Juniper L. Simonis, Ethan P. White, S. K. Morgan Ernest. 2021. Evaluating Probabilistic Ecological Forecasts. *Ecology*.

APPENDIX S1

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 $(\Omega, \mathcal{A})$. 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

$$E[s_n^{rm}] = S^r(H_n^m, G_n) \quad \text{S1}$$

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:

$$s_n^{rm} = S^r(E[H_n^m], y_n) \quad \text{S2}$$

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

$$S^r(G_n, G_n) \geq S^r(H_n^m, G_n) \text{ for all } M \in \mathcal{M} \text{ and } H_n^m, G_n \in \mathcal{P} \quad \text{S3}$$

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:

$$S^r(G_n, G_n) = S^r(H_n^m, G_n) \text{ if and only if } H_n^m = G_n \quad \text{S4}$$

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

$$S^2(H, y) = cS^1(H, y) + q(y) \quad \text{S5}$$

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:

$$d_n^{m=1,2} = s_n^{m=1} - s_n^{m=2} \quad \text{S6}$$

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 ($\bar{d}^{m=1,2}$) divided by an estimate of its standard deviation ($\hat{\sigma}_{\bar{d}^{m=1,2}}$) times the square root of the sample size ($N - n_o$):

$$DM^{m=1,2} = \sqrt{N - n_o}\,\frac{\bar{d}^{m=1,2}}{\hat{\sigma}_{\bar{d}^{m=1,2}}} \quad \text{S7}$$

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

$$\hat{\sigma}_{\bar{d}_{n_o+1:N}^{r;m=1,2}} = \sqrt{\frac{\hat{w}(0)}{N-n_o}} \quad \text{S8}$$

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 $\hat{w}(0)$:

$$\hat{w}(0) = \sum_{n=n_o+1}^{N} \left(S^r\left(H_n^{r;m=1}, y_n\right) - S^r\left(H_n^{r;m=2}, y_n\right)\right)^2 \quad \text{S9a}$$

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

$$\hat{w}(0) = \sum_{\tau=-(N-n_o-1)}^{N-n_o-1} l\left(\frac{\tau}{DM(N-n_o)}\right)\hat{\gamma}(\tau) \quad \text{S9b}$$

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

$$\hat{\gamma}(\tau) = \frac{1}{N-n_o}\sum_{n=|\tau|+1}^{N-n_o}\left(d_n^{r;m=1,2} - \bar{d}_{n_o+1:N}^{r;m=1,2}\right)\left(d_{n-|\tau|}^{r;m=1,2} - \bar{d}_{n_o+1:N}^{r;m=1,2}\right) \quad \text{S10}$$

(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

$$S^{rp}(H_n, y_n) = E_{H_n}|Y_n - y_n| - \frac{1}{2}E_{H_n}|Y_n - Y_n'| \quad \text{S11}$$

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, $Y_n^1 \ldots Y_n^S$, such as from MCMC (Gneiting and Raftery 2007, Krüger et al. 2019):

$$S^{rp}(H_n, y_n) = \frac{1}{D}\sum_{i=1}^{D}\left|Y_n^i - y_n\right| - \frac{1}{2D^2}\sum_{i,j=1}^{D}\left|Y_n^i - Y_n^j\right| \quad \text{S12}$$

Models with Characteristic Predictive Distributions

Figure S1 shows seven models with different characteristic predictive distributions and the resulting graphical consequences. Here we give a bit more detail about the models, and **Data S1** contains the relevant code for implementation.

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

$$y_n \sim Poisson\left(\lambda_n = 8 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right)\right) \quad \text{S13}$$

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:

$$y_n \sim Poisson\left(\lambda_n = 10 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right)\right) \quad \text{S14}$$

$$y_n \sim Poisson\left(\lambda_n = 6 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right)\right) \quad \text{S15}$$

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

$$y_n \sim Poisson(\lambda_n = \dot{y}_n) \quad \text{S16}$$

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:

$$y_n \sim Round\left(Normal\left(\mu_n = 8 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right), \sigma_n = \frac{\sqrt{\mu_n}}{1.6}\right)\right) \quad \text{S17}$$

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:

$$y_n \sim NegBinom\left(\mu_n = 8 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right), \omega = 1\right) \quad \text{S18}$$

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

$$y_n \sim Poisson\left(\lambda_n = \begin{cases} 3 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right) at\ p = 0.5 \\ 13 + 0.25x_n + 3\sin\left(\frac{2\pi x_n}{15}\right) at\ p = 0.5 \end{cases}\right) \quad \text{S19}$$

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 S1. Calculations of the Probability Integral Transform (PIT).

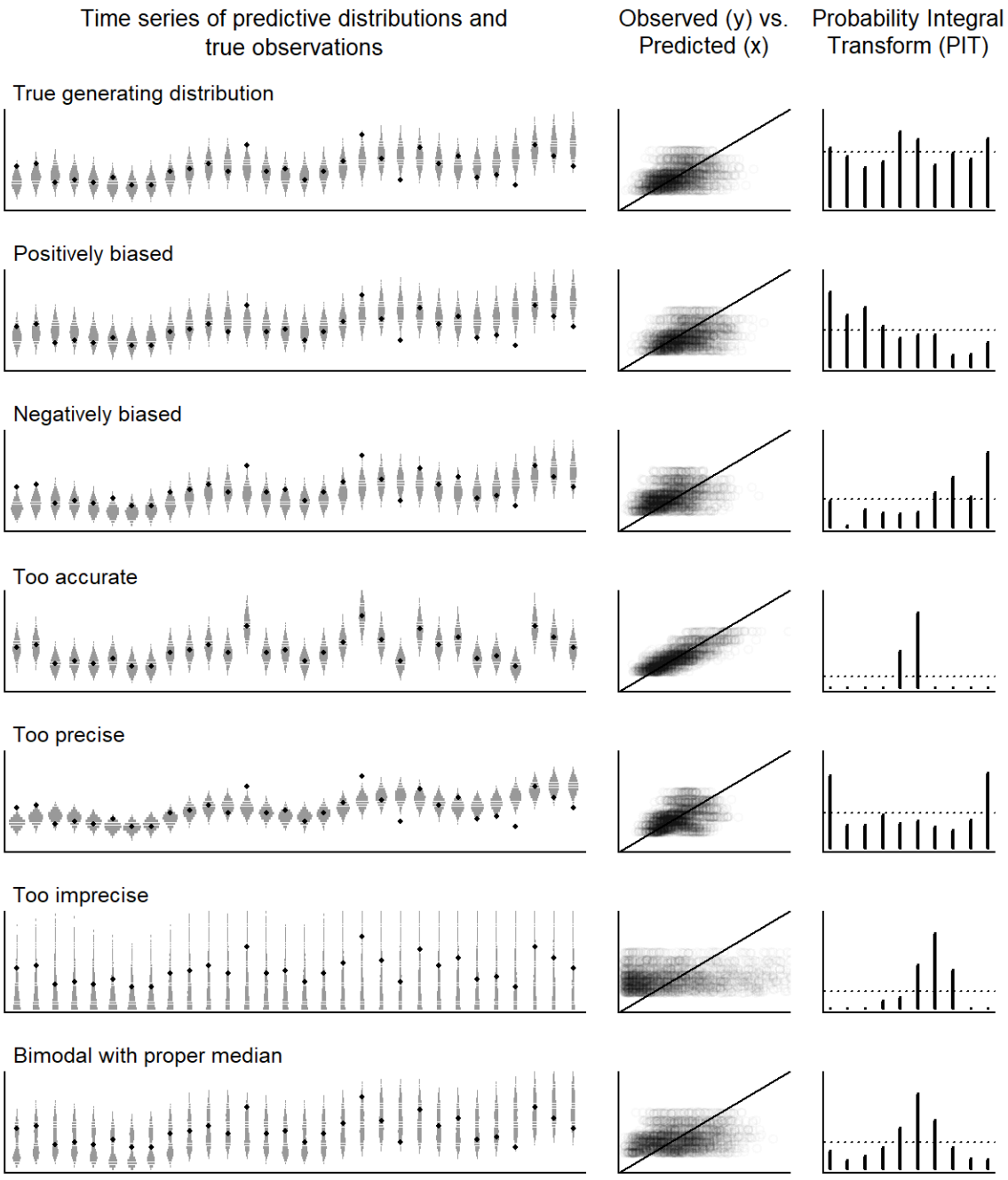
Type	Equation
Continuous Original	$PIT_n = F_{H_n}(y_n)$
Discrete Randomized	$rPIT_n = F_{H_n}(y_n - 1) + v\left(F_{H_n}(y_n) - F_{H_n}(y_n - 1)\right)$ where $F_{H_n}(y_n = -1) \equiv 0)$
Discrete Non-randomized	$F(rPIT_n \mid y_n) = \begin{cases} 0, & rPIT \le F_{H_n}(y_n - 1) \\ \dfrac{rPIT - F_{H_n}(y_n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

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 S1. 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”) 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 S1.


Time series of predictive distributions and true observations
Observed (y) vs. Predicted (x)
Probability Integral Transform (PIT)
True generating distribution
Positively biased
Negatively biased
Too accurate
Too precise
Too imprecise
Bimodal with proper median