

Localizing Strictly Proper Scoring Rules*

Ramon F. A. de Punder
Department of Quantitative Economics
University of Amsterdam and Tinbergen Institute

Cees G. H. Diks[†]
Department of Quantitative Economics
University of Amsterdam and Tinbergen Institute

Roger J. A. Laeven
Department of Quantitative Economics
University of Amsterdam, CentER and EURANDOM

Dick J. C. van Dijk
Department of Econometrics
Erasmus University Rotterdam and Tinbergen Institute

October 10, 2025

*We are very grateful to the Editor, Associate Editor and four referees for their comments and suggestions, which have significantly improved the paper. We are also grateful to Yacine Ait-Sahalia, Timo Dimitriadis, Tilmann Gneiting, Andrew Harvey, Alexander Jordan, Frank Kleibergen, Siem Jan Koopman, Rutger-Jan Lange, Sebastian Lerch, Xiaochun Meng, Marc-Oliver Pohle, Johannes Resin, Johanna Ziegel and participants at various seminars and conferences, including at the Heidelberg Institute for Theoretical Studies, Tinbergen Institute, University of Copenhagen, the 42nd International Symposium on Forecasting in Oxford (July 2022), the 10th International Workshop on Applied Probability in Thessaloniki (June 2023), the 5th Quantitative Finance and Financial Econometrics International Conference in Marseille (June 2023), the 12th ECB Conference on Forecasting Techniques in Frankfurt (June 2023), the 16th Meeting of the Netherlands Econometric Study Group in Rotterdam (June 2023), the International Association for Applied Econometrics Annual Conference in Oslo (June 2023), the 11th World Congress in Probability and Statistics in Bochum (August 2024) and the 76th European Meeting of the Econometric Society (August 2024), for their comments and suggestions. This research was supported in part by the Netherlands Organization for Scientific Research under grant NWO Vici 2020–2027 (Laeven). The authors report there are no competing interests to declare.

[†]Corresponding author. Mailing Address: PO Box 15867, 1001 NJ Amsterdam, The Netherlands. Phone: +31 (0) 20 525 4252. Email: C.G.H.Diks@uva.nl.

Abstract

When comparing predictive distributions, forecasters are typically not equally interested in all regions of the outcome space. To address the demand for focused forecast evaluation, we propose a procedure to transform strictly proper scoring rules into their localized counterparts while preserving the score divergence and strict propriety. This is accomplished by applying the original scoring rule to a censored distribution. Our procedure nests the censored likelihood score as a special case. Among a multitude of others, it also implies a class of censored kernel scores that offers a (possibly multivariate) alternative to the threshold weighted Continuously Ranked Probability Score (twCRPS), extending its local propriety to more general weight functions than single tail indicators. Within this localized framework, we obtain a generalization of the Neyman Pearson lemma, establishing the censored likelihood ratio test as uniformly most powerful. For other tests of localized equal predictive performance, results of Monte Carlo simulations and empirical applications to risk management, inflation and climate data consistently emphasize the excellent power properties of censoring versus other localization methods.

Keywords: Density forecast evaluation; Tests for equal predictive ability; Censoring; Likelihood ratio; CRPS.

1 INTRODUCTION

Over the past decades, probabilistic forecasts have garnered increasing attention across a variety of disciplines, primarily because they provide a more comprehensive understanding of the stochastic nature of a random variable under scrutiny than point forecasts (Dawid 1984). A cornerstone for the effective evaluation of such probabilistic forecasts is the use of strictly proper scoring rules (Gneiting and Raftery 2007; Brehmer and Gneiting 2020; Patton 2020), which have been widely advocated for their ability to ensure fair comparative assessments of different forecast methods. Scoring rules are inherently connected with divergence measures; under the restriction of strict propriety, these measures are subsumed under Bregman divergences (Dawid 2007; Ovcharov 2018; Painsky and Wornell 2020). While the usefulness of unweighted probabilistic forecasting is well-recognized and well-understood, various applications, such as the analysis of large financial portfolio losses, inflation targets or temperature ranges, require a focused, localized evaluation of predictive distributions.

In this paper, we introduce a natural localization mechanism for strictly proper scoring rules that preserves the score divergence and strict propriety. By censoring (Bernoulli 1760; Tobin 1958) the observation and distribution before applying the original scoring rule, we find a sweet spot between retaining and discarding information when focusing on a region of interest. Crucially, unlike existing approaches that employ conditional distributions, our method preserves the overall probability of receiving an observation in (or outside) the target region, obviously relevant when comparing various candidate distributions focused on the same area. Moreover, within the region of interest, our mechanism maintains the original distribution’s shape. This is particularly beneficial when evaluating functionals in this region, such as quantiles or conditional expectations. Our procedure can be used to generate a multitude of strictly locally proper scoring rules. These include the censored likelihood (CSL) score, proposed by Diks et al. (2011), and the threshold weighted Continuously Ranked Probability Score (twCRPS), put forward by Gneiting and Ranjan (2011), for weight functions for which Holzmann and Klar (2017a) have shown that the twCRPS is strictly locally proper. On the other hand, for weight functions for which the twCRPS is not strictly locally proper, our analysis provides a strictly locally proper alternative.

The information retained by our censoring approach translates into advantageous power properties of tests aimed at comparing density forecasts in regions of interest. We prove a generalization of the Neyman Pearson (1933) lemma, revealing that the censored likelihood ratio leads to a Uniformly Most Powerful (UMP) test. By contrast, we provide explicit evidence that the conditional likelihood (CL) score does not admit a UMP test. Monte Carlo simulations and empirical applications analyze the power properties of the Diebold and Mariano (2002) (DM) type test statistic, within the framework of Giacomini and White (2006), based on censored vis-à-vis alternative localized scoring rules. Censored scoring rules have competitive power properties in all Monte Carlo experiments conducted. In multiple empirical experiments, involving financial, macroeconomic and climate data,

we utilize the DM tests in the Model Confidence Set (MCS) procedure of Hansen et al. (2011). The MCSs resulting from censored scoring rules are typically smaller than those arising from alternative localization procedures, broadly aligning with the power properties displayed by the Monte Carlo results.

Our research contributes to the literature on focused scoring rules, initiated by the weighted likelihood score (WLS) of Amisano and Giacomini (2007). Diks et al. (2011) and Gneiting and Ranjan (2011) sought to correct the (regular) impropriety of this scoring rule by introducing the CL, CSL and twCRPS, respectively. Holzmann and Klar (2017a) substantially advanced focused scoring rules, using conditioning to construct proportionally locally proper scoring rules from unweighted scoring rules other than the logarithmic score. They also showed that strict local propriety of the ensuing scoring rules can be restored by adding an auxiliary weighted scoring rule, based on an arbitrary strictly proper rule for the probability of an observation landing in the region of interest. Our work differs importantly by opting for censoring rather than conditioning as localization mechanism. Through censoring, we enable the direct application of the original scoring rule to the localized measure, thereby avoiding the need for an auxiliary scoring rule and preserving the original Bregman divergence. As detailed by Brehmer and Gneiting (2020, Theorem 1), the conditional scoring rules of Holzmann and Klar (2017a) can also be viewed as an extension of the WLS refined through a ‘properization’ process. Consequently, properization is not a viable mechanism for retaining strict propriety of the original scoring rule. Recently, Allen et al. (2023) introduced a framework to create strictly locally proper scoring rules from the class of kernel scores. Furthermore, Mitchell and Weale (2023) proposed using censored density forecasts, to perform statistical inference based on a central region of the forecast distribution; they do *not* aim to evaluate competing candidate densities. We provide a detailed comparison of our censoring approach with these alternative localization procedures in Section 3.5.

Our research also rests upon a substantial body of research concerning unweighted strictly proper scoring rules and their associated divergence measures. Although the formalization of strict propriety was rigorously achieved by Gneiting and Raftery (2007), scoring rules satisfying this property date at least to the Quadratic Scoring rule of Brier (1950). The literature in this domain has evolved from an initial focus on discrete settings to a more general treatment. In this vein, we rely on the expanded frameworks of the Power (PowS_α) and PseudoSpherical (PsSphS_α) families as advocated by Gneiting and Raftery (2007) and Ovcharov (2018) rather than their discrete foundations.

Interest in targeting specific regions of predictive distributions has surged across diverse fields, including meteorology, climatology, hydrology, finance, and economics. In financial risk management, attention is particularly concentrated on the left tail of return distributions, according to mandated risk measures such as Value-at-Risk and Expected Shortfall (Cont et al. 2010; Fissler et al. 2015). Analogously, in macroeconomics, ‘Growth-at-Risk’ and ‘Inflation-at-Risk’ are emerging concepts, signifying values that deviate significantly from benchmarks established by institutions such as Central Banks (Adrian et al. 2019; Lopez-Salido and Loria 2020; Iacopini et al. 2023). In other scenarios, the emphasis might rest on the central region or on another specific region of the distribution, often dictated by external constraints or objectives. Examples range from optimizing growing conditions for specific crops such as tubers, to calibrating wind speeds for peak wind turbine performance, and regulating blood sugar levels for effective diabetes management. All these applications require region-specific performance evaluations aligned with the interest in particular outcomes. Accordingly, as illustrated by Lerch et al. (2017), it is crucial to distinguish between strict propriety and strict local propriety; failing to do so can result in misleading results.

The paper is organized as follows. Section 2 provides the foundational concepts. Section 3 introduces the censored scoring rule and establishes its strict local propriety. This section also contains guidance for the practical use of the censoring procedure, a general-

ization of the Neyman Pearson lemma, and a comparison with alternative weighted scoring rules. Section 4 discusses the empirical performance of our approach. Section 5 concludes. Proofs, derivations of theoretical properties, results of the Monte Carlo study, and additional empirical results are provided in the accompanying Supplementary Material. Data and codes are available from the paper’s GitHub: <https://github.com/rdpunder/LSPS/>.

2 SCORING RULES AND DIVERGENCES

2.1 Unweighted scoring rules and score divergences

Consider a random variable $Y : \Omega \rightarrow \mathcal{Y}$ from a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\mathcal{Y}, \mathcal{G})$. Denote by \mathcal{P} a convex class of probability distributions on $(\mathcal{Y}, \mathcal{G})$. A *scoring rule* S assigns numerical values (scores) to observations $y \in \mathcal{Y}$ and distributions $F \in \mathcal{P}$, through a mapping $S : \mathcal{P} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty\} =: \bar{\mathbb{R}}$. Following Holzmann and Klar (2017a), we assume that S is measurable w.r.t. \mathcal{G} and quasi-integrable w.r.t. all $P \in \mathcal{P}$, for all $F \in \mathcal{P}$, and such that $\mathbb{E}_P S(F, Y) < \infty$ and $\mathbb{E}_P S(P, Y) \in \mathbb{R}$. The latter condition guarantees that the *score divergence*, $\mathbb{D}_S(P||F) := \mathbb{E}_P S(P, Y) - \mathbb{E}_P S(F, Y)$, exists and maps onto $(-\infty, \infty]$. Adhering to Gneiting and Raftery (2007), a minimal requirement for S is that it is *strictly proper*.

Definition 1 (Strictly proper scoring rule). *A scoring rule $S : \mathcal{P} \times \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ is proper relative to \mathcal{P} if $\mathbb{D}_S(P||F) \geq 0, \forall P, F \in \mathcal{P}$, and strictly proper if, additionally, $\mathbb{D}_S(P||F) = 0$ if and only if $P = F, \forall P, F \in \mathcal{P}$.*

Equivalently, a score divergence is a *divergence measure* (see, e.g., Eguchi, 1985) if and only if S is strictly proper; here, a divergence measure $\mathbb{D} : \mathcal{P} \times \mathcal{P} \rightarrow (-\infty, \infty]$ satisfies (i) $\mathbb{D}(P||F) \geq 0, \forall P, F \in \mathcal{P}$, and (ii) $\mathbb{D}(P||F) = 0$ if and only if $P = F, \forall P, F \in \mathcal{P}$, by definition. For distributions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where $\mathcal{B}(\mathcal{Y})$ denotes the Borel σ -algebra on \mathcal{Y} ,

such a score divergence is known to be a Bregman (1967) divergence (Ovcharov 2018). This excludes f -divergences other than the Kullback-Leibler divergence (Kullback and Leibler 1951). Two remarks are in place. First, distributions $F \in \mathcal{P}$ are compared in terms of their P -expected score differences, so that uniqueness of members in \mathcal{P} should formally be interpreted in terms of P -a.s. equivalence classes of P . Similarly, $P = F$ is formally defined as $P(E) = F(E)$, $\forall E \in \mathcal{G}$. For ease of exposition, we henceforth omit technicalities about P -a.s. equivalence. Second, if there exists a σ -finite measure μ such that $F \ll \mu$, $\forall F \in \mathcal{P}$, with \ll denoting absolute continuity, then scoring rules and associated definitions and results can easily be formulated relative to the class of induced μ -densities $f := \frac{dF}{d\mu}$, also denoted by \mathcal{P} , like classes of distribution functions F .

Gneiting and Raftery (2007) provide an extensive list of strictly proper scoring rules, which can be divided into *local* scoring rules and *distance-sensitive* scoring rules (Ehm and Gneiting 2012). We use the same distinction when discussing examples, yet allowing local scoring rules to also depend on the density via a global norm of the density, and refer to these henceforth as *semi-local*. In this subcategory, our focus lies on the Logarithmic (LogS), Quadratic (QS) and Spherical (SphS) scoring rules, along with their extensions to the Power (PowS $_{\alpha}$) and PseudoSpherical (PsSphS $_{\alpha}$) families; see Table 1. For distance-sensitive scoring rules we confine ourselves to the rich class of kernel scores. This class has been shown to be strictly proper under known conditions (Gneiting and Raftery 2007; Steinwart and Ziegel 2021), nesting the multivariate Energy Score family, which in turn includes the univariate Continuously Ranked Probability Score (CRPS) as a special case.

2.2 Minimal localization and censoring

In this paper, we suppose the application at hand introduces a region of interest $A \subseteq \mathcal{Y}$, which is assumed to be measurable, i.e., $A \in \mathcal{G}$. Following Holzmann and Klar (2017a), we adopt the strict perspective that outcomes y in the complement $A^c \equiv \mathcal{Y} \setminus A$ are of

no interest. In terms of events $E \in \mathcal{G}$, we interpret focusing on A such that only the information generated by events intersecting with A are relevant. To formalize this, we consider the *smallest* σ -algebra on \mathcal{Y} containing all events intersecting with A : $\mathcal{G}_A^{\min} := \sigma(\{E \cap A : E \in \mathcal{G}\}) \subseteq \mathcal{G}$. Note that \mathcal{G}_A^{\min} also includes A^c since σ -algebras are closed under complements. We refer to the restriction of F to the minimal σ -algebra \mathcal{G}_A^{\min} , denoted $F|_{\mathcal{G}_A^{\min}}$, as the coarsest restriction or *minimal localization* of F to A .

Minimal localization naturally gives rise to *censoring*, formally via a pushforward measure of F , from \mathcal{G} to \mathcal{G}_A^{\min} . Censoring (Bernoulli 1760) refers to the statistical concept used to model a variable under scrutiny whose value, upon measurement or observation, is only partially known (Tobin 1958). Under censoring, for realizations of a random variable Y that occur in A^c , it is only known that they are not in A . Realizations in A^c are hence indistinguishable under censoring and ‘ A^c ’ may therefore be viewed as a single realization of the censored random variable. We censor all outcomes in A^c to a single abstract outcome ‘*’, which may be interpreted as ‘NaN’, uniquely identifying A^c . The importance of allowing * to be outside \mathcal{Y} becomes clear in Section 3.1. In Section 3.2, we discuss distance-sensitive scoring rules, which replace * by a suitable $y_0 \in \mathcal{Y}$. Formally, the *censored random variable* $Y_A^b : \mathcal{Y} \rightarrow \mathcal{Y}_A^b$ is defined as the $\mathcal{G}/\mathcal{G}_A^b$ -measurable function

$$Y_A^b \equiv Y_A^b(y) := \begin{cases} y, & y \in A, \\ *, & y \in A^c, \end{cases} \quad (1)$$

with $\mathcal{G}_A^b := \sigma(\{E \cap A : E \in \mathcal{G}\} \cup \{*\})$ on $\mathcal{Y}_A^b := A \cup \{*\}$. The distribution F_A^b of Y_A^b is referred to as the *censored distribution* of F . It is the pushforward measure of F by Y_A^b , equal to $F|_{\mathcal{G}_A^{\min}}$ up to relabeling the event $A^c \in \mathcal{G}_A^{\min}$ by $\{*\} \in \mathcal{G}_A^b$. In concise form,

$$F_A^b := F|_{\mathcal{G}_A^b} = F_A + \bar{F}_A \delta_*, \quad (2)$$

where $F_A(E) := F(A \cap E)$, $\forall E \in \mathcal{G}^* := \sigma(\{\mathcal{G}, *\})$, $\bar{F}_A := F(A^c)$ and δ_* denotes the Dirac measure at *, i.e., $\delta_*(E) = \mathbb{1}_E(*)$, $\forall E \in \mathcal{G}$. The indicator function $\mathbb{1}_A(y)$ equals unity if

$y \in A$ and zero otherwise. An illustration of minimal localization is given in Example 1.

Example 1 (Minimal localization). *Jim draws a ball from a vase containing six balls, one of which is silver, two are orange, and three are blue. He wins a car if, and only if, he draws the silver ball. This game can be described by a random variable Y , on the outcome space $\mathcal{Y} = \{s, o, b\}$ with power set σ -algebra $\mathcal{G} = \{\emptyset, \{s\}, \{o\}, \{b\}, \{s, o\}, \{s, b\}, \{o, b\}, \{s, o, b\}\}$, having probability mass function (pmf) $p(y) = \frac{1}{6}\mathbb{1}_{\{s\}}(y) + \frac{1}{3}\mathbb{1}_{\{o\}}(y) + \frac{1}{2}\mathbb{1}_{\{b\}}(y)$, $y \in \mathcal{Y}$. Suppose Jim only cares about whether he wins the car or not (and not how he loses). Thus, his region of interest is $A = \{s\}$, which induces $\mathcal{Y}_A^b = \{s, *\}$, where $\{*\}$ corresponds to $\{o, b\}$, and $\mathcal{G}_A^b = \sigma(\{s, *\}) = \{\emptyset, \{s\}, \{*\}, \{s, *\}\}$, corresponding to $\sigma(\{s\})$ on \mathcal{Y} . The pmf $p(y)$ localizes to $p_A^b(y) = \frac{1}{6}\mathbb{1}_{\{s\}}(y) + \frac{5}{6}\mathbb{1}_{\{*\}}(y)$, $y \in \{s, *\}$.*

2.3 Local and localized divergences and weighted scoring rules

Example 2 (The need to focus). *Continuing Example 1, suppose Jim and his friend Pam know that the vase contains six balls colored silver, orange, and blue, but do not know their exact numbers. Jim suspects one silver, four orange, and one blue ball, while Pam believes there are two silver, one orange, and three blue balls. Let Jim's and Pam's implied pmfs be f and g . One finds $\text{KL}(p\|f) - \text{KL}(p\|g) = \frac{1}{2}\log(\frac{3}{2}) > 0$, where $\text{KL}(p\|f) := \mathbb{E}_p(\log p(Y) - \log f(Y))$ denotes the Kullback-Leibler (KL) divergence from p to f . Hence, Pam's belief is statistically closer to the truth in absence of a region of interest. However, since Jim only cares about winning the car, Pam's accurate belief of the blue balls' count, with a relatively high true probability $p(\{b\}) = 1/2$, is irrelevant and even misleading. Her close fit outside the silver outcome obscures her inaccuracy where Jim is correct. This illustrates the need to localize the KL divergence to align with Jim's focus on winning.*

As demonstrated by Example 2, it is imperative to adapt the divergence when particular outcomes are of importance. Otherwise, an excellent fit in non-critical regions of the

outcome space may obscure a poor fit in regions of relevance. We describe the relative importance of outcomes $y \in \mathcal{Y}$ by a *weight function* $w \in \mathcal{W}$, where \mathcal{W} is a set of \mathcal{G} -measurable mappings $w : \mathcal{Y} \rightarrow [0, 1]$. Then the question arises how to accordingly transform the divergence and the scoring rule.

Censoring as defined in Section 2.2 pertains to the weight function $w(y) = \mathbb{1}_A(y)$. To simplify the notation, we often use the subscript A in place of $\mathbb{1}_A$ when referring to indicator functions. Censoring transforms the class of distributions from \mathcal{P} to \mathcal{P}_A^b where $\mathcal{P}_A^b := \{F_A^b, F \in \mathcal{P}\}$. As we will see in the next section, censoring induces a divergence on $\mathcal{P}_A^b \times \mathcal{P}_A^b$ that equals 0 if and only if $P_A^b = F_A^b$. This, in turn, is equivalent to the measures coinciding (only) locally on A , i.e., $P(A \cap E) = F(A \cap E)$, $\forall E \in \mathcal{G}$, for which we introduce the short-hand notation $P \stackrel{A}{=} F$.

Let us consider a general divergence measure \mathbb{D} , i.e., not necessarily a *score* divergence. In Definitions 2 and 3 below, we introduce a *local divergence* and (the more specific) *localized divergence*. Both definitions are given for general weight functions, corresponding to the region of interest A_w defined as

$$A_w := \{y \in \mathcal{Y} : w(y) > 0\}.$$

On A_w , we again use the short-hand notation $F \stackrel{A_w}{=} G$ for $F(A_w \cap E) = G(A_w \cap E)$, $\forall E \in \mathcal{G}$. Censoring will yield not just a local divergence but even a localized divergence.

Definition 2 (Local divergence). *A map $\mathbb{D}_w : \mathcal{P} \times \mathcal{P} \rightarrow (-\infty, \infty]$ is called a local divergence (w.r.t. A_w) if (i) $F \stackrel{A_w}{=} G$ implies $\mathbb{D}_w(P||F) = \mathbb{D}_w(P||G)$, $\forall P, F, G \in \mathcal{P}$, (ii) $\mathbb{D}_w(P||F) \geq 0$, $\forall P, F \in \mathcal{P}$, and (iii) $\mathbb{D}_w(P||F) = 0$ if and only if $P \stackrel{A_w}{=} F$, $\forall P, F \in \mathcal{P}$.*

Definition 3 (Localized divergence). *Let \mathcal{P}_w denote a class of distributions obtained by a map $[\cdot]_w : \mathcal{P} \rightarrow \mathcal{P}_w$ coinciding with the identity map $[F]_w = F$, $\forall F \in \mathcal{P}$ for $w = \mathbb{1}_{\mathcal{Y}}$. A local divergence $\mathbb{D}_w : \mathcal{P} \times \mathcal{P} \rightarrow (-\infty, \infty]$ is called a localized divergence of \mathbb{D} (w.r.t. A_w)*

if $\forall P_w, F_w \in \mathcal{P}_w, \exists P, F \in \mathcal{P}$:

$$\mathbb{D}_w(P\|F) = \mathbb{D}(P_w\|F_w).$$

Condition (i) in Definition 2 ensures invariance w.r.t. (information generated by the) events that are not intersecting with A_w , hence are irrelevant. Furthermore, condition (iii) applies only locally on A_w . Definition 3 introduces the subclass of local divergences that preserve the unweighted divergence measure \mathbb{D} by applying it to a weighted transformation of the distribution space.

Just as strictly proper scoring rules give rise to divergence measures, local divergences emerge naturally from *weighted scoring rules* that are *strictly locally proper* relative to some class of distributions \mathcal{P} and weight functions \mathcal{W} . For all $w \in \mathcal{W}$, we define a weighted scoring rule as the map $S_w : \mathcal{P} \times \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ such that $S_w(\cdot, \cdot)$ is a scoring rule. A weighted scoring rule is said to be *localizing* if measures coinciding on A_w receive the same score for any realization; formally, $\forall P, F \in \mathcal{P}, S_w(P, y) = S_w(F, y), \forall y \in \mathcal{Y}$, whenever $P \stackrel{A_w}{=} F$. Definition 4 extends strict propriety to localizing weighted scoring rules and is equivalent to that given by Holzmann and Klar (2017a, p. 2414). Here, $\mathbb{D}_{S_w}(P\|F) := \mathbb{E}_P(S_w(P, Y)) - \mathbb{E}_P(S_w(F, Y))$.

Definition 4 (Strictly locally proper scoring rule). *A weighted scoring rule $S_w : \mathcal{P} \times \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ is locally proper relative to $(\mathcal{P}, \mathcal{W})$ if it is localizing and $S_w(\cdot, \cdot)$ is proper for all $w \in \mathcal{W}$. Furthermore, it is strictly locally proper relative to $(\mathcal{P}, \mathcal{W})$ if, additionally, $P \stackrel{A_w}{=} F$ if and only if $\mathbb{D}_{S_w}(P\|F) = 0, \forall w \in \mathcal{W}$.*

Clearly, the score divergence $\mathbb{D}_{S_w}(P\|F)$ is a local divergence for all $w \in \mathcal{W}$ if and only if S_w is strictly locally proper relative to $(\mathcal{P}, \mathcal{W})$. As a general note, we remark that focusing on particular (subsets of) outcomes using a strictly locally proper scoring rule represents a strict modeling perspective; possibly non-localizing scoring rules, such as the twCRPS, can have their own merits in specific applications, by taking additional information into

account; see also Example D.3 in the Appendix and its discussion. As such, non-localizing scoring rules may statistically be competitive to strictly locally proper scoring rules as evidenced by our Monte Carlo simulations.

3 THE CENSORED SCORING RULE

In this section, we introduce the censored scoring rule. To this end, we first generalize the censored distribution in Equation (2) to general weight functions, and next to arbitrary nuisance distributions that suitably replace the Dirac measure in Equation (3) below, to accommodate distance-sensitive scoring rules.

For the definition of the censored distribution, it is helpful to first extend w and F to w^* and F^* on $(\mathcal{Y}^*, \mathcal{G}^*)$, where $\mathcal{Y}^* := \mathcal{Y} \cup \{*\}$, such that $w^*(y) = w(y)\mathbb{1}_{\mathcal{Y}}(y)$, $\forall y \in \mathcal{Y}^*$ and $F^*(E) = F(E \cap \mathcal{Y})$, $\forall E \in \mathcal{G}^*$. In the original measurable space, $*$ is then already an outcome, associated with the event $\{*\}$, albeit assigned with weight and measure zero. For ease of notation, we henceforth drop the superscripts $*$. The censored distribution in Equation (2) generalizes to

$$dF_w^b := dF_w + \bar{F}_w d\delta_*, \quad \text{where} \quad dF_w := w dF, \quad \bar{F}_w := 1 - \int_{\mathcal{Y}} w dF, \quad (3)$$

defining a probability measure on $(\mathcal{Y}_{A_w}^b, \mathcal{G}_{A_w}^b)$, with $\mathcal{G}_{A_w}^b := \sigma(\{E \cap A_w : E \in \mathcal{G}\} \cup \{*\})$ and $\mathcal{Y}_{A_w}^b := A_w \cup \{*\}$. In contrast to indicator functions, the probability of the censoring event \bar{F}_w is generally different from $F(A_w^c) = 1 - \int_{A_w} dF$, rendering the identification $Y_{A_w}^b(y) = * \iff y \in A_w^c$ infeasible. However, Appendix C shows that identification is recoverable by introducing an auxiliary random variable $Z|(Y = y)$, which is $B_{w(y)} \equiv \text{Bernoulli}(w(y))$ -distributed, to define $Y_w^b \equiv Y_w^b(y, z)$ as being y if $z = 1$ and $*$ otherwise, since then $Y_w^b(y, z) = * \iff z = 0$. The distribution F_w^b admits the $(\mu + \delta_*)$ -density $f_w^b(y) = w(y)f(y)\mathbb{1}_{A_w}(y) + \bar{F}_w\mathbb{1}_{\{*\}}(y)$, $y \in \mathcal{Y}_{A_w}^b$, provided that $F \ll \mu$; see Appendix B.1 for details.

For indicator weight functions, the density simplifies to $f_A^b(y) = f(y)\mathbf{1}_A(y) + \bar{F}_A\mathbf{1}_{A^c}(y)$, $y \in \mathcal{Y}$, similar to Borowska et al. (2020).

A critical difference from the (generalized) conditional distribution

$$dF_w^\sharp := \frac{1}{1 - \bar{F}_w} dF_w,$$

assuming $\bar{F}_w < 1$, is that $F_A^\sharp(A) \neq F(A) = F_A^b(A)$. In other words, conditioning does not preserve all probabilities of interest and does accordingly not coincide with the minimal localization of F to A . The symbols ‘sharp’ (\sharp) and ‘flat’ (b) reflect their respective operations: conditioning sharpens the density on A by a factor $1/(1 - \bar{F}_A)$, whereas censoring flattens the shape outside A into a point mass. The associated scoring rule $S_w^\sharp(F, y) := w(y)S(F_w^\sharp, y)$ fails to be strictly locally proper. Holzmann and Klar (2017a) remedy this by adding an auxiliary scoring rule for the missing information about A^c . However, by this addition, the corresponding score divergence generally fails to be a localized divergence; see Section 3.5.

3.1 Censored scoring

Ideally, the censored scoring rule would be given by the identity $S_A^b(F, y) = S(F_A^b, y_A^b)$, as this would fully respect the forecaster’s specific choice of the unweighted scoring rule S . The censored scoring rule given by Definition 5 below indeed reduces to this definition for the indicator weight function $w(y) = \mathbf{1}_A(y)$. It is also attractive for general weight functions, for which the randomization perspective based on the auxiliary random variable Z in Appendix C yields the similar identity $S_w^b(F, y) = \mathbb{E}_{B_{w(y)}} S(F_w^b, Y_w^b(y, Z))$.

Definition 5 (Censored scoring rule). *Let $S : \mathcal{P}^b \times \mathcal{Y} \rightarrow \bar{\mathbb{R}}$, $\mathcal{P}^b = \{F_w^b, F \in \mathcal{P}, w \in \mathcal{W}\}$, denote a scoring rule. Then, for all $w \in \mathcal{W}$, the corresponding censored scoring rule is given by the map $S_w^b : \mathcal{P} \times \mathcal{Y} \rightarrow \bar{\mathbb{R}}$,*

$$S_w^b(F, y) := w(y)S(F_w^b, y) + (1 - w(y))S(F_w^b, *),$$

where the censored distribution F_w^b is given in Equation (3).

Theorem 1 establishes that the censored scoring rule is strictly locally proper.

Theorem 1. *If the scoring rule S is strictly proper relative to \mathcal{P}^b , the censored scoring rule S_w^b in Definition 5 is strictly locally proper relative to $(\mathcal{P}, \mathcal{W})$. Moreover, its associated score divergence $\mathbb{D}_{S_w^b}$ is a localized divergence of \mathbb{D}_S , for all $w \in \mathcal{W}$.*

Theorem 1 is a special case of the more general Theorem 2 below, hence its proof is subsumed in the proof of Theorem 2. The assumption required in Theorem 1 ensures that the unweighted scoring rule is well-defined w.r.t. mixed continuous-discrete distributions on measurable spaces extended by $*$. For the PsSphS $_\alpha$ family, this is explicitly verified in Example 3; the same argument applies to all semi-local scoring rules considered, *mutatis mutandis*.

Let us provide some intuition for the result of Theorem 1, relying on the established mathematical identity $\mathbb{D}_{S_w^b}(P||F) = \mathbb{D}_S(P_w^b||F_w^b)$. Since S is assumed to be strictly proper relative to \mathcal{P}^b , its score divergence \mathbb{D}_S defines a divergence on $\mathcal{P}_w^b \times \mathcal{P}_w^b$, for all $w \in \mathcal{W}$. However, when viewed as a composite map, first censoring and then computing the divergence, it is no longer a divergence on the space of uncensored measures, as the censoring map is generally non-invertible. Nonetheless, by maintaining the probability of the censoring event, censoring preserves identifiability on A_w , i.e., $P \stackrel{A_w}{=} F$ if and only if $P_w^b = F_w^b$. Hence, under censoring, the strict propriety of the scoring rule and the associated divergence are retained locally. For the Logarithmic scoring rule, the censored scoring rule coincides with that of Diks et al. (2011).

Example 3 (Censored PsSphS). *Consider a class of μ -densities \mathcal{P}_α on $(\mathcal{Y}, \mathcal{G}, \mu)$ with finite L^α -norm, i.e., $\|f\|_\alpha := (\int_{\mathcal{Y}} f^\alpha d\mu)^{1/\alpha} < \infty, \forall f \in \mathcal{P}_\alpha$. The PseudoSpherical family $\text{PsSphS}_\alpha(f, y) = f(y)^{\alpha-1} / \|f\|_\alpha^{\alpha-1}$, $\alpha > 1$, as advocated by Gneiting and Raftery (2007), is strictly proper relative to \mathcal{P}_α . To verify its strict propriety relative to \mathcal{P}_α^b as required for*

Theorem 1, we write $\|f_w^b\|_\alpha^\alpha \leq 1 + \|f\|_\alpha^\alpha < \infty$, $\forall f \in \mathcal{P}_\alpha$, $\forall w \in \mathcal{W}$, hence $S_w^b(f, y)$ is strictly locally proper relative to $(\mathcal{P}, \mathcal{W})$. Notably, while $S_w^b(f, *) = \bar{F}_w^{\alpha-1}/(\|wf\|_\alpha + \bar{F}_w^\alpha)^{(\alpha-1)/\alpha}$ does not depend solely on \bar{F}_w , it holds that $S_w^b(f, *) = S_w^b(g, *)$, if $f = g$ a.s. on A_w .

In Theorem 1, we assume that the scoring rule S is well defined relative to a class of distributions \mathcal{P}^b with a point mass at $*$. Distance-sensitive scoring rules are generally incompatible with such measures, since the distance to $*$ is undefined. Section 3.2 shows that, in typical applications, $*$ can then be replaced by any $y_0 \in \mathcal{Y}$, provided that all $F \in \mathcal{P}$ assign zero mass to y_0 or all $w \in \mathcal{W}$ assign zero weight to y_0 . Then, $F_w(y_0) = 0$, which ensures identifiability of the censoring event. Here, we introduce the short-hand notation $F_w(y_0)$ to indicate the measure F_w of the event $\{y_0\}$. In Example 1, the above condition would prevent choosing $y_0 = s$, avoiding $(Y_A^b)^{-1}(s) = s \cup A^c = \mathcal{Y}$. When using semi-local scoring rules, labeling the censored outcome as $* \notin \mathcal{Y}$ ensures this identifiability by construction, even if all outcomes in \mathcal{Y} have positive mass under all distributions in \mathcal{P} and $w(y) > 0$ for all $y \in \mathcal{Y}$.

3.2 Generalized censored scoring

This subsection introduces a more flexible censoring framework. Suppose that a weight function introduces k pivotal points $r_1, \dots, r_k \in \mathcal{Y}$. Then a natural generalization of the censored distribution in Equation (3) reads

$$dF_{w, \mathcal{R}_k}^b := dF_w + \bar{F}_w \sum_{i=1}^k \gamma_i d\delta_{r_i}, \quad \gamma := (\gamma_1, \dots, \gamma_k)' \in \Delta(k), \quad (4)$$

where $\Delta(k)$ denotes the unit $(k-1)$ -simplex and $\mathcal{R}_k := \{r_i\}_{i=1}^k$, with $r_i \in \mathcal{Y}, \forall i$. Section 3.3 provides guidance on choosing (r_i, γ_i) . Definition 6 formalizes the adaptation of the censored scoring rule to generalized censored measures, nesting F_{w, \mathcal{R}_k}^b for $H = \sum_{i=1}^k \gamma_i \delta_{r_i}$. Here, we refer to H as a *nuisance* distribution since its sole role is to suitably allocate the probability mass \bar{F}_w .

Definition 6 (Generalized censored scoring rule). *Let $S : \mathcal{P}^b \times \mathcal{Y} \rightarrow \bar{\mathbb{R}}$, denote a scoring rule, where $\mathcal{P}^b = \{F_{w,H}^b, F \in \mathcal{P}, w \in \mathcal{W}, H \in \mathcal{H}\}$, in which $dF_{w,H}^b := dF_w + \bar{F}_w dH$ denotes the generalized censored distribution and $\mathcal{H} \subseteq \mathcal{P}$ a class of nuisance distributions. Then for all $w \in \mathcal{W}$ and $H \in \mathcal{H}$ the associated generalized censored scoring rule is given by the map $S_{w,H}^b : \mathcal{P} \times \mathcal{Y} \rightarrow \bar{\mathbb{R}}$,*

$$S_{w,H}^b(F, y) := w(y)S(F_{w,H}^b, y) + (1 - w(y))\mathbb{E}_H S(F_{w,H}^b, Q),$$

where H denotes the distribution of the random variable Q , distributed independently of y .

Since both F and H are defined relative to $(\mathcal{Y}, \mathcal{G})$, $F_{w,H}^b$ is defined relative to $(\mathcal{Y}, \mathcal{G})$ too, for all $w \in \mathcal{W}, H \in \mathcal{H}$. The dependence of \mathcal{P}^b on $(\mathcal{P}, \mathcal{W}, \mathcal{H})$ in Definition 6 is notationally suppressed.

Assumption 1. *The weight function $w \in \mathcal{W}$ and nuisance distribution $H \in \mathcal{H} \subseteq \mathcal{P}$ are such that $\exists E \in \mathcal{G} : F_w(E) = 0$ and $H(E) > 0, \forall F \in \mathcal{P}, H \in \mathcal{H}$.*

The following theorem, the proof of which is contained in Appendix A.1, establishes the strict local propriety of the generalized scoring rule.

Theorem 2. *Suppose that: (i) the unweighted scoring rule S in Definition 6 is strictly proper relative to \mathcal{P}^b , and (ii) \mathcal{W} and \mathcal{H} are such that Assumption 1 is satisfied. Then, the generalized censored scoring rule $S_{w,H}^b$ in Definition 6 is strictly locally proper relative to $(\mathcal{P}, \mathcal{W}, \mathcal{H})$. Moreover, its associated score divergence $\mathbb{D}_{S_{w,H}^b}$ is a localized divergence of \mathbb{D}_S , for all $w \in \mathcal{W}, H \in \mathcal{H}$.*

The intuition behind Theorem 2 builds on that of Theorem 1, by the extended identity

$$\mathbb{D}_{S_{w,H}^b}(P||F) = \mathbb{D}_S(P_{w,H}^b||F_{w,H}^b). \quad (5)$$

Assumption 1 implies that $P_{w,H}^b = F_{w,H}^b$ if and only if $P \stackrel{A_w}{=} F$. For any weight function, this holds trivially for $H = \delta_*$, while $H = \delta_{y_0}$, with $y_0 \in \mathcal{Y}$, requires the measure F to satisfy

$F_w(y_0) = 0$. For $H = \gamma\delta_{r_1} + (1 - \gamma)\delta_{r_2}$, $r_1, r_2 \in \mathcal{Y}$, and $\gamma \in [0, 1]$, Assumption 1 demands $F_w(r_i) = 0$ for at least one $i \in \{1, 2\}$; so, if both r_1 and r_2 belong to A_w , then at least one must not carry mass under F . This readily generalizes to finitely many points r_i . Indeed, Appendix C shows that by writing $S_{w,H}^b(F, y) = \mathbb{E}_{B_{w(y),H}}S(F_{w,H}^b, Y_w^b(y, Z, Q))$, identifiability of the censoring event under a single realization of Q suffices, since strict local propriety is retained if at least one scoring rule in a sum of proper scoring rules is strictly locally proper.

A rich class of scoring rules obtained by the generalization of the censored scoring rule is that of kernel scores (Gneiting and Raftery 2007). Kernel scores depend on a negative definite kernel $\rho : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$, i.e., a symmetric function satisfying $\sum_{i=1}^n \sum_{j=1}^n a_i a_j \rho(y_i, y_j) \leq 0$, for all $n \in \mathbb{N}$, $a_1, \dots, a_n \in \mathbb{R}$ that sum to 0 and all $y_1, \dots, y_n \in \mathcal{Y}$. These include popular examples such as the Euclidean distance on \mathbb{R}^d and the angular distance between two points on a circle. Example 4 displays the general expression for the generalized censored kernel score and corresponding localized divergence for $H = \delta_r$.

Example 4. Consider a class of distributions \mathcal{P}_r on some measurable space $(\mathcal{Y}, \mathcal{G})$, such that $F(r) = 0, \forall F \in \mathcal{P}_r$, where $r \in \mathcal{Y}$, including all (absolutely) continuous distributions on \mathcal{Y} . Consider the kernel score family $S_\rho(F, y) := \frac{1}{2}\mathbb{E}_F\rho(X, X') - \mathbb{E}_F\rho(X, y) + \frac{1}{2}\rho(y, y)$ with divergence $\mathbb{D}_{S_\rho}(P||F) = \mathbb{E}_{P,F}\rho(X, Y) - \frac{1}{2}\mathbb{E}_P\rho(X, X') - \frac{1}{2}\mathbb{E}_F\rho(Y, Y')$, nesting the CRPS(F, y) := $-\int_{\mathbb{R}} (F(s) - \mathbb{1}_{[y,\infty)}(s))^2 ds$ as a special case for $\rho(x, x') = |x - x'|$ on \mathbb{R} . The associated generalized censored scoring rule for $H = \delta_r$ reads $S_{\rho,w,r}^b(F, y) = \frac{1}{2}\mathbb{E}_{F_{w,r}^b}\rho(X, X') - w(y) \left(\mathbb{E}_{F_{w,r}^b}\rho(X, y) - \frac{1}{2}\rho(y, y) \right) - (1 - w(y)) \left(\mathbb{E}_{F_{w,r}^b}\rho(X, r) - \frac{1}{2}\rho(r, r) \right)$. Assumption 1 is clearly satisfied for all weight functions $w \in \mathcal{W}$ and distributions $F \in \mathcal{P}_r$. Therefore, the score divergence $\mathbb{D}_{S_{\rho,w,r}^b}(P||F) = \mathbb{E}_{P_{w,r}^b, F_{w,r}^b}\rho(X, Y) - \frac{1}{2}\mathbb{E}_{P_{w,r}^b}\rho(X, X') - \frac{1}{2}\mathbb{E}_{F_{w,r}^b}\rho(Y, Y')$ is a localized divergence if S_ρ is strictly proper relative to \mathcal{P}_r^b , which follows from the conditions under which S_ρ is strictly proper relative to \mathcal{P}_r . For one-sided indicator weight functions $\mathbb{1}_{(-\infty, r)}(y)$ and $\mathbb{1}_{(r, \infty)}(y)$, the censored CRPS coincides with the twCRPS(F, y) :=

$-\int_{\mathbb{R}} w(s)(F(s) - \mathbb{1}_{[y,\infty)}(s))^2 ds$ (Gneiting and Ranjan 2011). More details are given in Appendix E.4.

From Example 4, we have that $\text{CRPS}_{w,r}^b = \text{twCRPS}$ for all weight functions for which Holzmann and Klar (2017a, Theorem 5) proved that the twCRPS is strictly locally proper. For $A = (r_1, r_2)$, $A_1^c = (-\infty, r_1]$, $A_2^c = [r_2, \infty)$, $r_1, r_2 \in \mathbb{R}$, the twCRPS decomposes the mass \bar{F}_A into $F(A_1^c)$ and $F(A_2^c)$, hence would coincide with $F_{A,r_1,r_2}^b := F_A + \bar{F}_A(\gamma\delta_{r_1} + (1-\gamma)\delta_{r_2})$, $\gamma \in [0, 1]$, when $\gamma = F(A_1^c)/\bar{F}_A$. However, this choice of γ — which we do not allow — introduces a dependence on outcomes outside A , rendering the twCRPS non-localizing. For weight functions for which the twCRPS loses its strict local propriety due to its non-localizing nature, $\text{CRPS}_{w,\mathcal{R}_k}^b$ may serve as a strictly locally proper alternative.

3.3 Practical guidance

The censoring procedure proposed in this paper is sufficiently general to enable researchers who aim to compare competing forecasts to construct censored counterparts of their preferred scoring rules, for practically any chosen family of weight functions. With competing density forecasts at hand, the (generalized) censored scoring rules may be readily applied to obtain scores that can serve as input in a DM type test statistic; see Section 4. The GitHub page associated with this paper provides code for all 18 focused scoring rules considered in our simulations and empirical applications. In particular, for *semi-local scoring rules*, Definition 5 provides an analytic expression for the censored scoring rule for any choice of weight function. Table 1 lists the resulting (simplified) formulas for the unweighted, conditional and censored LogS, PowS $_{\alpha}$ and PsSphS $_{\alpha}$ families of scoring rules, as well as their localized divergences derived from the identity $\mathbb{D}_{S_w^b}(\mathbb{P}||\mathbb{F}) = \mathbb{D}_S(\mathbb{P}_w^b || \mathbb{F}_w^b)$. The table also displays the generalized censored scoring rules of Definition 6, which for semi-local scoring rules exhibit insensitivity to the nuisance distribution, as long as the nuisance distribution

is normalized to $\|h\|_\alpha = 1$.

For *distance-sensitive scoring rules*, such as the kernel scores discussed in Example 4, we use the generalized censoring approach in Definition 6 based on the censored measure in Equation (4). The localized versions of distance-sensitive scoring rules thereby depend on the choice of the pivotal points and associated weights. As illustrated by Examples D.1 and D.2 in Appendix D, weight functions often suggest natural choices for pivotal points, and it is these points we recommend incorporating into the censored measure in practice. Specifically, for the real-valued weight functions $I_L(y; r) := \mathbb{1}_{(-\infty, r)}(y)$, $I_R(y; r) := \mathbb{1}_{(r, \infty)}(y)$, $\Lambda_{a,L}(y; r) := \frac{1}{1 + \exp(a(y-r))}$, $a > 0$, the choice $r \in \mathbb{R}$ is considered pivotal. Similarly, for the weight functions on $\mathbf{y} \in \mathbb{R}^2$ given by $I_L^2(\mathbf{y}; \mathbf{r}) := I_L(y_1; r_1) \times I_L(y_2; r_2)$, $I_R^2(\mathbf{y}; \mathbf{r}) := I_R(y_1; r_1) \times I_R(y_2; r_2)$ and $\Lambda_{a,L}^2(\mathbf{y}; \mathbf{r}) := \Lambda_{a,L}(y_1; r_1) \times \Lambda_{a,L}(y_2; r_2)$, the use of $\mathbf{r} \in \mathbb{R}^2$ is considered natural. For the center indicator, $I_C(y; \ell, r) := \mathbb{1}_{(\ell-r, \ell+r)}(y)$, there are two pivotal points $r_{1,2} = \ell \pm r$. In Section 4, we use the fraction of observations smaller than r_1 to tune the weight γ in (4). For its complement, $I_C^c(y; \ell, r) := 1 - I_C(y; \ell, r)$, we adopt the single pivotal point ℓ .

As clarified in Section 3.2, Assumption 1 is easily satisfied by the censored distribution in Equation (4). For instance, if the underlying distributions are (absolutely) continuous, any value of r is valid; this includes Gaussian density and distribution functions. If the distribution is discrete or discrete-continuous, any r at which the distributions under consideration exhibit no point mass may be chosen.

3.4 Localized Neyman Pearson

Anticipating the applications in the next section, we now consider an explicit time-series context. Specifically, we consider a stochastic process $\{Y_t : \Omega \rightarrow \mathcal{Y}\}_{t=1}^T$ from a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\mathcal{Y}^T, \mathcal{G}^T)$, where \mathcal{Y}^T and \mathcal{G}^T denote the product outcome space and σ -algebra of the individual outcome spaces \mathcal{Y} and σ -algebras

Table 1: Examples of semi-local scoring rules.

Name	Logarithmic	Power family	PseudoSpherical family
Unweighted			
$S(f, y)$	$\text{LogS}(f, y) = \log f(y)$	$\text{PowS}_\alpha(f, y) = \alpha f(y)^{\alpha-1} - (\alpha-1)\ f\ _\alpha^\alpha, \quad \alpha > 1$	$\text{PSSphS}_\alpha(f, y) = \frac{f(y)^{\alpha-1}}{\ f\ _\alpha^{\alpha-1}}, \quad \alpha > 1$
<i>Special cases</i>			
	-	$\text{QS}(f, y) = \text{PowS}_2(f, y)$	$\text{SphS}(f, y) = \text{PSSphS}_2(f, y)$
		$\text{LogS}(f, y) = \lim_{\alpha \downarrow 1} \frac{\text{PowS}_\alpha(f, y) - 1}{\alpha - 1}$	$\text{LogS}(f, y) = \lim_{\alpha \downarrow 1} \frac{\text{PSSphS}_\alpha(f, y) - 1}{\alpha - 1}$
$\mathbb{D}_S(p\ f)$	$\text{KL}(p\ f) = \mathbb{E}_p \log \left(\frac{p}{f} \right)$	$\ p\ _\alpha^\alpha - \alpha \int f^{\alpha-1} (p-f) d\mu - \ f\ _\alpha^\alpha$	$\ p\ _\alpha - \frac{\int p f^{\alpha-1} d\mu}{\ f\ _\alpha^{\alpha-1}}$
$\alpha = 2$	-	$\ p - f\ _2^2$	$\ p\ _2 (1 - C(p, f))$
SP class	$\mathcal{P}_{\alpha=1}$	\mathcal{P}_α	\mathcal{P}_α
$\zeta(t)$	$t \log t$	t^α	-
$S(\tilde{f}, \tilde{y})$	$\log f(y) - \log b $	$\left(\frac{1}{ b }\right)^{\alpha-1} \text{PowS}_\alpha(f, y)$	$\left(\frac{1}{ b }\right)^\alpha \text{PSSphS}_\alpha(f, y)$
Focused			
$S_w^h(f, y)$	$w(y) \log \left(\frac{f(y)}{1-F_w} \right)$	$w(y) \left(\alpha \left(\frac{f_w(y)}{1-F_w} \right)^{\alpha-1} - (\alpha-1) \left\ \frac{f_w(y)}{1-F_w} \right\ _\alpha^\alpha \right)$	$w(y) \frac{f_w(y)^{\alpha-1}}{\ f_w\ _\alpha^{\alpha-1}}$
$S_w^b(f, y)$	$w(y) \log f(y) + (1-w(y)) \log \bar{F}_w$	$w(y) \alpha f_w(y)^{\alpha-1} + (1-w(y)) \alpha \bar{F}_w^{\alpha-1}$	$\frac{w(y) f_w(y)^{\alpha-1} + (1-w(y)) \bar{F}_w^{\alpha-1}}{(\ f_w\ _\alpha^\alpha + \bar{F}_w^\alpha)}$
$S_{w,h}^b(f, y)$	$w(y) \log f(y) + (1-w(y)) \log \bar{F}_w$	$-(\alpha-1) (\ f_w\ _\alpha^\alpha + \bar{F}_w^\alpha)$	$\frac{w(y) f_w(y)^{\alpha-1} + (1-w(y)) \bar{F}_w^{\alpha-1} \ h\ _\alpha^\alpha}{(\ f_w\ _\alpha^\alpha + \bar{F}_w^\alpha \ h\ _\alpha^\alpha)}$
$\mathbb{D}_{S_w}^b(p\ f)$	$f \log \left(\frac{p_w}{f_w} \right) p_w d\mu + \log \left(\frac{\bar{F}_w}{F_w} \right) \bar{P}_w$	$\ p_w\ _\alpha^\alpha + \bar{P}_w^\alpha - \int p_w f_w^{\alpha-1} d\mu - F_w \bar{F}_w^{\alpha-1}$	$(\ p_w\ _\alpha^\alpha + \bar{P}_w^\alpha) \frac{1}{\alpha} - \frac{\int p_w f_w^{\alpha-1} d\mu + F_w \bar{F}_w^{\alpha-1}}{(\ f_w\ _\alpha^\alpha + \bar{F}_w^\alpha)}$

NOTE: This table displays unweighted and focused scoring rules, divergences and associated properties based on two μ -densities, p and f , living on the measurable space $(\mathcal{Y}, \mathcal{G}, \mu)$, equipped with the L^α -norm $\|p\|_\alpha = (\int_{\mathcal{Y}} p^\alpha d\mu)^{1/\alpha}$. The common limiting case of PowS_α and PSSphS_α remains to hold for conditioning and censoring. $\mathbb{D}_S(p\|f)$ denotes the score divergence of f from p and $C(p, f) = \int p f d\mu / \sqrt{\int p^2 d\mu \int f^2 d\mu}$, the cosine similarity between p and f . The strict propriety class (SP class) is the class of probability measures relative to which the scoring rule is strictly proper. \mathcal{P}_α denotes the class of densities on $(\mathcal{Y}, \mathcal{G}, \mu)$ such that $\|p\|_\alpha < \infty, \forall p \in \mathcal{P}_\alpha$. The Bregman generator function $\zeta(t)$ parameterizes the subclass of separable Bregman divergences, consisting of the score divergences based on strictly proper scoring rules $S_\zeta: \mathcal{P} \times \mathcal{Y} \rightarrow \mathbb{R}$ of the form $S_\zeta(p, y) = \zeta'(p(y))p(y) - \int_{\mathcal{Y}} \zeta'(p(y))p(y) - \zeta(p(y))\mu(dy)$. $S(\tilde{f}, \tilde{y})$ denotes the score of the real-valued random variable $\tilde{Y} = bY + a$, where $a \in \mathbb{R}$ and $b \in \mathbb{R} \setminus \{0\}$, with density $\tilde{f}(\tilde{y}) = \frac{1}{|b|} f\left(\frac{\tilde{y}-a}{b}\right)$. The presented results for the focused scoring rules are equivalent in the sense that they yield the same expected score. The generalized censored scoring rule $S_{w,h}^b$ departs from a density h of which the support is a subset of $A_w^c \subseteq \mathcal{Y}$. The weight function is restricted accordingly. Appendix E details the derivations of the results presented in this table.

\mathcal{G} , respectively. The process generates the filtration $\{\mathcal{F}_t\}_{t=1}^T$, where $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$ is the information set at time t . The random variable of interest is Y_{t+1} conditional on \mathcal{F}_t , indicated by a subscript t to the (predictive) distributions, μ -densities and objects related to Q_{t+1} . The regions of interest $A_t \subseteq \mathcal{Y}$ are assumed to be \mathcal{F}_t -measurable.

This subsection aims to derive a uniformly most powerful (UMP) test for the hypotheses

$$\mathbb{H}_0 : p_t \mathbb{1}_{A_t} = f_{0t} \mathbb{1}_{A_t}, \quad \forall t \quad \text{vs.} \quad \mathbb{H}_1 : p_t \mathbb{1}_{A_t} = f_{1t} \mathbb{1}_{A_t}, \quad \forall t, \quad (6)$$

with f_{0t} and f_{1t} , fixed. Despite the densities f_{0t} and f_{1t} being fixed, the test concerns a multiple versus multiple hypothesis test due to the lacking specification of the densities outside the regions of interest A_t . Moreover, replacing A_t by A_{w_t} , for an \mathcal{F}_t -measurable weight function w_t , yields equivalent hypotheses in terms of w_t , specifically, $\mathbb{H}_0 : p_t w_t = f_{0t} w_t, \forall t$, versus $\mathbb{H}_1 : p_t w_t = f_{1t} w_t, \forall t$. Consequently, the UMP test for these hypotheses is equivalent to that for (6).

Theorem 3 reveals that (6) admits a UMP test, reducing to the Neyman and Pearson (1933) lemma when $A_t = \mathcal{Y}, \forall t$. A proof of this result is deferred to Appendix A.2.

Theorem 3 (Localized Neyman Pearson). *For any given $\alpha \in (0, 1)$, the UMP test of size α for testing problem (6) reads*

$$\phi_A^b(\mathbf{y}) = \begin{cases} 1, & \text{if } \lambda(\mathbf{y}) > c \\ \gamma, & \text{if } \lambda(\mathbf{y}) = c \\ 0, & \text{if } \lambda(\mathbf{y}) < c, \end{cases} \quad \lambda(\mathbf{y}) := \frac{f_{1,A}^b(\mathbf{y})}{f_{0,A}^b(\mathbf{y})}, \quad f_{j,A}^b(\mathbf{y}) := \prod_{t=0}^{T-1} f_{jt,A_t}^b(y_{t+1}), \quad j \in \{0, 1\},$$

where $\phi_A^b : \mathcal{Y}^T \rightarrow [0, 1]$ denotes a test function specifying the rejection probability, c is the largest constant such that $F_{0,A}^b(\lambda(\mathbf{y}) \geq c) \geq \alpha$ and $F_{0,A}^b(\lambda(\mathbf{y}) \leq c) \geq 1 - \alpha$, and $\gamma \in [0, 1]$ is such that $\alpha = F_{0,A}^b(\lambda(\mathbf{y}) > c) + \gamma F_{0,A}^b(\lambda(\mathbf{y}) = c)$.

For $T \equiv 1$, the test reduces to the UMP test of Holzmann and Klar (2017b). Corollary 1 reveals that it can alternatively be formulated in terms of the CSL introduced by Diks et al.

(2011). Corollary 2 endorses that conditioning does not bear a UMP test. The proofs of Corollaries 1 and 2 are deferred to Appendices B.2 and B.3, respectively.

Corollary 1. *An alternative formulation of the UMP test for testing problem (6) is given by the test defined in Theorem 3 with $\lambda(\mathbf{y})$ replaced by $\tilde{\lambda}^b(\mathbf{y}) := \sum_{t=0}^{T-1} (\text{Log}S_{A_t}^b(f_{1t}, y_{t+1}) - \text{Log}S_{A_t}^b(f_{0t}, y_{t+1}))$, i.e., in terms of the CSL.*

Corollary 2. *For testing problem (6), the test defined in Theorem 3 with $\lambda(\mathbf{y})$ replaced by $\tilde{\lambda}^\sharp(\mathbf{y}) := \sum_{t=0}^{T-1} (\text{Log}S_{A_t}^\sharp(f_{1t}, y_{t+1}) - \text{Log}S_{A_t}^\sharp(f_{0t}, y_{t+1}))$ is not UMP.*

Insisting on fully specified models under \mathbb{H}_0 and \mathbb{H}_1 as in the classical Neyman and Pearson (1933) framework, even if only on the restriction to A , can be too demanding in practice. In the empirical applications in Section 4, for instance, we consider a hypothesis where, under the null, two models are ‘equally wrong’. Then, generally, no UMP test is available, motivating the adoption of alternative scoring rules in addition to the CSL.

3.5 Related weighted scoring rules

In this subsection, we compare our approach to three alternative localization procedures. First, we consider Holzmann and Klar (2017a), who base their procedure on the conditional distribution F_w^\sharp . As $P_w^\sharp = F_w^\sharp$ if and only if $P \stackrel{A_w}{=} F$ does *not* hold, the score divergence $\mathbb{D}_{S_w^\sharp}(P||F) = (1 - \bar{P}_w)\mathbb{D}_S(P_w^\sharp||F_w^\sharp)$ generally fails to satisfy condition (ii) of Definition 2, unless $\bar{P}_w = \bar{F}_w$, hence is not a local divergence; see Example F.1 for a specific case. To resolve this, Holzmann and Klar (2017a) add an auxiliary scoring rule s , enforcing the score divergence to be zero if and only if $P \stackrel{A_w}{=} F$. Example 5 describes their composite scoring rule, for which the score divergence is a local divergence but not in general a localized divergence.

Example 5. *Holzmann and Klar (2017a) propose a class of weighted scoring rules defined as $\tilde{S}_{w,s}(F, y) := S_w^\sharp(F, y) + w(y)s(B_{1-\bar{F}_w}, 1) + (1 - w(y))s(B_{1-\bar{F}_w}, 0)$, assuming $\bar{F}_w < 1$, and where B_θ denotes the Bernoulli(θ) distribution with pmf $b(z; \theta) = \theta^z(1 - \theta)^{1-z}$, $z \in \{0, 1\}$.*

For any s , $\tilde{S}_{w,s}$ is strictly locally proper (Holzmann and Klar 2017a, Theorem 2) and hence $\mathbb{D}_{\tilde{S}_{(w,s)}}(\mathbb{P} \parallel \mathbb{F}) = (1 - \bar{P}_w) \mathbb{D}_S(\mathbb{P}_w^\# \parallel \mathbb{F}_w^\#) + \mathbb{D}_s(\mathbb{B}_{1-\bar{P}_w} \parallel \mathbb{B}_{1-\bar{F}_w})$ is a local divergence, but generally not a localized divergence due to the dependence on \mathbb{D}_s . Moreover, $\mathbb{D}_S(\mathbb{P}_w^\# \parallel \mathbb{F}_w^\#) = 0$ if \mathbb{P} and \mathbb{F} are proportional on A_w ; then, $\mathbb{D}_{\tilde{S}_{(w,s)}}(\mathbb{P} \parallel \mathbb{F}) = \mathbb{D}_s(\mathbb{B}_{1-\bar{P}_w} \parallel \mathbb{B}_{1-\bar{F}_w})$, which depends only on the auxiliary scoring rule.

Holzmann and Klar (2017a) provide two specific choices for s : (i) $\text{slog}(\mathbb{B}_{1-\bar{F}_w}, z) := z \log(1 - \bar{F}_w) + (1 - z) \log \bar{F}_w$ and (ii) $\text{sbar}(\mathbb{B}_{1-\bar{F}_w}, z) := z(\log(1 - \bar{F}_w) + 1) - (1 - \bar{F}_w)$. The combination $S = \text{LogS}$ and $s = \text{slog}$ recovers LogS_w^b and hence a localized divergence, whereas $S = \text{LogS}$ and $s = \text{sbar}$ leads to the weighted likelihood score by Pelenis (2014), for which the score divergence $\mathbb{D}_{\text{pwl}_w}(\mathbb{P} \parallel \mathbb{F}) = \mathbb{D}_{\text{LogS}}(\mathbb{P}_w^b \parallel \mathbb{F}_w^b) - \bar{P}_w \log \frac{\bar{P}_w}{\bar{F}_w} + (1 - \bar{P}_w) \log \frac{1 - \bar{P}_w}{1 - \bar{F}_w} + (\bar{P}_w - \bar{F}_w)$ is *not* a localized divergence of \mathbb{D}_{LogS} .

The arbitrariness of s allows for weighted scoring rules beyond the scope of (generalized) censored scoring rules. For example, $S = \text{QS}$ and $s = \text{slog}$, yields $\mathbb{D}_{\text{QS}_{(w,\text{slog})}}(p \parallel f) = (1 - \bar{P}_w) \|p_w^\# - f_w^\#\|_2^2 + \mathbb{D}_{\text{slog}}(\mathbb{B}_{1-\bar{P}_w} \parallel \mathbb{B}_{1-\bar{F}_w})$. A key distinction from $\mathbb{D}_{\text{QS}_{w,H}^b}(p \parallel f) = \|p_{w,H}^b - f_{w,H}^b\|_2^2$ is the role of the *auxiliary* KL-divergence, \mathbb{D}_{slog} in $\mathbb{D}_{\text{QS}_{(w,\text{slog})}}$, which will become *dominant* for large \bar{F}_w . Censoring is also not nested within the framework of Holzmann and Klar (2017a). To see this, note that unlike censoring (see Example 3), $\tilde{S}_{w,s}(\mathbb{F}, y)$ in Example 5 enforces $\tilde{S}_{w,s}(\mathbb{F}, y)$ to be only a function of \bar{F}_w if $y \in A_w^c$.

The second comparison concerns the threshold weighted kernel score $\text{tw}S_\rho$ introduced by Allen et al. (2023), generalizing the twCRPS by introducing the kernel $\rho(v(y), v(y'))$ based on a *measurable chaining function* $v : \mathcal{Y} \mapsto \mathcal{Y}$. With formalities presented in their Propositions 4.4 and 4.5, the $\text{tw}S_\rho$ is strictly locally proper if v is injective on A_w and $\rho(v(y), v(\cdot)) = \rho(v(y'), v(\cdot))$, $\forall y, y' \in A_w^c$. For indicator weight functions, both conditions are easily satisfied with $v(y) = y \mathbb{1}_{A_w}(y) + y_0 \mathbb{1}_{A_w^c}(y)$, for any $y_0 \in \mathcal{Y}$ such that $\mathbb{F}_w(y_0) = 0$. In that case, the $\text{tw}S_\rho$ reduces to the S_{ρ, y_0}^b score given in Example 4.

Our censoring approach also admits center indicator functions with two pivotal points,

as in Example D.2; these are not considered in the kernel score framework of Allen et al. (2023). For non-indicator weight functions, Allen et al. (2023) provide a specific multivariate example that meets their requirements for strict local propriety, and which can be extended to other non-indicator weight functions that are specified as a product of marginal weight functions, including the multivariate logistic weight functions considered here (see Appendix E.4 for details). For more general non-indicator weight functions, specifying the chaining function as required for the approach of Allen et al. (2023) is a less trivial task. Our procedure foregoes specifying the chaining function and, moreover, is not restricted to kernel scores.

Third and finally, we compare with Mitchell and Weale (2023), who, for real-valued, unimodal densities and with the center as the region of interest, also consider censored density forecasts based on LogS. However, unlike our setting, they do *not* aim to evaluate multiple candidate densities. As illustrated by Example 6, dependence of their region of interest on the candidate distribution renders the resulting scoring rule improper, thus not suitable for our aim of evaluating multiple density forecasts.

Example 6. *Mitchell and Weale (2023) consider the alternative censored likelihood score $\text{LogS}_\alpha^{\text{MW}}(f, y) := \log f(y)\mathbf{1}_{A(F; \alpha)}(y) + \log(\alpha)\mathbf{1}_{A(F; \alpha)^c}(y)$, where $A(F; \alpha)$ is the central region of interest. A key difference with the censored likelihood score $\log f_A^b(y)$ is the dependence of $A(F; \alpha)$ on f , by which $\text{LogS}_\alpha^{\text{MW}}(f, y)$ is improper. Indeed, for symmetric densities, $A(F; \alpha) = [F^{-1}(\alpha/2), F^{-1}(1 - \alpha/2)]$. Then, letting p and f be the $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, \frac{1}{2})$ density, respectively, we have $\mathbb{E}_p \text{LogS}_\alpha^{\text{MW}}(p, Y) - \mathbb{E}_p \text{LogS}_\alpha^{\text{MW}}(f, Y) < 0$, for all $\alpha > \alpha_0$ with $\alpha_0 \approx 0.052$.*

4 EMPIRICAL PERFORMANCE

We assess the empirical performance of the censoring approach to focus scoring rules on regions of interest by evaluating its ability to discriminate between different forecast methods. We compare the performance of censored scoring rules with conditional scoring rules and with the composite scoring rules proposed by Holzmann and Klar (2017a). For the latter, we augment the conditional scoring rule with the auxiliary rule sbar or slog, see Section 3.5. We consider applications in financial risk management, macroeconomics, and climate, evaluating the focused scoring rules in a similar manner, as described below.

Following Giacomini and White (2006), we treat all components underlying a density forecast, including its estimation procedure, as integral parts of the forecast method. Let \hat{f}_t and \hat{g}_t denote density forecasts resulting from competing methods, each estimated with a rolling window of length m . We test the null hypothesis of equal predictive ability, $\mathbb{H}_0 : \mathbb{E}_{p_t} S_w(\hat{f}_t, Y_{t+1}) = \mathbb{E}_{p_t} S_w(\hat{g}_t, Y_{t+1}), \forall t$, by means of the DM type test statistic $t_{m,n} := \frac{1}{n} \sum_{t=m}^{T-1} (S_w(\hat{f}_t, Y_{t+1}) - S_w(\hat{g}_t, Y_{t+1})) / \sqrt{\hat{\sigma}_{m,n}^2/n}$, where $n = T - m$ is the number of observations used for evaluation and $\hat{\sigma}_{m,n}^2$ is a heteroskedasticity and autocovariance-consistent (HAC) variance estimator.

The null hypothesis is equivalent to $\mathbb{D}_{S_w}(p_t || \hat{f}_t) = \mathbb{D}_{S_w}(p_t || \hat{g}_t)$ and is rejected if it is sufficiently unlikely that the weighted score divergences from p_t to \hat{f}_t and p_t to \hat{g}_t coincide. This null differs from that in (6), obstructing theoretical results on the power properties of the test. However, because censoring preserves more information than conditioning, we generally expect higher power for test statistics based on censored scoring rules. This is supported by the Monte Carlo results in Appendix G. In these simulation experiments, we also find that the composite scoring rules of Holzmann and Klar (2017a) generally result in test statistics with comparable power. However, further analysis shows that this cannot be attributed to the localization method but rather to the advantageous properties of the

logarithmic score that forms the basis for the auxiliary rules sbar and slog.

In practice, including the empirical applications discussed below, one commonly has more than two candidate forecast methods. We therefore start with a collection \mathcal{M}_0 of forecast methods, and then use the iterative procedure proposed by Hansen et al. (2011) to reduce \mathcal{M}_0 to a Model Confidence Set (MCS) of methods for which the null of equal predictive ability cannot be rejected. Elimination in round k is based on the statistic $\text{TR} := \max_{i,j \in \mathcal{M}_k} |t_{m,n}^{(i,j)}|$, where $t_{m,n}^{(i,j)}$ corresponds to the pairwise $t_{m,n}$ -statistic between forecast methods i and j introduced above. Favorable power properties of censoring in the pairwise tests intuitively accelerate elimination in the MCS procedure, resulting in smaller p -values and, consequently, reduced MCS cardinality. We present MCS results at the 0.90 confidence level, with results for the 0.75 confidence level deferred to Appendix I.1, using a block bootstrap with block length $b = 5$ and $B = 10,000$ replications, unless stated otherwise. Our results are robust to variations in these parameters, see Table I.3.

In each application below, the unweighted scoring rules are given by LogS, QS, SphS and S_{ρ_1} , with kernel $\rho_1(\mathbf{x}, \mathbf{x}') := \|\mathbf{x} - \mathbf{x}'\|$, where $\|\cdot\|$ the Euclidean norm, i.e., S_{ρ_1} is the Energy Score that reduces to the CRPS in univariate examples. These scoring rules are localized by (i) conditioning, (ii) censoring, (iii) conditioning with sbar and (iv) conditioning with slog. The twCRPS and tw S_{ρ_1} are included only in cases where they differ from CRPS^b. Hence, we consider 16 or 17 weighted scoring rules per application. There are $|\mathcal{M}_0| = 6$ candidate forecast methods, with specifications differing by application. For reproducibility, Appendix H includes details on the specification of the individual methods. Moreover, Appendix I reports the MCS p -values per individual scoring rule and weight function underlying the summary results presented in this section.

4.1 Financial risk management

Measuring and forecasting the downside risk of asset returns is crucial in risk management, particularly for compliance with regulatory requirements related to measures such as Value-at-Risk and Expected Shortfall. We evaluate density forecasts constructed for daily log-returns y_t on the S&P500 index over the period from January 2, 1996, to December 30, 2022 (6,777 observations), sourced from Yahoo Finance. To achieve the required focus on the left tail of the density forecast, we use the indicator weight function $I_L(y_t; \hat{r}_t^q)$, where \hat{r}_t^q denotes the q -th empirical quantile of y_t , based on the same fixed rolling window of length $m = 1,000$ used for estimation of the forecast methods.

All forecast methods used conform to $Y_t | \mathcal{F}_{t-1} \sim \mathcal{D}(\mu, \sigma_t^2, \boldsymbol{\vartheta})$, denoting a parametric family of distributions with constant mean μ , time-varying variance σ_t^2 , and any additional parameters collected in $\boldsymbol{\vartheta}$. Although we tested AR(1) and AR(5) models for the conditional mean, neither improved significantly over a constant mean. We consider three conditional variance models: GARCH, threshold GARCH (TGARCH) and realized GARCH (RGARCH), proposed by Bollerslev (1986), Glosten et al. (1993) and Hansen et al. (2012), respectively. We combine each of the volatility models with standard normal and Student- t_ν distributions. Density forecasts are constructed for horizons $\tau = 1$ and 5 days.

Table 2 reveals stark differences in the cardinality of MCS^b and MCS^\sharp , particularly at $\tau = 1$. In case no correction is applied to the conditional scoring rules, MCS^b is strictly smaller than MCS^\sharp in 75% of all replications, while MCS^\sharp contains more than twice the number of methods compared to MCS^b on average. These results moderate when the conditional scoring rules are appended with a Holzmann-Klar correction term. Nevertheless, MCS^b remains strictly smaller than MCS^\sharp in nearly 40% of the cases, with an average difference in cardinality of 20%. For $\tau = 5$, the differences become smaller but remain in favor of censoring, ranging between 10 and 30%.

We extend the univariate setting to the evaluation of bivariate density forecasts for the vector of log-returns $\mathbf{y}_t \in \mathbb{R}^2$ for the Energy Select Sector SPDR Fund (XLE) and Financial Select Sector SPDR Fund (XLF), for the period January 5, 1999 to December 29, 2023 (6,218 observations). We consider the approximated bivariate empirical q -th quantile of \mathbf{y}_t given by $\tilde{\mathbf{r}}_t^q := (\hat{r}_{1,t}^{q_2}, \hat{r}_{2,t}^{q_2})$, with $q_2 = \sqrt{q}$ to formulate the weight functions $I_{\mathbb{L}}^2(\mathbf{y}_t; \tilde{\mathbf{r}}_t^q)$ and $\Lambda_{a,\mathbb{L}}^2(\mathbf{y}_t; \tilde{\mathbf{r}}_t^q)$, with $a = 3$, while having verified stability of results for $a \in \{2, 4\}$. The individual mean (μ_i) and volatility ($\sigma_{i,t}^2$) specifications are as in the S&P500 models. We use the Dynamic Conditional Correlation (DCC) approach of Engle (2002) to map the univariate specifications into a bivariate conditional covariance matrix. The univariate distributions are replaced by bivariate standard normal and Student- t_ν distributions.

Table 2 shows that the results for the bivariate density forecasts corroborate the main findings for the univariate S&P500 application. First, the MCS obtained with censored scoring rules is not larger than the MCS resulting from conditional scoring rules in the large majority of cases, namely between 62% and 92%. For about one-third of cases, MCS^b is even strictly smaller than MCS[#]. Second, the former percentages are hardly affected by adding the Holzmann-Klar correction terms to the conditional scoring rule, while the latter decline but not substantially. The largest reduction occurs when using the slog correction term for the scoring rules focused with weight function $I_{\mathbb{L}}^2(\mathbf{y}_t; \tilde{\mathbf{r}}_t^q)$, from 33% to 17% for $\tau = 1$. Hence, the correction terms result in equally large MCSs for the censored and augmented conditional scoring rules more frequently. Closer inspection reveals that their compositions almost always are identical as well. Third, the results for the logistic weight function $\Lambda_{a,\mathbb{L}}^2(\mathbf{y}_t; \tilde{\mathbf{r}}_t^q)$ with $a = 3$ do not differ much from those for the indicator weight function $I_{\mathbb{L}}^2(\mathbf{y}_t; \tilde{\mathbf{r}}_t^q)$, which is recovered as $a \rightarrow \infty$. However, the discrepancies in cardinality are somewhat moderated, particularly for $\tau = 5$, aligning with the observation by Diks et al. (2011) that the score distribution of the weighted scoring rules becomes more alike for smaller values of a . Finally, in contrast to the univariate setting, the (relative)

performance of the censored scoring rules does not decline at longer forecast horizons. If anything, the percentages and average cardinality ratio improve for $\tau = 5$ compared to $\tau = 1$.

Table 2: MCS cardinality of censored and (un)corrected conditional scoring rules

Sec.	w_t	τ	no correction			sbar			slog		
			\leq	$<$	\sharp/b	\leq	$<$	\sharp/b	\leq	$<$	\sharp/b
4.1	$I_L(y_t; \hat{r}_t^q)$	1	96%	75%	2.38	71%	38%	1.20	71%	38%	1.20
		5	62%	29%	1.29	54%	25%	1.08	62%	25%	1.10
	$I_L^2(\mathbf{y}_t; \tilde{\mathbf{r}}_t^q)$	1	62%	33%	1.20	67%	25%	1.06	62%	17%	0.98
		5	92%	29%	1.23	92%	25%	1.28	92%	25%	1.26
	$\Lambda_{3,L}^2(\mathbf{y}_t; \tilde{\mathbf{r}}_t^q)$	1	62%	33%	1.06	62%	21%	0.95	62%	21%	0.95
		5	71%	33%	1.16	83%	25%	1.16	88%	25%	1.18
4.2	$I_C(y_t; 2, r_1)$	6	100%	92%	2.67	100%	83%	2.42	100%	67%	1.69
		24	92%	75%	2.27	58%	42%	1.24	67%	25%	1.23
	$I_C^c(y_t; 2, r_1)$	6	100%	92%	2.04	83%	50%	1.13	67%	33%	1.19
		24	92%	75%	2.86	50%	33%	1.18	75%	17%	1.04
	$I_R(y_t; \hat{r}_t^q)$	1	83%	58%	1.92	92%	67%	1.92	83%	33%	1.29
		3	75%	46%	1.54	79%	42%	1.40	75%	4%	0.96
$I_C(y_t; 18, r_2)$	1	100%	58%	2.21	100%	42%	1.42	100%	42%	1.42	
	3	100%	58%	1.58	100%	0%	1.00	100%	0%	1.00	
Total average			85%	56%	1.82	78%	37%	1.32	79%	27%	1.18

NOTE: This table presents changes in cardinality of the MCS in absolute and relative terms, at 0.90, across different forecast horizons τ , corresponding to the forecasting applications in risk management (Section 4.1), inflation (Section 4.2) and temperature (Section 4.3). sbar and slog refer to the correction terms for conditional scoring rules proposed by Holzmann and Klar (2017a). Columns labeled \leq ($<$) display the percentage of cases where MCS^b contains (strictly) fewer forecast methods than MCS[#] and the column labeled \sharp/b reports the ratio $|\text{MCS}^\#|/|\text{MCS}^b|$. Each result represents an average over a set of scoring rules $S \in \{\text{LogS}, \text{QS}, \text{SphS}, \text{CRPS}/S_{\rho_1}\}$ and quantile levels $q \in \{0.01, 0.05, 0.10, 0.15, 0.20, 0.25\}$ or levels $r_1 \in \{1, 1.5, 2\}$ and $r_2 \in \{1, 2, 4\}$. The empirical q -th quantiles \hat{r}_t^q of y_t are based on the parameter estimation window of m observations, and $\tilde{\mathbf{r}}_t^q := (\hat{r}_{1,t}^{q_2}, \hat{r}_{2,t}^{q_2})$, with $q_2 = \sqrt{q}$, approximates a bivariate empirical q -th quantile of \mathbf{y}_t . The p -values are obtained via a block bootstrap of $B = 10,000$ replications, with block length $b = 5$, or $b = 200$ for the climate data. Complete MCS details and associated p -values are provided in Appendix I.

4.2 Macroeconomics

We next consider forecasting inflation, a subject with a long history in macroeconomics that recently has regained prominence. Given that many central banks, including the

Federal Reserve System and the European Central Bank target an annual inflation rate of 2%, we focus on the central range $A_r = (2 - r, 2 + r)$, where $r > 0$, by using the weight function $I_C(y_t; 2, r)$. To address policymakers’ concerns for deviations beyond A_r , termed ‘Inflation at Risk’ (Lopez-Salido and Loria 2020), we additionally consider its complement $I_C^c(y_t; 2, r) = 1 - I_C(y_t; 2, r)$. For the CRPS, we adopt two pivotal points for $I_C(y_t; 2, r)$, while using $\ell = 2$ for its complement, i.e., treating non-tail observations to be on target. Following Stock and Watson (2002), among many others, we construct direct forecasts for annualized τ -month inflation rates $y_{t+\tau}^\tau = (1, 200/\tau) \log(P_{t+\tau}/P_t)$, where P_t denotes the U.S. consumer price index (CPI) in month t , for horizons $\tau = 6$ and 24. The sample period runs from January 1960 until December 2015 (672 observations), where density forecasts are obtained for the final 180 months in this time frame.

We consider forecast methods that aim to exploit the ‘data-rich environment’ in macroeconomic forecasting, with many potentially relevant predictors especially for inflation. Here we follow Medeiros et al. (2021) by using the same 122 variables from the FRED-MD database (\mathbf{x}_t). Each of the forecast methods can be represented as $y_{t+\tau}^\tau = \mu_{t+\tau}^\tau(\mathbf{x}_t) + u_{t+\tau}^\tau$, where we consider the following subset of methods listed by Medeiros et al. (2021) for the conditional mean $\mu_{t+\tau}^h$: Random Walk, Auto-Regressive model (AR), Bagging, Complete Subset Regression (CSR), Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest. The error $u_{t+\tau}^\tau$ is assumed to follow a two-piece normal distribution, congruent with the statistical model underlying the fan charts published by the Bank of England (Clements 2004; Mitchell and Hall 2005; Gneiting and Ranjan 2011).

The summary results presented in Table 2 reveal a distinct and pronounced preference for censoring, again especially when no correction is applied to the conditional scoring rules. In that case the cardinalities of MCS^b are almost always (weakly) smaller than those of MCS[#]. The relative increase in set cardinality when opting for conditioning over censoring is substantial at more than 100%. Interestingly, the censored scoring rules outperform the

conditional rules not only when focusing on the central range around the inflation target of 2%, but also when the interest is on the complementary ‘Inflation at Risk’ region of more extreme inflation rates. Finally, also in this application the Holzmann-Klar corrections to the conditional scoring rules improve their (relative) performance, although the MCS cardinality results largely remain favorable to censoring.

4.3 Climate

We generate density forecasts for Dutch daily average temperature data, focusing on high temperatures via the weight function $I_R(y_t; \hat{r}_t^q)$ and temperatures near the optimal temperature for tuber growth, approximately 18 degrees Celsius (Struik 2007, Section 18.5.5), using $I_C(y_t; 18, r)$. Extending the data and methodology of Franses et al. (2001) and Tol (1996), we focus on volatility clustering and asymmetries in the relationship between past temperature and volatility, along with seasonal variations in the mean and variance. We use daily observations for the period from February 1, 2003, to January 31, 2023, with a rolling estimation window of $m = 2,922$ days (or 8 years). Our volatility models closely follow the GARCH, QGARCH-I, and QGARCH-II specifications as in Franses et al. (2001), but with alterations in seasonal trend estimation. The GARCH-type models are combined with a standard normal and Student- t_ν distribution.

Using the right-tail weight function $I_R(y_t; \hat{r}_t^q)$ to focus on high daily temperatures, we find results exhibiting pronounced parallels with the left-tail risk management application. In particular, as seen in Table 2, the cardinalities of the censored MCSs are typically much smaller than their uncorrected conditional counterparts for $\tau = 1$ day-ahead forecasts; the differences diminish at the longer forecast horizon $\tau = 3$ or when a Holzmann-Klar correction is appended to the conditional scoring rule.

Focusing on the central range around 18 degrees Celsius with the weight function $I_C(y_t; 18, r)$, we find that there are no instances where conditioning leads to a smaller

MCS for $\tau = 3$ and almost no such cases for $\tau = 1$. Relative to inflation, there is a notable increase in cases where the MCSs possess identical cardinality, also reflected in the smaller ratios $|\text{MCS}^\#|/|\text{MCS}^b|$.

5 CONCLUSION

In this paper, we propose censoring as a focusing device to accommodate the fact that in many applications, forecasters are particularly interested in specific areas of the outcome space. We demonstrate that a key advantage of censoring is that applying scoring rules to censored distributions results in strictly locally proper scoring rules. To the best of our knowledge, we are the first to derive a transformation of the original scoring rule that preserves both the score divergence and strict propriety, and features high flexibility, being applicable across varied scoring rules, weight functions, and outcome spaces. For specific choices, the censored scoring rule yields intuitively appealing rules apt for practical use. For instance, we recover the twCRPS for tail indicators, while extending its strict local propriety to other weight functions. Our second theoretical contribution, a generalization of the Neyman Pearson lemma, revolves around the censored likelihood score. We have shown that the UMP test of the localized Neyman Pearson hypothesis is a censored likelihood ratio test, reducing to the original lemma if the weight function is positive for all outcomes. By contrast, the conditional likelihood ratio test is not UMP.

We demonstrate the practical relevance of censoring with empirical applications in financial risk management, macroeconomics, and climate. We use the size of the Model Confidence Set (MCS) to gauge the scoring rule's ability to discriminate between competing forecast methods. A common finding in the applications is that the censored MCS is (strictly) smaller than the conditional MCS in a large majority of cases, and often the difference in cardinality is substantial. This conclusion holds across different areas of interest,

in particular whether attention is focused on the central range of the distribution or on one (or both) of the tails.

SUPPLEMENTARY MATERIAL

All proofs and additional theoretical results, the Monte Carlo analysis, and full tables on the empirical performance are provided in an online supplementary document. (.pdf)

References

- Adrian, T., N. Boyarchenko, and D. Giannone (2019), “Vulnerable Growth”, *American Economic Review*, 109(4), 1263–1289.
- Allen, S., D. Ginsbourger, and J. Ziegel (2023), “Evaluating Forecasts for High-Impact Events Using Transformed Kernel Scores”, *SIAM/ASA Journal on Uncertainty Quantification*, 11(3), 906–940.
- Amisano, G. and R. Giacomini (2007), “Comparing Density Forecasts via Weighted Likelihood Ratio Tests”, *Journal of Business & Economic Statistics*, 25(2), 177–190.
- Bernoulli, D. (1760), “Essai d’une Nouvelle Analyse de la Mortalite Causee par la Petite Verole, et des Avantages de l’Inoculation Pour la Prevenir”, *Histoire de l’Acad., Roy. Sci.(Paris) avec Mem*, 1–45.
- Bollerslev, T. (1986), “Generalized Autoregressive Conditional Heteroskedasticity”, *Journal of Econometrics*, 31(3), 307–327.
- Borowska, A., L. Hoogerheide, S. J. Koopman, and H. K. Van Dijk (2020), “Partially Censored Posterior for Robust and Efficient Risk Evaluation”, *Journal of Econometrics*, 217(2), 335–355.
- Bregman, L. (1967), “The Relaxation Method of Finding the Common Point of Convex Sets and its Application to the Solution of Problems in Convex Programming”, *USSR Computational Mathematics and Mathematical Physics*, 7(3), 200–217.
- Brehmer, J. R. and T. Gneiting (2020), “Properization: Constructing Proper Scoring Rules via Bayes Acts”, *Annals of the Institute of Statistical Mathematics*, 72(3), 659–673.
- Brier, G. W. (1950), “Verification of Forecasts Expressed in Terms of Probability”, *Monthly Weather Review*, 78(1), 1–3.
- Clements, M. P. (2004), “Evaluating the Bank of England Density Forecasts of Inflation”, *The Economic Journal*, 114(498), 844–866.
- Cont, R., R. Deguest, and G. Scandolo (2010), “Robustness and Sensitivity Analysis of Risk Measurement Procedures”, *Quantitative Finance*, 10(6), 593–606.
- Dawid, A. P. (1984), “Statistical Theory: The Prequential Approach”, *Journal of the Royal Statistical Society. Series A (General)*, 147(2), 278–292.
- Dawid, A. P. (2007), “The Geometry of Proper Scoring Rules”, *Annals of the Institute of Statistical Mathematics*, 59(1), 77–93.

- Diebold, F. X. and R. S. Mariano (2002), “Comparing Predictive Accuracy”, *Journal of Business & Economic Statistics*, 20(1), 134–144.
- Diks, C., V. Panchenko, and D. Van Dijk (2011), “Likelihood-based Scoring Rules for Comparing Density Forecasts in Tails”, *Journal of Econometrics*, 163(2), 215–230.
- Eguchi, S. (1985), “A Differential Geometric Approach to Statistical Inference on the Basis of Contrast Functionals”, *Hiroshima Mathematical Journal*, 15(2), 341–391.
- Ehm, W. and T. Gneiting (2012), “Local Proper Scoring Rules of Order Two”, *The Annals of Statistics*, 40(1), 609–637.
- Engle, R. (2002), “Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models”, *Journal of Business & Economic Statistics*, 20(3), 339–350.
- Fissler, T., J. F. Ziegel, and T. Gneiting (2015). “Expected Shortfall is Jointly Elicitable with Value at Risk - Implications for Backtesting”. DOI: 10.48550/ARXIV.1507.00244. Available at <https://arxiv.org/abs/1507.00244>.
- Franses, P. H., J. Neele, and D. Van Dijk (2001), “Modeling Asymmetric Volatility in Weekly Dutch Temperature Data”, *Environmental Modelling & Software*, 16(2), 131–137.
- Giacomini, R. and H. White (2006), “Tests of Conditional Predictive Ability”, *Econometrica*, 74(6), 1545–1578.
- Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993), “On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks”, *The Journal of Finance*, 48(5), 1779–1801.
- Gneiting, T. and A. E. Raftery (2007), “Strictly Proper Scoring Rules, Prediction, and Estimation”, *Journal of the American Statistical Association*, 102(477), 359–378.
- Gneiting, T. and R. Ranjan (2011), “Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules”, *Journal of Business & Economic Statistics*, 29(3), 411–422.
- Hansen, P. R., Z. Huang, and H. H. Shek (2012), “Realized GARCH: A Joint Model for Returns and Realized Measures of Volatility”, *Journal of Applied Econometrics*, 27(6), 877–906.
- Hansen, P. R., A. Lunde, and J. Nason (2011), “The Model Confidence Set”, *Econometrica*, 79(2), 453–497.
- Holzmann, H. and B. Klar (2017a), “Focusing on Regions of Interest in Forecast Evaluation”, *The Annals of Applied Statistics*, 11(4), 2404–2431.
- Holzmann, H. and B. Klar (2017b). “Weighted Scoring Rules and Hypothesis Testing”. Available at <https://arxiv.org/abs/1611.07345v2>.
- Iacopini, M., F. Ravazzolo, and L. Rossini (2023), “Proper Scoring Rules for Evaluating Density Forecasts with Asymmetric Loss Functions”, *Journal of Business & Economic Statistics*, 41(2), 482–496.
- Kullback, S. and R. A. Leibler (1951), “On Information and Sufficiency”, *The Annals of Mathematical Statistics*, 22(1), 79–86.
- Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting (2017), “Forecaster’s Dilemma: Extreme Events and Forecast Evaluation”, *Statistical Science*, 32(1), 106–127.

- Lopez-Salido, D. and F. Loria (2020). “Inflation at Risk”. Finance and Economics Discussion Series 2020-013. Washington: Board of Governors of the Federal Reserve System. Available at <https://doi.org/10.17016/FEDS.2020.013>.
- Medeiros, M. C., G. F. R. Vasconcelos, A. Veiga, and E. Zilberman (2021), “Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods”, *Journal of Business & Economic Statistics*, 39(1), 98–119.
- Mitchell, J. and S. G. Hall (2005), “Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR ‘Fan’ Charts of Inflation”, *Oxford Bulletin of Economics and Statistics*, 67(s1), 995–1033.
- Mitchell, J. and M. Weale (2023), “Censored Density Forecasts: Production and Evaluation”, *Journal of Applied Econometrics*, 38(5), 714–734.
- Neyman, J. and E. Pearson (1933), “IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses”, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706), 289–337.
- Ovcharov, E. Y. (2018), “Proper Scoring Rules and Bregman Divergence”, *Bernoulli*, 24(1), 53–79.
- Painsky, A. and G. W. Wornell (2020), “Bregman Divergence Bounds and Universality Properties of the Logarithmic Loss”, *IEEE Transactions on Information Theory*, 66(3), 1658–1673.
- Patton, A. J. (2020), “Comparing Possibly Misspecified Forecasts”, *Journal of Business & Economic Statistics*, 38(4), 796–809.
- Pelenis, J. (2014). “Weighted scoring rules for comparison of density forecasts on subsets of interest”. Available at <https://sites.google.com/site/jpelenis/>.
- Steinwart, I. and J. F. Ziegel (2021), “Strictly proper kernel scores and characteristic kernels on compact spaces”, *Applied and Computational Harmonic Analysis*, 51, 510–542.
- Stock, J. H. and M. W. Watson (2002), “Macroeconomic Forecasting Using Diffusion Indexes”, *Journal of Business & Economic Statistics*, 20(2), 147–162.
- Struik, P. C. (2007). “Chapter 18 - Responses of the Potato Plant to Temperature”. In D. Vreugdenhil, J. Bradshaw, C. Gebhardt, F. Govers, D. K. L. Mackerron, M. A. Taylor, and H. A. Ross (Eds.), *Potato Biology and Biotechnology*, pp. 367–393. Amsterdam: Elsevier Science B.V.
- Tobin, J. (1958), “Estimation of Relationships for Limited Dependent Variables”, *Econometrica*, 26(1), 24–36.
- Tol, R. S. (1996), “Autoregressive Conditional Heteroscedasticity in Daily Temperature Measurements”, *Environmetrics*, 7(1), 67–75.