Assessing Monotone Dependence: Area Under the Curve Meets Rank Correlation

Eva-Maria Walz[1,2], Andreas Eberl[2], Tilmann Gneiting[1,2]

[1]Computational Statistics Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[2]Institute of Statistics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

October 22, 2025

Abstract

The assessment of monotone dependence between random variables X and Y is a classical problem in statistics and a gamut of application domains. Consequently, researchers have sought measures of association that are invariant under strictly increasing transformations of the margins, with the extant literature being splintered. Rank correlation coefficients, such as Spearman's Rho and Kendall's Tau, have been studied at great length in the statistical literature, mostly under the assumption that X and Y are continuous. In the case of a dichotomous outcome Y, receiver operating characteristic analysis and the asymmetric area under the curve (AUC) measure are used to assess monotone dependence of Y on a covariate X. Here we unify and extend thus far disconnected strands of literature, by developing common population level theory, estimators, and tests that bridge continuous and dichotomous settings and apply to all linearly ordered outcomes. In particular, we introduce *asymmetric grade correlation*, $\mathrm{AGC}(X, Y)$, as the covariance of the mid distribution function transforms, or grades, of X and Y, divided by the variance of the grade of Y. The *coefficient of monotone association* then is $\mathrm{CMA}(X, Y) = \frac{1}{2}(\mathrm{AGC}(X, Y) + 1)$. The AGC measure has range $[-1, 1]$, and the perfect monotone predictor property holds: $\mathrm{AGC}(X, Y) = 1$ if, and only if, there is a monotonically increasing function m such that $Y = m(X)$. When X and Y are continuous, AGC is symmetric and equals Spearman's Rho. When Y is dichotomous, CMA equals AUC. Based on recent results by Pohle et al. (2025), we establish central limit theorems for the sample versions of AGC and CMA and develop a test of DeLong et al. (1988) type for the equality of AGC or CMA values with a shared outcome Y. In case studies, we apply the new measures to assess progress in data-driven weather prediction, and to evaluate methods of uncertainty quantification for large language models.

Keywords: area under the curve, asymmetric grade correlation, asymmetric Kendall correlation, C index, coefficient of monotone association, concordance, DeLong test, dependence measure, rank correlation, receiver operating characteristic, Somers' D, Spearman's Rho

1 Introduction

There is a vast extant literature on quantitative measures of dependence between real-valued random variables X and Y (Embrechts et al., 2002). Generally, one distinguishes measures of complete dependence (e.g., Rényi, 1959; Nešlehová, 2007; Székely et al., 2007; Reshef et al., 2011) from measures of concordance or monotone dependence (e.g., Schweizer and Wolff, 1981; Scarsini, 1984). Despite the focus on symmetric measures of general or monotone dependence in the statistical literature, we follow Chatterjee (2021) in this paper and argue that asymmetric measures, which honor the specific roles of X as a covariate, feature, or marker, and of Y as the outcome of interest, have distinct benefits. A key argument in favor of asymmetric measures is that, while independence is a symmetric property of X and Y, the classical notions of complete dependence (Lancaster, 1963) and perfect monotone dependence (Kimeldorf and Sampson, 1978) fail to be symmetric.

Specifically, consider the pair (X, Y) of real-valued random variables with joint distribution $\mathbb{P}$. We generally assume that neither the covariate X nor the outcome Y is almost surely constant and refer to

this setting as nondegenerate. Let $F(x) = \mathbb{P}(X \leq x)$ denote the cumulative distribution function (CDF) of X, and let G denote the CDF of Y. When F and G are continuous, Spearman's rank correlation coefficient (Spearman, 1904) is defined as

$$\rho_{\mathrm{S}} = 12 \operatorname{cov}(F(X), G(Y)), \tag{1}$$

which equals the Pearson product moment correlation between the probability integral transforms $F(X)$ and $G(Y)$. Kendall's correlation coefficient τ_{K} (Kendall, 1938) is defined as

$$\tau_{\mathrm{K}} = \mathbb{E}\left(\operatorname{sgn}(X' - X'') \operatorname{sgn}(Y' - Y'')\right), \tag{2}$$

where here and in the remainder of the paper (X, Y), (X', Y'), and (X'', Y'') have distribution $\mathbb{P}$ and are independent. Evidently, Spearman's and Kendall's correlations are symmetric in the roles of X and Y, and they are measures of monotone dependence, in the classical sense that they are invariant under strictly increasing transformations and thus can be expressed in terms of copulas (Schweizer and Wolff, 1981).[1] These and other symmetric measures of monotone dependence can readily be extended, so that they apply to arbitrary distributions, without requiring continuity of X and Y. However, the extensions are subject to an attainability problem, as for non-continuous distributions the maximum value of the measure may not be reached even though a perfect monotone relationship holds (Nešlehová, 2007; Pohle et al., 2025). Considering that the notion of perfect monotone dependence fails to be symmetric, one needs to resort to asymmetric measures if attainability is to be preserved in discrete settings.

The most prominent and most pronounced case of discreteness arises under a dichotomous or binary outcome Y. We adopt a common convention in the dichotomous case and refer to $Y = 1$ as a positive outcome, and to $Y = 0$ as a negative outcome. In this setting, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) measure (Swets, 1988; Bradley, 1997; Fawcett, 2006) are widely used tools for the assessment of monotone dependence of Y on a real-valued covariate, feature, or marker X. Specifically, if Y is a nondegenerate dichotomous outcome, let F_1 and F_0 denote the cumulative distribution function (CDF) of X conditional on $Y = 1$ and $Y = 0$, respectively. The ROC curve is a plot of the hit rate $1 - F_1(x)$ versus the false alarm rate $1 - F_0(x)$ as the decision threshold x varies. It is well known that the area under the ROC curve (AUC) equals

$$\mathrm{AUC}(X, Y) = \mathbb{E}\left(s(X', X'') \mid Y' = 0, Y'' = 1\right), \tag{3}$$

where

$$s(x', x'') = \mathbb{1}\{x' < x''\} + \frac{1}{2}\mathbb{1}\{x' = x''\} \tag{4}$$

for $x', x'' \in \mathbb{R}$; for technical details, see, e.g., Gneiting and Vogel (2022). Traditionally, ROC analysis and the AUC measure have been limited to dichotomous outcomes. However, real-valued variables are ubiquitous, and researchers have felt compelled to dichotomize their outcomes when these popular tools are to be applied, with substantial adverse effects on the analysis (Altman and Royston, 2006; Huang et al., 2024). While generalizations of ROC analysis to ordinal or real-valued outcomes have been proposed in the extant literature (Pencina and D'Agostino, 2004; Gneiting and Walz, 2022), comprehensive extensions to general population settings and associated tools for statistical inference have been elusive.

To address the general setting, let $\bar{F}(x) = \mathbb{P}(X < x) + \frac{1}{2}\mathbb{P}(X = x)$ denote the mid distribution function (MDF) of X. Similarly, let $\bar{G}$ denote the MDF of Y. To introduce our key measures, we define the *asymmetric grade correlation* (AGC) of Y on X as

$$\mathrm{AGC}(X, Y) = \frac{\operatorname{cov}(\bar{F}(X), \bar{G}(Y))}{\operatorname{cov}(\bar{G}(Y), \bar{G}(Y))},$$

and the *coefficient of monotone association* (CMA) of Y on X as

$$\mathrm{CMA}(X, Y) = \frac{1}{2}\left(\mathrm{AGC}(X, Y) + 1\right),$$

[1] If X and Y are continuous, the copula of the bivariate distribution $\mathbb{P}$ is the bivariate distribution of $(F(X), G(Y))$. As F is strictly inceasing on the support of X, and G is strictly increasing on the support of Y, we see that a measure of monotone dependence can be expressed in terms of copulas only.

respectively. Despite their asymmetry, $\mathrm{AGC}(X,Y) \in [-1,-1]$ and $\mathrm{CMA}(X,Y) \in [0,1]$ are normalized measures, and the upper bound is attained if, and only if, X is a perfect monotone predictor of Y. When the outcome Y is dichotomous then $\mathrm{CMA}(X,Y) = \mathrm{AUC}(X,Y)$. When both X and Y are continuous then CMA and AGC become symmetric, and AGC equals Spearman's rank correlation ρ_{S}. In this sense, CMA and AGC, respectively, bridge the asymmetric AUC measure in the binary case and the symmetric ρ_{S} measure in the continuous case.

For a further connection to ROC classical analysis, suppose that Y is continuous and let $\mathrm{AUC}^{(\alpha)}(X,Y) = \mathrm{AUC}(X, \mathbb{1}\{Y \geq G^{-1}(\alpha)\})$ denote the AUC measure from (3) for the dichotomized outcome at the quantile level α. One of our key results is that

$$\mathrm{CMA}(X,Y) = 6 \int \alpha\,(1-\alpha)\,\mathrm{AUC}^{(\alpha)}(X,Y)\,\mathrm{d}\alpha.$$

These findings extend ROC analysis and the AUC measure from dichotomous to general real-valued outcomes, and afford the unification of thus far disparate theories and methodologies, both at the population level and for estimation and testing in empirical settings. In particular, we establish central limit theorems for the sample versions of CMA and AGC, and we develop a general test of DeLong et al. (1988) type for the comparison of CMA or AGC for competing predictors of a shared outcome Y. Our generalization nests tests for the equality of the asymmetric AUC measure for a dichotomous outcome, and tests for the equality of Spearman's rank correlation ρ_{S} in the continuous case, respectively.

The remainder of the paper is organized as follows. The key development is in Section 2 in the population setting and in Section 3 in empirical settings, where we introduce plug-in sample versions of AGC and CMA and draw on the recent work of Pohle et al. (2025) to devise estimators and tests, covering the full range from dichotomous to continuous outcomes. Case studies on revolutionary developments in the recent transition from physics-based to data-driven weather prediction, and on the assessment of uncertainty statements for large language models, are presented in Section 4. The paper closes with a discussion in Section 5. Proofs and technical details are deferred to Appendices A and B, respectively.

2 Population versions

In this section, we define and study asymmetric grade correlation (AGC) and the coefficient of monotone association (CMA) in the population setting, where the bivariate random vector (X,Y) has distribution $\mathbb{P}$. We let $F(x) = \mathbb{P}(X \leq x)$ be the cumulative distribution function (CDF) of X, and we let G denote the CDF of Y. We generally assume that X and Y are nondegenerate in the sense that neither of them is almost surely constant.

2.1 Asymmetric grade correlation (AGC) and coefficient of monotone association (CMA)

Let $\bar{F}(x) = \mathbb{P}(X < x) + \frac{1}{2}\,\mathbb{P}(X = x)$ denote the mid distribution function (MDF) of the covariate, feature, or marker X. Of course, if X is a continuous random variable then the MDF, $\bar{F}$, equals the CDF, F. Similarly, let $\bar{G}$ denote the MDF of the outcome Y. When X has CDF F and Y has CDF G, we refer to the random variables $\bar{F}(X)$ and $\bar{G}(Y)$ as the grades of X and Y, respectively. We define the *asymmetric grade correlation* (AGC) of Y on X as

$$\mathrm{AGC}(X,Y) = \frac{\mathrm{cov}(\bar{F}(X), \bar{G}(Y))}{\mathrm{cov}(\bar{G}(Y), \bar{G}(Y))} \tag{5}$$

$$= \mathrm{cor}(\bar{F}(X), \bar{G}(Y)) \left(\frac{1 - \mathbb{P}(X = X' = X'')}{1 - \mathbb{P}(Y = Y' = Y'')} \right)^{1/2}, \tag{6}$$

where the equality of the representations in (5) and (6) follows from results of Niewiadomska-Bugaj and Kowalczyk (2005) and Pohle et al. (2025). The associated *coefficient of monotone association* (CMA) measure,

$$\mathrm{CMA}(X,Y) = \frac{1}{2}\,(\mathrm{AGC}(X,Y) + 1)\,, \tag{7}$$

is an affine function of AGC. Perhaps surprisingly, the range of possible values is $[-1, 1]$ for AGC and $[0, 1]$ for CMA, as we demonstrate below. To quantify discreteness, we introduce the granularity of a random variable X as the quantity

$$\gamma(X) = \mathbb{P}\left(X = X' = X''\right), \tag{8}$$

which is prominent in (6). The minimal granularity of 0 arises under the finest possible resolution, namely, for continuous distributions, and the maximal granularity of 1 is reached for degenerate distributions with all probability mass in one point. We note that $\mathrm{AGC}(X, Y) = \mathrm{AGC}(Y, X)$ and $\mathrm{CMA}(X, Y) = \mathrm{CMA}(Y, X)$ if, and only if, $\gamma(X) = \gamma(Y)$.

Example 2.1. Let X be uniform on $(0, 1]$, let the function m be strictly increasing, and let the outcome

$$Y_k = \sum_{i=1}^{k} \mathbb{1}\left\{X \in \left(\frac{i-1}{k}, \frac{i}{k}\right]\right\} m\left(\frac{i}{k}\right)$$

be discrete, where $k \geq 2$ is an integer. Then $\mathrm{cor}(\bar{F}(X), \bar{G}(Y)) = (1 - \frac{1}{k^2})^{1/2}$, $\gamma(X) = 0$, and $\gamma(Y_k) = \frac{1}{k^2}$, so

$$\mathrm{AGC}(X, Y_k) = 1, \qquad \mathrm{AGC}(Y_k, X) = 1 - \frac{1}{k^2}.$$

The value of $\mathrm{AGC}(X, Y_k) = 1$ reflects the fact that the continuous covariate X is a perfect monotone predictor of the discrete outcome Y_k. Likewise, the value of $\mathrm{AGC}(Y_k, X) = 1 - \frac{1}{k^2}$ quantifies the suboptimality of Y_k as a predictor of X. In the limit as $k \to \infty$ the granularities of X and Y_k become equal and the difference between $\mathrm{AGC}(X, Y_k)$ and $\mathrm{AGC}(Y_k, X)$ vanishes.

2.2 Representations and properties

The following representation of CMA and AGC in terms of the similarity function s at (4) establishes that the measures are normalized. Specifically, CMA has range $[0, 1]$ and AGC attains values in the interval $[-1, 1]$.

Theorem 2.2. *For nondegenerate X and Y, it holds that*

$$\begin{aligned} \mathrm{CMA}(X, Y) &= \frac{1}{2}\left(\mathrm{AGC}(X, Y) + 1\right) \\ &= \frac{\mathbb{E}\left[(\bar{G}(Y'') - \bar{G}(Y'))\,\mathbb{1}\{Y' < Y''\}\, s(X', X'')\right]}{\mathbb{E}\left[(\bar{G}(Y'') - \bar{G}(Y'))\,\mathbb{1}\{Y' < Y''\}\right]}. \end{aligned} \tag{9}$$

The representation (9) demonstrates that CMA and AGC are measures of the amount of probability mass between outcomes that order in the same way as the covariate. If there are ties in the covariate, the corresponding probability mass is discounted by a factor of one half.

The next two results show that CMA and AGC, respectively, bridge the asymmetric AUC measure in the dichotomous case and Spearman's rank correlation coefficient ρ_{S} in continuous settings.

Corollary 2.3. *If Y is dichotomous, then* $\mathrm{CMA}(X, Y) = \mathrm{AUC}(X, Y)$.

Corollary 2.4. *If X and Y are continuous, then*

$$\begin{aligned} \mathrm{CMA}(X, Y) &= 6\,\mathbb{E}\left[(G(Y'') - G(Y'))\,\mathbb{1}\{Y' < Y''\}\,\mathbb{1}\{X' < X''\}\right] \\ &= \frac{1}{2}\left(\rho_{\mathrm{S}} + 1\right) \\ &= 6\,\mathbb{E}\left[(F(X'') - F(X'))\,\mathbb{1}\{X' < X''\}\,\mathbb{1}\{Y' < Y''\}\right] = \mathrm{CMA}(Y, X). \end{aligned}$$

As noted, while generalizations of ROC analysis to continuous outcomes have been developed in the extant literature (Pencina and D'Agostino, 2004; Gneiting and Walz, 2022), persuasive extensions to

general population settings have been elusive. In what follows, we tackle this problem and let Y now be a continuous real-valued random variable with CDF G. For $\alpha \in (0,1)$, we define

$$\mathrm{AUC}^{(\alpha)}(X,Y) = \mathrm{AUC}(X, \mathbb{1}\{Y \geq G^{-1}(\alpha)\}) \tag{10}$$

as the AUC measure (3) when the outcome is dichotomized at the α-quantile of its marginal distribution.

Theorem 2.5. *If Y is continuous, it holds that*

$$\mathrm{CMA}(X,Y) = 6 \int_0^1 \alpha\,(1-\alpha)\,\mathrm{AUC}^{(\alpha)}(X,Y)\,\mathrm{d}\alpha. \tag{11}$$

The representation (11) was hinted at by Gneiting and Walz (2022, equation (21)) in the special case when both X and Y are continuous, to which we tend now.

Corollary 2.6. *If X and Y are continuous, then*

$$\begin{aligned} \mathrm{CMA}(X,Y) &= 6 \int_0^1 \alpha\,(1-\alpha)\,\mathrm{AUC}^{(\alpha)}(X,Y)\,\mathrm{d}\alpha && (12)\\ &= \frac{1}{2}\,(\rho_{\mathrm{S}} + 1) \\ &= 6 \int_0^1 \alpha\,(1-\alpha)\,\mathrm{AUC}^{(\alpha)}(Y,X)\,\mathrm{d}\alpha = \mathrm{CMA}(Y,X). && (13) \end{aligned}$$

Example 2.7. Let (X,Y) be uniform on the triangle T with vertices $(0,0)$, $(\frac{1}{2},1)$, and $(1,0)$ in the Euclidean plane. Evidently, X and Y are dependent. However, the dependence of Y on X is not monotone; e.g., if $x < \frac{1}{2}$ then $X \leq x$ implies $Y < 2x$, but $X > 1-x$ also implies $Y < 2x$. On the other hand, the conditional distribution of X given $Y = y \in [0,1]$ is uniform on the interval $[\frac{y}{2}, 1-\frac{y}{2}]$. Not surprisingly, Spearman's correlation coefficient ρ_{S} vanishes. In Appendix A we show that

$$\mathrm{AUC}^{(\alpha)}(X,Y) = \frac{1}{2}$$

for $\alpha \in (0,1)$, whereas

$$\mathrm{AUC}^{(\alpha)}(Y,X) = 1 - \frac{\frac{2}{3}\sqrt{2\alpha} + \frac{\alpha}{6}}{1-\alpha}$$

and $\mathrm{AUC}^{(1-\alpha)}(Y,X) = 1 - \mathrm{AUC}^{(\alpha)}(Y,X)$ for $\alpha \in (0, \frac{1}{2})$. The common value of $\mathrm{CMA}(X,Y)$ and $\mathrm{CMA}(Y,X)$ in (12) and (13), respectively, is $\frac{1}{2}$, in accordance with $\rho_{\mathrm{S}} = 0$.

There is an expansive literature on desirable properties for measure of complete dependence (Rényi, 1959; Pohle and Wermuth, 2025) and monotone dependence (Schweizer and Wolff, 1981; Scarsini, 1984). Here we summarize key properties of AGC that carry over to the CMA measure in obvious ways.

Theorem 2.8. *Let (X,Y) be nondegenerate.*

(a) Range: $-1 \leq \mathrm{AGC}(X,Y) \leq 1$

(b) Symmetry under equal granularity: $\mathrm{AGC}(X,Y) = \mathrm{AGC}(Y,X)$ *if, and only if,* $\gamma(X) = \gamma(Y)$.

(c) Independence: If X and Y are independent, then $\mathrm{AGC}(X,Y) = 0$.

(d) Change of sign: $\mathrm{AGC}(-X,Y) = \mathrm{AGC}(X,-Y) = -\mathrm{AGC}(X,Y)$

(e) Perfect monotone predictor property: $\mathrm{AGC}(X,Y) = 1$ *if, and only if, there exists a nondecreasing function m such that $Y = m(X)$ almost surely.*

(f) Invariance under strictly increasing transformations: If f and g are strictly increasing, then $\mathrm{AGC}(f(X), g(Y)) = \mathrm{AGC}(X,Y)$.

While (a), (c), (d), and (f) are well established concepts, we note the modulation of symmetry in terms of the granularity γ at (8) in property (b). The perfect monotone predictor property (e) states that $\mathrm{AGC}(X, Y) = 1$ if, and only if, X is a perfect monotone predictor of Y in the sense of Kimeldorf and Sampson (1978). In this light, AGC and CMA are particularly relevant when one seeks to assess the extent to which an explanatory variable X can contribute to the prediction of an outcome variable Y, such as in variable selection or in the evaluation of the potential predictive performance of point forecasts. The perfect monotone predictor property is violated by Spearman's ρ_{S} and Kendall's τ_{K}, and indeed is incompatible with the traditional symmetry requirement, for if X is a perfect monotone predictor of Y, then Y may or may not be a perfect monotone predictor of X. However, the perfect monotone predictor property is compatible with the modulated symmetry property. Jointly with the change of sign property (d), the perfect monotone predictor property implies that $\mathrm{AGC}(X, Y) = -1$ if, and only if, there exists a nonincreasing function m such that $Y = m(X)$ almost surely.

In passing, the quantity $\mathrm{AGC}(X, Y)$ arises as the slope in rank-rank regression models, which have been developed for studies of intergenerational mobility (Chetverikov and Wilhelm, 2023), provided that the ranks take the form of mid ranks.

2.3 Relationships to other measures

Let us first consider a generalized version of the C index (CID) of Y on X as introduced by Harrell Jr et al. (1996) and studied by Pencina and D'Agostino (2004). Specifically, we define

$$\mathrm{CID}(X, Y) = \mathbb{E}(s(X', X'') \mid Y' < Y'') \tag{14}$$

as the tie-adjusted probability of concordance between the covariate and the outcome. Pencina and D'Agostino (2004, eq. (5)) restrict attention to survival data and assume that the covariate is continuous.[2] The assumption of continuity of X is unnecessarily restrictive, and we allow for any real-valued covariate. The rank graduation accuracy (RGA) measure of Y on X put forth by Raffinetti (2023) and Giudici and Raffinetti (2025) is

$$\mathrm{RGA}(X, Y) = \frac{1}{2}\left(\frac{\mathrm{cov}(Y, \bar{F}(X))}{\mathrm{cov}(Y, \bar{G}(Y))} + 1\right);$$

In general, RGA is not a measure of monotone dependence, as its value may change under strictly monotone transformations of the outcome Y. However, for a dichotomous outcome, CMA, CID, and RGA coincide and equal AUC.

Proposition 2.9. *If Y is dichotomous, then*

$$\mathrm{CMA}(X, Y) = \mathrm{CID}(X, Y) = \mathrm{RGA}(X, Y) = \mathrm{AUC}(X, Y). \tag{15}$$

The symmetric grade correlation measure ρ_{G} studied by Pohle et al. (2025) is defined as

$$\rho_{\mathrm{G}} = \mathrm{cor}(\bar{F}(X), \bar{G}(Y)) = \mathrm{AGC}(X, Y)\left(\frac{1 - \gamma(Y)}{1 - \gamma(X)}\right)^{1/2},$$

and we see that $\mathrm{AGC}(X, Y)\,\mathrm{AGC}(Y, X) = \rho_{\mathrm{G}}^2$. For continuous X and Y, our AGC measure becomes symmetric and equals ρ_{G} and ρ_{S}, respectively. Furthermore, the generalized CID measure at (14) relates to Kendall's τ_{K} in the same way that CMA relates to ρ_{G} and ρ_{S}, as follows.

Proposition 2.10. *If X and Y are continuous, then*

$$\mathrm{AGC}(X, Y) = \mathrm{AGC}(Y, X) = \rho_{\mathrm{G}} = \rho_{\mathrm{S}}, \tag{16}$$

$$\mathrm{CMA}(X, Y) = \mathrm{CMA}(Y, X) = \frac{1}{2}(\rho_{\mathrm{G}} + 1) = \frac{1}{2}(\rho_{\mathrm{S}} + 1), \tag{17}$$

[2] Note that Pencina and D'Agostino (2004) use the symbols Y and X for the covariate and the outcome, respectively.

and

$$\mathrm{CID}(X, Y) = \mathrm{CID}(Y, X) = \frac{1}{2}(\tau_{\mathrm{K}} + 1), \tag{18}$$

where τ_{K} *is defined at* (2).

Summarizing these relationships, the CID measure in its general form bridges AUC and Kendall's τ_{K} in the same way that AGC and CMA bridge AUC and Spearman's ρ_{S}. Mirroring the development of AGC at (5), we define the *asymmetric Kendall correlation* (AKC) of Y on X as

$$\mathrm{AKC}(X, Y) = 2\,\mathrm{CID}(X, Y) - 1 = \frac{\tau_{\mathrm{K}}}{1 - \mathbb{P}(Y = Y')}, \tag{19}$$

so that $\mathrm{AKC}(X, Y) = \mathrm{AKC}(Y, X) = \tau_{\mathrm{K}}$ if the outcome Y is continuous. This quantity has been known and studied under the label of Somers' D (Somers, 1962; Newson, 2002). Natural analogues of the properties in Theorem 2.8 hold for AKC and CID, respectively. However, we are unaware of analogues of the mixture representation for $\mathrm{CMA}(X, Y)$ at (11) in terms of $\mathrm{AUC}^{(\alpha)}(X, Y)$.

An ubiquitous special case arises in Gaussian populations (Kruskal, 1958).

Example 2.11. For a bivariate normal random vector (X, Y) with Pearson correlation $r \in (0, 1)$,

$$\mathrm{AGC}(X, Y) = \mathrm{AGC}(Y, X) = \rho_{\mathrm{G}} = \rho_{\mathrm{S}} = \frac{6}{\pi} \arcsin \frac{r}{2} < \frac{2}{\pi} \arcsin r = \tau_{\mathrm{K}} = \mathrm{AKC}(Y, X) = \mathrm{AKC}(X, Y).$$

While other, attractive symmetric measures of dependence that are invariant under strictly increasing transformations of X and Y have been developed, such as the maximum information coefficient (Reshef et al., 2011), Bergsma–Dassios sign correlation (Bergsma and Dassios, 2014), symmetric rank correlations (Weihs et al., 2018), Hellinger correlation (Geenens and Lafaye de Micheaux, 2022), and generalized correlations (Fissler and Pohle, 2023), we restrict a more detailed discussion to some striking commonalities between AGC and CMA and the recently proposed, asymmetric correlation coefficient of Chatterjee (2021).

2.4 Commonalities with Chatterjee correlation

In the population setting, the Chatterjee correlation coefficient ξ of Y on X (Chatterjee, 2021) is defined as

$$\xi(X, Y) = \frac{\int \mathrm{var}(\mathbb{E}(\mathbb{1}\{Y \leq y\} \mid X))\,\mathrm{d}G(y)}{\int \mathrm{var}(\mathbb{1}\{Y \leq y\})\,\mathrm{d}G(y)}.$$

Chatterjee's coefficient attains values in the unit interval $[0, 1]$, and it holds that $\xi(X, Y) = 0$ if, and only if, X and Y are independent. Furthermore, Chatterjee's coefficient has the complete predictor property: $\xi(X, Y) = 1$ if, and only if, Y is completely dependent on X in the sense of Lancaster (1963), i.e., if, and only if, there is a measurable function g such that $Y = g(X)$ almost surely.

Like the notion of perfect monotone dependence (Kimeldorf and Sampson, 1978), complete dependence fails to be symmetric. Therefore, like AGC and CMA, Chatterjee's coefficient is asymmetric, with Chatterjee (2021, p. 2010) noting that

> "this is intentional. We would like to keep it that way because we may want to understand if Y is a function of X, and not just if one of the variables is a function of the other."

However, while AGC and CMA are measures of monotone dependence, Chatterjee's coefficient is a measure of complete dependence. A further difference is that AGC and CMA become symmetric when the covariate X and the outcome Y have the same granularity. In contrast, it is possible that $\xi(X, Y) \neq \xi(Y, X)$ even when both X and Y are continuous.

Example 2.12. We revisit Example 2.7 and let (X, Y) be uniform on the triangle with vertices $(0, 0)$, $(\frac{1}{2}, 1)$, and $(1, 0)$ in the Euclidean plane. In Appendix A we show that $\xi(X, Y) = \frac{1}{24}$, whereas $\xi(Y, X) = \frac{1}{6}$.

Finally, we observe that when Y is continuous then $\mathrm{AGC}(X,Y)$, $\mathrm{CMA}(X,Y)$ and $\xi(X,Y)$ admit representations as mixtures of quantities that depend on the law of $(X, \mathbb{1}\{Y \le G^{-1}(\alpha)\})$ only. In the case of AGC and CMA, the representation (11) holds, where $\mathrm{AUC}^{(\alpha)}(X,Y)$ depends on (X,Y) via the law of $(X, \mathbb{1}\{Y \le G^{-1}(\alpha)\})$ only and vanishes if X and Y are independent. In the case of Chatterjee's correlation, we find that

$$\xi(X,Y) = 6 \int_0^1 \mathrm{var}\left(\mathbb{E}(\mathbb{1}\{Y \le G^{-1}(\alpha)\} \mid X)\right) \mathrm{d}\alpha,$$

where the integrand also depends on the law of $(X, \mathbb{1}\{Y \le G^{-1}(\alpha)\})$ only and vanishes if the covariate X and the outcome Y are independent. The same statement applies to the closely related, recently proposed integrated R^2 measure of Azadkia and Roudaki (2025).

3 Sampling theory and a general test of DeLong type

Following the lead of Pohle et al. (2025), we define the empirical asymmetric grade correlation, AGC_n, and the empirical coefficient of monotone association, CMA_n, in a simple yet efficient way. Specifically, we plug the empirical law, $\mathbb{P}_n$, of data

$$(X_1, Y_1), \ldots, (X_n, Y_n) \tag{20}$$

into the population representations at (5) and (7), respectively. Building on the far reaching results of Pohle et al. (2025), we then find asymptotic distributions for AGC_n and CMA_n and use the Gaussian limit distributions to develop estimators and tests.

3.1 Empirical versions of AGC and CMA

We introduce notation for a range of equivalent representations of AGC_n and CMA_n. Let $\bar{Q}_1, \ldots, \bar{Q}_n$ be the mid ranks (Woodbury, 1940) associated with the covariates $X_1, \ldots, X_n$, and let $\bar{R}_1, \ldots, \bar{R}_n$ be the mid ranks associated with the outcomes $Y_1, \ldots, Y_n$.[3] Let $Z_1 < \cdots < Z_M$ denote the $M \in \{2, \ldots, n\}$ unique values of $Y_1, \ldots, Y_n$, and define

$$N_c = \sum_{i=1}^n \mathbb{1}\{Y_i = Z_c\}, \quad c = 1, \ldots, M, \tag{21}$$

as the number of instances among the outcomes $Y_1, \ldots, Y_n$ that equal Z_c, so that $N_1 + \cdots + N_M = n$. Then the data at (20) can be written as

$$(X_{11}, Z_1), \ldots, (X_{1N_1}, Z_1), \ \ldots, \ (X_{M1}, Z_M), \ldots, (X_{MN_M}, Z_M), \tag{22}$$

where $X_{c1}, \ldots, X_{cN_c}$ are the covariate values that correspond to the unique outcome value Z_c. Let $\widetilde{R}_1, \ldots, \widetilde{R}_M$ be the unique sorted mid ranks associated with $Z_1, \ldots, Z_M$, namely, $\widetilde{R}_1 = \frac{1}{2}(N_1 + 1)$ and $\widetilde{R}_c = \sum_{j=1}^{c-1} N_j + \frac{N_c+1}{2}$ for $c = 2, \ldots, M$. Then the following representations hold.

Proposition 3.1. *The empirical asymmetric grade correlation equals*

$$\mathrm{AGC}_n = \frac{\sum_{i=1}^n (\bar{Q}_i - \frac{1}{2}(n+1))(\bar{R}_i - \frac{1}{2}(n+1))}{\sum_{i=1}^n (\bar{R}_i - \frac{1}{2}(n+1))^2}; \tag{23}$$

and the empirical coefficient of monotone association equals

$$\mathrm{CMA}_n = \frac{1}{2}\left(\mathrm{AGC}_n + 1\right) \tag{24}$$

$$= \frac{\sum_{k=1}^n \sum_{l=1}^n (\bar{R}_k - \bar{R}_l)\, \mathbb{1}\{Y_k < Y_l\}\, s(X_k, X_l)}{\sum_{k=1}^n \sum_{l=1}^n (\bar{R}_k - \bar{R}_l)\, \mathbb{1}\{Y_k < Y_l\}} \tag{25}$$

$$= \frac{\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} (\widetilde{R}_j - \widetilde{R}_i) \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} s(X_{ik}, X_{jl})}{\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} (\widetilde{R}_j - \widetilde{R}_i) N_i N_j}. \tag{26}$$

[3] Specifically, the mid rank $\bar{R}_j$ associated with Y_j is $\bar{R}_j = \sum_{i=1}^n \mathbb{1}\{Y_i < Y_j\} + \frac{1}{2}\sum_{i=1}^n \mathbb{1}\{Y_i = Y_j\} + \frac{1}{2}$.

By construction, the representations at (23) and (24) are plug-in empirical versions of the population quantities at (5) and (7), respectively. They demonstrate that AGC_n and CMA_n can be computed in $\mathcal{O}(n \log n)$ operations. The equivalent expression at (25) is the plug-in empirical version of the population level representation at (9).

The representation (26) is a variant of (25) that facilitates comparison with the empirical coefficient of predictive ability (CPA_n) proposed by Gneiting and Walz (2022). For data of the form (22),

$$\mathrm{CPA}_n = \frac{\sum_{i=1}^{M-1} \sum_{j=k+1}^{M} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} (j-i)\, s(X_{ik}, X_{jl})}{\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} (j-i)\, N_i N_j}. \tag{27}$$

However, Gneiting and Walz (2022) do not provide a population version of CPA_n, nor do we believe that such can be developed, except for settings where CPA_n equals CMA_n.

In general, CMA_n and CPA_n are distinct quantities, but they equal each other in important special cases. Specifically, when the outcome is dichotomous, then $M = 2$ and

$$\mathrm{CMA}_n = \mathrm{CPA}_n = \mathrm{AUC}_n = \frac{\sum_{k=1}^{N_1} \sum_{l=1}^{N_2} s(X_{1k}, X_{2l})}{N_1 N_2}, \tag{28}$$

where AUC_n is the empirical version of the AUC criterion at (3). Furthermore, when the outcomes $Y_1, \ldots, Y_n$ at (20) are pairwise distinct, then $M = n$ and $N_c = 1$ for $c = 1, \ldots, n$ at (21), and the following holds.

Corollary 3.2. *Suppose that the outcomes $Y_1, \ldots, Y_n$ for the data at (20) are pairwise distinct with order statistics $Z_1 < \cdots < Z_n$. For $i = 1, \ldots, n-1$, let*

$$\mathrm{AUC}_n^{((i+1)/n)} = \frac{\sum_{k=1}^{n} \sum_{l=1}^{n} s(X_k, X_l)\, \mathbb{1}\{Y_k < Z_{i+1} \leq Y_l\}}{i(n-i)}$$

denote AUC_n for the induced data $(X_1, \mathbb{1}\{Y_1 \geq Z_{i+1}\}), \ldots, (X_n, \mathbb{1}\{Y_n \geq Z_{i+1}\})$, where the outcome is dichotomized at the threshold Z_{i+1}. Then

$$\mathrm{CMA}_n = \mathrm{CPA}_n = \frac{6}{n} \sum_{i=1}^{n-1} \frac{i(n-i)}{(n-1)(n+1)}\, \mathrm{AUC}_n^{((i+1)/n)}. \tag{29}$$

We interpret the representation (29) for CMA_n as an empirical version of the relationship (11) for CMA at the population level. Essentially the same representation for CMA_n holds in slightly more general settings, namely, when all class sizes $N_1 = \cdots = N_M$ at (21) are equal.

3.2 Asymptotic distributions

We turn to the asymptotic distribution of AGC_n and CMA_n, with the development drawing heavily on theory recently developed by Pohle et al. (2025). We begin by showing that the univariate limit distribution of AGC_n and CMA_n for independent, identically distributed sequences is Gaussian. For clarity, we formalize the setting.

Scenario 3.3. Consider a sequence $(X_i, Y_i)_{i=1,2,\ldots}$ of independent random vectors, where each (X_i, Y_i) has the same law $\mathbb{P}$ as (X, Y), with X and Y having strictly positive variance.

We recall that $F(x) = \mathbb{P}(X \leq x)$ and $\bar{F}(x) = \mathbb{P}(X < x) + \frac{1}{2}\mathbb{P}(X = x)$ denote the cumulative distribution function (CDF) and the mid distribution function (MDF) of X, respectively. Let G and $\bar{G}$ be the CDF and MDF of Y, and let

$$\bar{H}(x, y) = \mathbb{P}(X < x, Y < y) + \frac{1}{2}\mathbb{P}(X = x, Y < y) + \frac{1}{2}\mathbb{P}(X < x, Y = y) + \frac{1}{4}\mathbb{P}(X = x, Y = y)$$

denote the bivariate MDF of (X, Y), where $x, y \in \mathbb{R}$. We let $\rho = 12\,\mathrm{cov}(\bar{F}(X), \bar{G}(Y))$, define $\widetilde{F}(x) = \mathbb{E}[\bar{H}(x, Y)]$ for $x \in \mathbb{R}$ and $\widetilde{G}(y) = \mathbb{E}[\bar{H}(X, y)]$ for $y \in \mathbb{R}$, and let

$$K^{(1)}(x, y) = 4\left(\widetilde{F}(x) + \widetilde{G}(y) + \bar{F}(x)\bar{G}(y) - \bar{F}(x) - \bar{G}(y)\right) + 1 - \rho \tag{30}$$

for $x, y \in \mathbb{R}$. With a view to the definition of granularity at (8), we write $\gamma_F = \gamma(X)$ and $\gamma_G = \gamma(Y)$, and we let

$$K^{(0)}(y) = (G(y) - G(y-))^2 - \gamma_G \tag{31}$$

for $y \in \mathbb{R}$. We are now ready to state the limit distribution of AGC_n, where we express the asymptotic variance in terms of the kernels at (30) and (31), respectively.

Theorem 3.4. *Under Scenario 3.3 it holds that*

$$\sqrt{n}\,(\mathrm{AGC}_n - \mathrm{AGC}) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right), \tag{32}$$

where

$$\sigma^2 = \frac{9}{(1-\gamma_G)^2}\,\mathbb{E}\left(K^{(1)}(X, Y) + \frac{\rho}{1-\gamma_G}K^{(0)}(Y)\right)^2. \tag{33}$$

In particular, if X and Y are independent then

$$\sqrt{n}\,\mathrm{AGC}_n \xrightarrow{d} \mathcal{N}\left(0, \frac{1-\gamma_F}{1-\gamma_G}\right). \tag{34}$$

Evidently, for $\mathrm{CMA}_n = \frac{1}{2}(\mathrm{AGC}_n + 1)$ the asymptotic variances at (33) and (34) need to be scaled by a factor of $\frac{1}{4}$. For a dichotomous outcome Y with success probability $\pi \in (0, 1)$, CMA equals AUC and our results supplement classical findings by Bamber (1975) and DeLong et al. (1988). In particular, in the case of independence between X and Y,

$$\sqrt{n}\left(\mathrm{AUC}_n - \frac{1}{2}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1-\gamma_F}{12\,\pi(1-\pi)}\right),$$

where AUC_n is defined at (28). If X and Y are continuous, AGC equals Spearman's rank correlation coefficient and grade correlation, and the asymptotic variance for AGC_n in (33) specializes to

$$\sigma^2 = 9\,\mathbb{E}[K^{(1)}(X, Y)^2],$$

reducing further to $\sigma^2 = 1$ in the case of independence.[4] Accordingly, Theorem 3.4 nests extant results for the sample versions of Spearman's rank correlation and grade correlation, as reported by Hoeffding (1948), Borkowf (2002), Genest et al. (2013), and Pohle et al. (2025), respectively.

Let us now turn to the case of m covariates, features, or markers $X^{(1)}, \ldots, X^{(m)}$, which we interpret as competing or complementary predictors for the real-valued outcome Y.

Scenario 3.5. Consider a sequence

$$\left(X_i^{(1)}, \ldots, X_i^{(m)}, Y_i\right)_{i=1,2,\ldots} \tag{35}$$

of independent random vectors, where each $(X_i^{(1)}, \ldots, X_i^{(m)}, Y_i)$ has the same distribution $\mathbb{P}$ as $(X^{(1)}, \ldots, X^{(m)}, Y)$, with $X^{(1)}, \ldots, X^{(m)}$, and Y all having strictly positive variance.

[4] For CMA, we obtain an asymptotic variance of $\frac{1}{4}$, compared to an asymptotic variance of $\frac{2}{5}$ for Chatterjee's correlation coefficient (Chatterjee, 2021), which has the same population range as CMA. We interpret the lesser variance for CMA as a reflection of its consideration of monotone (only) dependence.

As before, we let G, $\bar{G}$, and γ_G denote the CDF, the MDF, and the granularity of Y. For $k = 1, \ldots, m$, we let $\mathrm{AGC}^{(k)} = \mathrm{AGC}(X^{(k)}, Y)$ and write $\mathrm{AGC}_n^{(k)}$ for the corresponding empirical version. Furthermore, we write $\bar{F}^{(k)}$ for the univariate MDF of $X^{(k)}$ and $\bar{H}^{(k)}$ for the bivariate MDF of $(X^{(k)}, Y)$, and we define the nondecreasing functions $\widetilde{F}^{(k)}(x) = \mathbb{E}[\bar{H}^{(k)}(x, Y)]$ for $x \in \mathbb{R}$ and $\widetilde{G}^{(k)}(y) = \mathbb{E}[\bar{H}^{(k)}(X^{(k)}, y)]$ for $y \in \mathbb{R}$. Finally, we let

$$\rho^{(k)} = 12 \operatorname{cov}(\bar{F}^{(k)}(X^{(k)}), \bar{G}(Y)). \tag{36}$$

In the subsequent theorem we express the asymptotic covariance matrix in terms of expectations of the kernel functions

$$K^{(k)}(x, y) = 4 \left(\widetilde{F}^{(k)}(x) + \widetilde{G}^{(k)}(y) + \bar{F}^{(k)}(x)\bar{G}(y) - \bar{F}^{(k)}(x) - \bar{G}(y) \right) + 1 - \rho^{(k)} \tag{37}$$

for $k = 1, \ldots, m$, and $K^{(0)}(y)$ at (31), respectively.

Theorem 3.6. *Under Scenario 3.5 it holds that*

$$\sqrt{n} \left(\begin{pmatrix} \mathrm{AGC}_n^{(1)} \\ \vdots \\ \mathrm{AGC}_n^{(m)} \end{pmatrix} - \begin{pmatrix} \mathrm{AGC}^{(1)} \\ \vdots \\ \mathrm{AGC}^{(m)} \end{pmatrix} \right) \xrightarrow{d} \mathcal{N} \left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \Sigma = \left(\Sigma^{(k,l)} \right)_{k,l=1}^{m} \right),$$

where

$$\Sigma^{(k,l)} = \frac{9}{(1-\gamma_G)^2} \mathbb{E} \left[\left(K^{(k)}(X^{(k)}, Y) + \frac{\rho^{(k)}}{1-\gamma_G} K^{(0)}(Y) \right) \left(K^{(l)}(X^{(l)}, Y) + \frac{\rho^{(l)}}{1-\gamma_G} K^{(0)}(Y) \right) \right] \tag{38}$$

for $k, l = 1, \ldots, m$.

Evidently, the limit distribution translates to the CMA measure, for which $\Sigma^{(k,l)}$ at (38) needs to be multiplied by a factor of $\frac{1}{4}$, where $k, l = 1, \ldots, m$. As in the univariate special case of Theorem 3.4, our results nest extant findings when the outcome Y is dichotomous (DeLong et al., 1988) and when X and Y are continuous (Hoeffding, 1948; Gaißer and Schmid, 2010; Pohle et al., 2025).

If we wish to use these asymptotic results to generate confidence intervals or test hypotheses, we need to replace the population quantities at (38), (37), (36), and (31), respectively, by suitable estimates. For doing this, we follow recommendations in Pohle et al. (2025) and proceed stepwise, by evaluating the population quantities under the empirical law of the first n terms in (35). We use subscripts to distinguish sample quantities from population quantities, and to avoid double subscripts we write γ_n for the sample version of the granularity γ_G. We then find multi-step plug-in estimates as follows.

1. Compute the quantities γ_n and $G_n(Y_i)$, $\bar{G}_n(Y_i)$, $\bar{F}_n^{(k)}(X_i^{(k)})$, and $\bar{H}_n^{(k)}(X_i^{(k)}, Y_i)$, where $i = 1, \ldots, n$ and $k = 1, \ldots, m$.

2. Use the first step estimates to find $\rho_n^{(k)}$, $\widetilde{F}_n^{(k)}(X_i^{(k)})$, and $\widetilde{G}_n^{(k)}(Y_i)$, where $i = 1, \ldots, n$ and $k = 1, \ldots, m$.

3. Employ the second step estimates to compute the kernel values $K_n^{(0)}(Y_i)$ and $K_n^{(k)}(X_i^{(k)}, Y_i)$, where $i = 1, \ldots, n$ and $k = 1, \ldots, m$.

4. Plug the first and third step estimates into the sample version of (38).

We denote the resulting multi-step plug-in estimate by $\hat{\Sigma}_n$ and refer to its entries as $\hat{\Sigma}_n^{(k,l)}$, where $k, l = 1, \ldots, m$. When $m = 1$ we use more parsimonious notation and denote the estimate of the asymptotic variance by $\hat{\sigma}_n^2 = \hat{\Sigma}_n^{(1,1)}$. Under Scenario 3.5 it holds that $\hat{\Sigma}_n \longrightarrow \Sigma$ in probability as $n \longrightarrow \infty$ (Pohle et al., 2025).

Let us now discuss computational facets. Regarding the first step, we note from the second equality in Proposition 2.1(i) of Niewiadomska-Bugaj and Kowalczyk (2005) that

$$\gamma_n = 1 - \frac{12}{n^2} \widehat{\mathrm{var}}(\bar{R}_1, \ldots, \bar{R}_n) \tag{39}$$

is a linear function of the variance of the mid ranks of the outcomes and, therefore, can be computed from $\bar{R}_1, \ldots, \bar{R}_n$ in $\mathcal{O}(n)$ further operations. Consequently, the computational bottlenecks lie in the computation of $\bar{H}_n^{(k)}$ in the first step, and in the fourth step, which require up to $\mathcal{O}(mn^2)$ and $\mathcal{O}(m^2 n)$ operations, respectively.

For a dichotomous outcome, the CMA measure reduces to AUC, and an abundance of ties in the outcome can be leveraged. In particular, Sun and Xu (2014) devised a fast algorithm that computes the covariance matrix estimator of DeLong et al. (1988) for the AUC measure with a complexity of $\mathcal{O}(mn \log n) + \mathcal{O}(m^2 n)$ operations, as implemented in the R package `pROC` (Robin et al., 2011). The estimator of DeLong et al. (1988) differs from our multi-step plug-in estimator, but we leave an in-depth comparison to future work and note the connection to Demler et al. (2017), who relate U-statistics based variance estimates for differences in AUC to the estimate of DeLong et al. (1988).

In the continuous case, AGC reduces to Spearman's rank correlation coefficient, and Gaißer and Schmid (2010) propose a nonparametric bootstrap technique for variance estimation. Generally, bootstrap based variance estimators offer computationally efficient alternatives to our multi-step plug-in estimator $\hat{\Sigma}_n$, as they require $\mathcal{O}(bmn \log n)$ operations only, where b is the number of bootstrap replicates.

Finally, we note from Pohle et al. (2025) that a more general version of Proposition B.1 in Appendix B holds in dependent time series settings. Therefore, our Theorems 3.4 and 3.6 can be generalized as well. In practice, it suffices to replace our multi-step plug-in estimates of the population quantities at (38) by heteroskedasticity and autocorrelation consistent estimates (Newey and West, 1987), for which Pohle et al. (2025) provide implementation details and prove consistency. Computationally less demanding alternatives may lie in block bootstrap techniques (Künsch, 1989).

Thus far we have operated under Scenario 3.5 with interest in inference for the m quantities $\mathrm{AGC}^{(1)} = \mathrm{AGC}(X^{(1)}, Y), \ldots, \mathrm{AGC}^{(m)} = \mathrm{AGC}(X^{(m)}, Y)$ with shared second argument Y. More generally, we might consider a sequence

$$\left(X_i^{(1)}, \ldots, X_i^{(m)}\right)_{i=1,2,\ldots} \tag{40}$$

of independent random vectors, where each $(X_i^{(1)}, \ldots, X_i^{(m)})$ has the same distribution $\mathbb{P}$ as $(X^{(1)}, \ldots, X^{(m)})$, but now with interest in inference for the m^2 quantities

$$\mathrm{AGC}^{(i,j)} = \mathrm{AGC}(X^{(i)}, X^{(j)}), \qquad i, j = 1, \ldots, m.$$

The resulting multivariate normal limit distribution is degenerate[5] and has a singular covariance matrix Σ of dimension $m^2 \times m^2$ with components $\Sigma_{(i,j),(k,l)}$ for $i, j, k, l = 1, \ldots, m$. The details are tedious and we leave them to future work.

3.3 A general test of DeLong type

With variance estimates at hand, we can use the limit distributions from Theorems 3.4 and 3.6 to generate confidence intervals and test hypotheses about AGC and CMA. As CMA and AGC nest the AUC measure and Spearman's rank correlation coefficient, inference tools for these classical measures arise as special cases.

To generate confidence intervals, the Gaussian limit at (32) can be employed as usual. For testing, we start with a discussion of the most basic case, namely, tests of independence. Under independence of X and Y, it holds that $\mathrm{AGC}(X, Y) = 0$, and we use the test statistic

$$T_n = \sqrt{n}\, \frac{\mathrm{AGC}_n}{\hat{\sigma}_n}, \tag{41}$$

where AGC_n and $\hat{\sigma}_n$ are the estimates from Sections 3.1 and 3.2, respectively. By Theorem 3.4 and the consistency of $\hat{\sigma}_n$, T_n is asymptotically standard normal under independence. With Φ denoting the

[5]The degeneracy stems not only from the inclusion of $\mathrm{AGC}^{(i,i)}$, but may also arise in the submatrices for indices satisfying $k = j$ and $l = i$.

CDF of a standard normal distribution, we find a one-sided p-value, $p = \Phi(T_n)$, or a two-sided p-value, $p = 2\,(1 - \Phi(|T_n|)$, in the usual way. Similarly, we test simple hypotheses of the form AGC $= a_0 \in (-1, 1)$ or CMA $= b_0 \in (0, 1)$.

As noted, confidence intervals and tests of simple hypotheses for Spearman's rank correlation coefficient and AUC arise as special cases, for which there is considerable prior work. While comprehensive reviews of these two strands of literature are beyond the scope of our paper, we point at Edgeworth approximations (Best and Roberts, 1975) in the continuous case, as implemented in the `cor.test` function in the `stats` package for R (R Core Team, 2025), and at the aforementioned work of DeLong et al. (1988) and Sun and Xu (2014) for dichotomous outcomes, as implemented in the `pROC` package for R (Robin et al., 2011).

In what follows, we focus on tests of the equality of two or more AGC or CMA values for a shared outcome, as developed by DeLong et al. (1988) in the dichotomous case. Tests of this type address the ubiquitous task of the comparison of the potential predictive ability of competing covariates, scores, features, or markers, as frequently encountered in machine learning research (Rainio et al., 2024). Specifically, we consider the case $m = 2$ in Scenario 3.5, where we interpret $X^{(1)}$ and $X^{(2)}$ as predictors of the outcome Y and test the hypothesis

$$H_0 : \text{AGC}(X^{(1)}, Y) = \text{AGC}(X^{(2)}, Y). \tag{42}$$

For this we use the test statistic

$$\Delta_n = \sqrt{n}\,\frac{\text{AGC}_n^{(1)} - \text{AGC}_n^{(2)}}{\left(\hat{\Sigma}_n^{(1,1)} + \hat{\Sigma}_n^{(2,2)} - 2\hat{\Sigma}_n^{(1,2)}\right)^{1/2}}, \tag{43}$$

where $\hat{\Sigma}_n$ is the multi-step plug-in estimate from Section 3.2. By Theorem 3.6 and the consistency of $\hat{\Sigma}_n$, Δ_n is asymptotically standard normal under (42). Therefore, one-sided and two-sided tests yield p-values $1 - \Phi(\Delta_n)$ and $2\,(1 - \Phi(|\Delta_n|))$, respectively. In the same vein as before, tests of the equality of AUC and Spearman's rank correlation coefficient arise as special cases. In this light, we think of our test as a general test of DeLong et al. (1988) type that applies to all real-valued outcomes, covering the full range from dichotomous to continuous variables.

In the case of $m \geq 2$ predictors $X^{(1)}, \ldots, X^{(m)}$ for a shared outcome Y we adopt notation and terminology from Theorem 3.6 and test the hypothesis of equal AGC as follows. Let L be a fixed matrix with m columns, with each row representing a contrast. For example, when $m = 4$ the matrix

$$L = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

represents all six possible contrasts. Under the hypothesis of equal AGC, Theorem 2.5 in concert with the consistency of $\hat{\Sigma}_n$ implies that the test statistic

$$\chi_n = \text{AGC}_n\, L' \left(L \hat{\Sigma}_n L'\right)^{-1} L\, \text{AGC}_n'$$

is asymptotically chi-square distributed with degrees of freedom equal to the rank ℓ of the matrix $L\Sigma L'$. Accordingly, a p-value can be found as $p = 1 - F_\ell(\chi_n)$, where F_l is the CDF of the chi-square distribution with ℓ degrees of freedom. This test can also be considered a generalization of the DeLong et al. (1988) methodology that applies in the nested case when the outcome is dichotomous. Finally, when there are m distinct predictors applied to M distinct datasets, the methods described by Demšar (2006) can be used to check for statistically significant differences in AGC.

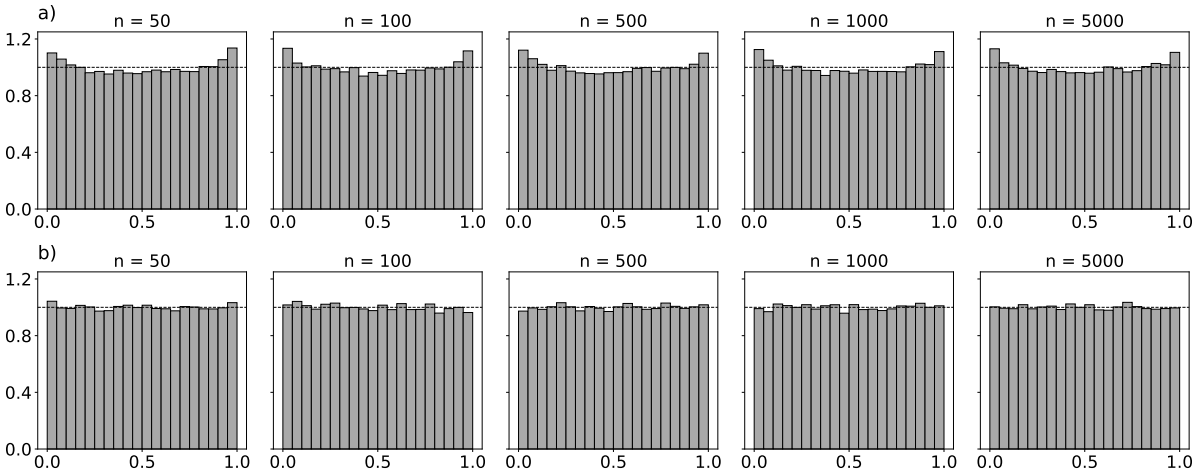

Figure 1: Histograms of one-sided p-values in tests of equal AGC in the first simulation example, where AGC equals Spearman's ρ_S, using a) the method of Myers and Sirois (2014) with the Meng et al. (1992) adjustment for correlation, and b) our general test of DeLong type with the test statistic at (43). The sample size n equals 50, 100, 500, 1000, and 5000 (left to right), and the number of Monte Carlo replicates is $100,000$.

3.4 Simulation examples

In this section, we study our general test of DeLong et al. (1988) type with test statistic at (43) in a number of simulation settings, where we compare against previously used tests. We operate under the null hypothesis and recall that a valid test generates p-values that are uniformly distributed between 0 and 1 (Lehmann and Romano, 2022, Lemma 3.3.1). Evidently, the uniformity of the p-value guarantees that the test attains its nominal size.

In our first simulation example, we address the continuous case, where AGC is symmetric and equals Spearman's rank correlation coefficient, ρ_S. Specifically, we let $X^{(0)}, Z^{(1)}$, and $Z^{(2)}$ be independent and standard normal. Conditionally on $X^{(0)}$, we let the outcome Y be normally distributed with mean $X^{(0)}$ and variance 1, and we seek to compare the correlated predictors $X^{(1)} = X^{(0)} + Z^{(1)}$ and $X^{(2)} = X^{(0)} + Z^{(2)}$. By construction, $(X^{(1)}, Y)$ and $(X^{(2)}, Y)$ have the same distribution, so the null hypothesis

$$H_0 : \mathrm{AGC}(X^{(1)}, Y) = \mathrm{AGC}(X^{(2)}, Y)$$

of equal AGC is satisfied. Perhaps surprisingly, tests targeted specifically at the comparison of two Spearman rank correlation coefficients do not appear to be available. Instead, the extant literature recommends that Spearman's ρ_S be treated as if it was a Pearson correlation, and then standard methods for the comparison of Pearson correlation coefficients be used (Myers and Sirois, 2014). Therefore, we compare our test to the test for the equality of correlated Pearson correlation coefficients developed by Meng et al. (1992),[6] as implemented in the `cocor` package (Diedenhofen and Musch, 2015) for R.

Figure 1 shows histograms of one-sided p-values for tests at sample size $n = 50$, 100, 500, 1000, and 5000, respectively. Our test statistic at (43) yields essentially uniform histograms even when $n = 50$. In contrast, the method of Myers and Sirois (2014) with the Meng et al. (1992) correction yields U-shaped histograms at all sample sizes, so at the levels typically used in practice, the null hypothesis gets rejected more often than warranted.

In a second example, we retain the above setting, but discretize $X^{(1)}$, $X^{(2)}$, and Y, by rounding to the nearest integer. The AGC measure now differs from Spearman's ρ_S, and we can no longer compare to previously used tests. Figure 2 shows histograms of two-sided p-values for our test at the aforementioned sample sizes. Again, the test yields essentially uniform histograms.

[6] Specifically, we apply Fisher's transformation (Fisher, 1915) to $\mathrm{AGC}_n^{(1)}$ and $\mathrm{AGC}_n^{(2)}$, to obtain $Z^{(1)}$ and $Z^{(2)}$, respec-

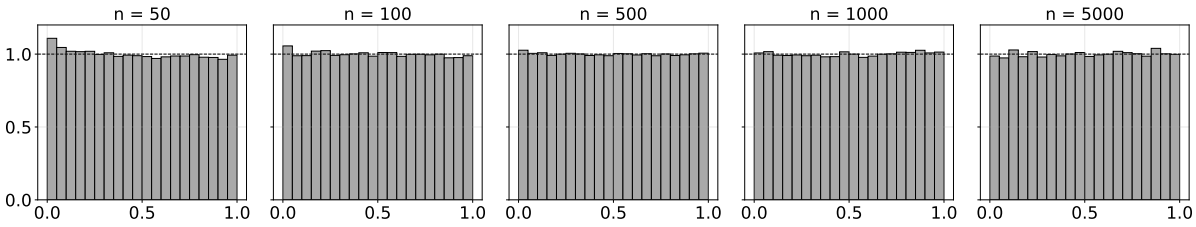

Figure 2: As in Figure 1b), but showing two-sided p-values in the second simulation example.

When the outcome Y is binary, the null hypothesis at (42) can be written equivalently as H_0 : $\text{AUC}(X^{(1)}, Y) = \text{AUC}(X^{(2)}, Y)$ in terms of the AUC measure at (3). As noted, our test and the classical test of DeLong et al. (1988) for the equality of two AUC values differ in the choice of the variance estimate only, and it is known that the DeLong et al. (1988) variance estimate is biased (Xu, 2024). In small sample simulation settings, both methods yield slightly non-uniform p-values under the null hypothesis, but the deviations generally are minor and difficult to interpret, and they disappear at larger sample sizes. We defer more detailed comparisons to future work and hint at the biostatistical literature on confidence intervals for differences in AUC values, where Demler et al. (2017) discuss the effects of nested models and adjustments for estimated parameters and relate to the theory of U-statistics.

4 Case studies

We now showcase the use of AGC and CMA in two case studies on topical scientific issues. First, we consider a recent study of Huang et al. (2024) and assess methods for uncertainty quantification in the output of large language models (LLMs). Then we follow up on the ongoing scientific revolution in weather prediction (Bi et al., 2023; Lam et al., 2023; Ben Bouallègue et al., 2024), where data-driven methods are replacing previously used physics-based approaches. Using the WeatherBench 2 benchmark dataset (Rasp et al., 2024), we shed new light on the predictive abilities of data-driven versus physics-based precipitation forecasts.

4.1 Evaluating uncertainty metrics for large language models

Recent breakthroughs in large language models (LLMs) have prompted the development of methods for uncertainty quantification for their outputs, aiming to mitigate hallucinations and improve interpretability and usability (Shorinwa et al., 2025). Uncertainty quantification for the output of LLMs bears challenges that hinder the direct application of evaluation methods developed for classification and regression tasks. To address the specific requirements in this setting, Huang et al. (2024) introduce a rank calibration framework and propose a new measure, the rank calibration error (RCE), to evaluate uncertainty or confidence measures for LLMs.

In their empirical study, Huang et al. (2024) consider five uncertainty metrics that map a query and the associated response generated by the LLM to a quantitative judgement of uncertainty or confidence — namely, the U_{Ecc}, U_{Deg}, U_{EigV}, U_{NLL}, and U_{SE} metrics, respectively — and use RCE to quantify the alignment of the measure with one of four considered a posteriori correctness scores — namely, BERT, METEOR, ROUGE-L, and ROUGE-1, respectively.[7] We adopt the experimental setup and data of Huang et al. (2024) and evaluate the five measures relative to each of the four scores, based on responses generated by the Llama-2-7b and Llama-2-7b-chat (Touvron et al., 2023) and GPT-3.5-turbo (Ouyang et al., 2021) models on NQ-open (Lee et al., 2019, $n = 3,600$), SQuAD (Rajpurkar et al., 2016, $n = 10,570$), and TriviaQA (Joshi et al., 2017, $n = 11,322$) queries, respectively. The key tenet in Huang et al. (2024) is that higher confidence (respectively, lower uncertainty) ought to imply higher

tively, and then use eq. (1) of Meng et al. (1992) to adjust for correlation.

[7]For detailed information on the uncertainty metrics and correctness scores, we refer to Huang et al. (2024) and references therein.

Table 1: Results in terms of CMA for experimental configurations in the format of Table 2 in Huang et al. (2024), with temperature values set at .6 for Llama-2-7b and Llama-2-7b-chat and 1.0 for GPT-3.5, respectively. For every value larger than .500 in the table, our one-sided test rejects the hypothesis of a CMA value less than or equal to $\frac{1}{2}$. The highest CMA value in each row is shown in green color. The subscripts to these values indicate the result (CMA_+: rejection; CMA_0: no rejection) of a one-sided test of the hypothesis of equality with the second-highest value in the row. All tests are at level .01.

Model	Queries	Correctness	U_{Ecc}	U_{Deg}	U_{EigV}	U_{NLL}	U_{SE}
Llama-2-7b	NQ-open	BERT	.526	$.653_+$	.650	.545	.565
		METEOR	.522	$.633_+$	.626	.521	.565
		ROUGE-L	.530	$.656_+$	.649	.528	.578
		ROUGE-1	.530	$.656_+$	.649	.528	.578
	SQuAD	BERT	.510	.624	$.626_+$	.605	.530
		METEOR	$.504_+$	.458	.458	.411	.443
		ROUGE-L	$.497_+$	.453	.452	.426	.455
		ROUGE-1	$.497_+$	.452	.452	.426	.455
	TriviaQA	BERT	.531	$.718_+$	.715	.558	.646
		METEOR	.515	.638	.631	.529	$.653_+$
		ROUGE-L	.528	$.700_+$	.694	.547	.672
		ROUGE-1	.528	$.700_+$	.694	.546	.672
Llama-2-7b-chat	NQ-open	BERT	.610	.684	$.686_+$	.576	.673
		METEOR	.606	.656	.658	.549	$.672_+$
		ROUGE-L	.620	.677	.680	.554	$.680_0$
		ROUGE-1	.620	.677	.680	.554	$.680_0$
	SQuAD	BERT	.547	.578	.584	$.614_+$	.586
		METEOR	.493	.433	.437	$.553_+$	.520
		ROUGE-L	.475	.423	.429	$.559_+$	.513
		ROUGE-1	.475	.423	.429	$.558_+$	.513
	TriviaQA	BERT	.719	.769	.767	.767	$.775_+$
		METEOR	.670	.706	.704	.671	$.712_+$
		ROUGE-L	.721	.765	.763	.750	$.777_+$
		ROUGE-1	.721	.765	.763	.750	$.777_+$
GPT-3.5	NQ-open	BERT	.689	.765	.765	$.775_+$	.761
		METEOR	.660	.716	$.717_0$	.683	.707
		ROUGE-L	.689	$.755_0$	.755	.739	.752
		ROUGE-1	.689	$.755_0$	.755	.739	.752
	SQuAD	BERT	.521	.531	.533	$.576_+$	.562
		METEOR	.544	.567	.567	$.618_+$	.612
		Rouge-L	.549	.578	.578	$.650_+$	.640
		Rouge-1	.544	.571	.571	$.640_+$	.629
	TriviaQA	BERT	.682	.705	.704	$.715_0$	.713
		METEOR	.662	$.681_+$	.681	.620	.645
		Rouge-L	.749	.780	.778	.783	$.791_+$
		Rouge-1	.750	.781	.779	.782	$.790_+$

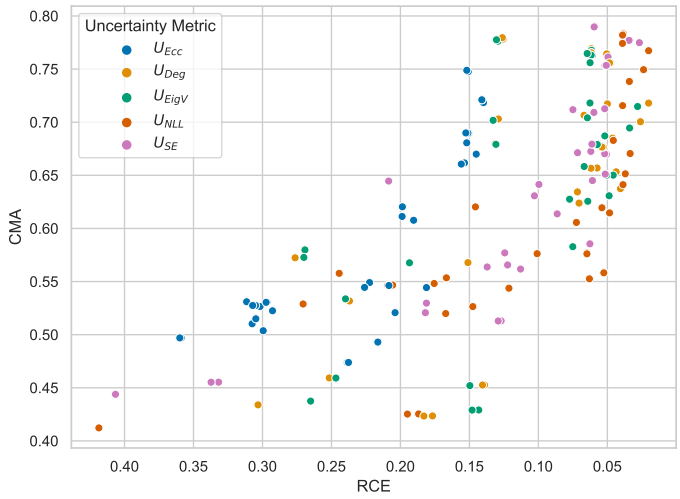

Figure 3: Scatterplot of CMA and RCE in the 180 experimental configurations from Table 1. Following Huang et al. (2024) we show values averaged over 20 bootstrap samples, which (for CMA) deviate slightly from the values in Table 1. Note the reversed orientation of RCE for improved interpretability.

generation quality. Given an uncertainty metric and a correctness score, the negatively oriented RCE measure (the lower, the better) quantifies deviations from this key assumption.

Huang et al. (2024) argue that RCE avoids thresholding of the correctness score and applies to any type of output range for the uncertainty metric. The positively oriented CMA measure shares these desirable properties. Furthermore, if the correctness values are dichotomous at the outset, CMA reduces to the AUC measure at (3), which has been widely used in the setting at hand (Kuhn et al., 2023; Lin et al., 2024). For any type of outcome, a value of CMA $= \frac{1}{2}$ corresponds to random assignments.

While RCE was developed specifically for the task at hand, we now compare to our general purpose measure CMA that quantifies monotone association. Specifically, Table 1 shows CMA$(-X, Y)$ in 180 experimental configurations, where X is one of the five uncertainty metrics and Y is one of the four correctness scores, for each of the three language models and three query sources. We note a certain robustness to the choice of the correctness score; in particular, the results under the ROUGE-L and the ROUGE-1 scores are very similar. For the Llama-2-7b and Llama-2-7b-chat models and responses to SQuAD queries various CMA values are below $\frac{1}{2}$, whereas for NQ-open and TriviaQA queries the CMA value is above $\frac{1}{2}$ for all $3 \times 2 \times 5 \times 4 = 120$ configurations of LLM, query source, uncertainty metric, and correctness score, respectively. The markedly different pattern observed for SQuAD with many CMA values near or even below $\frac{1}{2}$ may stem from fundamental differences in task structure compared to the open-domain question answering format of NQ-open and TriviaQA. The reading comprehension format of SQuAD creates uncertainty not only about factual knowledge, but also related to response granularity, as multiple correct answers of varying exactitude might co-exist. We encourage the future exploration of task-dependent performance patterns to inform evaluation practices in the LLM research community.

Based on the tests developed in Section 3 statistical significance can be assessed. Specifically, we test the hypothesis that CMA equals the null value of $\frac{1}{2}$ based on the test statistic at (41), and we consider pairwise comparisons based on the test statistic at (43). In tests of H_0 : CMA $\leq \frac{1}{2}$ at level .01 every single CMA value larger than .500 in Table 1 entails a rejection. The results of one-sided tests of equal CMA between the uncertainty metrics with the highest and the second-highest CMA value, respectively, are documented in the table.

The scatterplot in Figure 3 shows CMA and RCE, where we follow the computation of the RCE values in Huang et al. (2024) and average over 20 bootstrap samples. We note that CMA and RCE quantify

distinct facets of predictive performance. The all-purpose measure CMA shares desirable properties of RCE and might be easier to use and interpret, due to the natural threshold of $\frac{1}{2}$ that corresponds to randomly assigned judgments of uncertainty, and the reduction to AUC when the correctness values are dichotomous in the first place.

4.2 WeatherBench 2: Data-driven vs physics-based weather prediction

The past few years have witnessed dramatic change in the practice of weather prediction. While until a few years ago, weather forecasts had relied on physics-based numerical weather prediction (NWP) models, recent advances in neural network methodologies have enabled the rise of purely data-driven, artificial intelligence (AI) based weather prediction (AIWP; Bi et al., 2023; Lam et al., 2023) models. Arguably, by now AIWP models are outperforming NWP models (Ben Bouallègue et al., 2024; Gneiting et al., 2025). However, in contrast to temperature, pressure, and wind speed, forecasts of precipitation accumulations have received scant attention only in the comparison of physics-based and data-driven weather forecasts, with the notable exception of Radford et al. (2025). The reasons for the neglect of precipitation forecasts are two-fold. On the one hand, researchers typically rely on the ERA5 product (Hersbach et al., 2020) for ground truth data, and the quality thereof can be poor for precipitation, particularly towards polar regions, where precipitation subsumes rainfall, hail, and snow (Lavers et al., 2022). Furthermore, precipitation is a mixed discrete-continuous variable with a point mass at zero (when there is no precipitation) and a skewed and heavy right tail, which challenges the application of standard evaluation measures. However, our AGC and CMA measures adapt naturally to the mixed discrete-continuous character of precipitation accumulations.

In this light, we now consider the WeatherBench 2 benchmark dataset (Rasp et al., 2024) and use the CMA measure to compare NWP and AIWP forecasts of 24-hour precipitation accumulation at a prediction horizon of a day ahead. We restrict the analysis to the premier NWP model, namely, the European Centre for Medium Range Weather Forecasts (ECMWF) high-resolution (HRES) model, and a leading AIWP model, namely the GraphCast model (Lam et al., 2023) in its operational version,[8] which are run with the same initial conditions, thus allowing for a fair comparison. The spatial resolution is 1.5 degrees, so each latitude band comprises 240 grid cells, for which forecasts were generated twice daily (in retrospect, for 2020) with initialization times 12 hours apart. For reference, we compare the HRES and GraphCast models to the static WeatherBench 2 climatology forecast, which does not invoke numerical, statistical, or machine learning methods, and to the simplistic persistence forecast, which projects the most recent observed precipitation forward. The aforementioned ERA5 product (Hersbach et al., 2020) serves to provide ground truth.

We follow up on the analysis in Section 5.3 of Gneiting and Walz (2022) for the earlier WeatherBench 1 dataset (Rasp et al., 2020), where NWP forecasts outperformed AIWP forecasts by huge margins. As in the earlier study, we use a range of performance metrics, which includes CMA skill in the form of the percentage improvement of CMA for the forecast at hand over CMA for the climatology forecast. Furthermore, we consider root mean squared error (RMSE) skill relative to the climatology forecast, stable equitable error in probability space (SEEPS; Rodwell et al., 2010) skill relative to the climatology forecast, and the anomaly correlation coefficient (ACC) in the form described by Rasp et al. (2024, Section 4.3.2). All four measures are positively oriented — the larger, the better — and by definition their value equals zero for the climatology forecast.[9]

Figure 4 shows the performance measures in their dependence on latitude bands at a resolution of 1.5 degrees. Each latitude band has 240 grid cells with forecasts twice daily in the 2020 evaluation period, for a total of 732 forecasts per grid cell. We compute the performance measures grid cell by grid cell and then average across the 240 grid cells in the latitude band. The climatology forecast is unable

[8] These models are denoted IFS HRES and GraphCast (oper.) in the WeatherBench 2 dataset (Rasp et al., 2024), which is available at https://sites.research.google/gr/weatherbench/.

[9] In passing, we note that the ideas behind the representation of AGC and CMA in (5), (6), and (9), respectively, resemble those put forth by meteorologists in the context of forecast evaluation for precipitation forecasts, where the linear error in probability space (LEPS) and SEEPS measures have been proposed (Rodwell et al., 2010). However, unlike SEEPS, AGC and CMA do not require researchers to artificially discretize their data. We also note that RMSE is a measure of actual predictive performance, whereas ACC, AGC, and CMA can be interpreted as measures of potential predictive ability.

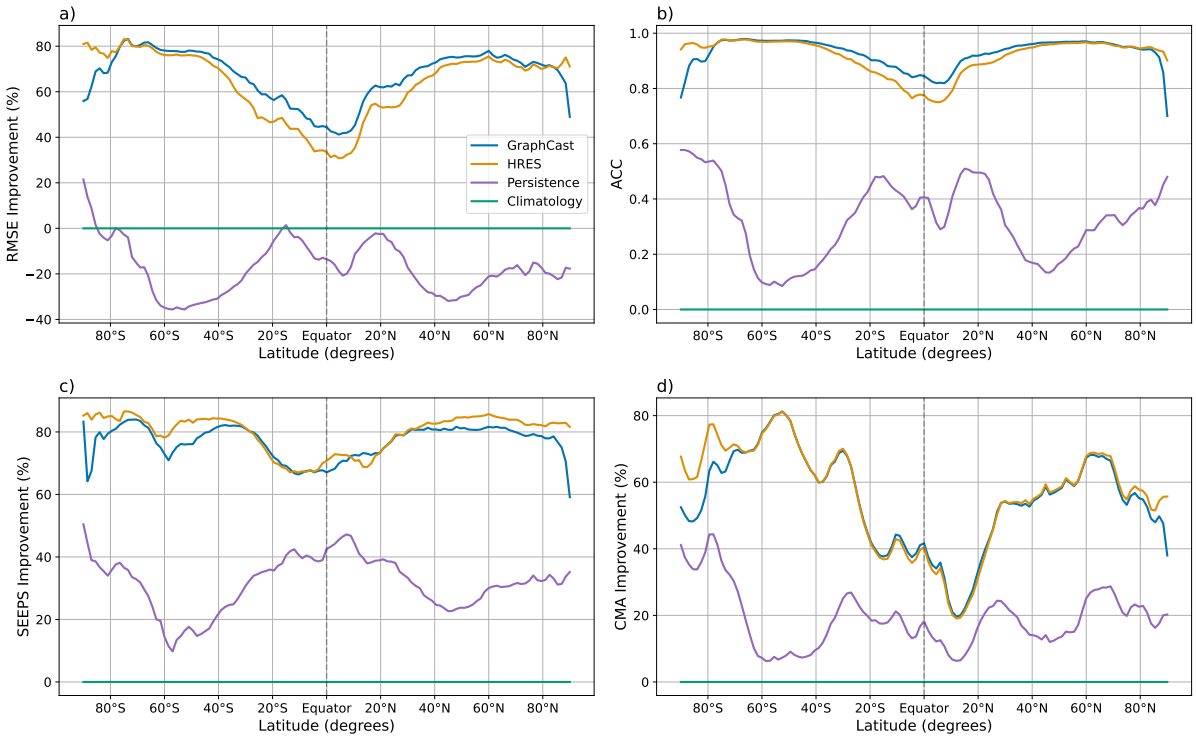

Figure 4: Predictive ability of WeatherBench 2 HRES, WeatherBench 2 GraphCast, and persistence forecasts of 24-hour precipitation accumulation in terms of a) RMSE skill relative to the WeatherBench 2 climatology forecast, b) ACC, c) SEEPS skill, and d) CMA skill. The performance measures are averaged over the 240 grid cells in each latitude band.

to capture day-to-day variability but tends to have smaller errors than the persistence forecast, which extrapolates the precipitation patterns at the grid cell at hand, but magnifies magnitude errors when weather systems move or intensify swiftly. Therefore, climatology outperforms the persistence forecast in terms of RMSE, and vice versa in terms of ACC, SEEPS, and CMA. The GraphCast and HRES forecasts outperform the reference forecasts by huge margins. Interestingly, the data-driven GraphCast forecast has a competitive edge over the physics-based HRES forecast in the tropics near the equator, and vice versa towards the poles. However, details depend on the performance measure used, and the differences between the GraphCast and HRES forecasts pale relative to the differences to the reference forecasts.

In this light, Figure 5 provides a more nuanced comparison of the predictive performance of the WeatherBench 2 GraphCast and HRES forecasts in terms of CMA. As noted, each latitude band comprises 240 grid cells, and for each latitude band we show a boxplot of the respective differences in CMA, with positive values indicating superiority of the GraphCast model. WeatherBench 2 comprises two daily runs with initialization times 12 hours apart, so successive forecasts and outcomes for 24-hour precipitation accumulation must be assumed to be dependent. Therefore, we use HAC estimates in the denominator of the test statistic at (43), with implementation details as proposed by Pohle et al. (2025, Section 5.3). The boxplots for the respective one-sided p-values show that around the equator GraphCast has a competitive edge over HRES, whereas HRES outperforms GraphCast towards the poles. These findings are striking in view of the emerging consensus that the latest data-driven weather models outperform physics-based models. However, they are broadly compatible with a note in Ben Bouallègue et al. (2024, p. 869), who report that for pressure the otherwise competitive, data-driven Pangu-Weather model of Bi et al. (2023) performs much worse over the central Arctic than the HRES model. One possible explanation is the use of cosine-latitude weighted scores when training the data-driven models (Ben Bouallègue et al., 2024). A complementary or alternative explanation lies in a possible neglect

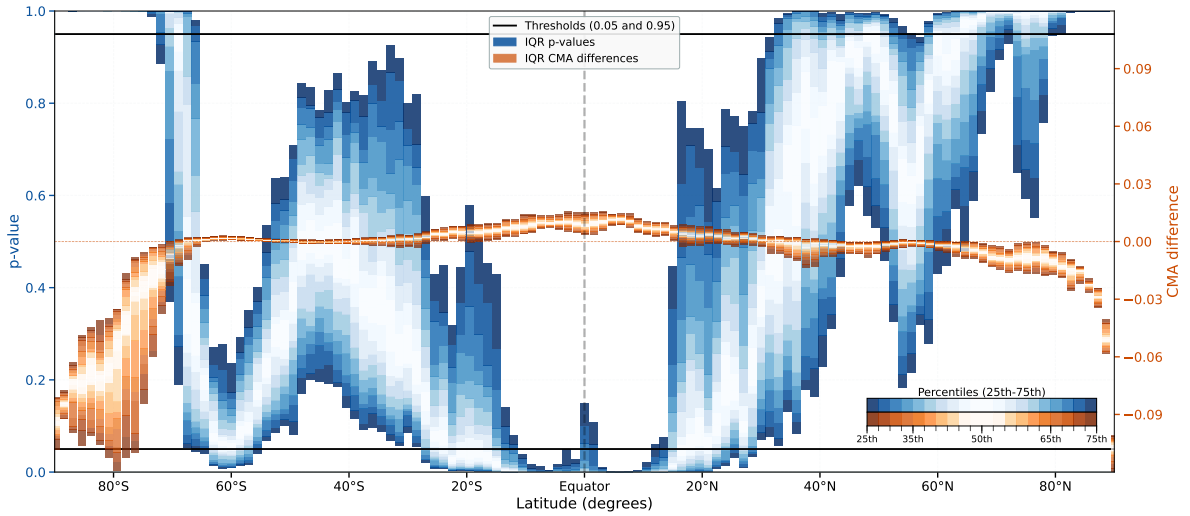

Figure 5: Comparison of WeatherBench 2 GraphCast and HRES forecasts of 24-hour precipitation accumulation by latitude bands. The boxplots show the interquartile range (IQR) of the difference in CMA in shades of red, and the associated one-sided p-value for the test statistic at (43) in shades of blue, considering the 240 grid cells in each latitude band, respectively. Positive CMA differences and small p-values indicate superiority of the GraphCast model over the HRES model, and vice versa.

of the tropics in the decade-long development of physics-based NWP models, with almost exclusive involvement of researchers residing at mid-latitude locations. We find it tempting to speculate that, if further development of NWP models were to target precipitation in the tropics, while the training of AIWP models was to give more emphasis to subpolar and polar regions, then the predictive ability of AIWP models and NWP models would be roughly on par globally, and rather close to predictability limits, given current observational assets.

5 Discussion

We have introduced asymmetric grade correlation (AGC) and the coefficient of monotone association (CMA) for the quantification of monotone dependence between random variables X and Y. The two measures relate linearly as $\mathrm{AGC}(X,Y) = 2\,\mathrm{CMA}(X,Y) - 1$. If the outcome Y is binary, then $\mathrm{CMA}(X,Y)$ equals the asymmetric area under the receiver operating characteristic (ROC) curve measure, $\mathrm{AUC}(X,Y)$. If X and Y are continuous, then AGC is symmetric, equals Spearman's rank correlation coefficient, and can be represented as a weighted average of the AUC values for the induced binary problems. Alternatively, CMA can be interpreted as a weighted probability of concordance.

Therefore, AGC and CMA provide a natural bridge from AUC to Spearman's rank correlation, for a unified way of quantifying monotone dependence for dichotomous, ranked categorical, mixed discrete-continuous, and continuous types of outcomes. The associated large sample theory extends, bridges, and unifies previously disjoint literatures and methodologies for inference about monotone dependence. In particular, our generalization of the DeLong et al. (1988) test for pairwise comparisons of AUC values for binary outcomes allows for pairwise comparisons of CMA values for all types of linearly ordered outcomes. Software is available in Python (Python Software Foundation, 2025), and we refer the reader to `https://github.com/evwalz/paper_monotone_dependence` for code and reproduction materials.

The AGC and CMA measures address a longstanding issue with symmetric correlation coefficients in discrete settings, namely, that even under perfect monotone dependence their optimal values may not be attainable, as discussed by Nešlehová (2007), Pohle et al. (2025), and Pohle and Wermuth (2025), among others. The attainability problem is unavoidable if symmetric measures of dependence are used. In this light, we join Chatterjee (2021) and Azadkia and Roudaki (2025) in arguing that, in much of statistical

practice and data analytics, a reorientation from symmetric to asymmetric measures of dependence might be warranted.

To suggest a further such measure, the asymmetric Kendall correlation (AKC) or Somers' D at (19) and the associated C index (CID) at (14) provide a natural bridge from AUC to Kendall's rank correlation, while avoiding the attainability problem with the symmetric version of Kendall correlation. The recently developed large sample theory of Pohle et al. (2025) allows for a similar generalization of the DeLong et al. (1988) test as in our paper, but now based on AKC and CID, in lieu of AGC and CMA, respectively. Another challenge is the further development of the limit laws and the associated tools for statistical inference to the general setting at (40). We leave these and related developments to future work.

Acknowledgements

We thank Marc-Oliver Pohle for numerous discussions and generously sharing a preliminary version of Pohle et al. (2025), which supplied the technical tools in the proofs of the asymptotic results in our Section 3. We have furthermore benefited from discussions with Alexander Jordan, Florian Kalinke, Sebastian Lerch, Jan Stühmer, and Lu Yang. Eva-Maria Walz and Tilmann Gneiting are grateful for support by the Klaus Tschira Foundation.

A Proofs for Section 2

Proof of Theorem 2.2. The definition of the AGC measure at (5) implies that

$$\mathrm{CMA}(X,Y) = \frac{1}{2}\left(\frac{\mathrm{cov}(\bar{F}(X),\bar{G}(Y))}{\mathrm{cov}(\bar{G}(Y),\bar{G}(Y))}+1\right) = \frac{\mathbb{E}\big(\bar{F}(X)\bar{G}(Y)\big)+\mathbb{E}\big((\bar{G}(Y))^2\big)-\frac{1}{2}}{2\mathbb{E}\big((\bar{G}(Y))^2\big)-\frac{1}{2}}. \tag{A.1}$$

To establish the equivalence of the representations in (9) and (A.1), respectively, we consider enumerator and denominator separately. For the enumerator,

$$\begin{aligned}
&\mathbb{E}\big((\bar{G}(Y'')-\bar{G}(Y'))\mathbb{1}\{Y'<Y''\}s(X',X'')\big)\\
&\quad=\mathbb{E}\big((\bar{G}(Y'')-\bar{G}(Y'))s(Y',Y'')s(X',X'')\big)\\
&\quad=\mathbb{E}\big(\bar{G}(Y')s(Y'',Y')s(X'',X')\big)-\mathbb{E}\big(\bar{G}(Y')s(Y',Y'')s(X',X'')\big)\\
&\quad=\mathbb{E}\big(\bar{F}(X)\bar{G}(Y)\big)+\mathbb{E}\big(\bar{G}(Y)^2\big)-\tfrac{1}{2},
\end{aligned}$$

where the final equality follows upon conditioning on (X',Y') and simplifying. For the denominator,

$$\mathbb{E}\big((\bar{G}(Y'')-\bar{G}(Y'))\mathbb{1}\{Y'<Y''\}\big)=\mathbb{E}\big(\bar{G}(Y'')G(Y''-)\big)-\mathbb{E}\big(\bar{G}(Y')(1-G(Y'))\big)=2\,\mathbb{E}\big(\bar{G}(Y)\big)^2\big)-\tfrac{1}{2},$$

where $G(y-)=\lim_{z\uparrow y}G(z)$, with the first equality following upon conditioning on Y'' and Y', respectively, and the second equality using the fact that $G(y-)+G(y)=2\,\bar{G}(y)$ for $y\in\mathbb{R}$. □

Proof of Corollary 2.3. Immediate from (3) and (9). □

Proof of Corollary 2.4. Immediate from (1), (5), (7), and (9), respectively. □

Proof of Theorem 2.5. Since G is continuous, $\bar{G}=G$, and we conclude from (9) and (10) in concert with

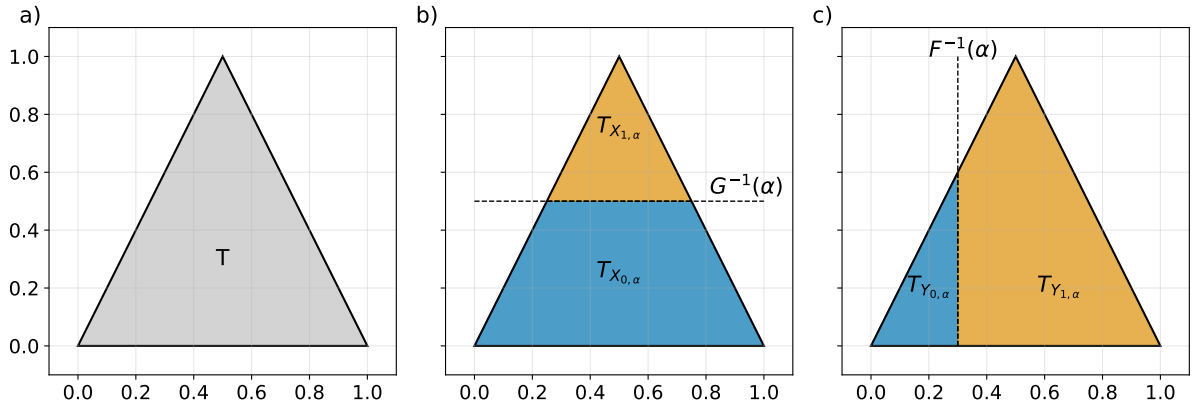

Figure A.1: The random vector (X, Y) in Example 2.7 is uniformly distributed on a) the triangle T. The b) horizontal and c) vertical divisions support the associated computations in this appendix.

Fubini's theorem that

$$\begin{aligned}
\mathrm{CMA}(X,Y) &= \frac{\mathbb{E}\left[(G(Y'') - G(Y'))\,\mathbb{1}\{Y' < Y''\}s(X', X'')\right]}{\mathbb{E}\left[(G(Y'') - G(Y'))\,\mathbb{1}\{Y' < Y''\}\right]} \\
&= \frac{\int \mathbb{E}\left(s(X', X'')\,\mathbb{1}\{Y' < Y''\}\,\mathbb{1}\{Y' < y \le Y''\}\right) \mathrm{d}G(y)}{\int \mathbb{E}\left(\mathbb{1}\{Y' < Y''\}\,\mathbb{1}\{Y' < y \le Y''\}\right) \mathrm{d}G(y)} \\
&= \frac{\int \mathbb{E}\left(s(X', X'')\,\mathbb{1}\{Y' < y\}\,\mathbb{1}\{Y'' \ge y\}\right) \mathrm{d}G(y)}{\int \mathbb{E}\left(\mathbb{1}\{Y' < y\}\,\mathbb{1}\{Y'' \ge y\}\right) \mathrm{d}G(y)} \\
&= \frac{\int_0^1 \mathbb{E}\left(s(X', X'')\,\mathbb{1}\{Y' < G^{-1}(\alpha)\}\,\mathbb{1}\{Y'' \ge G^{-1}(\alpha)\}\right) \mathrm{d}\alpha}{\int_0^1 \mathbb{E}\left(\mathbb{1}\{Y' < G^{-1}(\alpha)\}\,\mathbb{1}\{Y'' \ge G^{-1}(\alpha)\}\right) \mathrm{d}\alpha} \\
&= \frac{\int_0^1 \mathbb{E}\left(\mathbb{1}\{Y' < G^{-1}(\alpha), Y'' \ge G^{-1}(\alpha)\}\right) \mathbb{E}\left(s(X', X'') \mid Y' < G^{-1}(\alpha), Y'' \ge G^{-1}(\alpha)\right) \mathrm{d}\alpha}{\int_0^1 \alpha(1-\alpha)\,\mathrm{d}\alpha} \\
&= 6 \int_0^1 \alpha(1-\alpha)\,\mathrm{AUC}^{(\alpha)}(X,Y)\,\mathrm{d}\alpha,
\end{aligned}$$

where the fourth equality follows upon substituting $\alpha = G(y)$. □

Proof of Corollary 2.6. Immediate from Corollary 2.4 and Theorem 2.5. □

Details for Example 2.7. The joint distribution $\mathbb{P}$ of (X, Y) is uniform on the triangle T, as illustrated in Figure A.1a). As $\mathbb{P}$ and the marginal distributions $F = \mathcal{L}(X)$ and $G = \mathcal{L}(Y)$ are continuous, we may write

$$\mathrm{AUC}^{(\alpha)}(X,Y) = \mathbb{P}(X_{0,\alpha} < X_{1,\alpha}),$$

where $X_{0,\alpha}$ and $X_{1,\alpha}$ are independent with distribution equal to $\mathcal{L}(X \mid Y < G^{-1}(\alpha))$ and $\mathcal{L}(X \mid Y \ge G^{-1}(\alpha))$, respectively.

The quantity $\mathrm{AUC}^{(\alpha)}(X,Y)$ is straightforward to find for reasons of symmetry. From Figure A.1b), the distributions of $X_{0,\alpha}$ and $X_{1,\alpha}$ are symmetric about $\frac{1}{2}$, so

$$\mathbb{P}(X_{0,\alpha} < X_{1,\alpha}) = \mathbb{P}(1 - X_{0,\alpha} < 1 - X_{1,\alpha}) = \mathbb{P}(X_{0,\alpha} > X_{1,\alpha}).$$

Therefore, $\mathrm{AUC}^{(\alpha)}(X,Y) = \mathbb{P}(X_{0,\alpha} < X_{1,\alpha}) = \frac{1}{2}$, as claimed.

Similarly, we may write

$$\mathrm{AUC}^{(\alpha)}(Y,X) = \mathbb{P}(Y_{0,\alpha} < Y_{1,\alpha}),$$

where $Y_{0,\alpha}$ and $Y_{1,\alpha}$ are independent with distribution equal to $\mathcal{L}(Y \mid X < F^{-1}(\alpha))$ and $\mathcal{L}(Y \mid X \geq F^{-1}(\alpha))$, respectively (Figure A.1c). It suffices to consider $\alpha \in (0, \frac{1}{2})$, since

$$\mathrm{AUC}^{(1-\alpha)}(Y, X) = \mathbb{P}(Y_{0,1-\alpha} < Y_{1,1-\alpha}) = 1 - \mathbb{P}(Y_{0,\alpha} < Y_{1,\alpha}) = 1 - \mathrm{AUC}^{(\alpha)}(Y, X) \tag{A.2}$$

for reasons of symmetry. If we let $x_\alpha = F^{-1}(\alpha)$, the densities of $Y_{0,\alpha}$ and $Y_{1,\alpha}$ are given by

$$g_{0,\alpha}(y) = \frac{1}{x_\alpha^2}\left(x_\alpha - \frac{y}{2}\right) \mathbb{1}\{0 \leq y \leq 2x_\alpha\}$$

and

$$g_{1,\alpha}(y) = \frac{1}{1-2x_\alpha^2}\left[\left(1 - x_\alpha - \frac{y}{2}\right) \mathbb{1}\{0 \leq y \leq 2x_\alpha\} + (1-y)\,\mathbb{1}\{2x_\alpha \leq y \leq 1\}\right],$$

respectively. Therefore,

$$\begin{aligned}
\mathrm{AUC}^{(\alpha)}(Y, X) &= \mathbb{P}(Y_{0,\alpha} < Y_{1,\alpha}) \\
&= \int_0^1 \int_0^1 \mathbb{1}\{y_0 < y_1\}\, g_{0,\alpha}(y_0)\, g_{1,\alpha}(y_1)\, \mathrm{d}y_0\, \mathrm{d}y_1 \\
&= \int_0^{2x_\alpha} \int_0^{y_1} \frac{1}{x_\alpha^2}\left(x_\alpha - \frac{y_0}{2}\right) \mathrm{d}y_0\, \frac{2}{1-2x_\alpha^2}\left(1 - x_\alpha - \frac{y_1}{2}\right) \mathrm{d}y_1 + \int_{2x_\alpha}^1 \frac{2}{1-2x_\alpha^2}(1-y_1)\,\mathrm{d}y_1 \\
&= \frac{2}{1-2x_\alpha^2}\left[\frac{1}{x_\alpha^2}\int_0^{2x_\alpha}\left(x_\alpha y_1 - \frac{y_1^2}{4}\right)\left(1 - x_\alpha - \frac{y_1}{2}\right)\mathrm{d}y_1 + \int_{2x_\alpha}^1 (1-y_1)\,\mathrm{d}y_1\right] \\
&= \frac{2}{1-2x_\alpha^2}\left[\frac{1}{x_\alpha^2}\int_0^{2x_\alpha}\left((x_\alpha - x_\alpha^2)\, y_1 - \frac{1+x_\alpha}{4} y_1^2 + \frac{1}{8} y_1^3\right)\mathrm{d}y_1 + (1-2x_\alpha) - \frac{1-4x_\alpha^2}{2}\right] \\
&= 1 - \frac{\frac{4}{3}x_\alpha + \frac{1}{3}x_\alpha^2}{1-2x_\alpha^2}.
\end{aligned}$$

To find $x_\alpha = F^{-1}(\alpha)$ for $\alpha \in (0, \frac{1}{2})$, we note that the density and CDF of X on $[0, \frac{1}{2}]$ are given by $f(x) = 4x$ and $F(x) = 2x^2$, respectively. Hence, $x_\alpha = F^{-1}(\alpha) = \sqrt{\alpha}/\sqrt{2}$ and

$$\mathrm{AUC}^{(\alpha)}(Y, X) = 1 - \frac{\frac{2}{3}\sqrt{2\alpha} + \frac{\alpha}{6}}{1-\alpha} \tag{A.3}$$

for $\alpha \in (0, \frac{1}{2})$, as claimed. For reasons of symmetry, $\mathrm{AUC}^{(1/2)}(Y, X) = \frac{1}{2}$, whereas $\mathrm{AUC}^{(\alpha)}(Y, X)$ for $\alpha \in (\frac{1}{2}, 1)$ is obtained from (A.2) and (A.3). □

Proof of Theorem 2.8. Immediate from (6) and (9), respectively. □

Proof of Proposition 2.9. If Y is dichotomous and nondegenerate, then

$$\mathrm{CID}(X, Y) = \mathbb{E}(s(X', X'') \mid Y' < Y'') = \mathbb{E}(s(X', X'') \mid Y' = 0, Y'' = 1) = \mathrm{AUC}(X, Y)$$

and

$$\mathrm{RGA}(X, Y) = \frac{1}{2}\left(\frac{\mathrm{cov}(Y, \bar{F}(X))}{\mathrm{cov}(Y, \bar{G}(Y))} + 1\right) = \frac{1}{2}\left(\frac{\mathrm{cov}(\bar{G}(Y), \bar{F}(X))}{\mathrm{cov}(\bar{G}(Y), \bar{G}(Y))} + 1\right) = \mathrm{CMA}(X, Y);$$

furthermore, $\mathrm{CMA}(X, Y) = \mathrm{AUC}(X, Y)$ by Corollary 2.6. □

Proof of Proposition 2.10. If X and Y are continuous, (16) and (17) are immediate from (6) and (7), and (18) follows readily from the definitions of $\mathrm{CID}(X, Y)$ at (14) and τ_{K} at (2), respectively. □

Details for Example 2.12. The conditional distribution of Y given $X = x$, where $x \in [0,1]$, is uniform on $[0, 1 - 2|x - \frac{1}{2}|]$. Therefore,

$$\begin{aligned}\xi(X,Y) &= 6\int_0^1 \operatorname{var}\left(\mathbb{E}(\mathbb{1}\{Y \le G^{-1}(\alpha)\} \mid X)\right) \mathrm{d}\alpha \\ &= 6\int_0^1 \int_0^1 \mathbb{P}(Y \le G^{-1}(\alpha) \mid X = x)^2 \, f(x) \, \mathrm{d}x \, \mathrm{d}\alpha - 2 \\ &= \frac{1}{6},\end{aligned}$$

where $f(x) = 4x$ for $x \le \frac{1}{2}$ and $f(x) = 4(1-x)$ for $x > \frac{1}{2}$, and

$$\mathbb{P}\left(Y \le G^{-1}(\alpha) \mid X = x\right) = \begin{cases} \frac{1-\sqrt{1-\alpha}}{1-2|x-\frac{1}{2}|}, & |x - \frac{1}{2}| < \frac{\sqrt{1-\alpha}}{2}, \\ 1, & |x - \frac{1}{2}| \ge \frac{\sqrt{1-\alpha}}{2}. \end{cases}$$

Similarly, the conditional distribution of X given $Y = y$, where $y \in [0,1]$, is uniform on the interval $[\frac{y}{2}, 1 - \frac{y}{2}]$, so

$$\begin{aligned}\xi(Y,X) &= 6\int_0^1 \operatorname{var}\left(\mathbb{E}(\mathbb{1}\{X \le F^{-1}(\alpha)\} \mid Y)\right) \mathrm{d}\alpha \\ &= 6\int_0^1 \left(\mathbb{E}\left(\mathbb{P}\left(X \le F^{-1}(\alpha) \mid Y\right)^2\right) - \alpha^2\right) \mathrm{d}\alpha \\ &= 6\int_0^1 \int_0^1 \mathbb{P}\left(X \le F^{-1}(\alpha) \mid Y = y\right)^2 \, 2(1-y) \, \mathrm{d}y \, \mathrm{d}\alpha - 2 \\ &= 6\int_0^{1/2} \left(\int_0^{2\sqrt{\alpha/2}} \left(\frac{\sqrt{\alpha/2} - y/2}{1-y}\right)^2 2(1-y)\,\mathrm{d}y + \int_{2(1-\sqrt{\alpha/2})}^1 2(1-y)\,\mathrm{d}y \right) \mathrm{d}\alpha \\ &\quad + 6\int_{1/2}^1 \int_0^{2\sqrt{(1-\alpha)/2}} \left(\frac{(1-\sqrt{(1-\alpha)/2}) - y/2}{1-y}\right)^2 2(1-y)\,\mathrm{d}y\,\mathrm{d}\alpha - 2 \\ &= 6\left(\frac{37}{288} + \frac{61}{288}\right) - 2 \\ &= \frac{1}{24},\end{aligned}$$

where we use that

$$\mathbb{P}\left(X \le F^{-1}(\alpha) \mid Y = y\right) = \begin{cases} \frac{F^{-1}(\alpha) - y/2}{1-y}, & 0 \le y < 2m_\alpha, \\ \mathbb{1}\{F^{-1}(\alpha) > \frac{1}{2}\}, & 2m_\alpha \le y < 2M_\alpha, \\ \mathbb{1}\{F^{-1}(\alpha) < \frac{1}{2}\}, & 2M_\alpha \le y \le 1. \end{cases}$$

with $m_\alpha := \min\{F^{-1}(\alpha), 1 - F^{-1}(\alpha)\}$ and $M_\alpha := \max\{F^{-1}(\alpha), 1 - F^{-1}(\alpha)\}$ and $F^{-1}(\alpha) = \sqrt{\alpha/2}\,\mathbb{1}\{\alpha \le \frac{1}{2}\} + (1 - \sqrt{(1-\alpha)/2})\,\mathbb{1}\{\alpha > \frac{1}{2}\}$, respectively. □

B Proofs for Section 3

Proof of Proposition 3.1. For data of the form (20), let $\bar{F}_n$ and $\bar{G}_n$ denote the empirical MDF of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively. The mid rank $\bar{R}_j$ associated with Y_j equals $\bar{R}_j = \sum_{i=1}^n \mathbb{1}\{Y_i < Y_j\} + \frac{1}{2}\sum_{i=1}^n \mathbb{1}\{Y_i = Y_j\} + \frac{1}{2}$. Therefore,

$$\bar{G}_n(Y_i) = \frac{1}{n}\left(\bar{R}_i - \frac{1}{2}\right), \qquad \mathbb{E}\left[\bar{G}_n(Y_i)\right] = \frac{1}{2}$$

for $i = 1, \ldots, n$, and analogously for $\bar{F}_n(X_i) = \frac{1}{n}\bar{Q}_i$. Therefore, (23) and (24) follow from (5) and (7) by plugging in and simplifying. Similarly, (25) follows from (9) by plugging in and simplifying. The transition to (26) then is straightforward. □

Proof of Corollary 3.2. If $Y_1, \ldots, Y_n$ are pairwise distinct, we find from (26) and (27) that $\mathrm{CPA}_n = \mathrm{CMA}_n$. The second equality then follows readily from Theorem 1 in concert with Definition 3 in Gneiting and Walz (2022). □

Our proof of the central limit laws in Theorems 3.4 and 3.6 relies heavily on theory recently developed by Pohle et al. (2025), and we review their results. We operate under Scenario 3.5, which nests Scenario 3.3, and consider the sequence at (35).

Therefore, each $(X_i^{(1)}, \ldots, X_i^{(m)}, Y_i)$ has the same law $\mathbb{P}$ as the random vector $(X^{(1)}, \ldots, X^{(m)}, Y)$, with $X^{(1)}, \ldots, X^{(m)}$, and Y all having strictly positive variance. Let us recall that $\bar{G}$ and γ_G denote the univariate MDF and the granularity (8) of the outcome Y. For $k = 1, \ldots, m$, we write $\bar{F}^{(k)}$ for the univariate MDF of $X^{(k)}$ and $\bar{H}^{(k)}$ for the bivariate MDF of $(X^{(k)}, Y)$, and we define the nondecreasing functions $\widetilde{F}^{(k)}(x) = \mathbb{E}[\bar{H}^{(k)}(X^{(k)}, Y)]$ for $x \in \mathbb{R}$ and $\widetilde{G}^{(k)}(y) = \mathbb{E}[\bar{H}^{(k)}(X^{(k)}, y)]$ for $y \in \mathbb{R}$, respectively. Furthermore, we let $\rho^{(k)} = 12\,\mathrm{cov}(\bar{F}^{(k)}(X^{(k)}), \bar{G}(Y))$ as defined at (36).

Switching from population quantities to sample quantities, let $\rho_n^{(1)}, \ldots, \rho_n^{(m)}$ and γ_n, respectively, denote the coefficients at (36) and the granularity (8) under the empirical law of the first n random vectors in the sequence at (35). We note that the sample coefficients and $1 - \gamma_n$ are asymptotically equivalent to versions of the U-statistic in eq. (12) of Pohle et al. (2025). Therefore, Proposition A.3 and Lemma B.1 of Pohle et al. (2025) yield the following result.

Proposition B.1. *Under Scenario 3.5 it holds that*

$$\sqrt{n}\left(\begin{pmatrix}\rho_n^{(1)}\\ \vdots \\ \rho_n^{(m)} \\ \gamma_n\end{pmatrix} - \begin{pmatrix}\rho^{(1)}\\ \vdots \\ \rho^{(m)} \\ \gamma_G\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0\\ \vdots \\ 0 \\ 0\end{pmatrix}, \Sigma_\circ = \left(\Sigma_\circ^{(k,l)}\right)_{k,l=1}^{m+1}\right),$$

where

$$\Sigma_\circ^{(k,l)} = 9\,\mathbb{E}\Big(K^{(k)}(X^{(k)}, Y)\,K^{(l)}(X^{(l)}, Y)\Big)$$

for $k, l = 1, \ldots, m$,

$$\Sigma_\circ^{(m+1,k)} = \Sigma_\circ^{(k,m+1)} = 9\,\mathbb{E}(K^{(k)}(X^{(k)}, Y)\,K^{(0)}(Y))$$

for $k = 1, \ldots, m$, *and*

$$\Sigma_\circ^{(m+1,m+1)} = 9\,\mathbb{E}(K^{(0)}(Y)^2),$$

with

$$K^{(k)}(x, y) = 4\,\widetilde{F}^{(k)}(x) + 4\,\widetilde{G}^{(k)}(y) + \bar{F}^{(k)}(x)\bar{G}(y) - \bar{F}^{(k)}(x) - \bar{G}(y) + 1 - \rho^{(k)}$$

for $k = 1, \ldots, m$ *and*

$$K^{(0)}(y) = (G(y) - G(y-))^2 - \gamma_G,$$

respectively.

Towards the proof of Theorem 3.6, we remain in Scenario 3.5 and note that

$$\mathrm{AGC}^{(k)} = \mathrm{AGC}(X^{(k)}, Y) = \frac{\mathrm{cov}(\bar{F}^{(k)}(X^{(k)}), \bar{G}(Y))}{\mathrm{var}(\bar{G}(Y))} = \frac{\rho^{(k)}}{1 - \gamma_G} \tag{B.1}$$

for the population quantities and analogously $\mathrm{AGC}_n^{(k)} = \rho_n^{(k)}/(1 - \gamma_n)$ for the sample quantities, where $k = 1, \ldots, m$. Therefore, we can apply the delta method to find the Gaussian limit distribution of the vector $(\mathrm{AGC}_n^{(1)}, \ldots, \mathrm{AGC}_n^{(m)})'$. The details are as follows.

Proof of Theorem 3.6. In view of Proposition B.1 and the equality at (B.1), we may apply the delta method to show that

$$\sqrt{n}\left(\begin{pmatrix}\mathrm{AGC}_n^{(1)}\\ \vdots \\ \mathrm{AGC}_n^{(m)}\end{pmatrix} - \begin{pmatrix}\mathrm{AGC}^{(1)}\\ \vdots \\ \mathrm{AGC}^{(m)}\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}, \Sigma = \left(\Sigma^{(k,l)}\right)_{k,l=1}^{m}\right),$$

where $\Sigma = \Lambda\Sigma_\circ\Lambda'$ with the matrix $\Sigma_\circ = \left(\Sigma_\circ^{(k,l)}\right)_{k,l=1}^{m}$ as specified in Proposition B.1 and

$$\Lambda = \begin{pmatrix} 1/(1-\gamma_G) & 0 & \cdots & 0 & \rho^{(1)}/(1-\gamma_G)^2 \\ 0 & 1/(1-\gamma_G) & \cdots & 0 & \rho^{(2)}/(1-\gamma_G)^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1/(1-\gamma_G) & \rho^{(m)}/(1-\gamma_G)^2 \end{pmatrix} \in \mathbb{R}^{m\times(m+1)}.$$

Straightforward computations yield

$$\Sigma^{(k,l)} = \frac{\Sigma_\circ^{(k,l)}}{(1-\gamma_G)^2} + \frac{\rho^{(k)}\,\Sigma_\circ^{(l,m+1)} + \rho^{(l)}\,\Sigma_\circ^{(k,m+1)}}{(1-\gamma_G)^3} + \frac{\rho^{(k)}\rho^{(l)}\,\Sigma_\circ^{(m+1,m+1)}}{(1-\gamma_G)^4}$$

for $k, l = 1, \ldots, m$, from which (38) follows readily. □

Proof of Theorem 3.4. Evidently, the first part is a special case of Theorem 3.6. If X and Y are independent, then $\rho = 0$, $\widetilde{F}(x) = \frac{1}{2}\bar{F}(x)$ for $x \in \mathbb{R}$, and $\widetilde{G}(y) = \frac{1}{2}\bar{G}(y)$ for $y \in \mathbb{R}$, so the kernel at (30) simplifies to

$$K^{(1)}(x,y) = 4\left(\bar{F}(x) - \frac{1}{2}\right)\left(\bar{G}(y) - \frac{1}{2}\right).$$

The proof is completed by noting from Niewiadomska-Bugaj and Kowalczyk (2005) that $\mathbb{E}[\bar{F}(X)] = \frac{1}{2}$ and $\mathrm{var}[\bar{F}(X)] = \frac{1}{12}(1-\gamma_F)$, whence (33) reduces to

$$\sigma^2 = \frac{9}{(1-\gamma_G)^2}\,\mathbb{E}[K^{(1)}(X,Y)^2] = \frac{9\cdot 16}{(1-\gamma_G)^2}\,\mathrm{var}[\bar{F}(X)]\,\mathrm{var}[\bar{G}(Y)] = \frac{1-\gamma_F}{1-\gamma_G},$$

which is the asymptotic variance at (34). □

References

Altman, D. G. and Royston, P. (2006). The cost of dichotomising continuous variables. *British Medical Journal*, 332:1080.

Azadkia, M. and Roudaki, P. (2025). A new measure of dependence: Integrated R^2. Preprint, https://arxiv.org/abs/2505.18146.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415.

Ben Bouallègue, Z., Clare, M. C., Magnusson, L., Gascon, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F. (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, 105:E864–E883.

Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20:1006–1028.

Best, D. and Roberts, D. (1975). Algorithm AS 89: The upper tail probabilities of Spearman's rho. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 24:377–379.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2023). Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619:533–538.

Borkowf, C. B. (2002). Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation. *Computational Statistics & Data Analysis*, 39:271–286.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159.

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, 116:2009–2022.

Chetverikov, D. and Wilhelm, D. (2023). Inference for rank-rank regressions. Preprint, https://arxiv.org/abs/2310.15512.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845.

Demler, O. V., Pencina, M. J., Cook, N. R., and D'Agostino Sr, R. B. (2017). Asymptotic distribution of ΔAUC, NRIs, and IDI based on theory of U-statistics. *Statistics in Medicine*, 36:3334–3360.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Diedenhofen, B. and Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, 10:e0121945.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. In Dempster, M. A. H., editor, *Risk Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874.

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

Fissler, T. and Pohle, M.-O. (2023). Generalised covariances and correlations. Preprint, https://arxiv.org/abs/2307.03594.

Gaißer, S. and Schmid, F. (2010). On testing equality of pairwise rank correlations in a multivariate random vector. *Journal of Multivariate Analysis*, 101:2598–2615.

Geenens, G. and Lafaye de Micheaux, P. (2022). The Hellinger correlation. *Journal of the American Statistical Association*, 117:639–653.

Genest, C., Nešlehová, J. G., and Rémillard, B. (2013). On the estimation of Spearman's rho and related tests of independence for possibly discontinuous multivariate data. *Journal of Multivariate Analysis*, 117:214–228.

Giudici, P. and Raffinetti, E. (2025). RGA: A unified measure of predictive accuracy. *Advances in Data Analysis and Classification*, 19:67–93.

Gneiting, T., Biegert, T., Kraus, K., Walz, E.-M., Jordan, A. I., and Lerch, S. (2025). Probabilistic measures afford fair comparisons of AIWP and NWP model output. Preprint, https://doi.org/10.48550/arXiv.2506.03744.

Gneiting, T. and Vogel, P. (2022). Receiver operating characteristic (ROC) curves: Equivalences, beta model, and minimum distance estimation. *Machine Learning*, 111:2147–2159.

Gneiting, T. and Walz, E.-M. (2022). Receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and coefficient of predictive ability (CPA). *Machine Learning*, 111:2769–2797.

Harrell Jr, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146:1999–2049.

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

Huang, X., Li, S., Yu, M., Sesia, M., Hassani, H., Lee, I., Bastani, O., and Dobriban, E. (2024). Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*. Association for Computational Linguistics.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30:81–93.

Kimeldorf, G. and Sampson, A. R. (1978). Monotone dependence. *Annals of Statistics*, 6:895–903.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53:814–861.

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Eleventh International Conference on Learning Representations*.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17:1217–1241.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382:1416–1421.

Lancaster, H. (1963). Correlation and complete dependence of random variables. *Annals of Mathematical Statistics*, 34:1315–1321.

Lavers, D. A., Simmons, A., Vamborg, F., and Rodwell, M. J. (2022). An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, 148:3152–3165.

Lee, K., Chang, M.-W., and Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Lehmann, E. L. and Romano, J. P. (2022). *Testing Statistical Hypotheses*. Springer, 4th edition.

Lin, Z., Trivedi, S., and Sun, J. (2024). Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Meng, X.-L., Rosenthal, R., and Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, 111:172–175.

Myers, L. and Sirois, M. J. (2014). Spearman correlation coefficients, differences between. In *Wiley StatsRef: Statistics Reference Online*. Wiley Online Library, https://doi.org/10.1002/9781118445112.stat02802.

Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis*, 98:544–567.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal*, 2(1):45–64.

Niewiadomska-Bugaj, M. and Kowalczyk, T. (2005). On grade transformation and its implications for copulas. *Brazilian Journal of Probability and Statistics*, 19:125–137.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2021). Training language models to follow instructions with human feedback. In *35th Conference on Neural Information Processing Systems*.

Pencina, M. J. and D'Agostino, R. B. (2004). Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Statistics in Medicine*, 23:2109–2123.

Pohle, M.-O. and Wermuth, J.-L. (2025). Proper correlation coefficients for discrete random variables. Unpublished manuscript.

Pohle, M.-O., Wermuth, J.-L., and Weiß, C. H. (2025). Statistical inference for rank correlations. Unpublished manuscript.

Python Software Foundation (2025). Python language reference. Available at http://www.python.org.

R Core Team (2025). R: A language and environment for statistical computing. Available at http://www.r-project.org.

Radford, J. T., Ebert-Uphoff, I., and Stewart, J. Q. (2025). A comparison of AI weather prediction and numerical weather prediction models for 1–7-day precipitation forecasts. *Weather and Forecasting*, 40:561–575.

Raffinetti, E. (2023). A rank graduation accuracy measure to mitigate artificial intelligence risks. *Quality & Quantity: International Journal of Methodology*, 57:131–150.

Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14:6086.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. Preprint, https://doi.org/10.48550/arXiv.1606.05250.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12:e2020MS002203.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallègue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F. (2024). WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16:e2023MS004019.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10:441–451.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334:1518–1524.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:1–8.

Rodwell, M. J., Richardson, D. S., Hewson, T. D., and Haiden, T. (2010). A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 136:1344–1363.

Scarsini, M. (1984). On measures of concordance. *Stochastica*, 8:201–218.

Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics*, 9:879–885.

Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., and Majumdar, A. (2025). A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):63.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27:799–811.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

Sun, X. and Xu, W. (2014). Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21:1389–1393.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35:2769–2794.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. Preprint, https://doi.org/10.48550/arXiv.2307.09288.

Weihs, L., Drton, M., and Meinshausen, N. (2018). Symmetric rank covariances: A generalized framework for nonparametric measures of dependence. *Biometrika*, 105:547–562.

Woodbury, M. A. (1940). Rank correlation when there are equal variates. *Annals of Mathematical Statistics*, 11:358–362.

Xu, J. (2024). On the bias in the AUC variance estimate. *Pattern Recognition Letters*, 178:62–68.