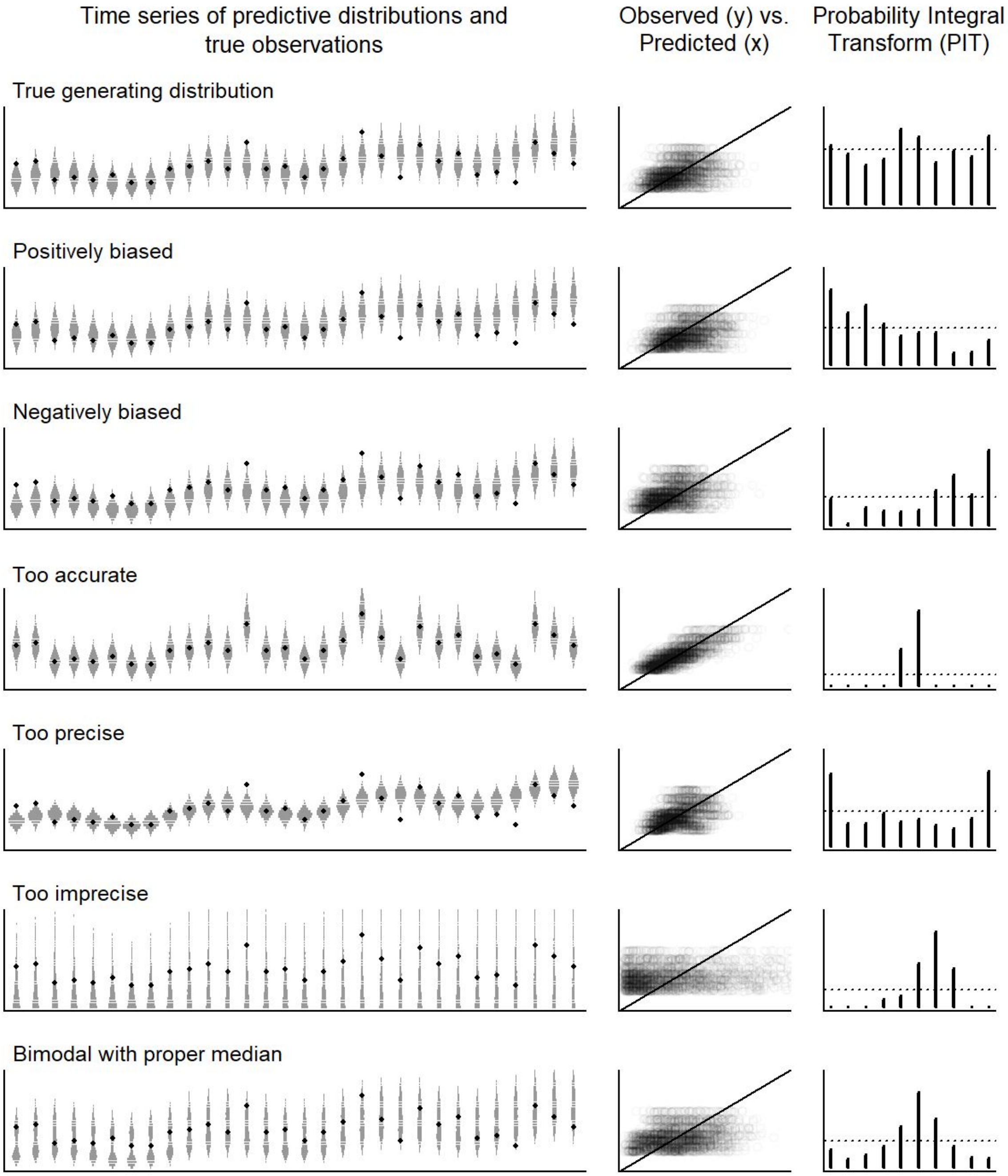
Table A1. Calculations of the Probability Integral Transform (PIT).

Type	Equation
Continuous Original	$PIT_n = F_{H_n}(y_n)$
Discrete Randomized	$rPIT_n = \; F_{H_n}(y_n - 1) + v(F_{H_n}(y_n) - F_{H_n}(y_n - 1))$ where $F_{H_n}(y_n = -1) \equiv 0)$
Discrete Non-randomized	$F(rPIT_n \mid y_n) = \begin{cases} 0, & rPIT \le F_{H_n}(y_n - 1) \\ \dfrac{rPIT - F_{H_n}(y_n - 1)}{F_{H_n}(y_n) - F_{H_n}(y - 1)}, & F_H(y_n - 1) \le rPIT \le F_{H_n}(y_n) \\ 1, & rPIT \ge F_{H_n}(y) \end{cases}$ $nrPIT = \sum_{n=1}^{N} F(rPIT_n \mid y_n)$

n: sample, H_n: predictive distribution, y_n: observed value, F: 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 ($F(rPIT_n \mid y_n)$) over the observed values (Czado et al. 2009).

Figure A1. Distributional predictive time series, observed-predicted scatter plots, and Probability Integral Transform (PIT) histograms (columns) for seven models (rows) with characteristic predictive distributions (headers, e.g. “Too precise”; Appendices A and C) evaluated against a time series of 50 Poisson-distributed data points. In the PIT histograms, the horizontal dashed line represents a uniform distribution. Note that the y axes scales vary among PIT histograms.

Figure A1.



540

APPENDIX B

Additional details of the desert pocket mouse (*Chaetodipus penicillatus*) example.

Pocket Mouse Data and Summary Statistics

There are 24 50 m^2 (50 × 50 m) plots at Portal, each of which contains 49 permanent stations in a 7 × 7 grid that are sampled with Sherman live traps every lunar month. Four of the plots have always been available to rodents except for kangaroo rats, and the focal plot for the example is one of these four (plot 19). *C. penicillatus* has always been at the Portal site, but did not become prevalent in this plot until the 1990s, since when it has dominated the samples (Ernest et al. 2009, Ernest et al. 2016, Ernest et al. 2019). We accessed the data as version 1.110.0 on 2019-06-04 using R version 3.5.1 (R Core Team 2018) scripts (**Appendix C**) leveraging version 0.2.5 of the portlar package (Christensen et al. 2019, Yenni et al. 2019).

We start our training data at sample 200 in the time series, corresponding to the date 1993-08-17, after which *C. penicillatus* has constituted 41.9% (729 of 1,740) of rodents trapped in the plot across the 290 complete surveys (out of 319 possible) through 2019-06-04 (Ernest et al. 2019). The next most abundant species during that time frame was 33.6% of the observations and all other species were less than 5% each (Ernest et al. 2019). Across those observations, *C. penicillatus* counts in the plot have cycled seasonally, ranging from 0 to 17 with a median of 1, a mean of 2.51, a variance of 8.77, and positive skew (skewness measures as 1.50 using the method of moments population estimate); the samples were 0-heavy (32.8%) and 45.9% of the samples contained 1 to 4 individuals (Fig. 4 in the main text; Ernest et al. 2019).

Fit and Analysis Details

Models were fit under a Bayesian framework via the Just Another Gibbs Sampler (JAGS, v.4.2; Plummer 2003, Plummer 2016) software, run from R (v3.5.1; R Core Team 2018) using the run.jags function in the runjags package (v2.0.4-2; Denwood 2016). Each model was fit using three separate chains, each of which was initialized at a random starting location then run for adaptation, burn-in, and sampling phases of 1,000, 5,000, and 10,000 steps, respectively. The 30,000 sampling steps were used without thinning to estimate parameters and the true count for each sample during the test period. We assessed chain convergence using autocorrelation, sample size adjusted for autocorrelation, and potential scale reduction factors (psrf, a.k.a. Gelman-Rubin statistic; Gelman and Rubin 1992).

Summary, analysis, and presentation were facilitated using custom R (v3.5.1; R Core Team 2018) scripts (**Appendix C**). Portal data were accessed using the summarize_rodent_data function in the portalr package (v0.2.5; Christensen et al. 2019, Yenni et al. 2019). We processed the MCMC output using the as.mcmc.list, combine.mcmc, and as.mcmc functions in the coda package (v 0.19-2; Plummer et al. 2006). Calculation of the rank probability score was conducted via the crps_sample function in the scoringRules package (v0.9.5; Jordan et al. 2018a, Jordan et al. 2018b). We measured skewness of distributions using the skewness function in the e1071 package (v1.7-1; Meyer et al. 2019). The non-randomized PIT values were calculated using code based on that provided in Czado et al. (2009) (see **Appendix C**).

Literature Cited

Christensen, E. M., G. M. Yenni, H. Ye, J. L. Simonis, E. K. Bledsoe, R. Diaz, S. D. Taylor, E. P. White, and S. K. M. Ernest. 2019. portalr: an R package for summarizing and using the Portal Project Data. *Journal of Open Source Software* **4**(33):1098.

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

Denwood, M. J. 2016. runjags: an R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software* **71**:1-25.

Ernest, S. K. M., T. J. Valone, and J. H. Brown. 2009. Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona, USA. *Ecology* **90**, 1708.

Ernest, S. K. M., G. M. Yenni, G. Allington, E. M. Christensen, K. Geluso, J. R. Goheen, M. R. Schutzenhofer, S. R. Supp, K. M. Thibault, J. H. Brown and T. J. Valone. 2016. Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona (1977–2013). *Ecology* **97**:1082.

Ernest, S. M., G. M. Yenni, G. Allington, E. Bledsoe, E. Christensen, R. Diaz, K. Geluso, J. R. Goheen, Q. Guo, E. Heske, D. Kelt, J. M. Meiners, J. Munger, C. Restrepo, D. A. Samson, M. R. Schutzenhofer, M. Skupski, S. R. Supp, K. Thibault, S. Taylor, E. White, D. W. Davidson, J. H. Brown, and T. J. Valone. 2018. The portal project: A long-term study of a Chihuahuan Desert ecosystem. bioRxiv. 10.1101/332783

Ernest, S. M., G. M. Yenni, G. Allington, E. Bledsoe, E. Christensen, R. Diaz, K. Geluso, J. R. Goheen, Q. Guo, E. Heske, D. Kelt, J. M. Meiners, J. Munger, C. Restrepo, D. A. Samson, M. R. Schutzenhofer, M. Skupski, S. R. Supp, K. Thibault, S. Taylor, E. White, D. W. Davidson, J. H. Brown, and T. J. Valone. 2019. Portal Project Data. v1.110.0. zenodo 10.5281/zenodo.3238678

Gelman, A. and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* **7**:457-511.

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. https://CRAN.R-project.org/package=scoringRules.

Meyer, D. E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch. 2019. e1071: Misc. functions of the Department of Statistics, Probability Theory Group (formerly: E1071), TU Wien. R package version 1.7-1. https://CRAN.R-project.org/package=e1071

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.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**: 7-11.

Plummer, M. 2016. JAGS: Just Another Gibbs Sampler. v4.2.0. https://sourceforge.net/projects/mcmc-jags/files/JAGS/4.x/

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/.

Yenni, G. M., H. Ye, E. M. Christensen, J. L. Simonis, E. K. Bledsoe, R. M. Diaz, S. D. Taylor, E. P. White, and S. K. M. Ernest. 2019. portalr: Create useful summaries of the Portal data. R package version 0.2.5. https://CRAN.R-project.org/package=portalr.

Table B1. Component equations of the models used with the pocket mouse example.

Model name	Equations
Random Walk [RW]	$x_0 = \mu_0$ $\mu_n = x_{n-1}$ $\tau = \tau$ $x_n \sim \mathcal{N}(\mu_n, \tau)$
First-order autoregressive [AR(1)]	$x_0 = \mu_0$ $\mu_n = \varphi x_{n-1}$ $\tau = \tau$ $x_n \sim \mathcal{N}(\mu_n, \tau)$
Cyclic first-order autoregressive [cAR(1)]	$x_0 = \mu_0$ $\mu_n = \varphi x_{n-1} + \beta_1 \cos 2\pi j_n + \beta_2 \sin 2\pi j_n$ $\tau = \tau$ $x_n \sim \mathcal{N}(\mu_n, \tau)$

x: log-scale density, μ_0: log-scale density at time 0 (prior: Normal with mean log (mean(y)), precision 0.25), $\mathcal{N}$ normal distribution (time varying mean μ_n and constant precision τ), φ: auto-regressive parameter (prior: Normal with mean 0, precision 1, and truncated at -1 and 1), β_1 and β_2: cyclic parameters (prior: Normal with mean 0, precision 0.16), j_n: fraction of the year at n, τ: precision (prior: Gamma with shape 1, rate 0.1). Samples n in $1 \ldots N$ are evenly spaced but an observation need not occur at every sample (allowing for missing observations).

APPENDIX C

Software used and written.

Overview

Custom written scripts for use in R (v3.5.1; R Core Team 2018) with runjags (v2.0.4-2; Denwood 2016), coda (v 0.19-2; Plummer et al. 2006), scoringRules (v0.9.5; Jordan et al. 2018a, 2018b), e1071 (v1.7-1; Meyer et al. 2019), and portalr (v0.2.5; Yenni et al. 2019, Christensen et al. 2019) packages, based on interface to JAGS (Just Another Gibbs Sampler, v4.2.0) (Plummer 2003, Plummer 2016), are available at https://www.github.com/weecology/forecast_evaluation.

Literature Cited

Christensen, E. M., G. M. Yenni, H. Ye, J. L. Simonis, E. K. Bledsoe, R. Diaz, S. D. Taylor, E. P. White, and S. K. M. Ernest. 2019. portalr: an R package for summarizing and using the Portal Project Data. *Journal of Open Source Software* **4**(33):1098.

Denwood, M. J. 2016. runjags: an R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software* **71**:1-25.

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. https://CRAN.R-project.org/package=scoringRules.

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.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**: 7-11.

Plummer, M. 2016. JAGS: Just Another Gibbs Sampler. v4.2.0. https://sourceforge.net/projects/mcmc-jags/files/JAGS/4.x/

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/.

Yenni, G. M., H. Ye, E. M. Christensen, J. L. Simonis, E. K. Bledsoe, R. M. Diaz, S. D. Taylor, E. P. White, and S. K. M. Ernest. 2019. portalr: Create useful summaries of the Portal data. R package version 0.2.5. https://CRAN.R-project.org/package=portalr.