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 and Tinbergen Institute

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

October 13, 2023

Abstract

When comparing predictive distributions, forecasters are typically not equally interested in all regions of the outcome space. However, extant methods to focus strictly proper scoring rules on a region of interest, do not maintain strict propriety. We propose a natural procedure for transforming strictly proper scoring rules into their strictly locally proper counterparts, nesting the censored likelihood score as a special case. Our procedure also implies a class of censored kernel scores, offering a multivariate alternative to the weighted Continuously Ranked Probability Score (twCRPS), additionally solving its local impropriety for weight functions other than single tail indicators, for which the twCRPS is recovered. Within this framework, we obtain a generalization of the Neyman Pearson lemma. For less restricted hypotheses, the results of Monte Carlo simulations and applications to risk management, inflation and climate data, confirm our intuition that censoring generally leads to higher power than conditioning.

Keywords: Censoring; Forecast Evaluation; Hypothesis Testing; Density Forecasts

*Corresponding author. Mailing Address: PO Box 15867, 1001 NJ Amsterdam, The Netherlands. Phone: +31 (0) 20 525 4252. Email: r.f.a.depunder@uva.nl

1 INTRODUCTION

Over the past decades, probabilistic forecasts have garnered increasing attention across a variety of disciplines, primarily as 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 employment of strictly proper scoring rules (Gneiting and Raftery, 2007; Brehmer and Gneiting, 2020; Patton, 2020), which have been widely advocated for their capacity to ensure unbiased comparative assessments of different forecasting methods. While the utility of probabilistic forecasting is well-recognized, specialized applications, such as the assessment of extreme financial portfolio losses, require a localized evaluation of predictive distributions.

In this paper, we introduce a localization mechanism for strictly proper scoring rules that preserves strict propriety. Leveraging the concept of censoring (Bernoulli, 1760) from the Tobit model (Tobin, 1958), the proposed transformation finds a sweet spot between retaining and excluding information when focusing the original distribution to a region of interest. Specifically, unlike approaches that employ conditional distributions, our method maintains the overall probability of receiving an observation in (or outside) the target region, obviously informative when comparing various candidate distributions focused on the same area. Moreover, within the region of interest, our mechanism replicates the original distribution's shape, which is particularly beneficial when evaluating functionals specific to this region, such as Value-at-Risk (VaR) or Expected Shortfall (ES). Our procedure encompasses well-established strictly locally proper scoring rules, including the censored likelihood (csl) score, proposed by Diks et al. (2011) and the weighted Continuously Ranked Probability Score (twCRPS, proposed by Gneiting and Ranjan (2011)), for weight functions for which Holzmann and Klar (2017) 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, we show that the twCRPS suffers from a *localization bias* because it retains too much information.

We prove a generalization of the Neyman Pearson 1933 lemma, revealing that the censored likelihood ratio leads to a Uniformly Most Powerful (UMP) test. In contrast, we provide explicit evidence that the conditional likelihood (cl) score does not admit a UMP test. This insight suggests that the additional information retained by our censoring approach translates into advantageous power properties. In applied work, the comparative performance between candidate distributions is often evaluated using the framework developed by Giacomini and White (2006), utilizing a Diebold and Mariano (2002) type test statistic, henceforth referred to as DM test. However, conducting a power analysis for this test becomes intricate, as the null hypothesis, asserting that the expected score difference between the candidate distributions with respect to the underlying distribution is zero, depends on the scoring rule being employed. Resorting to a Monte Carlo study based on specific scoring rules, candidate distributions and weight functions, we obtain strong evidence in favor of our censoring approach. Additionally, we revisit the size experiment originally proposed by Diks et al. (2011), corroborating that all evaluated tests are size correct.

The empirical part of our paper focuses on three different domains: finance, specifically extreme portfolio losses of the S&P500; macroeconomics, centering on inflation rates both far from and near the target; and meteorology, examining high and agriculturally-optimal temperatures. In each of these studies, we employ a collection of candidate methods, which are subjected to the Model Confidence Set (MCS) procedure as delineated by Hansen et al. (2011). Notably, a higher power in the DM test corresponds to a smaller MCS; therefore, we

use the size of the MCS as a proxy for power in our experiments. The results overwhelmingly favor the censored scoring rules, as evidenced by the typically (much) smaller MCS across the different applications.

Our research primarily contributes to the literature on focused scoring rules, which starts with the weighted likelihood score of Amisano and Giacomini (2007), which simply multiplies the unweighted logarithmic scoring rule by a weight function. However, as pointed out by Diks et al. (2011) and Gneiting and Ranjan (2011), this method leads to improper scoring rules because it favors distributions that allocate more mass to regions with higher weights, irrespective of the true underlying distribution. To address this, Gneiting and Ranjan (2011) formulated the twCRPS, while Diks et al. (2011) introduced the cl and csl rules. Holzmann and Klar (2017, Theorem 1) expanded upon the idea of using the conditional likelihood by offering a general procedure for focusing regular scoring rules, applying the regular scoring rule to a conditioning-type transformation of the original distribution. Our work diverges from theirs in the specific transformation applied to the original distribution: we utilize a censored distribution as opposed to a conditional one. This distinction bears significant impact: our censoring-based mechanism is the sole approach guaranteed to yield strictly locally proper scoring rules. Moreover, we note that the conditioning framework put forth by Holzmann and Klar (2017, Theorem 1) can also be established from a generalization of the weighted log-likelihood scoring rule proposed by Amisano and Giacomini (2007), modified by a 'properization' transformation as delineated by Brehmer and Gneiting (2020, Theorem 1). Consequently, 'properization' is not a viable mechanism for achieving strict local propriety either.

Our research also rests upon a substantial body of research concerning strictly proper scoring rules and their associated divergence measures. While the formalization of strict

propriety was rigorously achieved by Gneiting and Raftery (2007), scoring rules satisfying this property date back to at least the Quadratic Scoring rule by Brier (1950). Literature in this domain has evolved from an initial focus on discrete settings (Good, 1952; Toda, 1963; Roby, 1964; Good, 1971; Shuford et al., 1966; Savage, 1971; Selten, 1998; Jose, 2009), to the more general scope of Gneiting and Raftery (2007). 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. Additionally, scoring rules are inherently interconnected with divergence measures; under the banner of strict propriety, these measures are subsumed under Bregman divergences (Dawid, 2007; Gneiting and Raftery, 2007; Ovcharov, 2018; Painsky and Wornell, 2019). This effectively excludes f-divergences other than Kullback-Leibler divergence (Kullback and Leibler, 1951), distinguished for its favorable properties (Liese and Vajda, 2006).

Interest in targeting specific regions of predictive distributions has surged across diverse fields, underscored by analyses of extreme events in disciplines such as meteorology, climatology, hydrology, finance, and economics (Lerch et al., 2017). In the sphere of financial risk management, attention is particularly concentrated on the left tail of loss distributions, conforming to mandated risk metrics like Value-at-Risk (VaR) and Expected Shortfall (ES) (Fissler et al., 2015; Nieto and Ruiz, 2016). Analogously, in macroeconomic frameworks, concepts such as 'Inflation at Risk' and 'GDP at Risk' are emerging, signifying values that deviate significantly from benchmarks established by institutions like Central Banks (Lopez-Salido and Loria, 2022; Iacopini et al., 2023). In other scenarios, the emphasis might rest on the central region or on a specific subset of the distribution, often dictated by external constraints or objectives. Examples range from optimizing growing conditions for specific crops like tubers, to calibrating wind speeds for peak wind turbine performance,

maintaining optimal reservoir levels for hydroelectric power generation, managing queue lengths in retail settings for enhanced customer service, and regulating blood sugar levels for effective diabetes management. 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 forecast outcomes.

The remainder of this paper is organized as follows. Section 2 lays the groundwork by introducing the foundational concepts essential for the subsequent analysis. Section 3 formally defines the Generalized Censored Scoring Rule and establishes the conditions under which it is strictly locally proper. This section also showcases a variety of examples and debuts the Z-Q-Randomization procedure, proven to be equivalent to the Generalized Censored Scoring Rule. It concludes with a generalization of the Neyman-Pearson Lemma. Section 3.5 contains Monte Carlo studies comparing the size and power of tests evaluating the equal predictive ability of conditional and censored scoring rules. Section 4 discusses the results of our empirical applications. Section 5 concludes.

2 SCORING RULES

2.1 Regular scoring rules

Consider a random variable $Y : \Omega \to \mathcal{Y}$ 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 distributions on $(\mathcal{Y}, \mathcal{G})$. A *scoring rule* S assigns numerical values (scores) to observations $y \in \mathcal{Y}$ and distributions $\mathrm{F} \in \mathcal{P}$, through a mapping $S : \mathcal{Y} \times \mathcal{P} \to \mathbb{R} \cup \{-\infty\}$. Following Holzmann and Klar (2017), we assume that any scoring rule S is measurable with respect to $\mathcal{G}$ and quasi-integrable with respect to all $\mathrm{P} \in \mathcal{P}$, for all $\mathrm{F} \in \mathcal{P}$, and such that $\mathbb{E}_{\mathrm{P}} S(\mathrm{F}, Y) < \infty$

and $\mathbb{E}_{\mathrm{P}} S(\mathrm{P}, Y) \in \mathbb{R}, \forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$. The latter condition guarantees that the *score divergence* $\mathbb{D}_S(\mathrm{P}\|\mathrm{F}) := \mathbb{E}_{\mathrm{P}} S(\mathrm{P}, Y) - \mathbb{E}_{\mathrm{P}} S(\mathrm{F}, Y)$, 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).

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

Equivalently, a score divergence is a divergence measure (see e.g. Eguchi (1985)) if and only if S strictly proper. For distributions on $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$, where $\mathcal{B}(\mathcal{Y})$ denotes the Borel σ-algebra on $\mathcal{Y}$, this divergence is known to be a Bregman divergence (Bregman, 1967) under the conditions listed by Ovcharov (2018). Two remarks are in place. First, comparisons of candidates $\mathrm{F} \in \mathcal{P}$ are in terms of P-expectations, whence it follows that uniqueness of members in $\mathcal{P}$ should formally be interpreted as a P-a.s. equivalence class of P. For clarity, we omit technicalities about P-a.s. equivalence. Second, if there exists a σ-finite measure $(\mathcal{Y}, \mathcal{G})$ such that $\mathrm{F} \ll \mu,\ \forall \mathrm{F} \in \mathcal{P}$ then scoring rules and associated definitions and results can easily be formulated relative to the class of induced μ-densities $f = \frac{\mathrm{dF}}{\mathrm{d}\mu}$, also denoted by $\mathcal{P}$.

In their review paper, Gneiting and Raftery (2007) provide an abundant list of strictly proper scoring rules, which can be divided into two categories: *local* scoring rules and *distance sensitive* scoring rules (Ehm and Gneiting, 2012). We use the same structure when discussing examples, yet allowing local scoring rules, henceforth called *semi-local*, to also depend on the density via a global norm of the density. Within this subcategory, our focus lies on the Logarithmic (LogS) (Good, 1952; Toda, 1963), Quadratic (QS) (Brier, 1950) and Spherical (SphS) (Roby, 1964; Good, 1971) scoring rules, along with their extensions to the Power (PowS_α) and PseudoSpherical (PsSphS_α) families. Our choice of distance-sensitive

scoring rules is confined to the Energy Scores (ES) subfamily, a subclass of the class of strictly proper scoring rules given by Theorem 5 of Gneiting and Raftery (2007), nesting the real-valued Continuously Ranked Probability Score (CRPS) (Matheson and Winkler, 1976; Hersbach, 2000) as a special case.

2.2 Weighted scoring rules

Example 1 (The need to focus)**.** *Let Y be a random variable that follows a piecewise uniform distribution across the intervals $A = [0, 1)$, $B = [1, 2)$ and $C = [2, 3]$, with probabilities π_A, π_B and π_C, respectively. The parameters of the true distribution* P *and two candidates* F *and* G *are detailed in Table 1, with corresponding CDFs displayed in Figure 1. Consider the CRPS, which is strictly proper and has score divergence $\mathbb{D}_{CRPS}(\mathrm{F}||\mathrm{G}) = \int_0^3 \big(F(s) - G(s)\big)^2 \mathrm{d}s$. From Figure 1 it is obvious that $\mathbb{D}_{CRPS}(\mathrm{P}||\mathrm{F}) > \mathbb{D}_{CRPS}(\mathrm{P}||\mathrm{G})$. But then $\mathbb{D}_{CRPS}$ becomes useless if only observations in B are pertinent, as* F *coincides with* P *on B, that is, $\mathrm{P}(E \cap B) = \mathrm{F}(E \cap B), \forall E \in \mathcal{G}$, in contrast to* G*.*

	π_A	π_B	π_C
P	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$
F	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$
G	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$

Table 1: Parameters

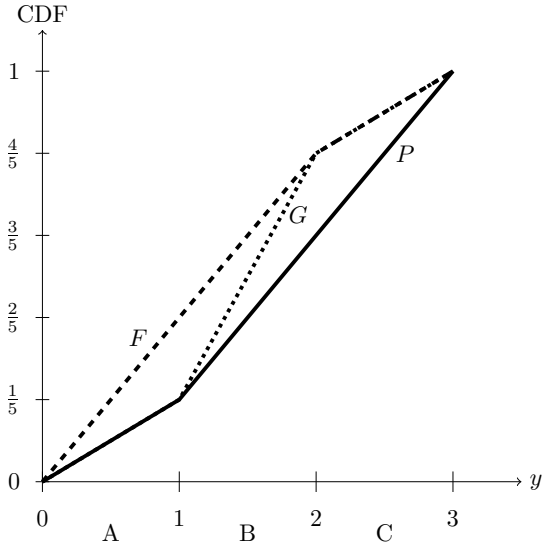

Figure 1: Distribution functions

As demonstrated in Example 1, it becomes imperative to adapt the scoring rule when

particular outcomes are of particular importance. Without such adaptation, an excellent fit in non-critical regions of the outcome space may obscure a poor fit in regions of actual relevance. Modeling the relative importance of outcomes $y \in \mathcal{Y}$ by a *weight function* $w \in \mathcal{W}$, defined as any $\mathcal{G}$-measurable mapping $w : \mathcal{Y} \to [0, 1]$, the question arises how to the transform the original scoring rule S by this weight function. We concur with the arguments put forward by Holzmann and Klar (2017) that the weighted scoring rule S_w must be *localizing*. Specifically, for all outcomes, the variation in S_w should be solely dependent on changes in the distribution within the region of interest $\{w > 0\} := \{y \in \mathcal{Y} : w(y) > 0\}$. This concept is formalized in Definition 2, borrowed from Holzmann and Klar (2017). If a weighted scoring rule is non-localising, this may cause a so-called localization bias, as illustrated by Example 2.

Definition 2 (Localizing weighted scoring rule)**.** *A weighted scoring rule S, that is, a map $S : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$ such that $S_w(\cdot, \cdot)$ is a scoring rule for each $w \in \mathcal{W}$, is localizing if for any P, $\mathrm{F} \in \mathcal{P}$, $w \in \mathcal{W}$, it holds that*

$$\forall E \in \mathcal{G} : \mathrm{P}(\{w > 0\} \cap E) = \mathrm{F}(\{w > 0\} \cap E) \implies S_w(\mathrm{P}, y) = S_w(\mathrm{F}, y), \ \forall y \in \mathcal{Y}.$$

Example 2 (Localization bias)**.** *Revisit the case of the random variable Y with actual distribution P and candidates F and G as described in Example 1. Assume the region of interest is B, modeled by the weight function $w(y) = \mathbb{1}_B(y)$. A prevalent weighted extension of the CRPS is given by $twCRPS(F, y) = \int_B \big(F(s) - \Delta_y(s)\big)^2 \mathrm{d}s$, with score divergence $\mathbb{D}_{twCRPS}(\mathrm{P}\|\mathrm{F}) = \int_B \big(F(s) - G(s)\big)^2 \mathrm{d}s$. This weighted variant of the CRPS is clearly non-localizing. For instance, because its value is influenced by $\mathrm{F}(A)$, while $\mathrm{F}(A)$ is not implied by $\mathrm{F}(B)$; only the sum $\mathrm{F}(A) + \mathrm{F}(C)$ is. Consequently, the scoring rule depends on the distribution F outside B in a way that is not implied by F restricted to B. Its failure to be localizing introduces a bias in evaluating distributions over B. Indeed, by suggesting that*

G *is statistically closer to* P *than* F *is, while* F *coincides with* P *on* B *(see Figure 1), the twCRPS inappropriately favors* G.

Example 3 (Locally improper)**.** *Returning to the context of Example 1, we examine the weighted likelihood score* $wl(f, y) = \log f(y)\mathbb{1}_B(y)$ *proposed by Amisano and Giacomini (2007). Although this scoring rule is localizing and the unweighted logarithmic scoring rule strictly proper, it still inappropriately favors* G. *Specifically, we have* $\log g(y) > \log f(y), \forall y \in B$, *irrespective of* $p(y)$, *leading to* $\mathbb{D}_{wl}(\mathrm{F}\|\mathrm{G}) > \mathbb{D}_{wl}(\mathrm{P}\|\mathrm{G})$.

Example 3 illustrates that localizing versions of strictly proper scoring rules are not automatically proper for all weight functions. In light of this, we focus on the specialized subclass of localizing scoring rules that consistently maintain this property. By construction, a localizing weighted scoring cannot be strictly proper, unless $w(y) > 0, \forall y \in \mathcal{Y}$. This is because any distribution $\tilde{\mathrm{P}}$ equivalent to P on $\{w > 0\}$ but different on $\{w = 0\}$ will receive an identical score. Nonetheless, as emphasized by Example 4 some notion of strictness remains advantageous.

Example 4 (Proportionally locally proper)**.** *Consider the family of weighted scoring rules*

$$S_w^\sharp(\mathrm{F}, y) := w(y)S(\mathrm{F}_w^\sharp, y), \qquad \mathrm{dF}_w^\sharp := \frac{1}{1 - \bar{F}_w}\mathrm{dF}_w,$$

proposed by Holzmann and Klar (2017), where S *is a regular scoring rule,* $\mathrm{dF}_w := w\mathrm{dF}$ *the weighted kernel of distribution* F *and* $\bar{F}_w = \int_{\mathcal{Y}}(1 - w)\mathrm{dF}$. *This scoring rule is localizing and proper for weight functions for which it remains a scoring rule (see Section 2.1). Yet, when revisiting the setup of Example 1 for* $w(y) = \mathbb{1}_B(y)$, *we trivially have that* $S_B^\sharp(\mathrm{F}, y) = S_B^\sharp(\mathrm{G}, y) = S_B^\sharp(\mathrm{P}, y), \forall y \in B$, *since* $S_w^\sharp$ *cannot discriminate between distributions that are proportional to each other on* $\{w > 0\}$. *Accordingly,* $\mathbb{D}_{S_B^\sharp}(\mathrm{P}\|\mathrm{F}) = \mathbb{D}_{S_B^\sharp}(\mathrm{P}\|\mathrm{G}) = 0$, *while only* F *coincides with* P *on* B. *In other words, the score divergence of a candidate from*

P *is properly zero if (but not only if) the candidate coincides with* P *on* B*, as is the case for* F.

Motivated by Examples 2, 3 and 4, this paper posits the necessity for weighted scoring rules to be *strictly locally proper*, as articulated in Definition 3. Compared to the definition of strict propriety (Definition 1), strictness is only required locally. More precisely, equivalent distributions on $\{w > 0\}$ must have weighted score divergence zero and, vice versa, distributions at zero weighted score divergence of each other must be equivalent on $\{w > 0\}$. The latter ruling out the ambiguities highlighted in Example 4.

Definition 3 ((Strictly) locally proper scoring rule)**.** *A weighted scoring rule* $S : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$ *is locally proper relative to* $(\mathcal{P}, \mathcal{W})$ *if it is localizing and* $S_w(\cdot, \cdot)$ *is proper for each* $w \in \mathcal{W}$*. Furthermore, it is strictly locally proper relative to* $(\mathcal{P}, \mathcal{W})$ *if, additionally,*

$$\mathrm{P}(\{w > 0\} \cap E) = \mathrm{F}(\{w > 0\} \cap E), \forall E \in \mathcal{G} \iff \mathbb{D}_{S_w}(\mathrm{P} \| \mathrm{F}) = 0, \ \forall w \in \mathcal{W}.$$

3 THE CENSORED SCORING RULE

To overcome issues like the non-strictness and non-locality of the weighted scoring rules discussed above, we propose to use censoring as focusing mechanism. Censoring (Bernoulli, 1760) is a statistical concept that is used in econometrics to model a dependent variable whose value is only partially known (Tobin, 1958). More specifically, for realisations in A^c, the complement of A, it is only known that they are not in A. Events in A^c are hence indistinguishable after censoring and 'A^c' could therefore be viewed as a single outcome of the censored random variable. To avoid confusion, we label observations in A^c by '$*$' rather than 'A^c' itself, which is nothing but an abstract event for which one can alternatively read

'NaN'. The censored random variable

$$Y_A^\flat = \begin{cases} Y, & Y \in A, \\ \\ *, & Y \in A^c, \end{cases}$$

is defined relative to the extended measurable space $(\mathcal{Y}^*, \mathcal{G}^*)$, where $\mathcal{Y}^* = \mathcal{Y} \cup \{*\}$ and $\mathcal{G}^* = \sigma(\{\mathcal{G}, *\})$, that is, the smallest σ-algebra containing the collection $\{\mathcal{G}, *\}$. Similar to the conditional distribution in Example 4, we extend the definition of the distribution of $Y_A^\flat$ to general weight functions $w \in \mathcal{W}$. Specifically, we define the *censored distribution* as

$$\mathrm{dF}_w^\flat = \mathrm{dF}_w + \bar{\mathrm{F}}_w \mathrm{d}\delta_*, \qquad \bar{F}_w := \int_{\mathcal{Y}} (1 - w) \mathrm{dF}, \qquad w \in \mathcal{W},\ \mathrm{F} \in \mathcal{P}, \tag{1}$$

where δ_* denotes the Dirac measure at $*$, i.e. $\delta_*(E) = \mathbb{1}_E(*)$. To make this change of measure well-defined, we consider the original measures $\mathrm{F} \in \mathcal{P}$ relative to the extended measurable space $(\mathcal{Y}^*, \mathcal{G}^*)$, by defining $\mathrm{F}(*) = 0$ and taking some value for $w(*)$. In case $\mathrm{F} \ll \mu, \forall \mathrm{F} \in \mathcal{P}$, we are invited to work with the μ-densities $f \in \mathcal{P}$ instead, and their associated $(\mu + \delta_*)$-densities

$$f_w^\flat = w f \mathbb{1}_{y \neq *} + \bar{F}_w \mathbb{1}_{y = *}, \qquad w \in \mathcal{W},\ f \in \mathcal{P}. \tag{2}$$

A detailed proof of this result is deferred to Appendix B.1. Albeit restricted to $w(y) = \mathbb{1}_A(y)$, Borowska et al. (2020) also work with an explicit formulation of the censored density, coinciding with $f_A^\flat$ in the context of maximum likelihood. To ease notation, we consistently use the subscript A instead of $\mathbb{1}_A$ in indicator function references.

Ideally, the censored scoring rule would be given by $S_A^\flat(\mathrm{F}, y) = S(\mathrm{F}_A^\flat, y_A^\flat)$, as this would fully respect the forecaster's specific choice of the regular scoring rule S. The censored scoring rule given by Definition 4 reduces to this definition for the indicator weight function $w(y) = \mathbb{1}_A(y)$. The censored scoring rule is also attractive for general weight functions, but this will be particularly clear from the randomization perspective taken in Section 3.2,

which yields a similar identity for general weight functions; see Equation (4). According to Theorem 1, the censored scoring rule is strictly locally proper. Since Theorem 1 is a corollary of Theorem 2, we have sustainably omitted a proof for this result.

Definition 4 (Censored scoring rule)**.** *Let* $S : \mathcal{P}^\flat \times \mathcal{Y} \to \bar{\mathbb{R}}$, $\mathcal{P}^\flat = \{\mathrm{F}_w^\flat, \mathrm{F} \in \mathcal{P}, w \in \mathcal{W}\}$, *denote a scoring rule. Then, the corresponding censored scoring rule is given by the map* $S^\flat : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$,

$$S_w^\flat(\mathrm{F}, y) := w(y)S(\mathrm{F}_w^\flat, y) + \big(1 - w(y)\big)S(\mathrm{F}_w^\flat, *),$$

where the censored distribution $\mathrm{F}_w^\flat$ *is defined in Equation* (1)*.*

Theorem 1. *Suppose that the regular scoring rule* S *is strictly proper relative to* $\mathcal{P}^\flat$*. Then, the censored scoring rule* $S^\flat$ *in Definition 4 is strictly locally proper relative to* $(\mathcal{P}, \mathcal{W})$*.*

The assumption in Theorem 1 ensures that the scoring rule is well-defined with respect to mixed continuous-discrete distributions on $(\mathcal{Y}^*, \mathcal{G}^*)$. We will verify that this assumption holds in the examples discussed in Subsection 3.3.

Let us conclude this section by providing some intuition for the result of Theorem 1. Given some weight function $w \in \mathcal{W}$, it should be clear that censoring maintains a one-to-one connection with the original distribution on $\{w > 0\}$. This relation can be harmed by conditioning due to the additional normalisation of the weighted kernel. This difference is even clearer for indicator weight functions since $\mathrm{F}_A^\flat = \mathrm{F}$, while $\mathrm{F}_A^\sharp \neq \mathrm{F}$, on A. Because of this, only the censored scoring rule allows for identifying the original distributions on $\{w > 0\}$ when comparing two candidates F and G. This additionally requires disentanglabity of the weighted kernels and discrete probabilities in the censored measures, implied by $\mathrm{F}_w(*) = \mathrm{G}_w(*) = 0$. Consequently, the assumed strict propriety of the original rule localizes to $\{w > 0\}$ for the censored scoring rule.

3.1 Generalized censored scoring rule

Given the intuition at the end of the previous section, it is not entirely surprising that one can perform other transformations to the distribution on $\{w > 0\}$ as long as the transformation is independent of the distribution and traceable when comparing two candidate distributions. The latter requirement is formalized by Assumption 1, under which the *generalized censored scoring rule* in Definition 5 is still strictly locally proper. Appendix A.1 details a proof for this result, summarized by Theorem 2.

Definition 5 (Generalized censored scoring rule)**.** *Let $S : \mathcal{P}^\flat \times \mathcal{Y} \to \bar{\mathbb{R}}$ denote a scoring rule. The associated generalized censored scoring rule is given by the map $S^\flat : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \times \mathcal{H} \to \bar{\mathbb{R}}$,*

$$S^\flat_{w,\mathrm{H}}(\mathrm{F}, y) = w(y)S(\mathrm{F}^\flat_{w,\mathrm{H}}, y) + \big(1 - w(y)\big)\mathbb{E}_{\mathrm{H}} S(\mathrm{F}^\flat_{w,\mathrm{H}}, \cdot), \quad \mathrm{dF}^\flat_{w,\mathrm{H}} = \mathrm{dF}_w + \bar{F}_w \mathrm{dH},$$

where $\mathrm{F}^\flat_{w,\mathrm{H}}$ is referred to as the generalized censored distribution of F*.*

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

Theorem 2. *Suppose that (i) the regular scoring rule S in Definition 5 is strictly proper relative to $\mathcal{P}$, and (ii) $\mathcal{W}$ and $\mathcal{H}$ are such that Assumption 1 is satisfied. Then, the generalized censored scoring rule $S^\flat$ in Definition 5 is strictly locally proper relative to $(\mathcal{P}, \mathcal{W}, \mathcal{H})$.*

Finally, a corollary of Lemma 2 in the proof of Theorem 2 in Appendix A.1 is that

$$\mathbb{D}_{S^\flat_{w,\mathrm{H}}}(\mathrm{F} \| \mathrm{G}) = \mathbb{D}_S(\mathrm{F}^\flat_{w,\mathrm{H}} \| \mathrm{G}^\flat_{w,\mathrm{H}}), \tag{3}$$

i.e. the censored score divergence from F to G is the score divergence of the corresponding censored distributions. In particular, this means that we have identified a family of so-called *localized divergence measures*, satisfying the properties of a divergence measure (see

Subsection 2.1) on $\{w > 0\}$. Indeed, if S is strictly proper, such that $\mathbb{D}_S$ is a divergence measure, it follows that $\mathbb{D}_{S^{\flat}_{w,\mathrm{H}}}(\mathrm{F}\|\mathrm{G}) \geq 0$, with strict equality if and only if $\mathrm{F}(E \cap \{w > 0\}) = \mathrm{G}(E \cap \{w > 0\})$, $\forall E \in \mathcal{G}$.

3.2 Z, Q-Randomization

The (generalized) censored scoring rule in Definition 4 (5) of the previous section can alternatively be formulated in terms of a randomization procedure. This procedure relies on an auxiliary random variable Z_w, indicating, conditional on the realisation y, whether the observation is censored or not. More specifically, we let

$$y^{\flat}_{Z_w} = \varphi(y, Z_w), \qquad \varphi(y, Z_w) := \begin{cases} y, & Z_w = 1, \\ *, & Z_w = 0, \end{cases}$$

where $Z_w|(Y = y) \sim \mathrm{BIN}\big(1, w(y)\big)$. By working out the conditional expectation, it is obvious that $Y^{\flat}_w = \mathbb{E}_{Z_w|(Y=Y)}\varphi(Y, Z_w)$ coincides with the specification of the censored random variable in Equation (1). For $w(y) = \mathbb{1}_A(y)$, the random variable Z_A degenerates to being one if $y \in A$ and zero otherwise, so that $Y^{\flat}_{Z_A} = Y^{\flat}_A$ with probability one. Correspondingly, the *Z-randomization* definition of the censored scoring rule reads

$$S^{\flat}_w(\mathrm{F}, y) = \mathbb{E}_{Z_w|Y=y} S(\mathrm{F}^{\flat}_w, y^{\flat}_{Z_w}), \tag{4}$$

which is equivalent to the censored scoring rule defined by Definition 4.

A similar line of reasoning holds for the generalized censored scoring rule. In addition to the auxiliary random variable Z_w, we introduce an independent random variable Q with distribution H. Rather than labeling the observation as censored, we now take a random

draw from Q if $Z_w = 1$, i.e. we define

$$y^\flat_{\mathrm{H},w} := \varphi_{w,\mathrm{H}}(y, Z_w, Q), \qquad \varphi_{w,\mathrm{H}}(y, Z_w, Q) := \begin{cases} Y, & \text{if } Z_w = 1, \\ Q, & \text{if } Z_w = 0. \end{cases}$$

As anticipated, the distribution of $Y^\flat_{\mathrm{H},w} = \mathbb{E}_{Z_w|(Y=Y),\mathrm{H}}\varphi(Y, Z_w)$ coincides with the specification of $\mathrm{F}^\flat_{w,\mathrm{H}}$ in Equation (1). Additionally, the generalized censored scoring rule of Definition 5 admits the *Z, Q-randomization* representation

$$S^\flat_{\mathrm{H},w}(\mathrm{F}, y) = \mathbb{E}_{Z_w|(Y=y),\mathrm{H}} S\big(\mathrm{F}^\flat_{w,\mathrm{H}}, y^\flat_{\mathrm{H},w}\big).$$

The randomization perspective further clarifies why $S^\flat_{\mathrm{H},w}$ generalizes $S^\flat_w(\mathrm{F}, y)$. Indeed, by choosing a degenerate distribution for Q at $*$, each 'random draw' from Q will be precisely the censoring label $*$ of the Z-randomization procedure. Put differently, $S^\flat_{\mathrm{H},w} = S^\flat_w(\mathrm{F}, y)$ for $\mathrm{H} = \delta_*$.

3.3 Examples

3.3.1 Semi-local scoring rules

We will now apply our censoring framework to the regular scoring rules introduced in Subsection 2.1. Following the classification into semi-local and distance-sensitive scoring rules, we start with localizing the former class. Together with the main characteristics of the LogS, PowS_α and PsSphS_α families, Table 2 presents the localized versions of these families based on conditioning, censoring and generalized censoring. Given the strict propriety classes in Table 2, one can easily verify their strict propriety with respect to $\mathcal{P}^\flat_\alpha$ since $\|f^\flat_w\|^\alpha_\alpha \leq 1 + \|f\|^\alpha_\alpha < \infty$, $\forall f \in \mathcal{P}_\alpha$, $\forall w \in \mathcal{W}$, where $\alpha = 1$ for LogS. Furthermore, the Bregman generator functions $\zeta(t)$ refer to the well-known subclass of *separable Bregman divergences*, consisting of the score divergences based on strictly proper scoring rules S_ζ :

$\mathcal{P}(\mathcal{Y}, \mathcal{G}) \times \mathcal{Y} \to \mathbb{R}$ of the form

$$S_\zeta(p, y) = \zeta'(p(y)) - \int_{\mathcal{Y}} \zeta'(p(y))p(y) - \zeta(p(y))\mu(\mathrm{d}y).$$

Comparing the censored and conditioned versions of the rules, we notice that the censored variants bear an isolated $\bar{F}_w$-dependent term, preserving the coverage probability of $\{w = 0\}$. While preserving the likelihood $\bar{F}_w$ of being censored, Table 2 also shows that the censored scoring rules are independent of $*$, the label of a censored observation. Hence, for this selection of scoring rules, one could alternatively work with an actual number like r for the location of the residual probability $\bar{F}_w$. Strictly speaking, we need to require $\mathrm{F}_w(r) = 0$ in that case, to keep the censored scoring rule strictly locally proper (see Assumption 1), but this is trivially met by restricting to either continuous measures or weight functions satisfying $w(r) = 0$, or both. The generalized censored scoring rules in Table 2 show that the invariance with respect to the location of the discrete probability mass holds more generally. In particular, the generalized censored scoring rules turn out to be entirely invariant to the choice of the nuisance density on $\{w = 0\}$ upon normalisation by the α-norm of h, i.e. to the class of densities $\tilde{h} = h/\|h\|_\alpha$, where $\alpha = 1$ for LogS. Since $\|h\|_1 = 1$, the latter means that LogS is invariant to the unnormalised choice of h, as can be seen from Table 2. Finally, Table 2 includes the localized divergence measures $\mathbb{D}_{S_w^\flat}$, which are all localized Bregman divergences since all regular divergences $\mathbb{D}_S$ in this table are Bregman.

3.3.2 Distance sensitive scoring rules

A rich class of distance-sensitive scoring rules is the Energy Score family

$$\mathrm{ES}_\beta(\mathrm{F}, y) = \frac{1}{2}\mathbb{E}_\mathrm{F}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|_2^\beta - \mathbb{E}_\mathrm{F}\|\mathbf{Y} - \mathbf{y}\|_2^\beta, \qquad \beta \in (0, 2),$$

known to be strictly proper to the class of Borel probability measures on $\mathbb{R}^d$ such that $\mathbb{E}_F\|\mathbf{Y}\|_2^\beta < \infty$ (Gneiting and Raftery, 2007). From this expression, it is immediately clear

Table 2: Examples semi-local scoring rules

Name	Logarithmic	Power family	PseudoSpherical family
		Regular	
$S(f,y)$	$\mathrm{LogS}(f,y)=\log f(y)$	$\mathrm{PowS}_\alpha(f,y)=\alpha f(y)^{\alpha-1}-(\alpha-1)\|f\|_\alpha^\alpha$	$\mathrm{PsSphS}_\alpha(f,y)=\frac{f(y)^{\alpha-1}}{\|f\|_\alpha^{\alpha-1}}$
Special cases		$\mathrm{QS}(f,y)=\mathrm{PowS}_2(f,y)$	$\mathrm{SphS}(f,y)=\mathrm{PsSphS}_2(f,y)$
		$\mathrm{LogS}(f,y)=\lim_{\alpha\downarrow 1}\mathrm{PowS}_\alpha(f,y)$	$\mathrm{LogS}(f,y)=\lim_{\alpha\downarrow 1}\mathrm{PsSphS}_\alpha(f,y)$
$\mathbb{H}_S(f)$	$\mathbb{E}_f\log f$	$\|f\|_\alpha^\alpha$	$\|f\|_\alpha$
$\mathbb{D}_S(f\|g)$	$\mathrm{KL}(f\|g)=\mathbb{E}_f\log\left(\frac{f}{g}\right)$	$\|f\|_\alpha^\alpha-\alpha\int g^{\alpha-1}(f-g)\mathrm{d}\mu-\|g\|_\alpha^\alpha$	$\|f\|_\alpha-\frac{\int fg^{\alpha-1}\mathrm{d}\mu}{\|g\|_\alpha^{\alpha-1}}$
$\alpha=2$	–	$\|f-g\|_2^2$	$\|f\|_2\,(1-C(f,g))$
SP class	$\mathcal{P}_1$	$\mathcal{P}_\alpha$	$\mathcal{P}_\alpha$
$\zeta(t)$	$t\log t$	t^α	–
$S(\tilde{f},\tilde{y})$	$\log f(y)-\log\|b\|$	$\left(\frac{1}{\|b\|}\right)^{\alpha-1}\mathrm{PowS}_\alpha(f,y)$	$\left(\frac{1}{\|b\|}\right)^{\frac{\alpha-1}{\alpha}}\mathrm{PsSphS}_\alpha(f,y)$
		Focused	
$S_w^\sharp(f,y)$	$w(y)\log\left(\frac{f(y)}{1-\bar{F}_w}\right)$	$w(y)\left(\alpha\left(\frac{f_w(y)}{1-\bar{F}_w}\right)^{\alpha-1}-(\alpha-1)\left\|\frac{f_w(y)}{1-\bar{F}_w}\right\|_\alpha^\alpha\right)$	$w(y)\frac{f_w(y)^{\alpha-1}}{\|f_w\|_\alpha^{\alpha-1}}$
$S_w^\flat(f,y)$	$w(y)\log f(y)+\big(1-w(y)\big)\log\bar{F}_w$	$w(y)\alpha f_w(y)^{\alpha-1}+\big(1-w(y)\big)\alpha\bar{F}_w^{\alpha-1}$ $-(\alpha-1)\left(\|f_w\|_\alpha^\alpha+\bar{F}_w^\alpha\right)$	$\frac{w(y)f_w(y)^{\alpha-1}+(1-w(y))\bar{F}_w^{\alpha-1}}{\left(\|f_w\|_\alpha^\alpha+\bar{F}_w^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$
$S_{w,h}^\flat(f,y)$	$w(y)\log f(y)+\big(1-w(y)\big)\log\bar{F}_w$	$w(y)\alpha f_w(y)^{\alpha-1}+\big(1-w(y)\big)\alpha\bar{F}_w^{\alpha-1}\|h\|_\alpha^\alpha$ $-(\alpha-1)\left(\|f_w\|_\alpha^\alpha+\bar{F}_w^\alpha\|h\|_\alpha^\alpha\right)$	$\frac{w(y)f_w(y)^{\alpha-1}+\big(1-w(y)\big)\bar{F}_w^{\alpha-1}\|h\|_\alpha^\alpha}{\left(\|f_w\|_\alpha^\alpha+\bar{F}_w^\alpha\|h\|_\alpha^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$
$\mathbb{H}_{S_w^\flat}(f)$	$\int\log(f)f_w\mathrm{d}\mu+\log\left(\bar{F}_w\right)\bar{F}_w+\int\log(w)f_w\mathrm{d}\mu$	$\|f_w\|_\alpha^\alpha+\bar{F}_w^\alpha$	$\left(\|f_w\|_\alpha^\alpha+\bar{F}_w^\alpha\right)^{\frac{1}{\alpha}}$
$\mathbb{D}_{S_w^\flat}(f\|g)$	$\int\log\left(\frac{f}{g}\right)f_w\mathrm{d}\mu+\log\left(\frac{\bar{F}_w}{\bar{G}_w}\right)\bar{F}_w$	$\|f_w\|_\alpha^\alpha+\bar{F}_w^\alpha-\int f_w g_w^{\alpha-1}\mathrm{d}\mu-\bar{F}_w\bar{G}_w^{\alpha-1}$ $-(\alpha-1)\big(\|g_w\|_\alpha^\alpha+\bar{G}_w^\alpha\big)$	$\left(\|f_w\|_\alpha^\alpha+\bar{F}_w^\alpha\right)^{\frac{1}{\alpha}}-\frac{\int f_w g_w^{\alpha-1}\mathrm{d}\mu+\bar{F}_w\bar{G}_w^{\alpha-1}}{\left(\|g_w\|_\alpha^\alpha+\bar{G}_w^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$

NOTE: $C(f,g)=\int fg\mathrm{d}\mu/\sqrt{\int f^2\mathrm{d}\mu\int g^2\mathrm{d}\mu}$ and $\mathbb{D}_S(f||g)$ denote the cosine similarity and the score divergence between f and g, respectively. $\mathbb{H}_S(f)$ denotes the negative entropy and the PowS_α and PsSphS_α families are restricted to $\alpha>1$. Furthermore, $\mathcal{P}_\alpha$ denotes the space for which the L^α-norm is finite, where μ is measure relative to which the densities p and f are defined, i.e. $\frac{\mathrm{dF}}{\mathrm{d}\mu}$, with $\mathrm{F}\ll\mu$. The common limiting case of the PowS_α and PsSphS_α remains to hold for conditioning and censoring, i.e. $\lim_{\alpha\downarrow1}\frac{1}{\alpha-1}\mathrm{PsSphS}_{\alpha,w}^x(f,y)=\lim_{\alpha\downarrow1}\frac{1}{\alpha-1}\mathrm{PowS}_{\alpha,w}^x(f,y)=\mathrm{LogS}^x(f,y)$, $x\in\{\sharp,\flat\}$. The generalized censored distribution $S_{w,h}^\flat$ departs from a density h which support is a subset of $\{w=0\}\subseteq\mathcal{Y}$. The weight function is restricted accordingly.

that the corresponding censored ES family depends, in contrast to the semi-local scoring rules, on $*$, or more particularly, the distance $d(\mathbf{y}) = \|\mathbf{y} - *\|_2$. Specifically,

$$S^{\flat}_{w,d}(\mathrm{F}, \mathbf{y}) = \frac{1}{2}\mathbb{E}_{\mathrm{F}^{\flat}_w}\|\mathbf{Y} - \tilde{\mathbf{Y}}\|_2^{\beta} - \mathbb{E}_{\mathrm{F}^{\flat}_w}\Big(w(\mathbf{y})\|\mathbf{Y} - \mathbf{y}\|_2^{\beta} + \big(1 - w(\mathbf{y})\big)d(\mathbf{Y})^{\beta}\Big).$$

It is unsurprising that distance-sensitive scoring rules are sensitive to the location of the discrete probability $\bar{F}_w$. An easy way to define $d(\mathbf{y})$ is to simply add the location of $\bar{F}_w$ by choosing $* \in \mathbb{R}^d$. It is important, however, to keep in mind that the censored scoring rule is not invariant with respect to this additional piece of information. More precisely, the selected value for $*$, say $\mathbf{r}$, is now not only representing the event of being censored but also the value an observation gets after being censored. In the subsequent discussion, we do not extend the generality by introducing a new distance measure. Instead, we allocate the residual mass $\bar{F}_w$ across a set of 'pivotal points,' denoted as $\mathcal{A} := \{\mathbf{a}_i\}_{i=1}^{n_a}$. This approach is motivated by the empirical observation that weight functions often possess pivotal points, such as the edges of an indicator function or the center of a kernel (Gneiting and Ranjan, 2011).

If the weight function at hand has one pivotal point a_1, i.e. $w(y) = \mathbb{1}_{(-\infty, a_1)}(y)$, then we thus propose to use the censored scoring rule $S^{\flat}_w$ in conjunction with the censored distribution $\mathrm{dF}^{\flat}_w = \mathrm{dF}_w + \bar{F}_w \mathrm{d}\delta_{a_1}$ and if $n_a > 1$, being the special case ($\gamma_1 = 1$) of the general solution to use the generalized censored measure with the censored distribution

$$\mathrm{dF}^{\flat}_w = \mathrm{dF}_w + \bar{F}_w \sum_{i=1}^{n_a} \gamma_i \mathrm{d}\delta_{a_i}, \qquad (\gamma_1, \ldots, \gamma_{n_a})' \in \Delta(n_a) \tag{5}$$

where $\Delta(n_a)$ denotes the unit simplex.

Our approach to distance-sensitive scoring rules has some interesting implications for the CRPS. First of all, for all left- and right-tail indicator functions, $\mathrm{CRPS}^{\flat}_w$ coincides with twCRPS. In other words, $\mathrm{CRPS}^{\flat}_w = \mathrm{twCRPS}$, for all weight functions for which (Holzmann

and Klar, 2017, Theorem 5) proved that the twCRPS is strictly locally proper. Second, for other weight functions like the center indicator functions for which the twCRPS twCRPS loses its strict local propriety due to its non-localizing nature, $\mathrm{CRPS}_w^\flat$ serves as a strictly locally proper alternative. This alternative bears an additional parameter γ, the selection of which is contingent on the specific application. For the indicator function $w(y) = \mathbb{1}_{[-r,r]}$ it makes sense to choose $\gamma = \frac{1}{2}$ if one aims to compare the predictive ability of two candidates that are both symmetric around zero. Moreover, in applications where empirical data are available to estimate residual probabilities based on the DGP, using such data to set γ facilitates a more equitable comparison of candidate performance on A. It is important to note that using the data instead of the candidates to estimate γ, sets a level playing field for the candidates in terms of their performance on A. After all, this approach ensures that the relative performance of the candidates on A is not obscured by the performance outside A (for the part that is not entirely implied by the distribution on A).

Mathematically, we can illustrate the difference between the generalized censored scoring rule based on the censored measure in Equation (5) and the twCRPS as follows. Consider the center indicator function $w(y) = \mathbb{1}_A(y)$, where $A = [a_1, a_2]$. The twCRPS and the censored CRPS are related as follows

$$\mathrm{twCRPS}(\mathrm{F}, y) = \mathrm{CRPS}(\mathrm{F}_w^\dagger, y_w^\dagger), \quad \mathrm{dF}_w^\dagger = \mathrm{dF}_w + \bar{F}_w\big(\gamma_\mathrm{F}\mathrm{d}\delta_{a_1} + (1-\gamma_\mathrm{F})\mathrm{d}\delta_{a_2}\big)$$

where $\gamma_\mathrm{F} = F_{w\mathrm{L}}/\bar{F}_w$, $F_{w\mathrm{L}} = \mathrm{F}(A_\mathrm{L})$, $A_\mathrm{L}^c = (-\infty, a_1)$. Furthermore, $y_w^\dagger = y\mathbb{1}_A(y) + a_1\mathbb{1}_{A_\mathrm{L}^c}(y) + a_2\mathbb{1}_{A_\mathrm{R}^c}(y)$, with $A_\mathrm{R}^c = (a_2, \infty)$, allowing the twCRPS to assign different scores to observations in A_L^c and A_R^c. One critical difference between the generalized censored measure and $\mathrm{F}_w^\dagger$ is that the latter candidate's reference distribution depends on the candidate itself, namely through the dependence of the proportion parameter on F. In expectation, the difference between the twCRPS and the generalized censored scoring rule reduces to

precisely this difference between $\gamma = \mathrm{P}_{w\mathrm{L}}/\bar{P}_w$, where P denotes the underlying distribution of Y, and γ_{F}. Specifically,

$$\mathbb{E}_{\mathrm{P}}\mathrm{twCRPS}(\mathrm{F}, Y) = \mathbb{E}_{\mathrm{P}}\mathrm{CRPS}^{\dagger}_{w}(\mathrm{F}, Y),$$

where the only difference between $\mathrm{CRPS}^{\dagger}_{w}$ and $\mathrm{CRPS}^{\flat}_{w}$ is the dependence on $\mathrm{F}^{\dagger}_{w}$ rather than $\mathrm{F}^{\flat}_{w}$, i.e.

$$\mathrm{CRPS}^{\dagger}_{w}(\mathrm{F}, y) = \begin{cases} \mathrm{CRPS}(\mathrm{F}^{\dagger}_{w}, y), & \text{if } y \in A \\ \gamma\mathrm{CRPS}(\mathrm{F}^{\dagger}_{w}, a_1) + (1-\gamma)(\mathrm{F}^{\dagger}_{w}, a_2), & \text{if } y \in A^c. \end{cases}$$

Unlike the twCRPS, this scoring rule does not depend on whether an observation is in A_{L} or A_{R}.

For the center indicator case, for which the twCRPS is not strictly locally proper and hence not a generalized censored scoring rule, we have now derived the alternative (close to censoring) procedure, which is helpful in two ways. (i) By revealing the recipe for obtaining the twCRPS, we uncovered the multivariate twCPRS for practitioners that are despite the localization bias still willing to use the twCRPS in a multivariate setting. (ii) We have uncovered precisely the difference between the twCRPS and the generalized censored scoring rule, i.e. γ versus γ_{F} in the definition of the focused measure.

3.4 Localized Neyman–Pearson

In anticipation of our favorite applications, we now switch to an explicit time series context. In particular, consider a stochastic process $\{Y_t : \Omega \to \mathcal{Y}\}_{t=1}^{T}$ 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 $\mathcal{G}$, respectively. The process generates the filtration $\{\mathcal{F}_t\}_{t=1}^{T}$, in which $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$ is the

information set at time t, satisfying $\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F},\ \forall t$. We denote predictive distributions of Y_{t+1} based on $\mathcal{F}_t$ by F_t, predictive distribution functions by F_t and predictive μ_t-densities by f_t. The existence of the sequence of densities f_t is implied by the existence of a sequence of measures $\{\mu_t\}$ such that $\mathrm{F}_t \ll \mu_t, \forall t$. Furthermore, the regions of interest $A_t \subseteq \mathcal{Y}$ are always assumed to be $\mathcal{F}_t$-measurable.

The aim of this section is to derive a uniformly most powerful (UMP) test for the following null and alternative hypothesis

$$\mathbb{H}_0 : p_{0t}\mathbb{1}_{A_t} = f_{0t}\mathbb{1}_{A_t},\ \forall t \qquad \text{vs} \qquad \mathbb{H}_1 : p_{1t}\mathbb{1}_{A_t} = f_{1t}\mathbb{1}_{A_t}, \quad \forall t. \tag{6}$$

Although the predictive densities $f_{jt} = \frac{\mathrm{F}_{jt}}{\mathrm{d}\mu_t}$, $j \in \{0,1\}$, are assumed to be known, the testing problem remains a multiple versus multiple hypothesis test due to the lacking specification of the density outside the regions of interest A_t. Theorem 3 reveals that this setting nevertheless admits a UMP test, reducing to the Neyman and Pearson (1933) lemma for $w(y_t) = 1,\ \forall t$. A detailed proof of this result is deferred to Appendix A.2.

Theorem 3 (Localized Neyman-Pearson)**.** *The UMP test for testing problem* (6) *reads*

$$\phi_A^\flat(\mathbf{y}) = \begin{cases} 1, & \text{if } \lambda(\mathbf{y}) > c \\ \gamma & \text{if } \lambda(\mathbf{y}) = c \\ 0, & \text{if } \lambda(\mathbf{y}) < c \end{cases} \qquad \lambda(\mathbf{y}) = \frac{[f_1]_A^\flat(\mathbf{y})}{[f_0]_A^\flat(\mathbf{y})}, \qquad [f_j]_A^\flat(\mathbf{y}) = \prod_{t=0}^{T-1} [f_{jt}]_{A_t}^\flat(y_{t+1}),$$

where $\phi_A^\flat : \mathcal{Y}^T \to [0,1]$ *denote a test function determining which values should be included in the critical region,* $j \in \{0,1\}$ *and* c *is the largest constant such that* $[\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) \geq c\big) \geq \alpha$ *and* $[\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) \leq c\big) \geq 1-\alpha$*, and* $\gamma \in [0,1]$ *is such that* $\alpha = [\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) > c\big) + \gamma [\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) = c\big)$.

We close this section with two corollaries of Theorem 3, the proofs of which are deferred to the Online Supplementary Material. Corollary 1 reveals that, unsurprisingly, the localized NP test given by Theorem 3 can alternatively be formulated by the censored likelihood

score of Diks et al. (2011). Corollary 2 ensures that the conditional operator does not bear a UMP test too, making the censored operator strictly preferred over the conditional one in the current setting.

Corollary 1. *Another formulation of the UMP test for testing problem* (6) *is given by the test defined in Theorem 3, with* $\lambda(\mathbf{y})$ *replaced by* $\tilde{\lambda}(\mathbf{y}) = \sum_{t=0}^{T-1} \big(S_{A_t}^{csl}(f_{1t}, y_{t+1}) - S_{A_t}^{csl}(f_{0t}, y_{t+1})\big)$, *where* $S_{A_t}^{csl}$ *denotes the censored likelihood score (csl) proposed by Diks et al. (2011).*

Corollary 2. *For testing problem* (6)*, the test*

$$\phi_A^\sharp(\mathbf{y}) = \begin{cases} 1, & \text{if } \lambda^\sharp(\mathbf{y}) > c \\ \gamma & \text{if } \lambda^\sharp(\mathbf{y}) = c \\ 0, & \text{if } \lambda^\sharp(\mathbf{y}) < c \end{cases} \qquad \lambda_A^\sharp(\mathbf{y}) = \frac{[f_1]_A^\sharp(\mathbf{y})}{[f_0]_A^\sharp(\mathbf{y})} \mathbb{1}_A(\mathbf{y}), \qquad [f_j]_A^\sharp(\mathbf{y}) = \prod_{t=1}^{T} [f_{jt}]_{A_t}^\sharp(y_t),$$

where $j \in \{0, 1\}$ *and* c *is the largest constant such that* $[\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) \geq c\big) \geq \alpha$ *and* $[\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) \leq c\big) \geq 1-\alpha$*, and* $\gamma \in [0, 1]$ *is such that* $\alpha = [\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) > c\big) + \gamma [\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) = c\big)$*, is not UMP.*

3.5 Monte Carlo Study

Employing a simulation design similar to Diks et al. (2011), Lerch et al. (2017) and Holzmann and Klar (2016), simulation results in Appendix E, show the size and power properties of the Giacomini and White (2006) test based on conditional and censored scoring rules. This test relies on the score difference series of two candidates F and G, that is, realisations of $D = S(\mathrm{F}, Y) - S(\mathrm{G}, Y)$, in testing the null hypothesis

$$\mathbb{H}_0 : \mathbb{E}_\mathrm{P} S(\mathrm{F}, Y) = \mathbb{E}_\mathrm{P} S(\mathrm{G}, Y),$$

by means of the Diebold Mariano-type statistic $t_T = \frac{\frac{1}{T}\sum_{t=1}^{T} d_t}{\sqrt{\hat{\sigma}_t^2/T}}$, where $\hat{\sigma}_t$ should be a heteroskedasticity and autocorrelation-consistent (HAC) variance estimator in non-i.i.d. set-

tings. This null hypothesis, which is equivalent to $\mathbb{H}_0 : \mathbb{D}_S(\mathrm{P}\|\mathrm{F}) = \mathbb{D}_S(\mathrm{F}\|\mathrm{G})$, is rejected if it is unlikely enough that quoting F instead of P leads to the same information loss as quoting G instead of P.

A natural conjecture is that strictly locally proper scoring rules generally lead to higher power since they are sensible with respect to all measurable aspects of the distribution. Yet, the dependence of the null hypothesis on the scoring rule makes the null and rejection set dependent on the scoring rule too, obstructing theoretical results like Theorem 3. Nevertheless, the results listed in Appendix E, are clearly in favor of censoring. In the left-tail application for a standard Normal and Student-t candidate the differences are less monotonic than the in the other experiments due to the fact that the scores intersect by construction by the selection of candidates. In this application, censoring is particularly powerful for regions of interest quite far into the tail.

4 Empirical Applications

In this section, we assess the empirical impact of censoring versus conditioning by comparing the MCS implied conditional and censored scoring rules, extending the power analysis from Section 3.5. As delineated by Hansen et al. (2011), the MCS procedure expands the GW hypothesis to larger sets of H_0-equivalent methods, employing an iterative elimination procedure using either the TR or Tmax equivalence tests. Optimal power properties of censoring in the GW environment intuitively accelerate elimination in the MCS procedure, resulting in smaller MCS p-values and, consequently, reduced cardinality. We present results at the 0.90 and 0.75 confidence levels, utilizing the TR statistic with a block bootstrap with $B = 10,000$ replications and block length $k = 5$, unless stated otherwise. Our results are robust to variations in these parameters. When $\mathrm{CRPS}^\flat$ and twCRPS differ, we include

twCRPS for reference. We quantify differences in cardinality in absolute terms, framed as the proportion of cases wherein the number of methods in $\text{MCS}^{\flat}$ is strictly smaller and larger than $\text{MCS}^{\sharp}$. Additionally, we provide the factor by which the cardinality of the MCS expands when a conditioning is adopted in lieu of censoring.

4.1 Risk management

Evaluating the downside risk of asset returns is a crucial task in risk management, particularly for compliance with regulatory requirements related to risk measures like the Value-at-Risk ($\text{VaR}^q_{\hat{f}_t}$), which represents the q-th quantile of the model-based estimated density forecast $\hat{f}_t$ and the more recently mandated Expected Shortfall $\text{ES}^q_{\hat{f}_t}$, which quantifies expected losses conditional on those losses exceeding $\text{VaR}^q_{\hat{f}_t}$. To achieve this, we opt for a weight function of $w_t(y_t) = \mathbb{1}_{\left(-\infty, r_t^q\right)}(y_t)$ and choose for the variable of interest y_t the log-returns of the S&P500, that is, $y_t = \log(P_t/P_{t-1})$, where P_t is the adjusted closing price on day t. The dataset used for this study consists of 6,777 observations in total, spanning from January 2, 1996, to December 30, 2022, sourced from Yahoo Finance.

All selected forecasting methods conform to $Y_t|\mathcal{F}_{t-1} \sim \mathcal{D}(\mu, \sigma_t^2, \boldsymbol{\vartheta})$, denoting a parametric family of distributions with mean μ, variance σ_t^2 and other parameters $\boldsymbol{\vartheta}$. While we evaluated AR(1) and AR(5) models for the conditional mean, they did not yield significant improvements over a constant mean specification. We consider two conditional variance models: the GARCH(1,1) model by Bollerslev (1987), defined as

$$\sigma_t^2 = \omega + \alpha(y_t - \mu)^2 + \beta\sigma_{t-1}^2, \tag{7}$$

and the RGARCH(1,1) model proposed by Hansen et al. (2012), given by

$$\sigma_t^2 = \omega + \alpha x_{t-1} + \beta \sigma_{t-1}^2,$$

$$x_t = \xi + \phi \sigma_t^2 + \tau z_t + \kappa(z_t^2 - 1) + u_t,$$

where x_t represents the realized measure, $z_t = (y_t - \mu)/\sigma_t$ and u_t denotes a white noise process with variance σ_u^2. The realized measure is downloaded from the Risklab page of Dacheng Xiu's website: `dachxiu.chicagobooth.edu/#risklab`. Parameters are estimated via maximum likelihood on a rolling window of length $T_{\text{est}} = 1,000$.

Table 3 summarizes the cardinality differences of $\text{MCS}^\flat$ and $\text{MCS}^\sharp$, revealing stark differences, particularly for $h = 1$. At a 0.90 confidence level and $h = 1$, $\text{MCS}^\sharp$ is smaller in only one case across the examined quantiles and scoring rules, namely for $q = 0.25$ and $S = \text{QS}$, (see Table F.1.b). Equality in MCS size occurs mainly for higher quantile values, where information scarcity with respect to the distributions on $(-\infty, r_t^q)$ is less critical. For $h = 1$, $\text{MCS}^\sharp$ contains more than twice the number of methods compared to $\text{MCS}^\flat$ on average. For $h = 5$, the differential reduces but remains substantial, averaging around a factor 1.7. This attenuation in differences from $h = 1$ to $h = 5$ is largely attributed to the enhanced performance of $\text{CRPS}^\sharp$ relative to $\text{CRPS}^\flat$, except for $q = 0.15$.

Examining the composition of the MCS reveals that the censored MCS is often a subset of the conditional MCS, when $|\text{MCS}^\flat| \leq |\text{MCS}^\sharp|$. The significance of reductions due to censoring is further emphasised by the fact that the resulting MCS encompass more complex model specifications, which would be the optimal choices in the absence of parameter and forecasting uncertainty. Robustness checks, pertaining to k and T_{est} confirm the stability with respect to these parameters (see Table F.1.b). Additionally, the use of the TR statistic tends to expedite model elimination, yielding smaller MCS p-values compared to Tmax; this acceleration, however, is consistent across both censoring and conditioning.

Table 3: Changes in MCS cardinality between censored and conditional scoring rules.

	Tail(s)						Interval					
	$\lvert\mathbf{MCS}_{0.90}\rvert$			$\lvert\mathbf{MCS}_{0.75}\rvert$			$\lvert\mathbf{MCS}_{0.90}\rvert$			$\lvert\mathbf{MCS}_{0.75}\rvert$		
h	$<$	$>$	**Ratio**	$<$	$>$	**Ratio**	$<$	$>$	**Rel.**	$<$	$>$	**Ratio**
Risk Management												
1	**71%**	4%	**2.28**	**63%**	8%	**2.04**						
5	**38%**	25%	**1.69**	**50%**	42%	**1.72**						
Inflation												
6	**92%**	0%	**2.00**	**83%**	8%	**2.93**	**83%**	0%	**2.56**	**100%**	0%	**3.06**
12	**50%**	25%	**1.86**	**58%**	33%	**2.38**	**50%**	25%	**2.38**	**58%**	33%	**2.75**
24	**75%**	8%	**2.86**	**58%**	8%	**3.31**	**67%**	0%	**2.31**	**92%**	0%	**3.27**
Climate												
1	**54%**	17%	**2.67**	**63%**	13%	**2.41**	**42%**	8%	**1.50**	**42%**	17%	**1.46**
2	**54%**	4%	**1.67**	**46%**	4%	**1.55**	**67%**	0%	**1.67**	**58%**	0%	**1.58**
3	**42%**	21%	**1.36**	**38%**	17%	**1.38**	**58%**	0%	**1.58**	**25%**	0%	**1.25**

NOTE: The table presents changes in cardinality of the MCS in absolute and relative terms, at confidence level 0.75 and 0.90, across different forecast horizons h, based on $B = 10,000$ bootstrap replications, with block length $k = 5$, or $k = 200$ for climate data. Columns labeled $<$ ($>$) display the percentage of cases where $\text{MCS}^\flat$ contains strictly fewer (more) forecasting methods than $\text{MCS}^\sharp$, averaged over a set of levels or quantiles q and scoring rules $S \in \{\text{LogS}, \text{QS}, \text{SphS}, \text{CRPS}\}$. The "Ratio" column reports the factor $|\text{MCS}^\sharp|/|\text{MCS}^\flat|$. Specifically, the regions of interest for inflation are defined as $A_q = [2-q, 2+q]$ and its complement, where $q \in \{1, 2, 3\}$. For the climate data, $A_q = (r_q, \infty)$, where r_q is the empirical q-th quantile of the estimation window, with $q \in \{0.75, 0.80, 0.85, 0.90, 0.95, 0.99\}$ or $A_q = [18-q, 18+q]$ for $q \in \{1, 2, 4\}$. Complete MCS details and associated p values are provided in Appendix F of the Supplementary Material. Bolded numbers indicate strictly smaller ($<$ and Ratio column) or larger ($>$ column) $\text{MCS}^\flat$.

Beyond the statistical assessment of forecast methods, we compute their 1- and 5-step ahead Value at Risk ($\text{VaR}^q_{\hat{f}_t}$) and Expected Shortfall ($\text{ES}^q_{\hat{f}_t}$). These measures provide only partial insight into the forecasts, since the tail component of the density forecast carries more comprehensive information than a single quantile ($\text{VaR}^q_{\hat{f}_t}$) or conditional moment $\text{ES}^q_{\hat{f}_t} = \mathbb{E}_{\hat{f}_t}\left(Y_{t+h}|Y_{t+h} \leq \text{VaR}^q_{\hat{f}_t}\right)$. Notably, the conditioning in $\text{ES}^q_{\hat{f}_t}$ is a quantile of the density forecast itself rather than $\hat{r}^q_t$, a.s. implying a discrepancy between the operational region of $\text{ES}^q_{\hat{f}_t}$ and the focused scoring rules introduced above. Additionally, if the $\text{VaR}^q_{\hat{f}_t}$ is quite off, then the 'risk' indicated by $\text{ES}^q_{\hat{f}_t}$ can become quite detached from the true risk ES^q_p, where p denotes the density of the DGP. Hence, the $\text{ES}^q_{\hat{f}_t}$ is (particularly) useful when the $\text{VaR}^q_{\hat{f}_t}$ is accurate, i.e. we preferably have a good fit for the pair $(\text{VaR}^q_{\hat{f}_t}, \text{ES}^q_{\hat{f}_t})$, rather than just $\text{ES}^q_{\hat{f}_t}$ itself.

We highlight a corollary before discussing results. Given a fixed level q, let r be such that $\text{VaR}^q_{\hat{f}_t} \vee \text{VaR}p^q \leq r$. A property of the censored scoring rule is its ability to render the true $(\text{VaR}^q_p, \text{ES}^q_p)$ pair, since

$$\mathbb{D}_{S^\flat_w}(p\|f) = 0 \implies (\text{VaR}^q_p, \text{ES}^q_p) = (\text{VaR}^q_{\hat{f}_t}, \text{ES}^q_{\hat{f}_t}), \tag{8}$$

where $w(y) = \mathbb{1}_{(-\infty,r)}(y)$. This is a direct consequence of (3), i.e. another corollary of Lemma 1, and holds also more generally for any functional on distributions on $\{w > 0\}$. In (sharp) contrast, $\mathbb{D}_{S^\natural_w}(p\|f) = 0$ implies that $p \propto f$ on $(-\infty, r)$ and hence $(\text{VaR}^q_p, \text{ES}^q_p) \neq (\text{VaR}^q_{\hat{f}_t}, \text{ES}^q_{\hat{f}_t})$, unless $\bar{F}_w = \bar{P}_w$. Therefore, model selection based on censored scoring rules aligns more effectively with backtesting of functionals of the distribution compared to model selection based on conditional scoring rules.

Thus, censoring is designed to generate MCS containing forecast models that produce $(\text{VaR}^q_{\hat{f}_t}, \text{ES}^q_{\hat{f}_t})$ pairs closer to the true pair. Support for this conjecture is found in Table F.1.b. Despite often being smaller, the censored MCS contains well-fitted $(\text{VaR}^q_{\hat{f}_t}, \text{ES}^q_{\hat{f}_t})$

pairs, defined as 0% mismatches for both VaR and ES, more than twice as often (9 versus 4). If we accept up to 4% mismatches, the comparison remains favorable: 14 versus 6, endorsing censored MCS as a superior selection mechanism prior to VaR and ES calculations.

4.2 Inflation

In our second case, we focus on forecasting inflation, a subject recently magnified in the literature. Guided by the European Central Bank's target of 2% (`ecb.europa.eu/mopo`), we center our study on the range $A_q = [2 - q, 2 + q]$, where $q > 0$, employing the weight function $w(y_t) = \mathbb{1}_{A_q}(y_t)$. Simultaneously, we consider policymakers' concerns for deviations beyond A, termed 'Inflation at Risk' (Lopez-Salido and Loria, 2022), utilizing the complement weight function $w(y_t) = \mathbb{1}_{A_q^c}(y_t)$.

While the evaluation ingredients remain almost exactly the same, the unique characteristics of the inflation time series necessitate an adapted set of forecasting methods. We closely align with the methodology presented by Medeiros et al. (2021), using the same 122 variables from the FRED-MD database ($\mathbf{x}_t$), spanning January 1960 to December 2015. This timeframe encompasses a total of 672 observations, with the final 180 being out-of-sample relative to the initial estimation window. While using the same baseline data $P_t = \text{CPI}_t$ as Medeiros et al. (2021), we follow Stock and Watson (2002) and Borup et al. (2022) by analyzing the h-step ahead forecasts of the accumulated series $y_{t+h}^h = (1200/h) \log\big(P_{t+h}/P_t\big)$, instead of the accumulation of the individual h-step ahead forecasts of the monthly rate. This direct approach is standard in the literature and especially advantageous for density forecasts, as accumulating densities is more complex than aggregating point forecasts.

Each of the forecasting methods under consideration can be represented as

$$y_{t+h}^h = \mu_{j,t+h}^h(\mathbf{x}_t) + u_{t+h}^h, \qquad u_{t+h}^h|\mathcal{F}_t \sim \mathcal{N}_{\text{TP}}(0, \sigma_1, \sigma_2), \qquad \sigma_1, \sigma_2 > 0,$$

where $\mathcal{N}_{\text{TP}}(0, \sigma_1, \sigma_2)$ denotes the two-piece normal distribution distribution. For the conditional mean $\mu_{j,t+h}^h$, we take the following subset of models listed by Medeiros et al. (2021): Random Walk, Auto-Regressive model (AR), Bagging, Complete Subset Regression (CSR), Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest models. The implementation specifics of these models are elaborated upon in Section 4 of Medeiros et al. (2021). The density of the two-piece normal distribution reads

$$f(y; \mu, \sigma_1, \sigma_2) = \frac{2}{\sigma_1 + \sigma_2} \left(\phi\left(\frac{y-\mu}{\sigma_1}\right) \mathbb{1}_{y<\mu} + \phi\left(\frac{y-\mu}{\sigma_2}\right) \mathbb{1}_{y\geq\mu} \right), \qquad \sigma_1, \sigma_2 > 0,$$

where $\phi(z)$ denotes the density of the standard normal distribution. This distributional choice is congruent with the underlying statistical model employed in the fan charts published by the Monetary Policy Committee of the Bank of England (Clements, 2004; Mitchell and Hall, 2005; Gneiting and Ranjan, 2011).

The summary results presented in Table 3 show the difference between the cardinality of the $\text{MCS}^\flat$ and $\text{MCS}^\sharp$, averaged over $q \in \{1, 1.5, 2\}$. A glance at Table 3 reveals a distinct and pronounced preference for censoring. Notably, the cardinalities of $\text{MCS}^\flat$ are generally –with 'generally' here not seldom verging on unanimity– smaller than those of $\text{MCS}^\sharp$. This is especially salient in the Center case, where the $\text{MCS}^\flat$ are almost always weakly smaller than the corresponding $\text{MCS}^\sharp$. While it is unsurprising, given these results, that the relative increase in set cardinality when opting for conditioning over censoring is positive, the specific magnitudes of these increases even (substantially) exceed 100%. This is a striking finding: it effectively indicates that $\text{MCS}^\sharp$ consistently encompasses more than twice the number of methods compared to $\text{MCS}^\flat$, thereby making any defence of the use of $\text{MCS}^\sharp$ tenuous at best.

The differences between the MCS variants are clearly highlighted by the p-values presented in Table F.2.b, which also offers more detailed insights. To begin, for $q = 1$ the cardinality of $\text{MCS}^{\sharp}_{0.90}$ consistently exceeds or equals that of $\text{MCS}^{\flat}_{0.90}$ with the sole exceptions occurring in tail cases predicated on the CRPS for $h = 12$ and $h = 24$, and QS for $h = 12$. These exceptions feature a marginal difference of one. At a confidence level of 0.75, a similar trend is observed, albeit without the QS exception the tail case but with two additional exceptions for the center case at $h = 12$ in both the QS and CRPS rules.

Finally, a closer look at the differences between the twCRPS and $\text{CRPS}^{\flat}$ is in place. In the Center panel, we observe that the $\text{CRPS}^{\flat}$ is quite competitive with the twCRPS, leading to preferable MCS sizes for $h = 6$ and $h = 24$ (particularly for $h = 24$), but not for $h = 12$. Instances yielding comparable scores can be understood by recalling that the twCRPS and $\text{CRPS}^{\flat}$ coincide when the distribution of the remaining probability $\bar{F}_w$ based on the ratio implied by the empirical distribution, aligns with that of the candidate-implied ratio. This is different for the tails case, where observations falling outside the interval [1,3] are censored to a value of 2. In the current example, the adopted method of censoring does not manifest in enhanced discriminating ability, as suggested by the more favorable p-values associated with twCRPS.

4.3 Climate

In our third application, we generate density forecasts for Dutch daily average temperature data, extending the data and methodology of Franses et al. (2001) and Tol (1996). We maintain focus on volatility clustering and changing asymmetries in past temperature to volatility relations, along with accounting for seasonal variations in the mean and variance. Our data set, spanning February 1, 2003, to January 31, 2023, uses daily observations rather

than weekly averages. The first $T_{\text{est}} = 2922$ days form the initial rolling estimation window. Our models closely follow the GARCH, QGARCH-I, and QGARCH-II specifications as in Franses et al. (2001), but with alterations in seasonal trend estimation. Specifically, we use local day averages for the mean and a sine function for volatility, as opposed to a quadratic function. The models can be formalized as: $Y_t|\mathcal{F}_{t-1} \sim \mathcal{D}(\mu_t, \sigma_t^2, \boldsymbol{\vartheta})$, where $\mu_t = m_{t|t-1} + \phi y_{t-1}$ and

$$\sigma_t^2 = \varphi(t; \omega_0, \omega_1) + \alpha\big(y_{t-1} - \mu_{t-1} - \varphi(t; \gamma_0, \gamma_1)\big) + \beta\sigma_{t-1}^2,$$

Here, $m_{t|t-1}$ is the day's local average temperature in the estimation window and $\varphi(t; \theta_0, \theta_1) = \theta_0 + \theta_1 \sin(\pi/365 \cdot \tilde{T}_t)$, with $\tilde{T}_t = \min(T_t, 365)$, where T_t is the day number, with $T_t = 1$ on the first of February. For GARCH, QGARCH-I, and QGARCH-II, the restrictions are $\gamma_0 = \gamma_1 = 0$, $\gamma_1 = 0$ and no restrictions, respectively. These models are combined with both Normal and Student-t distributions to produce six forecasting methods.

The summary findings are presented in Table 3, with an emphasis on the right tail (r_t^q, ∞) and the interval $[18 - q, 18 + q]$. The latter interval has its roots in the agricultural literature, corresponding to the optimal temperature for tuber growth, agreed to be approximately 18 degrees Celsius (Struik, 2007, Section 18.5.5). Across both panels, the analyses demonstrate a clear preference for censoring methods, particularly for lower values of h. Consistent with the inflation scenario, the interval-based analyses yield the most unequivocal results: for $h = 2$ and $h = 3$, none of twelve MCS favor conditioning and for $h = 1$ only one or two.

5 Conclusion

In many applications, forecasters are particularly interested in particular areas of the outcome space. Addressing this, we champion censoring as focusing device, demonstrating

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 strict propriety. Our approach stands out in its flexibility, 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, applying our approach to the logarithmic scoring rule results in the well-established censored likelihood score. We accord specific attention to the censored CRPS, emerging as a strictly locally proper alternative to the twCRPS. Conveniently, the censored CRPS reduces to the twCRPS for left and right tail indicators, which are the only weight functions for which the twCRPS is established to be strictly locally proper.

Our second theoretical contribution, a generalization of the famous 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 one for all outcomes. In contrast, the conditional likelihood ratio test is not UMP. Monte Carlo simulations incorporate the Giacomini and White test to assess the power properties of conditional versus censored scoring rules based on the score differences between two candidates. The findings endorse the superior power properties of censoring, extending beyond the stylised scenario in which the candidates' tails are close to proportional. Both conditional and censored scoring rules maintain size correctness.

In the empirical analysis, we use the size of the Model Confidence Set (MCS) as an indicator of power. Notably, in our inflation example –where the number of observations is characteristically low, akin to many macro-applications– the frequency with which the censored MCS is strictly smaller than the conditional MCS strikes, as does the difference

in cardinality. These observations hold across different horizons, whether centered on the 2% target or its complement. In density forecast assessments of the S&P500, a comparable trend emerges, though the difference narrows at $h = 5$. The application to climate data corroborates the enhanced power of the censored approach, revealing that one could better not entirely forget about the winter when growing potatoes.

SUPPLEMENTARY MATERIAL

All proofs, additional lemmas, tables and figures are made available in the supplementary document.

References

Amisano, G. and R. Giacomini (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics 25*, 177–190.

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, 1–45.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics 69*(3), 542–547.

Borowska, A., L. Hoogerheide, S. Koopman, and H. van Dijk (2020). Partially censored posterior for robust and efficient risk evaluation. *Journal of Econometrics 217*, 335–355.

Borup, D., P. G. Coulombe, D. Rapach, E. C. M. Schütte, and S. Schwenk-Nebbe (2022). The anatomy of out-of-sample forecasting accuracy. Working paper 2022-16, FRB Atlanta.

Bregman, L. M. (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. and T. Gneiting (2020). Properization: Constructing proper scoring rules via bayes acts. *Annals of the Institute of Statistical Mathematics 72*, 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.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General) 147*, 278–290.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics 59*, 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*, 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. *Annals of Statistics 40*(1), 609–637.

Fissler, T., J. F. Ziegel, and T. Gneiting (2015). Expected shortfall is jointly elictable with value at risk-implications for backtesting. *arXiv preprint arXiv: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.

Gneiting, T. and A. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

Gneiting, T. and R. Ranjan (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics 29*, 411–422.

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society 14*, 107–114.

Good, I. (1971). Comment on "measuring information and uncertainty". *Foundation of Statistical Inference*, 265–273.

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, SI), 877–906.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*(5), 559–570.

Holzmann, H. and B. Klar (2016). Weighted scoring rules and hypothesis testing. Arxiv, ArXiv E-Prints.

Holzmann, H. and B. Klar (2017). Focusing on regions of interest in forecast evaluation. *Annals of Applied Statistics 11*, 2404–2431.

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.

Jose, V. R. (2009). A characterization for the spherical scoring rule. *Theory and Decision 66*, 263–281.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*, 79–86.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science 59*(1), 106–127.

Liese, F. and I. Vajda (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory 52*, 4394–4412.

Lopez-Salido, D. and F. Loria (2022). Inflation at risk. *Available at SSRN 4002673*.

Matheson, J. and R. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*(10), 1087–1096.

Medeiros, M. C., G. F. Vasconcelos, Á. 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*, 995–1033.

Neyman, J. and E. S. 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*, 289–337.

Nieto, M. R. and E. Ruiz (2016). Frontiers in Var forecasting and backtesting. *International Journal of Forecasting 32*(2), 475–501.

Ovcharov, E. (2018). Proper scoring rules and Bregman divergence. *Bernoulli 24*, 53–79.

Painsky, A. and G. W. Wornell (2019). Bregman divergence bounds and universality properties of the logarithmic loss. *IEEE Transactions on Information Theory 66*, 1658–1673.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics 38*(4), 796–809.

Roby, T. B. (1964). Belief states: A preliminary empirical study. Technical report, Tufts University Medford MA.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66*, 783–801.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics 1*, 43–61.

Shuford, E. H., A. Albert, and H. E. Massengill (1966). Admissible probability measurement procedures. *Psychometrika 31*, 125–145.

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). Responses of the potato plant to temperature. In *Potato Biology and Biotechnology*, pp. 367–393. Elsevier, North-Holland.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*(1), 24–36.

Toda, M. (1963). *Measurement of subjective probability distribution*. Division of Mathematical Psychology, Institute for Research.

Tol, R. S. (1996). Autoregressive conditional heteroscedasticity in daily temperature measurements. *Environmetrics 7*(1), 67–75.