THE PROBABILISTIC CONCORDANCE CORRELATION COEFFICIENT: A CHANCE-CORRECTED PERFORMANCE METRIC FOR CLASSIFICATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluation of probabilistic classifiers commonly considers discrimination and calibration. The area under the receiver operating characteristic curve summarizes ranking performance but is insensitive to calibration, whereas calibration measures such as expected calibration error do not quantify case-specific predictive information. The Brier score provides a strictly proper assessment of overall probabilistic accuracy, but its magnitude depends on class prevalence, case mix, and irreducible uncertainty, limiting its interpretation across prediction problems. We propose the *Probabilistic Concordance Correlation Coefficient* (P-CCC), ρ_P, a metric for evaluating probabilistic classifiers through chance-corrected agreement between predicted probability vectors and realized class labels. P-CCC extends the repeated-measures concordance correlation coefficient, originally developed for repeated continuous measurements, to prediction-label agreement on the probability simplex. Equivalently, P-CCC is one minus the Brier score divided by the error the same forecaster would incur against independent labels. This representation yields a bounded scale and a prior-independent zero point, with P-CCC interpreted as the proportion of model-specific chance-level error removed through prediction-label dependence. We derive an exact $O(nK)$ U-statistic estimator, establish asymptotic normality, and provide confidence intervals. Theoretical analysis shows that P-CCC penalizes uninformative predictive variation but may favor overconfident forecasts and is therefore not a proper scoring rule. Simulations and benchmark experiments demonstrate that P-CCC provides information not captured by AUC, expected calibration error, or the Brier score. We recommend reporting P-CCC alongside the Brier score as complementary measures of chance-corrected prediction-label agreement and overall probabilistic accuracy, respectively.

1 INTRODUCTION

Probabilistic classifiers are evaluated along two central dimensions. *Discrimination* concerns whether predictions distinguish observations with different outcomes and is commonly summarized by the area under the receiver operating characteristic curve (AUC). *Calibration* concerns whether predicted probabilities agree with observed outcome frequencies and is often assessed using the expected calibration error (ECE). Both are important when predicted risks inform thresholds, treatment decisions, or expected costs (Dawid, 1982). Neither AUC nor ECE alone provides a comprehensive assessment of probabilistic prediction. AUC depends only on ranking and is therefore unchanged by strictly monotone transformations of the scores (Hanley & McNeil, 1982); models with substantially different calibration can therefore share the same AUC (Guo et al., 2017; Minderer et al., 2021). Conversely, calibration alone does not measure case-specific predictive information. A classifier that reports the marginal class prevalence for every observation may achieve an optimal ECE while providing no discrimination (Naeini et al., 2015).

The Brier score (BS) addresses both aspects by computing the mean squared distance between predicted probabilities and realized outcomes (Brier, 1950). As a strictly proper scoring rule, it evaluates overall probabilistic accuracy and is uniquely optimized in expectation by the true conditional probabilities (Gneiting & Raftery, 2007). In this sense, it reflects both calibration and the ability of predictions to vary informatively across observations. Its magnitude, however, depends on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

prevalence, case mix, and irreducible outcome uncertainty. A raw Brier score therefore has no fixed chance-referenced interpretation and may be difficult to compare across prediction problems.

A standard statistical remedy for this interpretability problem is to compare observed agreement with that expected under a reference mechanism. Cohen's κ applies this principle to nominal ratings (Cohen, 1960), while Lin's concordance correlation coefficient (CCC) combines correlation with departures in location and scale for continuous measurements (Lin, 1989). We extend this idea to probability vectors by applying the trace-ratio repeated-measures CCC (King et al., 2007) to the probability simplex, yielding the *Probabilistic Concordance Correlation Coefficient* (P-CCC), ρ_P.

The P-CCC equals one minus the Brier score divided by the loss the same forecaster would incur when paired with an independent copy of the label. This reference preserves the forecaster's marginal mean and predictive variance while removing its dependence on the outcome. The resulting coefficient is bounded and has a chance-corrected zero point, with a direct interpretation as the proportion of model-specific chance-level squared error removed by prediction-label dependence. Because the reference denominator depends on the evaluated model, P-CCC can also distinguish models that the conventional climatological Brier skill score cannot.

P-CCC is an agreement coefficient, not a proper scoring rule. Since one-hot labels are more variable than their generating conditional probabilities, maximizing concordance may favor overdispersed, overconfident forecasts. We do not propose P-CCC as a replacement for the Brier score or as a stand-alone model-selection or training criterion. Rather, we recommend reporting the pair $(\mathrm{BS}, \hat{\rho}_P)$: the Brier score summarizes overall probabilistic accuracy, while P-CCC quantifies how much model-specific chance-level error is removed through agreement between predictions and labels.

Contributions. We introduce the multiclass P-CCC and identify it as an independence-baseline Brier skill score with a bounded scale and prior-independent zero point (Theorem 1). We establish its main properties, including a decomposition into prediction–label association and a penalty for location and scale mismatch, and a trace-ratio construction that remains well defined on the simplex (Proposition 1). We further derive an exact $O(nK)$ U-statistic estimator, asymptotic normality, and jackknife-based Fisher-z intervals (Theorems 2–3). We characterize its incentives, develop a prevalence-weighted extension, and use simulations and benchmarks to show how P-CCC rankings differ from those of AUC, ECE, and the Brier score. These results support reporting $(\mathrm{BS}, \hat{\rho}_P)$ as complementary measures of probabilistic accuracy and chance-corrected prediction–label agreement.

2 Related Work

Agreement coefficients. Lin (1989) introduced the CCC to measure reproducibility between two continuous measurements, penalizing deviation from the 45° line rather than from the best-fit line. Multivariate and multi-rater generalizations followed (King & Chinchilli, 2001; Carrasco & Jover, 2003; Barnhart et al., 2002), including the trace-based repeated-measures CCC (King et al., 2007) used here and a matrix-based form using the Frobenius norm (Hiriote & Chinchilli, 2011); for nominal outcomes the analogous chance correction is Cohen's κ (Cohen, 1960). To our knowledge, this literature has not been applied to the evaluation of probabilistic classifiers, nor has it exploited the constraint that predictions lie on the simplex.

Proper scoring rules and skill scores. The Brier score is strictly proper (Brier, 1950; Gneiting & Raftery, 2007). Its decomposition into reliability, resolution, and uncertainty clarifies how overall probabilistic accuracy reflects both calibration and informative variation across observations (Murphy, 1973; DeGroot & Fienberg, 1983). Skill scores normalize a score by a reference forecast, whose choice governs their interpretation; the conventional Brier skill score uses the class prior.

Calibration of classifiers. Post-hoc methods such as Platt scaling, isotonic regression, and temperature scaling are widely used to improve probability calibration (Platt, 1999; Zadrozny & Elkan, 2002; Guo et al., 2017). Calibration remains important for modern classifiers and under distribution shift (Ovadia et al., 2019; Minderer et al., 2021). ECE summarizes the gap between confidence and observed frequency (Naeini et al., 2015), but calibration alone measures neither discrimination nor resolution.

3 THE PROBABILISTIC CONCORDANCE CORRELATION COEFFICIENT

3.1 SETUP

Let (p, q) denote a random pair, where $p \in \Delta^{K-1} = \{v \in \mathbb{R}^K_{\geq 0} : \sum_k v_k = 1\}$ is a predicted probability vector and $q \in \{e_1, \ldots, e_K\}$ is the one-hot encoding of the realized class, with e_k denoting the kth standard basis vector. Write $\mu_p = \mathbb{E}[p]$, $\mu_q = \mathbb{E}[q] = \pi$, $\Sigma_p = \mathrm{Cov}(p)$, $\Sigma_q = \mathrm{Cov}(q)$ and $\Sigma_{pq} = \mathrm{Cov}(p, q)$. The (multiclass, summed) Brier score is the expected squared Euclidean distance on the simplex, $\mathrm{BS} = \mathbb{E}\|p - q\|^2$. Throughout, $\mathrm{BS}(p, q)$ and $\rho_P(p, q)$ are functionals of the *joint distribution* of (p, q) rather than functions of a single observation.

Definition 1 (P-CCC)**.** The probabilistic concordance correlation coefficient is

$$\rho_P \;=\; 1 - \frac{\mathbb{E}\|p - q\|^2}{\mathrm{tr}(\Sigma_p) + \mathrm{tr}(\Sigma_q) + \|\mu_p - \mu_q\|^2} \;=\; \frac{2\,\mathrm{tr}(\Sigma_{pq})}{\mathrm{tr}(\Sigma_p) + \mathrm{tr}(\Sigma_q) + \|\mu_p - \mu_q\|^2}, \tag{1}$$

that is, an extension of the repeated-measures CCC of King et al. (2007), originally developed for repeated continuous measurements, to agreement between predicted probability vectors and realized class labels on the probability simplex.

The second equality follows from $\mathbb{E}\|p - q\|^2 = \mathrm{tr}(\Sigma_p) + \mathrm{tr}(\Sigma_q) - 2\,\mathrm{tr}(\Sigma_{pq}) + \|\mu_p - \mu_q\|^2$. The denominator is strictly positive unless p and q are both degenerate at the same point.

3.2 WHY THE TRACE RATIO, AND NOT A MATRIX NORM

Hiriote & Chinchilli (2011) summarize multivariate concordance using a Frobenius-norm functional that whitens by $\Sigma^{-1/2}$. This formulation is not directly applicable on the simplex: $\mathbf{1}^\top p = 1$ almost surely implies $\Sigma_p \mathbf{1} = 0$, so Σ_p has rank at most $K - 1$ and is singular for every K; the same holds for Σ_q. Covariance-whitening formulations thus require regularization. The trace ratio requires no inversion, aggregates variance additively across coordinates, and links (1) directly to the Brier score. This link gives P-CCC its skill-score interpretation rather than that of a generic similarity measure.

3.3 WHAT THE COEFFICIENT IS

Proofs of all results are collected in Appendix A.

Theorem 1 (Independence-baseline skill score)**.** *Let q' be an independent copy of q, independent of p and with the same marginal law. Then $\mathbb{E}\|p - q'\|^2 = \mathrm{tr}(\Sigma_p) + \mathrm{tr}(\Sigma_q) + \|\mu_p - \mu_q\|^2$, and hence*

$$\rho_P = 1 - \frac{\mathrm{BS}(p, q)}{\mathrm{BS}_{\mathrm{indep}}(p, q)}, \qquad \mathrm{BS}_{\mathrm{indep}}(p, q) = \mathbb{E}\|p - q'\|^2. \tag{2}$$

The independence-baseline error, $\mathrm{BS}_{\mathrm{indep}}(p, q)$, pairs the forecaster with an independent copy of the label, preserving their marginal distributions while removing their dependence. Analogous to the chance correction underlying Cohen's κ, P-CCC is the fraction of this baseline squared error removed through prediction-label dependence. This baseline is a reference benchmark, not an attainable error bound: a forecaster may perform either better or worse than it. A P-CCC of zero indicates no net first-order agreement, although the predictions may still contain outcome information; the location of this zero does not change with class prevalence or outcome uncertainty.

3.4 ELEMENTARY PROPERTIES

Proposition 1 (Elementary properties)**.** *Let $\pi = \mu_q$ be the class prior. Then*

(a) (Target variance.) $\mathrm{tr}(\Sigma_q) = 1 - \|\pi\|^2$.

(b) (Factorization and range.) *If* $\mathrm{tr}(\Sigma_p) > 0$ *and* $\mathrm{tr}(\Sigma_q) > 0$ *then*

$$\rho_P = r_{\mathrm{tr}}\, C_b, \qquad r_{\mathrm{tr}} = \frac{\mathrm{tr}(\Sigma_{pq})}{\sqrt{\mathrm{tr}(\Sigma_p)\,\mathrm{tr}(\Sigma_q)}}, \qquad C_b = \frac{2\sqrt{\mathrm{tr}(\Sigma_p)\,\mathrm{tr}(\Sigma_q)}}{\mathrm{tr}(\Sigma_p) + \mathrm{tr}(\Sigma_q) + \|\mu_p - \mu_q\|^2}, \tag{3}$$

with $r_{\mathrm{tr}} \in [-1, 1]$ and $C_b \in (0, 1]$, so that $|\rho_P| \le |r_{\mathrm{tr}}| \le 1$ and $\rho_P \in [-1, 1]$. Equality $C_b = 1$ holds if and only if $\mu_p = \mu_q$ and $\mathrm{tr}(\Sigma_p) = \mathrm{tr}(\Sigma_q)$. If either trace vanishes and the denominator of (1) *is positive, then $\mathrm{tr}(\Sigma_{pq}) = 0$ and $\rho_P = 0$.*

(c) (Domination of the climatological skill score.) *With $\mathrm{BSS} = 1 - \mathrm{BS}/\,\mathrm{tr}(\Sigma_q)$ the Brier skill score against the constant forecast $p \equiv \pi$, we have $\rho_P \ge \mathrm{BSS}$, with equality if and only if $\mathrm{BS} = 0$ or $\mathrm{tr}(\Sigma_p) + \|\mu_p - \mu_q\|^2 = 0$.*

Part (b) is the simplex analogue of Lin's precision–accuracy decomposition: r_{tr} measures prediction–label association, whereas C_b penalizes the remaining location and scale mismatch. We state the result as a factorization rather than through the chain $\rho_P \le 2\,\mathrm{tr}(\Sigma_{pq})/\{\mathrm{tr}(\Sigma_p) + \mathrm{tr}(\Sigma_q)\} \le r_{\mathrm{tr}}$, because this chain holds only when $\mathrm{tr}(\Sigma_{pq}) \ge 0$. When $\mathrm{tr}(\Sigma_{pq}) < 0$, removing $\|\mu_p - \mu_q\|^2$ from the denominator makes the ratio more negative, and the corresponding inequalities are reversed. Model H of Table 1 illustrates this case, with $\rho_P = -0.295$ and $r_{\mathrm{tr}} = -0.415$. The factorization remains valid for either sign and establishes the $[-1, 1]$ range required for the Fisher-z transform in §4.

Part (c) has a direct implication for model comparison. Within a fixed evaluation set, BSS is an affine transformation of BS and therefore produces exactly the same model ranking. Thus, climatological normalization changes the scale of the Brier score but provides no additional information for within-dataset ranking. By contrast, the independence baseline used by P-CCC is model-specific because its denominator incorporates the model's predictive variance and marginal bias. P-CCC can therefore rank models differently from the Brier score, as investigated in §7.2. It does so only through miscalibration. For any calibrated forecaster, $\mathbb{E}[q \mid p] = p$ gives $\Sigma_{pq} = \Sigma_p$, $\mu_p = \mu_q$ and $\mathrm{BS} = \mathrm{tr}(\Sigma_q) - \mathrm{tr}(\Sigma_p)$, so that

$$\rho_P = \frac{2\,\mathrm{BSS}}{1 + \mathrm{BSS}}, \tag{4}$$

a strictly increasing function of BSS; the Bayes forecaster of Table 1, with $\mathrm{BSS} = 0.173$ and $\rho_P = 0.295$, illustrates it. Among calibrated forecasters, P-CCC therefore ranks models exactly as the Brier score does. What P-CCC adds is thus interpretive and diagnostic: a bounded, chance-referenced scale and, through (4), a check on dispersion. Calibrated forecasters lie on the curve $\rho_P = 2\,\mathrm{BSS}/(1 + \mathrm{BSS})$; a forecaster with $\mathrm{BSS} > -1$ lies above it exactly when $\mathrm{tr}(\Sigma_p) + \|\mu_p - \mu_q\|^2 > \mathrm{tr}(\Sigma_q) - \mathrm{BS}$, the dispersion of a calibrated forecaster with the same Brier score, and below it when it falls short. In Table 1, A and B lie on the curve, D, E and F above it, and G below it.

Remark 1 (Interpretation of zero)**.** Because the denominator of (1) is positive, $\rho_P = 0$ if and only if $\mathrm{tr}(\Sigma_{pq}) = 0$. Independence is sufficient but not necessary, because class-specific covariances may cancel. For example, consider a three-class problem with uniform class probabilities and a forecaster that reports the true class with probability $1/3$ and a fixed cyclic permutation otherwise. Then $\mathrm{tr}(\Sigma_{pq}) = 0$ exactly and $\rho_P = 0$, although the forecast remains informative: observing $p = e_k$ excludes one class and increases the posterior probability of another to $2/3$. Thus, zero indicates no net first-order agreement, not necessarily an absence of predictive information. If all observations have the same label, then $\mathrm{tr}(\Sigma_q) = 0$ and $\rho_P = 0$ for every forecaster; an implementation should report this degeneracy explicitly.

4 Estimation and Inference

Let $\{(p_i, q_i)\}_{i=1}^n$ be an i.i.d. sample, let $\bar{p}$ and $\bar{q}$ denote the sample means, and let S_p, S_q and S_{pq} denote the unbiased $(1/(n-1))$ sample covariance matrices. Write

$$A = \mathrm{tr}(S_p) + \mathrm{tr}(S_q), \qquad C = \mathrm{tr}(S_{pq}), \qquad d = \|\bar{p} - \bar{q}\|^2. \tag{5}$$

Plug-in estimator. $\hat{\rho}_P^{\mathrm{plug}} = 2C/(A + d)$.

U-statistic estimator. Following the concordance literature we estimate the two expectations in Theorem 1 separately,

$$\hat{\Delta}_{\mathrm{dep}} = \frac{1}{n}\sum_{i=1}^{n} \|p_i - q_i\|^2, \qquad \hat{\Delta}_{\mathrm{indep}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j \ne i} \|p_i - q_j\|^2, \tag{6}$$

and set $\hat{\rho}_P^U = 1 - \hat{\Delta}_{\mathrm{dep}}/\hat{\Delta}_{\mathrm{indep}}$. Direct evaluation of the double sum requires $O(n^2 K)$ operations, but an exact linear-time form is available.

Theorem 2 (Exact $O(nK)$ estimation)**.** *For every $n \geq 2$, and writing $\hat{\Delta}_{\mathrm{indep}}$ for the double sum above,*

$$\hat{\Delta}_{\mathrm{indep}} = A + \frac{n}{n-1}\, d - \frac{1}{n-1}\hat{\Delta}_{\mathrm{dep}}, \qquad \textit{so that} \qquad \hat{\rho}_P^U = \frac{2C}{\big(1-\frac{1}{n}\big)A + \frac{2}{n}C + d}. \tag{7}$$

If in addition $\hat{\Delta}_{\mathrm{indep}} > 0$, then $\hat{\rho}_P^U = 2C/\big(A + d - \frac{1}{n}(A - 2C)\big)$ with $A - 2C = \mathrm{tr}(S_{p-q}) \geq 0$, so $\hat{\rho}_P^U \geq \hat{\rho}_P^{\mathrm{plug}}$ whenever $C \geq 0$, and $\hat{\rho}_P^U - \hat{\rho}_P^{\mathrm{plug}} = O_p(n^{-1})$ in general.

The closed form eliminates the quadratic computation without approximation. Thus, the U-statistic and plug-in estimators both require $O(nK)$ operations and differ only by a term of order n^{-1}.

Theorem 3 (Asymptotic normality)**.** *Let $\theta = (\mu_p, \mu_q, \mathrm{tr}\,\Sigma_p, \mathrm{tr}\,\Sigma_q, \mathrm{tr}\,\Sigma_{pq})$ and let f denote the map $\theta \mapsto \rho_P$ in* (1)*. If the denominator of* (1) *is strictly positive at θ, then both estimators satisfy $\sqrt{n}(\hat{\rho}_P - \rho_P) \xrightarrow{d} \mathcal{N}(0, \sigma_\rho^2)$ with $\sigma_\rho^2 = \nabla f(\theta)^\top \Sigma_\theta \nabla f(\theta)$, where Σ_θ is the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta} - \theta)$.*

Confidence intervals. We estimate σ_ρ by a delete-one jackknife, computed in $O(nK)$ by downdating the four sufficient moments, and form confidence intervals on the Fisher scale $z = \mathrm{arctanh}\,\rho_P$ (Fisher, 1915) with $\widehat{\mathrm{se}}_z = \widehat{\mathrm{se}}_\rho/(1-\hat{\rho}_P^2)$, mapping back through $\tanh$ so the endpoints stay in $[-1, 1]$. Because $|C| \leq A/2$, the U-statistic lies in $[-1/(1-2/n), 1]$ and can fall below -1 in small, strongly anti-concordant samples; the released implementation truncates it to $[-1, 1]$ and, when $|\hat{\rho}_P| = 1$, where the transform is undefined, reports a Wald interval truncated to $[-1, 1]$. A percentile bootstrap (Efron, 1979; Efron & Tibshirani, 1994) is provided as an alternative. Coverage is verified in Appendix C. Because estimates for two models on the same test set are positively correlated, separate intervals are conservative for comparing them; a paired interval for $\rho_{P,A} - \rho_{P,B}$ should be used instead (Appendix B.1).

5 What P-CCC Rewards

Because an evaluation metric may also be used as an optimization criterion, it is important to understand what it rewards. The following results show that P-CCC penalizes uninformative predictive variation but may favor overconfident forecasts, whose predicted probabilities are more extreme than the corresponding calibrated probabilities. This behavior motivates reporting P-CCC alongside a strictly proper score.

Proposition 2. *(a)* (Uninformative sharpness is penalized.) *Let $\tilde{p} = p + \varepsilon$, where $\mathbb{E}[\varepsilon \mid p, q] = 0$ and $\mathrm{tr}(\mathrm{Cov}(\varepsilon)) > 0$, and where $\mathbf{1}^\top \varepsilon = 0$ and $p + \varepsilon \in \Delta^{K-1}$ almost surely so that $\tilde{p}$ is again a probability forecast. Then $\rho_P(\tilde{p}, q) < \rho_P(p, q)$ whenever $\mathrm{tr}(\Sigma_{pq}) > 0$.*

(b) (Affine supremum.) *Fix a binary forecaster s with $\mathrm{Var}(s) > 0$ and $\mathrm{Cov}(s, y) > 0$, and let $r = \mathrm{Cor}(s, y)$. Over the unconstrained affine family $s_{a,b} = as + b$ with $a, b \in \mathbb{R}$,*

$$\sup_{a,b}\ \rho_P(s_{a,b}, y) = r, \qquad \textit{attained at} \quad a^* = \frac{\sigma_y}{\sigma_s}, \quad b^* = \mathbb{E}[y] - a^*\mathbb{E}[s]. \tag{8}$$

(c) (Truthful forecasts are not optimal.) *Let $p^* = \Pr(y = 1 \mid x)$ be a binary Bayes-optimal forecaster with $\Pr\{p^* \notin \{0, \pi, 1\}\} > 0$, which implies a positive Bayes risk. For $\phi(p) = (p - \pi)\,p(1-p)$, the forecast $s_\varepsilon = p^* + \varepsilon\phi(p^*)$ lies in $[0, 1]$ for every $\varepsilon \in [0, 1]$, and*

$$\frac{d}{d\varepsilon}\,\rho_P(s_\varepsilon, y)\Big|_{\varepsilon=0} = \frac{2\,\mathbb{E}\big[(p^* - \pi)^2 p^*(1-p^*)\big]\,\mathbb{E}\big[p^*(1-p^*)\big]}{\{\mathrm{Var}(p^*) + \mathrm{Var}(y)\}^2} > 0. \tag{9}$$

Part (a) shows that P-CCC penalizes uninformative predictive variation: adding noise that is conditionally mean-zero given the prediction and the label increases predictive variance without increasing agreement with the outcome, thereby lowering the coefficient. The simplex conditions ensure that the perturbed predictions remain valid probability vectors; in the binary case they hold for $\varepsilon = p(1-p)\eta$ with η independent of (p, q), $\mathbb{E}\eta = 0$ and $|\eta| \leq 1$. Requiring ε to be independent of (p, q) would be too strong: when predictions approach both 0 and 1, as for $p^* = \mathrm{expit}(z)$, independence and feasibility together force $\varepsilon = 0$. Part (b) identifies a competing incentive. Within the unconstrained

affine family, P-CCC is maximized by matching predictive variance to label variance, which generally differs from the calibrated target. Because realized labels are more dispersed than their generating conditional probabilities, this variance-matching objective may favor predictions that are more extreme than the calibrated probabilities. The affine optimizer may fall outside $[0, 1]$ and therefore need not define a valid probability forecast; for the process in §7, this occurs for 38% of predictions. Remark 2 provides further details. For the Bayes forecaster, $\sigma_y^2 = \mathrm{Var}(p^*) + \mathbb{E}[p^*(1-p^*)]$, so $a^* = \sigma_y/\sigma_{p^*} > 1$ whenever the Bayes risk is positive. Part (c) shows that the same preference holds among valid probability forecasts: a small sharpening of the Bayes forecast away from the prevalence stays in $[0, 1]$ and strictly increases P-CCC. Proposition 2(c) shows that P-CCC is not maximized by the true conditional probabilities. We therefore describe it as improper, extending the terminology of Gneiting & Raftery (2007) from scoring rules to a functional of the joint distribution of predictions and labels. The dispersion identity of §3 explains why: a calibrated forecaster satisfies $\mathrm{tr}(\Sigma_p) = \mathrm{tr}(\Sigma_q) - \mathrm{BS}$, so it is less dispersed than the realized labels whenever $\mathrm{BS} > 0$, and Proposition 1(b) then gives $C_b < 1$. For the process in §7, $\mathrm{tr}(\Sigma_p) = 0.0868$, $\mathrm{tr}(\Sigma_q) = 0.5000$, and $\mathrm{BS} = 0.4138$, giving $C_b = 0.710$ at the Bayes forecast. P-CCC therefore favors agreement with realized labels by moving predictive variance toward label variance, whereas probabilistic accuracy targets the less dispersed conditional probabilities. The effect is quantified in §7.3 and §8.

6 Extensions: Weighted P-CCC for Class Imbalance

Under the unweighted Euclidean geometry, frequent classes may dominate the coefficient. For a positive-definite W, replacing $\|\cdot\|^2$ by $\|x\|_W^2 = x^\top W x$ throughout gives

$$\rho_P^{(W)} = \frac{2\,\mathrm{tr}(W\Sigma_{pq})}{\mathrm{tr}(W\Sigma_p) + \mathrm{tr}(W\Sigma_q) + (\mu_p - \mu_q)^\top W(\mu_p - \mu_q)}. \tag{10}$$

The proofs of Theorem 1, Theorem 2, Proposition 1(b) and (c), Proposition 2(a) and Theorem 3 use only bilinearity and positive-definiteness of $\langle x, y\rangle_W$, so all of them carry over with traces replaced by W-traces. Proposition 1(a) is the exception, being specific to the Euclidean trace; Appendix B gives its W-form and explains what the choice $W = \mathrm{diag}(1/\pi_k)$ does and does not achieve. Appendix C shows a case where the unweighted coefficient is indifferent between a model that never detects the rare class and one with rare-class recall 0.821, as is the Brier score, whereas (10) separates them.

7 Simulation Study

Table 1: Eight forecasters of $y \sim \mathrm{Bernoulli}(\mathrm{expit}(z))$, $z \sim \mathcal{N}(0, 1)$, scored on $n = 10^4$ observations; means over 500 replications, with Brier in the summed form. AUC cannot separate A, D, E and G, ECE is minimized by the uninformative B, and the P-CCC ranks the overconfident D and E above the Bayes-optimal A.

	Model	Forecast $s(z)$	Acc.	AUC	ECE	Brier	BSS	P-CCC
A	Oracle	$\mathrm{expit}(z)$	0.675	0.739	0.009	0.413	0.173	0.295
B	Lazy	$\hat{\pi}$	0.504	0.500	0.000	0.500	0.000	0.000
C	Random	$\mathrm{Bern}(0.5)$	0.500	0.500	0.500	1.000	-1.000	0.000
D	Overconfident	$\mathrm{expit}(z/0.2)$	0.675	0.739	0.220	0.523	-0.047	0.384
E	Sharpened	$\mathrm{expit}(z/0.3)$	0.675	0.739	0.175	0.484	0.032	0.386
F	Noisy	$\mathrm{expit}(z) + \varepsilon$	0.644	0.696	0.063	0.455	0.089	0.270
G	Underconfident	$\mathrm{expit}(z/2.5)$	0.675	0.739	0.097	0.439	0.122	0.154
H	Anti-correlated	$\mathrm{expit}(-z)$	0.325	0.261	0.349	0.760	-0.520	-0.295

This section asks a practical question: if a results table already carries accuracy, AUC, a proper score and a calibration error, does the P-CCC add anything? We show that the standard metrics fail on forecasters the P-CCC separates (§7.1), how its rankings relate to theirs (§7.2), and that its own failure mode is specific and predictable (§7.3).

Simulation design. All three studies use $z \sim \mathcal{N}(0, 1)$, $p^* = \mathrm{expit}(z)$ and $y \sim \mathrm{Bernoulli}(p^*)$, a balanced problem with positive Bayes risk whose Bayes forecaster has $\rho_P = 0.295$. The P-CCC is

computed with the U-statistic of Theorem 2 and ECE with 15 equal-width bins. The Brier score is the summed form $\mathbb{E}\|p - q\|^2$ used throughout, twice the familiar binary score, so the constant forecast $s \equiv 0.5$ scores 0.5. Estimator accuracy, coverage and a class-imbalance study are in Appendix C.

7.1 Standard metrics fail on forecasters the P-CCC separates

Table 1 defines eight forecasters, each isolating one way a forecast can go wrong. A is the Bayes forecaster $\mathrm{expit}(z)$. D, E and G keep its ranking but rescale confidence by a temperature, $\mathrm{expit}(z/T)$: D ($T = 0.2$) and E ($T = 0.3$, near the P-CCC optimum of §7.3) are overconfident, and G ($T = 2.5$) is underconfident. F adds noise, $\varepsilon \sim \mathcal{N}(0, 0.15^2)$ clipped to $(0, 1)$, and H reverses the signal. B and C use no information: B reports the sample prevalence $\hat{\pi}$ for every case; C issues hard $0/1$ predictions independent of z. Each is scored on $n = 10^4$ observations, averaged over 500 replications; the Monte Carlo standard deviation of the P-CCC never exceeds 0.011.

What the standard metrics miss. *AUC cannot see calibration*: A, D, E and G rank every pair of cases identically, so all four score 0.739, and AUC reports the Bayes forecaster and three distortions of it as one model. *ECE rewards a model that has learned nothing*: B ignores its input, yet its ECE of 0.000 is the best in the table, because a constant equal to the prevalence is calibrated by construction. *The Brier score lacks a fixed scale, and the usual remedy measures something else*: A's 0.413, against 0.500 for the constant forecast, cannot be judged without the prior and the noise level, while the climatological skill score, which does supply a scale, asks whether a model beats the base rate rather than whether it carries information, and gives -1.000 to C and 0.000 to B although neither uses its input.

What the P-CCC adds. It places B and C together at its no-information point (0.000), gives the reversed forecaster H a negative value (-0.295), and ranks every positively informative forecaster above zero, on a scale whose zero does not move with the prior. Where $\mathrm{BS} = 0.413$ needs context, $\rho_P = 0.295$ says directly that the Bayes forecaster removes about 30% of the error that chance-level agreement leaves. Plotting the eight in the (BS, ρ_P) plane (Figure A1) shows that neither coordinate separates them alone: B and C share a P-CCC of 0.000 while their Brier scores differ twofold, and D and E have worse Brier scores than A but higher P-CCC. No two coincide in the plane, the first indication that the two numbers are best read together.

7.2 How the P-CCC ranking relates to existing metrics

A new metric is most informative when it orders models differently from those already reported. We drew 400 forecasters $s = \mathrm{expit}\{(gz + b)/T\} + \varepsilon$, clipped to $(0, 1)$, varying the signal strength $g \sim U(0.3, 1)$, bias $b \sim \mathcal{N}(0, 0.6^2)$, temperature T log-uniform on $[0.2, 4]$ and noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = |\mathcal{N}(0, 0.25^2)|$. All are scored on one common sample of $n = 20{,}000$, so differences reflect the forecasters rather than sampling, and Figure 1 shows Kendall's τ between the rankings each metric induces.

Two things follow. First, Brier and BSS agree perfectly, $\tau = 1.00$: on a fixed evaluation set the climatological skill score is an affine function of the Brier score, so a baseline that does not depend on the model being scored can never reorder anything. Second, the P-CCC agrees only moderately with everything else, at $\tau = 0.28$ with Brier, $\tau = 0.57$ with AUC and $\tau = 0.08$ with ECE. It sits between discrimination and proper scoring without coinciding with either, which is the empirical content of the claim that it answers a third question. Because the generating family varies calibration, these differences reflect miscalibration, and their size depends on how much miscalibration the family contains: by (4), calibrated forecasters would be ranked identically by the P-CCC and the Brier score.

7.3 What it rewards, and where it fails

The same experiments locate the weakness precisely, and a metric whose failure mode is known is safer than one whose is not. We sweep the temperature T of the Bayes forecaster, $p_T = \mathrm{expit}(z/T)$, over 80 values in $[0.1, 3]$ on a single sample of $n = 400{,}000$, so that $T = 1$ is exactly calibrated and $T < 1$ is overconfident (Figure 2). The Brier score and ECE are both optimized essentially at $T = 1$ ($T^* = 0.98$ and 1.02), as the properness of the Brier score and the calibration of the $T = 1$ forecast imply. The P-CCC peaks at $T^* = 0.28$, a $3.5\times$ oversharpening, reaching 0.387 against 0.294 at

calibration, by which point the ECE has degraded to 0.182. This is Proposition 2(c) made visible. The maximum is interior rather than at $T \to 0$ because, along the temperature path, the covariance with the labels saturates as predictions harden while predictive variance keeps growing. At T^* the predictive dispersion is $\mathrm{tr}(\Sigma_p) = 0.2975$, still below the label dispersion $\mathrm{tr}(\Sigma_q) = 0.5000$, so the affine variance-matching point of Proposition 2(b) is not reached within this family.

Table 1 shows the same effect in one column: the overconfident model D scores 0.384 and the sharpened oracle E scores 0.386, both above the Bayes-optimal A at 0.295. The implication is direct: the P-CCC should not be optimized and should not select a model on its own. Alongside a proper score it is safe, because the proper score is minimized at $T = 1$ exactly where the P-CCC is not, so the pair cannot be moved in a favorable direction by sharpening alone.

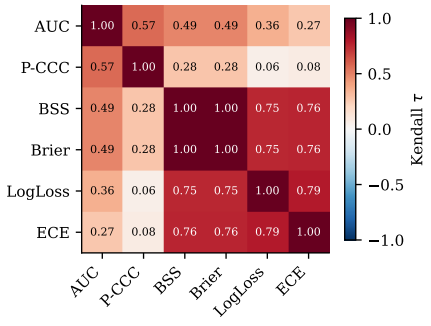

Figure 1: Kendall's τ between metric rankings over randomly generated forecasters, all oriented so higher is better. Brier and BSS are rank-identical; over this miscalibrated family the P-CCC agrees only moderately with the others, reordering models rather than restating an existing ranking.

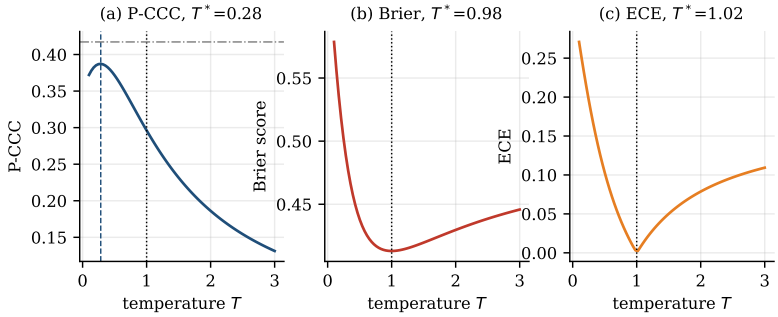

Figure 2: Metric as a function of forecast temperature T ($T = 1$ is calibrated, dotted line). The Brier score (b) and ECE (c) are optimized at calibration; the P-CCC (a) is maximized at $T^* = 0.28$ (dashed), a substantially overconfident forecast. Gray line in (a): the affine-family supremum $r = 0.417$ (Proposition 2(b)).

8 Benchmark Evidence: Does It Matter on Real Models?

We next examine whether these findings extend to models fitted to public datasets.

Setup. We train 6 classifier families, namely logistic regression, Gaussian naive Bayes, a random forest, histogram gradient boosting, a multilayer perceptron and an RBF SVM, on 5 public datasets (adult, breast-cancer, credit-g, digits, wine) using scikit-learn (Pedregosa et al., 2011). Each dataset is split $60/20/20$ into training, calibration and test parts, stratified by class. Every fitted model is evaluated raw and under three recalibration methods, namely temperature scaling (Guo et al., 2017), Platt scaling (Platt, 1999) and isotonic regression (Zadrozny & Elkan, 2002), giving 116 model and treatment combinations. Preprocessing details and per-dataset results are in Appendix C.

P-CCC provides complementary ranking information. Averaging Kendall's τ within datasets, the P-CCC correlates 0.79 with the Brier score, 0.59 with AUC and 0.34 with ECE (Table A5). Agreement is higher than in §7.2, which is expected, since fitted models span a narrower range of sharpness than a synthetic sweep. Unlike the simulation, where the P-CCC agrees most with AUC, here it agrees most with the Brier score; in both settings it agrees least with ECE.

Recalibration and the predicted incentive. Across all 87 recalibrations, the Brier score improves in 63.2% of cases (mean change -0.0134), log-loss in 70.1%, ECE in 56.3% and the P-CCC in 54.0% (Table A6, Appendix C). Two caveats qualify this comparison. Platt and isotonic calibration use `CalibratedClassifierCV`, which refits the base model on the combined training and calibration data, changing the model as well as its calibration (AUC changes in 88% of those cases); under temperature scaling alone, the only arm that recalibrates the fitted model, the Brier score improves in 65.5% of cases and the P-CCC in 69.0%. The difference between the aggregate rates is also not significant (exact McNemar test, $p = 0.13$). Proposition 2(c) is instead visible in the direction of the temperature change: all 13 of the 13 fits that sharpened the model ($T < 1$) raised the P-CCC, while the 16 that softened it lowered it in 9 cases; MLP on credit-g is the clearest such case, its Brier score improving from 0.476 to 0.373 while the P-CCC falls from 0.339 to 0.272. The two changes agree in direction in 65 of 87 cases (Spearman -0.45; the near-zero Pearson $r = -0.07$ reflects pooling datasets on different scales), so the metrics usually agree about recalibration and disagree, in a minority of cases, in the predicted direction. P-CCC should therefore not be used alone.

9 Discussion and Conclusion

We introduced P-CCC by formulating the evaluation of probabilistic classifiers as an agreement problem between predicted probability vectors and realized outcomes on the simplex. It normalizes the Brier score by the error the same forecaster would incur against independent labels, yielding a bounded, chance-corrected measure with a prior-independent zero point. An exact $O(nK)$ estimator and accompanying confidence intervals permit efficient estimation with uncertainty quantification. P-CCC complements rather than replaces established measures. AUC summarizes ranking performance but is insensitive to calibration changes that preserve ordering; ECE assesses calibration but may be optimized by an uninformative prevalence-only forecast. The Brier score is strictly proper and evaluates overall probabilistic accuracy, although its magnitude depends on prevalence, case mix, and irreducible uncertainty. P-CCC adds a model-relative interpretation: the proportion of chance-level squared error removed by dependence between predictions and labels.

Why report $(\mathrm{BS}, \hat{\rho}_P)$**.** The Brier score measures probabilistic error and encourages truthful probability reporting, whereas P-CCC expresses prediction-label agreement on a bounded, chance-referenced scale. This pairing is important because P-CCC is not proper and may favor forecasts more dispersed than the true conditional probabilities. We therefore recommend reporting P-CCC with a confidence interval alongside a strictly proper score, comparing models through a paired interval for their difference (§4), and not using it alone for model selection or optimization. It adds interpretation and a dispersion diagnostic (4) rather than a new ranking of calibrated models. Simulations and benchmarks support this complementary interpretation: P-CCC is not rank-equivalent to AUC or ECE, and it differs from the Brier score for miscalibrated forecasts while ranking calibrated forecasts as the Brier score does; its response to temperature scaling agrees with the theoretical preference for dispersion. The weighted extension accommodates class-specific importance.

Limitations and future work. P-CCC is a functional of the joint distribution over an evaluation population and therefore remains case-mix dependent. Although its zero point does not depend on prevalence, its magnitude does: by (4), for calibrated forecasters it is a monotone function of the Brier skill score and so inherits that score's dependence on prevalence, as Cohen's κ does. A zero value denotes no net first-order agreement, not necessarily absence of predictive information, and the coefficient is uninformative when only one class is observed. Its model-relative baseline provides the chance correction but also underlies its improperness. Evaluation under distribution shift remains needed. Further work may consider baselines tied to irreducible uncertainty, structured outcome geometries through W, and a calibration-resolution decomposition.

Reproducibility Statement

All results are produced by the released code (`pccc.py`, `metrics.py`, `simulation.py`, `benchmark.py`) under fixed seeds. `test_pccc.py` numerically checks the formal results of this paper, including Theorem 2 against the brute-force $O(n^2)$ U-statistic and the gradient (11) against finite differences; Theorem 3 is checked through interval coverage in a single design. Benchmark datasets are obtained through scikit-learn and OpenML.

Large Language Models Usage Statement

Large language models (LLMs) were employed to polish the language, improve grammar, and enhance the readability of this manuscript. All conceptual ideas, technical contributions, analyses, and conclusions presented in this work are the authors' own. For software development, we utilized a code editor with generative AI functionalities, primarily to correct coding errors. We have carefully verified all statements to ensure that the final version of the manuscript accurately reflects our intended meaning and contributions.

References

Huiman X Barnhart, Michael Haber, and Jingli Song. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, 58(4):1020–1027, 2002.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Josep L Carrasco and Lluis Jover. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, 59(4):849–858, 2003.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Bradley Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Sasiprapa Hiriote and Vernon M Chinchilli. Matrix-based concordance correlation coefficient for repeated measures. *Biometrics*, 67(3):1007–1016, 2011.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Tonya S King and Vernon M Chinchilli. A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine*, 20(14):2131–2147, 2001.

Tonya S King, Vernon M Chinchilli, and Josep L Carrasco. A repeated measures concordance correlation coefficient. *Statistics in Medicine*, 26(16):3095–3113, 2007.

Lawrence I-K Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45 (1):255–268, 1989.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15682–15694, 2021.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12 (4):595–600, 1973.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

A PROOFS

A.1 PROOF OF THEOREM 1

Let q' be independent of p with the law of q. Then $\mathbb{E}[p - q'] = \mu_p - \mu_q$ and, by independence, $\mathrm{Cov}(p - q') = \Sigma_p + \Sigma_q$. For any random vector v, $\mathbb{E}\|v\|^2 = \mathrm{tr}(\mathrm{Cov}(v)) + \|\mathbb{E}v\|^2$. Therefore

$$\mathbb{E}\|p - q'\|^2 = \mathrm{tr}(\Sigma_p + \Sigma_q) + \|\mu_p - \mu_q\|^2 = \mathrm{tr}(\Sigma_p) + \mathrm{tr}(\Sigma_q) + \|\mu_p - \mu_q\|^2,$$

which is the denominator of (1). Dividing the numerator $\mathbb{E}\|p - q\|^2 = \mathrm{BS}$ by this quantity and subtracting from one gives the claim. □

A.2 PROOF OF PROPOSITION 1

(a) Each q_k is Bernoulli(π_k), so $\mathrm{tr}(\Sigma_q) = \sum_k \pi_k(1 - \pi_k) = 1 - \|\pi\|^2$.

(b) Write $A_p = \mathrm{tr}(\Sigma_p)$, $A_q = \mathrm{tr}(\Sigma_q)$, $C = \mathrm{tr}(\Sigma_{pq})$ and $d = \|\mu_p - \mu_q\|^2$, with $A_p, A_q > 0$. The identity is immediate:

$$r_{\mathrm{tr}}\, C_b = \frac{C}{\sqrt{A_p A_q}} \cdot \frac{2\sqrt{A_p A_q}}{A_p + A_q + d} = \frac{2C}{A_p + A_q + d} = \rho_P.$$

For the range of r_{tr}, apply Cauchy–Schwarz coordinatewise and then to the resulting sums:

$$|C| = \Big|\sum_k \mathrm{Cov}(p_k, q_k)\Big| \le \sum_k \sqrt{\mathrm{Var}(p_k)\,\mathrm{Var}(q_k)} \le \sqrt{\sum_k \mathrm{Var}(p_k)}\sqrt{\sum_k \mathrm{Var}(q_k)} = \sqrt{A_p A_q},$$

so $|r_{\mathrm{tr}}| \le 1$. For C_b, positivity is clear and the inequality of arithmetic and geometric means (AM–GM inequality) gives $A_p + A_q + d \ge A_p + A_q \ge 2\sqrt{A_p A_q}$, so $C_b \le 1$ with equality exactly when $d = 0$ and $A_p = A_q$. Combining, $|\rho_P| = |r_{\mathrm{tr}}| C_b \le 1$. If $A_p = 0$ then p is almost surely constant, so $C = 0$ and $\rho_P = 0$ whenever $A_p + A_q + d > 0$; likewise if $A_q = 0$.

The chain $\rho_P \le 2C/(A_p + A_q) \le r_{\mathrm{tr}}$ that one might write instead is valid only for $C \ge 0$. For $C < 0$, removing d from the denominator makes a negative ratio more negative and the AM–GM step reverses, so both inequalities run the other way; Model H of Table 1 realizes this.

(c) Both coefficients have the form $1 - \mathrm{BS}/D$. For the P-CCC, $D_P = A_p + A_q + d$; for the climatological skill score, $D_{\mathrm{BSS}} = \mathbb{E}\|\pi - q\|^2 = A_q$ by part (a). Since $A_p \ge 0$ and $d \ge 0$ we have $D_P \ge D_{\mathrm{BSS}} > 0$, hence $\mathrm{BS}/D_P \le \mathrm{BS}/D_{\mathrm{BSS}}$ and $\rho_P \ge \mathrm{BSS}$. Writing the gap as $\rho_P - \mathrm{BSS} = \mathrm{BS}\,(1/D_{\mathrm{BSS}} - 1/D_P)$, the second factor vanishes only when $D_P = D_{\mathrm{BSS}}$, that is when $A_p + d = 0$, which forces p to be almost surely constant and equal to π. The product therefore vanishes if and only if $\mathrm{BS} = 0$ or that condition holds, and both branches occur: a forecaster with $p = q$ almost surely has $\mathrm{BS} = 0$ and $\rho_P = \mathrm{BSS} = 1$ while $A_p + d = A_q > 0$. □

A.3 Proof of Theorem 2

Write $\hat{\Delta}_{\mathrm{dep}} = \frac{1}{n}\sum_i \|p_i - q_i\|^2$. Expanding the full double sum,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}\|p_i - q_j\|^2 &= n\sum_i \|p_i\|^2 + n\sum_j \|q_j\|^2 - 2\Big(\sum_i p_i\Big)^{\top}\Big(\sum_j q_j\Big)\\ &= n\big[(n-1)\,\mathrm{tr}(S_p) + n\|\bar{p}\|^2\big] + n\big[(n-1)\,\mathrm{tr}(S_q) + n\|\bar{q}\|^2\big] - 2n^2\bar{p}^{\top}\bar{q}\\ &= n(n-1)A + n^2\|\bar{p} - \bar{q}\|^2 = n(n-1)A + n^2 d,\end{aligned}$$

using $\sum_i \|p_i\|^2 = (n-1)\,\mathrm{tr}(S_p) + n\|\bar{p}\|^2$ and its analogue for q. Subtracting the diagonal $\sum_i \|p_i - q_i\|^2 = n\hat{\Delta}_{\mathrm{dep}}$ and dividing by $n(n-1)$,

$$\hat{\Delta}_{\mathrm{indep}} = A + \frac{n}{n-1}d - \frac{1}{n-1}\hat{\Delta}_{\mathrm{dep}},$$

which is the first claim. Note the factor $n/(n-1)$ multiplies d alone (scaling the whole bracket by $n/(n-1)$ instead would inflate the baseline by exactly $A/(n-1)$). For the second claim, the standard identity $\sum_i \|u_i\|^2 = (n-1)\,\mathrm{tr}(S_u) + n\|\bar{u}\|^2$ applied to $u_i = p_i - q_i$ gives $\hat{\Delta}_{\mathrm{dep}} = \frac{n-1}{n}(A - 2C) + d$. Substituting,

$$\hat{\Delta}_{\mathrm{indep}} = A + \frac{n}{n-1}d - \frac{1}{n-1}\Big[\frac{n-1}{n}(A - 2C) + d\Big] = \Big(1 - \tfrac{1}{n}\Big)A + \tfrac{2}{n}C + d.$$

Finally $\hat{\Delta}_{\mathrm{indep}} - \hat{\Delta}_{\mathrm{dep}} = \big[(1 - \frac{1}{n})A + \frac{2}{n}C + d\big] - \big[(1 - \frac{1}{n})A - \frac{2(n-1)}{n}C + d\big] = 2C$, so $\hat{\rho}_P^U = 1 - \hat{\Delta}_{\mathrm{dep}}/\hat{\Delta}_{\mathrm{indep}} = 2C/\hat{\Delta}_{\mathrm{indep}}$, as claimed. □

Second display and the comparison with the plug-in. From (7), $(1 - \frac{1}{n})A + \frac{2}{n}C + d = A + d - \frac{1}{n}(A - 2C)$. Now $A - 2C = \mathrm{tr}(S_p) + \mathrm{tr}(S_q) - 2\,\mathrm{tr}(S_{pq}) = \mathrm{tr}(S_{p-q}) \ge 0$ since S_{p-q} is a sample covariance matrix. Hence the U-statistic denominator is no larger than the plug-in denominator $A + d$, and if $C \ge 0$ the ratio is no smaller. The difference of the two estimators is $2C \cdot \frac{1}{n}(A - 2C)/[(A + d)(A + d - \frac{1}{n}(A - 2C))] = O_p(n^{-1})$, because the denominator converges in probability to the positive population value. □

A.4 Proof of Theorem 3

All coordinates of p and q lie in $[0, 1]$, so moments of every order exist; the sample moments are averages (or U-statistics of degree two (Hoeffding, 1948)) and are jointly asymptotically normal by the multivariate CLT. f is a ratio of polynomials with non-vanishing denominator, hence continuously differentiable at θ, and the delta method applies (van der Vaart, 2000, Ch. 3). The two estimators share the limit because Theorem 2 makes their difference $O_p(n^{-1}) = o_p(n^{-1/2})$. Appendix A.5 gives ∇f in closed form. □

A.5 Gradient in Theorem 3

With $N = 2\,\mathrm{tr}(\Sigma_{pq})$ and $D = \mathrm{tr}(\Sigma_p) + \mathrm{tr}(\Sigma_q) + \|\mu_p - \mu_q\|^2$, so that $\rho_P = N/D$,

$$\frac{\partial \rho_P}{\partial\,\mathrm{tr}(\Sigma_{pq})} = \frac{2}{D}, \quad \frac{\partial \rho_P}{\partial\,\mathrm{tr}(\Sigma_p)} = \frac{\partial \rho_P}{\partial\,\mathrm{tr}(\Sigma_q)} = -\frac{N}{D^2}, \quad \frac{\partial \rho_P}{\partial \mu_p} = -\frac{2N(\mu_p - \mu_q)}{D^2} = -\frac{\partial \rho_P}{\partial \mu_q}.$$

Substituting into $\sigma_\rho^2 = \nabla f^\top \Sigma_\theta \nabla f$ gives the asymptotic variance; in practice we use the jackknife, which is consistent here because f is smooth at θ.

A.6 Proof of Proposition 2

(a) Since $\mathbb{E}[\varepsilon \mid p, q] = 0$, we have $\mathbb{E}\varepsilon = 0$ and $\mathrm{Cov}(\varepsilon, p) = \mathrm{Cov}(\varepsilon, q) = 0$. Hence $\mu_{\tilde{p}} = \mu_p$ and $\mathrm{tr}(\Sigma_{\tilde{p}q}) = \mathrm{tr}(\Sigma_{pq})$, so the numerator of (1) is unchanged, while $\mathrm{tr}(\Sigma_{\tilde{p}}) = \mathrm{tr}(\Sigma_p) + \mathrm{tr}(\mathrm{Cov}(\varepsilon))$, so the denominator increases by $\mathrm{tr}(\mathrm{Cov}(\varepsilon)) > 0$; the ratio therefore falls strictly when the numerator is positive.

(b) Write $\sigma_s^2 = \mathrm{Var}(s)$, $\sigma_y^2 = \mathrm{Var}(y)$, $\sigma_{sy} = \mathrm{Cov}(s, y)$. For $s_{a,b} = as + b$ we have $\mathrm{Var}(s_{a,b}) = a^2\sigma_s^2$, $\mathrm{Cov}(s_{a,b}, y) = a\sigma_{sy}$ and $\mathbb{E}s_{a,b} = a\mathbb{E}s + b$. For $K = 2$, writing $p = (1 - s, s)$ and $q = (1 - y, y)$, each trace in (1) is twice its scalar counterpart and $\|\mu_p - \mu_q\|^2 = 2(\mathbb{E}s - \mathbb{E}y)^2$, so the factor 2 cancels and $\rho_P(s, y)$ is Lin's scalar CCC between s and y. Hence

$$\rho_P(s_{a,b}, y) = \frac{2a\sigma_{sy}}{a^2\sigma_s^2 + \sigma_y^2 + (a\mathbb{E}s + b - \mathbb{E}y)^2}.$$

For fixed a the denominator is minimized (and, when $\sigma_{sy} > 0$, the ratio maximized) by $b^* = \mathbb{E}y - a\mathbb{E}s$, killing the last term. It remains to maximize $g(a) = 2a\sigma_{sy}/(a^2\sigma_s^2 + \sigma_y^2)$ over $a > 0$. Setting $g'(a) = 0$ gives $2\sigma_{sy}(\sigma_y^2 - a^2\sigma_s^2) = 0$, so $a^* = \sigma_y/\sigma_s$, and

$$g(a^*) = \frac{2(\sigma_y/\sigma_s)\sigma_{sy}}{\sigma_y^2 + \sigma_y^2} = \frac{\sigma_{sy}}{\sigma_s\sigma_y} = r.$$

The restriction to $a > 0$ is without loss: $\mathrm{Cov}(s, y) > 0$ by hypothesis, so $g(a) \le 0$ for $a \le 0$ while $g(a^*) > 0$. Since $g(a) \to 0$ as $a \to 0$ or $a \to \infty$, this stationary point is the global maximum of the unconstrained problem. It is a supremum rather than a maximum over admissible forecasters, because Definition 1 requires $s_{a,b} \in [0, 1]$ almost surely and s_{a^*,b^*} generally violates this (Remark 2).

(c) Use the binary form of ρ_P from part (b): $\rho_P(s, y) = N/D$ with $N = 2\,\mathrm{Cov}(s, y)$ and $D = \mathrm{Var}(s) + \mathrm{Var}(y) + (\mathbb{E}s - \mathbb{E}y)^2$. At $\varepsilon = 0$ the forecast p^* is calibrated, so with $v = \mathrm{Var}(p^*)$ we have $\mathrm{Cov}(p^*, y) = v$, $\mathbb{E}p^* = \mathbb{E}y = \pi$, $N = 2v$ and $D = v + \mathrm{Var}(y)$. Because $\mathbb{E}[y \mid p^*] = p^*$, the derivatives at $\varepsilon = 0$ are $N' = 2\,\mathrm{Cov}(\phi(p^*), y) = 2\,\mathrm{Cov}(\phi(p^*), p^*)$ and $D' = 2\,\mathrm{Cov}(p^*, \phi(p^*))$, the mean term vanishing at $\varepsilon = 0$. Hence $N'D - ND' = 2\,\mathrm{Cov}(\phi(p^*), p^*)\{\mathrm{Var}(y) - v\} = 2\,\mathrm{Cov}(\phi(p^*), p^*)\,\mathbb{E}[p^*(1 - p^*)]$, and $\mathrm{Cov}(\phi(p^*), p^*) = \mathbb{E}[(p^* - \pi)^2 p^*(1 - p^*)]$, which is positive unless $p^* \in \{0, \pi, 1\}$ almost surely. For feasibility, $|s_\varepsilon - p^*| = \varepsilon|p^* - \pi|\,p^*(1 - p^*) \le p^*(1 - p^*) \le \min(p^*, 1 - p^*)$ for $\varepsilon \le 1$, so $s_\varepsilon \in [0, 1]$. For the process of §7, $\varepsilon = 1$ keeps s_ε in $[0.004, 0.994]$ and raises ρ_P from 0.295 to 0.325. □

B Algorithm and Extensions

Because the proofs use only bilinearity and positive-definiteness of $\langle x, y\rangle_W$, the following carry over with traces replaced by W-traces: Theorem 1 (the independence-baseline identity), Theorem 2 (estimation), Proposition 1(b) (the factorization, hence the $[-1, 1]$ range) and (c), Proposition 2(a), and Theorem 3. Proposition 1(a) is the exception, since it is specific to the Euclidean trace. For diagonal $W = \mathrm{diag}(w_k)$, its W-form is $\mathrm{tr}(W\Sigma_q) = \sum_k w_k\pi_k(1 - \pi_k)$, which for $W = \mathrm{diag}(1/\pi_k)$ equals $K - 1$ and so no longer depends on the prior at all. The effect of that choice is not to equalize the classes: class k contributes $1 - \pi_k$ to the baseline, and for the priors of Appendix C these are $(0.45, 0.70, 0.88, 0.97)$. What it does is let a rare class contribute almost as much as a common one, instead of being down-weighted in proportion to its prevalence.

Remark 2 (The affine optimizer is not a forecast)**.** Definition 1 confines forecasts to the simplex, and s_{a^*,b^*} generally violates that. For the data-generating process of §7 we have $a^* = 2.40$ and $b^* = -0.70$, so s_{a^*,b^*} ranges over $[-0.68, 1.67]$ and 38% of its values fall outside $[0, 1]$. The supremum in (8) is therefore unattained over admissible forecasters, which is consistent with the temperature sweep of §7.3 peaking at 0.387 below $r = 0.417$. Proposition 2(c) accordingly establishes the result along a direction that never leaves the simplex.

Algorithm 1 states the estimator of §4 in the form the released implementation takes.

Algorithm 1 P-CCC with a Fisher-z interval, $O(nK)$ time and $O(K)$ memory

input predictions $P \in \mathbb{R}^{n\times K}$, one-hot targets $Q \in \mathbb{R}^{n\times K}$, level α, optional weight matrix W
$\bar{p} \leftarrow \text{colMean}(P)$; $\bar{q} \leftarrow \text{colMean}(Q)$; $\tilde{P} \leftarrow P - \bar{p}$; $\tilde{Q} \leftarrow Q - \bar{q}$
$A \leftarrow \frac{1}{n-1}\big[\langle \tilde{P}W, \tilde{P}\rangle + \langle \tilde{Q}W, \tilde{Q}\rangle\big]$; $\quad C \leftarrow \frac{1}{n-1}\langle \tilde{P}W, \tilde{Q}\rangle$; $\quad d \leftarrow (\bar{p} - \bar{q})^\top W(\bar{p} - \bar{q})$
$\hat{\rho}_P \leftarrow 2C/\big[(1 - 1/n)A + (2/n)C + d\big]$
$\widehat{\text{se}} \leftarrow \textsc{JackknifeSE}(P, Q, W)$
return $\hat{\rho}_P$ and $\tanh\big(\operatorname{arctanh}\hat{\rho}_P \pm z_{1-\alpha/2}\widehat{\text{se}}/(1 - \hat{\rho}_P^2)\big)$

B.1 Comparing two models

Let models A and B be evaluated on the same test set. Stacking the moments of both models computed on the same observations, the argument of Theorem 3 applies to the pair $(\hat{\rho}_{P,A}, \hat{\rho}_{P,B})$, so their difference is asymptotically normal. Its standard error follows from a paired delete-one jackknife that removes the same observation from both models, reusing the $O(nK)$ downdating of Algorithm 1 for each, and yields a Wald interval and test for $\rho_{P,A} - \rho_{P,B}$. Because the two estimates are positively correlated, this interval is narrower than the comparison of separate intervals suggests. A paired bootstrap that resamples observations jointly is an alternative.

B.2 A differentiable batch objective

For a batch of size B, $\mathcal{L}_{\text{P-CCC}} = 1 - \hat{\rho}_P^{\text{plug}}$ has the exact gradient

$$\frac{\partial \mathcal{L}}{\partial p_i} = -\frac{1}{D^2}\Big[D \cdot \frac{2(q_i - \bar{q})}{B - 1} - N\Big(\frac{2(p_i - \bar{p})}{B - 1} + \frac{2(\bar{p} - \bar{q})}{B}\Big)\Big], \quad N = 2C,\ D = A + d, \tag{11}$$

verified against central finite differences to a maximum relative error below 10^{-8} in the released tests. The derivative is taken in the ambient $\mathbb{R}^K$ and so does not respect the simplex constraint on its own; in practice one composes it with a softmax layer, whose Jacobian projects the gradient onto the tangent space of Δ^{K-1}. The $2(p_i - \bar{p})/(B - 1)$ term is a variance penalty, which is the mechanism behind Proposition 2(a). We note the objective is batch-coupled (it is not a sum of per-example losses) and, by Proposition 2(c), is not minimized at the true conditional distribution; we therefore present (11) as a building block for composite objectives rather than as a drop-in replacement for cross-entropy, and we do not claim training benefits we have not measured.

C Additional Experimental Results

This appendix collects the tables and figures supporting §7 and §8. Every number in them is generated by the released code under the seeds recorded there.

C.1 Simulation study

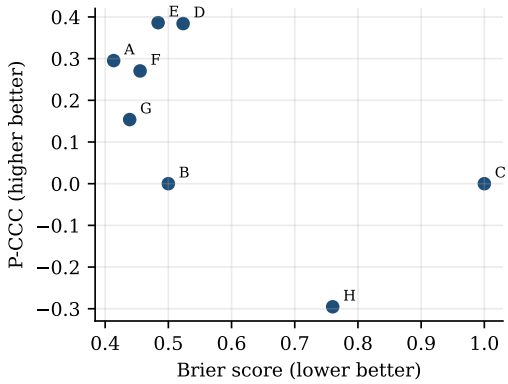

Figure A1: The eight archetypal forecasters of Table 1 in the (BS, ρ_P) plane. Neither coordinate separates them on its own: models B and C share a P-CCC of 0.000 while their Brier scores differ by a factor of two, and models D and E sit at worse Brier scores than the Bayes-optimal model A while scoring higher on the P-CCC. No two of the eight coincide in the plane.

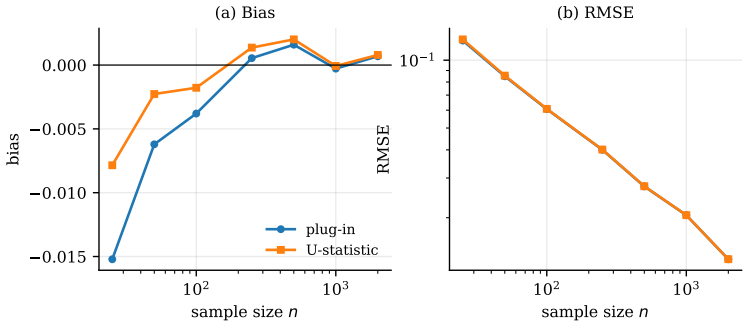

Figure A2: Bias (a) and RMSE (b) of the plug-in and U-statistic estimators against the population P-CCC, over repetitions per sample size.

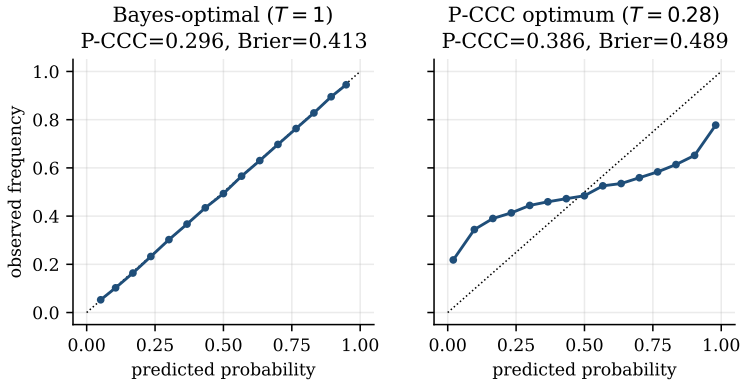

Figure A3: Reliability diagrams for the calibrated Bayes forecaster and for the forecaster that maximizes the P-CCC (§7.3). The latter is visibly off the diagonal.

Table A1: Empirical coverage and mean width of nominal 95% intervals.

	Fisher-z (jackknife)		Percentile bootstrap	
n	coverage	width	coverage	width
50	0.948	0.347	0.937	0.336
100	0.967	0.245	0.962	0.240
200	0.946	0.173	0.940	0.170
500	0.949	0.109	0.948	0.108
1000	0.934	0.077	0.931	0.077

Table A2: Class imbalance ($K = 4$, rare-class prior 0.03). Unweighted P-CCC and Brier are both indifferent between the majority specialist and the rare-class-aware model; the weighted P-CCC is not. The weighted P-CCC also ranks the rare-class-aware model, which is miscalibrated, above the Bayes forecaster, an instance of the dispersion incentive of §5.

Model	Acc.	AUC	Brier	ECE	Rare recall	P-CCC	P-CCC (W)
Oracle (Bayes)	0.893	0.987	0.175	0.103	0.194	0.794	0.678
Majority specialist	0.888	0.986	0.187	0.087	0.000	0.780	0.591
Rare-class aware	0.900	0.987	0.188	0.153	0.821	0.778	0.694

C.2 Benchmark study

Categorical features are ordinally encoded and numeric features standardized, with median and modal imputation for the missing values in `adult`. Platt and isotonic calibration use `CalibratedClassifierCV` with three-fold internal cross-validation on the combined training and calibration split, which refits the base model, so these two methods change the model as well as its calibration. The RBF SVM uses `probability=True`, whose outputs are already Platt-scaled internally. Temperature scaling fits a single scalar by minimizing calibration-split negative log-likelihood. ECE uses 15 equal-width bins on the top-label confidence. Calibrated variants and the weighted coefficient are in the released CSV files.

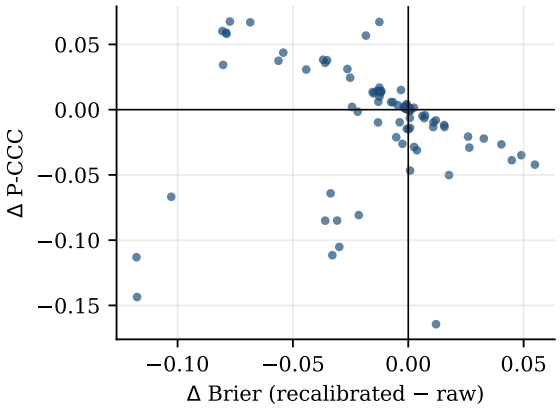

Figure A4: Change in Brier score against change in P-CCC for all 87 recalibrated models relative to their raw counterparts; negative ΔBS and positive $\Delta\hat{\rho}_P$ are improvements. The two agree in direction of improvement in 65 of 87 cases (Spearman correlation -0.45); the near-zero Pearson correlation ($r = -0.07$) reflects pooling datasets on different scales. Recalibration improves the Brier score in 55 cases, of which 15 lower the P-CCC.

Table A3: Kendall's τ between metric rankings over randomly generated forecasters (all oriented so higher is better).

	AUC	P-CCC	BSS	Brier	LogLoss	ECE
AUC	1.00	0.57	0.49	0.49	0.36	0.27
P-CCC	0.57	1.00	0.28	0.28	0.06	0.08
BSS	0.49	0.28	1.00	1.00	0.75	0.76
Brier	0.49	0.28	1.00	1.00	0.75	0.76
LogLoss	0.36	0.06	0.75	0.75	1.00	0.79
ECE	0.27	0.08	0.76	0.76	0.79	1.00

Table A4: Benchmark results, raw (uncalibrated) models.

Model	Acc.	AUC	LogLoss	Brier	BSS	ECE	P-CCC
adult ($n_{\text{test}} = 9769$, $K = 2$)							
LogReg	0.825	0.854	0.383	0.243	0.333	0.009	0.498
GaussianNB	0.808	0.856	0.783	0.317	0.130	0.146	0.440
RandomForest	0.860	0.911	0.323	0.195	0.464	0.009	0.644
HistGBM	0.873	0.927	0.280	0.178	0.510	0.011	0.673
MLP	0.850	0.901	0.343	0.208	0.430	0.018	0.626
SVM-RBF	0.871	0.922	0.283	0.166	0.452	0.010	0.692
breast-cancer ($n_{\text{test}} = 114$, $K = 2$)							
LogReg	0.991	1.000	0.041	0.018	0.961	0.027	0.980
GaussianNB	0.921	0.990	0.839	0.155	0.667	0.078	0.834
RandomForest	0.956	0.996	0.095	0.055	0.881	0.044	0.935
HistGBM	0.947	0.996	0.117	0.076	0.837	0.041	0.918
MLP	0.982	0.999	0.045	0.030	0.935	0.020	0.967
SVM-RBF	0.982	0.999	0.053	0.028	0.941	0.043	0.969
credit-g ($n_{\text{test}} = 200$, $K = 2$)							
LogReg	0.760	0.727	0.546	0.354	0.156	0.082	0.288
GaussianNB	0.740	0.765	0.590	0.366	0.129	0.083	0.434
RandomForest	0.745	0.756	0.521	0.344	0.182	0.041	0.310
HistGBM	0.755	0.703	0.738	0.384	0.086	0.124	0.378
MLP	0.725	0.718	1.427	0.476	-0.133	0.234	0.339
SVM-RBF	0.765	0.749	0.522	0.341	0.187	0.087	0.282
digits ($n_{\text{test}} = 360$, $K = 10$)							
LogReg	0.964	0.999	0.106	0.052	0.943	0.019	0.970
GaussianNB	0.883	0.981	1.965	0.220	0.756	0.110	0.877
RandomForest	0.978	0.999	0.319	0.116	0.871	0.211	0.920
HistGBM	0.969	0.999	0.133	0.046	0.948	0.021	0.974
MLP	0.964	0.999	0.106	0.051	0.944	0.014	0.971
SVM-RBF	0.975	0.999	0.141	0.046	0.949	0.067	0.973
wine ($n_{\text{test}} = 36$, $K = 3$)							
LogReg	0.972	0.997	0.102	0.053	0.920	0.048	0.958
GaussianNB	1.000	1.000	0.015	0.005	0.993	0.014	0.996
RandomForest	0.972	0.994	0.197	0.097	0.853	0.123	0.914
HistGBM	0.917	0.990	0.223	0.125	0.810	0.084	0.903
MLP	1.000	1.000	0.018	0.005	0.993	0.017	0.997
SVM-RBF	1.000	1.000	0.067	0.013	0.981	0.063	0.990

Table A5: Mean within-dataset Kendall's τ between metric rankings on the benchmark suite.

	AUC	P-CCC	BSS	Brier	LogLoss	ECE
AUC	1.00	0.59	0.70	0.70	0.68	0.21
P-CCC	0.59	1.00	0.79	0.79	0.65	0.34
BSS	0.70	0.79	1.00	1.00	0.78	0.39
Brier	0.70	0.79	1.00	1.00	0.78	0.39
LogLoss	0.68	0.65	0.78	0.78	1.00	0.47
ECE	0.21	0.34	0.39	0.39	0.47	1.00

Table A6: Effect of recalibration relative to the raw model, over all model×dataset×method combinations. Platt and isotonic calibration also refit the base model; the difference between the Brier and P-CCC improvement rates is not statistically significant (McNemar $p = 0.13$).

Metric	improves	mean change
P-CCC	54.0%	0.0074
Brier	63.2%	-0.0134
ECE	56.3%	-0.0155
LogLoss	70.1%	-0.1179
AUC	47.1%	+0.0024