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

March 27, 2025

*We are very grateful to the Editor, Associate Editor and three referees for their comments and suggestions, which have significantly improved the paper. We are also grateful to Timo Dimitriadis, Tilmann Gneiting, Alexander Jordan, Frank Kleibergen, Siem Jan Koopman, 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) 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 2020–2025 (Laeven).

†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 strict propriety. This is accomplished by applying the original scoring rule to a censored distribution, acknowledging that censoring emerges as a natural localization device due to its ability to retain precisely all relevant information of the original 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 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 superior power properties of censoring.

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), thus excluding f-divergences other than the Kullback and Leibler (1951) divergence. 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 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 the latter, this holds only 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 delineates the adverse consequences arising from this failure in localization, and provides a strictly locally proper alternative.

The additional information retained by our censoring approach also translates into advantageous power properties of tests aimed to compare density forecasts on 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 conditional scoring rules. Censored scoring rules enhance power in all Monte Carlo experiments conducted. Substantial spurious power is observed solely for conditional scoring rules, which also falter in terms of power when tails become proportional. 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 much smaller than their conditional counterparts, aligning with the power enhancements due to censoring 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 show 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.

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 a generalization of the Neyman Pearson lemma, a comparison with alternative weighted scoring rules, and guidance for the practical use of the censoring procedure. Section 4 discusses the empirical performance of our approach. Section 5 concludes. Proofs, derivations of theoretical properties, results of the Monte Carlo study, and detailed empirical results are provided in the accompanying Supplementary Material.

2 FOCUSED DIVERGENCES

2.1 Unweighted scoring rules and divergences

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 probability 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 $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty\} =: \bar{\mathbb{R}}$. Following Holzmann and Klar (2017a), we assume that 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 $\mathbb{E}_{\mathrm{P}} S(\mathrm{P}, Y) \in \mathbb{R}$. 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 (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, $\mathbb{D}_S(\mathrm{P}\|\mathrm{F}) = 0$ if and only if $\mathrm{P} = \mathrm{F},\ \forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$.*

Equivalently, a score divergence is a divergence measure (see, e.g., Eguchi, 1985) if and only if S is 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 (1967) divergence under the conditions listed by Ovcharov (2018). Strictly proper scoring rules easily facilitate the construction of such divergence measures. Two remarks are in place. First, distributions $\mathrm{F} \in \mathcal{P}$ 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, $\mathrm{P} = \mathrm{F}$ is formally defined as $\mathrm{P}(E) = \mathrm{F}(E),\ \forall E \in \mathcal{G}$. For ease of exposition, we omit technicalities about P-a.s. equivalence throughout. Second, if there exists a σ-finite

measure μ such that $\mathrm{F} \ll \mu,\ \forall \mathrm{F} \in \mathcal{P}$, 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 $f = \frac{\mathrm{dF}}{\mathrm{d}\mu}$, also denoted by $\mathcal{P}$, like classes of distributions 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 them 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_α) and PseudoSpherical (PsSphS_α) families. 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 Censoring

We consider the case where 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 in the complement $A^c \equiv \mathcal{Y} \backslash A$ are of no interest. The region of interest A is used to transform a class of distributions $\mathcal{P}$ to a localized class $\mathcal{P}_A$, transforming the random variable Y to $Y_A : (\mathcal{Y}, \mathcal{G}, \mathrm{F}) \to (\mathcal{Y}_A, \mathcal{G}_A)$. However, not all transformations are equally desirable. In terms of events $E \in \mathcal{G}$, focusing on A implies that only events intersecting with A are of interest. To formalize this, consider the *smallest* σ-algebra on $\mathcal{Y}$ containing *all* events of interest: $\mathcal{G}_A^\flat := \sigma(\{E \cap A : E \in \mathcal{G}\}) \subseteq \mathcal{G}$, which

also includes A^c because it is closed under complements. We refer to the restriction of F to the minimal σ-algebra $\mathcal{G}_A^\flat$, denoted $\mathrm{F}_A^\flat \equiv \mathrm{F}_{|\mathcal{G}_A^\flat}$, as the *minimal localization* of F to A.

The distribution $\mathrm{F}_A^\flat$ is associated with the censored random variable $Y_A^\flat : (\mathcal{Y}, \mathcal{G}, \mathrm{F}) \to (\mathcal{Y}_A^\flat, \mathcal{G}_A^\flat)$, where $\mathcal{Y}_A^\flat := A \cup \{*\}$, defined by the transformation

$$Y_A^\flat = \begin{cases} Y, & Y \in A, \\ *, & Y \in A^c, \end{cases} \tag{1}$$

where $* = A^c$ represents the complement of A treated as a single outcome. Since A^c serves both as a subset of outcomes of Y and a single outcome of $Y_A^\flat$, we use the symbol $*$ in the latter case to avoid confusion. 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). In applications, we typically work on the measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, where $\mathcal{B}(\mathcal{Y})$ denotes the Borel σ-algebra on $\mathcal{Y}$. Then, $Y_A^\flat : (\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mathrm{F}) \to (\mathcal{Y}_A^\flat, \mathcal{B}(\mathcal{Y}_A^\flat))$; see Example 1, in which $\mathbb{1}_A(y)$ equals unity if $y \in A$ and zero otherwise.

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, with outcome space $\mathcal{Y} = \{s, o, b\}$, σ-algebra $\mathcal{G} = \{\emptyset, \{s\}, \{o\}, \{b\}, \{s, o\}, \{s, b\}, \{o, b\}, \{s, o, b\}\}$, and probability mass function (pmf) $p(y) = \frac{1}{6}\mathbb{1}_{\{s\}}(y) + \frac{1}{3}\mathbb{1}_{\{o\}}(y) + \frac{1}{2}\mathbb{1}_{\{b\}}(y)$, $y \in \mathcal{Y}$. Because Jim only cares about whether he wins the car or not (and not how he loses), his region of interest is $A = \{s\}$, which induces $\mathcal{Y}_A^\flat = \{s\} \cup \{*\}$, where $\{*\} \equiv \{o, b\}$, and $\mathcal{G}_A^\flat = \sigma(\{s, *\}) = \{\emptyset, \{s\}, \{*\}, \{s, *\}\}$, equivalent to $\sigma(\{s\})$ on $\mathcal{Y}$. The pmf $p(y)$ localizes to $p_A^\flat(y) = \frac{1}{6}\mathbb{1}_{\{s\}}(y) + \frac{5}{6}\mathbb{1}_{\{o,b\}}(y)$, $y \in \{s, *\} \equiv \{s, \{o, b\}\}$.*

2.3 Local and localized divergences

Example 2. *Revisiting Example 1, suppose Jim and his friend Pam know that the vase contains six balls colored silver, orange and blue, but not their exact numbers. Jim suspects one silver, four orange, and one blue ball, while Pam guesses two silver, one blue, and three orange. Let Jim's and Pam's implied pmfs be f and g, respectively. A straightforward calculation shows that $\mathsf{KL}(p\|f) - \mathsf{KL}(p\|g) = \frac{1}{2}\log(\frac{3}{2}) > 0$, where $\mathsf{KL}(p\|f) := \mathbb{E}_p(\log p(Y) - \log f(Y))$ denotes the Kullback Leibler (KL) divergence from p to f. Hence, Pam's guess is statistically closer to the truth if the region of interest is ignored. However, since Jim only cares about winning the car, Pam's accurate guess of the orange balls' count, with a relatively high true probability $p(\{o\}) = 1/2$, is irrelevant and even misleading. Her close fit outside the silver outcome obscures her inaccuracy where Jim was 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. Describing the relative importance of outcomes $y \in \mathcal{Y}$ by a *weight function* $w : \mathcal{Y} \to [0, 1]$, assumed to be $\mathcal{G}$-measurable, the question arises how to accordingly transform the divergence $\mathbb{D}$. As the relation to a scoring rule is currently not required, we work with a general divergence $\mathbb{D}$, satisfying (i) $\mathbb{D}(\mathrm{P}\|\mathrm{F}) \geq 0, \forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$, and (ii) $\mathbb{D}(\mathrm{P}\|\mathrm{F}) = 0$ if and only if $\mathrm{P} = \mathrm{F}, \forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$.

The minimal localization $\mathrm{F}^\flat_A$ is considered the optimal transformation for $w(y) = \mathbb{1}_A(y)$ since the restriction to $\mathcal{G}_A$ eliminates all probabilities of no interest while preserving all probabilities of interest. It changes the researcher's world from $(\mathcal{P}, \mathbb{D})$ to $(\mathcal{P}^\flat_A, \mathbb{D})$, where $\mathcal{P}^\flat_A := \{\mathrm{F}^\flat_A, \mathrm{F} \in \mathcal{P}\}$, provided that $\mathbb{D}$ is well-defined on $\mathcal{P}^\flat_A \times \mathcal{P}^\flat_A$, in which case $\mathbb{D}^\flat_A(\mathrm{P}\|\mathrm{F}) := \mathbb{D}(\mathrm{P}^\flat_A\|\mathrm{F}^\flat_A)$ is well-defined for all $\mathrm{P}, \mathrm{F} \in \mathcal{P}$. Since $\mathbb{D}$ is a divergence measure, $\mathbb{D}(\mathrm{P}^\flat_A\|\mathrm{F}^\flat_A) = 0$

if and only if $\mathrm{P}_A^\flat = \mathrm{F}_A^\flat$. The latter is, in turn, equivalent to the measures coinciding on A, i.e. $\mathrm{P}(A \cap E) = \mathrm{F}(A \cap E), \forall E \in \mathcal{G}$, for which we introduce the short-hand notation $\mathrm{P} \overset{A}{=} \mathrm{F}$. Hence, $\mathbb{D}_A^\flat : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ satisfies the conditions of a localized divergence in Definition 3, which is a specific type of local divergence; see Definition 2. Both definitions are given for general weight functions, generalizing the region of interest to $A_w := \{y \in \mathcal{Y} : w(y) > 0\}$.

Definition 2 (Local divergence)**.** *A map* $\mathbb{D}_w : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ *is a* local divergence *(to* A_w*) if (i)* $\mathrm{F} \overset{A_w}{=} \mathrm{G}$ *implies that* $\mathbb{D}_w(\mathrm{P}\|\mathrm{F}) = \mathbb{D}_w(\mathrm{P}\|\mathrm{G})$*,* $\forall \mathrm{P}, \mathrm{F}, \mathrm{G} \in \mathcal{P}$*, (ii)* $\mathbb{D}_w(\mathrm{P}\|\mathrm{F}) \geq 0, \forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$*, and (iii)* $\mathbb{D}_w(\mathrm{P}\|\mathrm{F}) = 0$ *if and only if* $\mathrm{P} \overset{A_w}{=} \mathrm{F}, \forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$*.*

Definition 3 (Localized divergence)**.** *Let* $[\cdot]_w : \mathcal{P} \to \mathcal{P}_w$ *be a surjective map such that* $[\mathrm{F}]_{\mathbb{1}_\mathcal{Y}} = \mathrm{F}, \forall \mathrm{F} \in \mathcal{P}$*. A local divergence* $\mathbb{D}_w : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ *is called a* localized divergence *of* $\mathbb{D} \equiv \mathbb{D}_{\mathbb{1}_\mathcal{Y}}$ *(to* A_w*) if* $\forall \mathrm{P}_w, \mathrm{F}_w \in \mathcal{P}_w, \exists \mathrm{P}, \mathrm{F} \in \mathcal{P} : \mathbb{D}_w(\mathrm{P}\|\mathrm{F}) = \mathbb{D}(\mathrm{P}_w\|\mathrm{F}_w)$*.*

Condition (i) in Definition 2 ensures invariance with respect to events that are irrelevant as they are not intersecting with A_w. Dependence on such events could lead to what we refer to as *localization bias*, where a strong fit on these events obscures a poor fit on events of interest. Definition 3 introduces a subclass of local divergences that preserve the unweighted divergence measure $\mathbb{D}$ by applying it to a weighted transformation of the distribution space, denoted by $[\cdot]_w$. Due to the required surjectivity of this map, F_w exists for all $\mathrm{F} \in \mathcal{P}$, validating $\mathbb{D}_w(\mathrm{P}\|\mathrm{F}) = \mathbb{D}(\mathrm{P}_w\|\mathrm{F}_w)$. While local divergences can also be specified on an *ad hoc* basis, a localized divergence is inherently linked to a given unweighted setting $(\mathcal{P}, \mathbb{D})$ via the transformation $[\cdot]_w$. This allows a researcher with a rationale for using a divergence $\mathbb{D}$ to localize it to a region of interest. Mathematically, reusing $\mathbb{D}$ is appealing because any desirable properties naturally carry over to the weighted setting.

Just as strictly proper scoring rules give rise to divergences, local divergences emerge naturally from *weighted scoring rules* that are *strictly locally proper* with respect to some

class of distributions $\mathcal{P}$ and weight functions $\mathcal{W}$. Holzmann and Klar (2017a) define a weighted scoring rule S_w for weight function w as a map $S_\cdot : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$ such that $S_w(\cdot, \cdot)$ is a scoring rule for each $w \in \mathcal{W}$. A weighted scoring rule is said to be *localizing* if measures coinciding on A_w receive the same score for any realization. Specifically, $S_w(\mathrm{P}, y) = S_w(\mathrm{F}, y), \forall y \in \mathcal{Y}$, whenever $\mathrm{P} \overset{A_w}{=} \mathrm{F}$, $\forall \mathrm{P}, \mathrm{F} \in \mathcal{P}$. Definition 4 extends strict propriety to localizing weighted scoring rules and is equivalent to that given by Holzmann and Klar (2017a, p.2414). Clearly, the score divergence $\mathbb{D}_{S_w}(\mathrm{P}\|\mathrm{F}) := \mathbb{E}_{\mathrm{P}}(S_w(\mathrm{P}, Y)) - \mathbb{E}_{\mathrm{P}}(S_w(\mathrm{F}, Y))$ is a local divergence if and only if S_w is strictly locally proper.

Definition 4 (Strictly locally proper scoring rule)**.** *A weighted scoring rule $S_\cdot : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$ 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,* $\mathrm{P} \overset{A_w}{=} \mathrm{F}$ *if and only if* $\mathbb{D}_{S_w}(\mathrm{P}\|\mathrm{F}) = 0$, $\forall w \in \mathcal{W}$.

3 THE CENSORED SCORING RULE

For indicator weight functions $w(y) = \mathbb{1}_A(y)$, censoring is the underlying focusing method of the minimal localization $\mathrm{F}^\flat_A$. In this section, we generalize the censored distribution to general weight functions and to scoring rules that are incompatible with the outcome $*$. Similar to the conditional approach of Holzmann and Klar (2017a), we extend the definition of the distribution of $Y^\flat_A$ in Section 2.2. Specifically, we consider the *censored random variable* $Y^\flat_w : (\mathcal{Y}^*, \mathcal{G}^*, \mathrm{F}^*) \to (\mathcal{Y}^\flat_w, \mathcal{G}^\flat_w)$, with *censored distribution*

$$\mathrm{dF}^\flat_w := \mathrm{dF}^*_w + \bar{F}_w \mathrm{d}\delta_*, \qquad \bar{F}_w = \int_{\mathcal{Y}} (1 - w)\mathrm{dF}, \tag{2}$$

where $\mathrm{dF}^*_w := w^* \mathrm{dF}^*$ and δ_* denotes the Dirac measure at $*$, i.e., $\delta_*(E) = \mathbb{1}_E(*)$. Moreover, $\mathcal{Y}^\flat_w := A_w \cup \{*\}$, $\mathcal{G}^\flat_w := \sigma(\{E \cap A_w : E \in \mathcal{G}\} \cup \{*\})$, $\mathcal{Y}^* := \mathcal{Y} \cup \{*\}$, $\mathcal{G}^* := \sigma(\{\mathcal{G} \cup \{*\}\})$. The

distribution $\mathrm{F}^* : \mathcal{G}^* \to [0,1]$ coincides with F on $\mathcal{G}$, i.e. $\mathrm{F}^*(*) = 0$, and $w^* : \mathcal{Y}^* \to [0,1]$ is $\mathcal{G}^*$-measurable and such that $w^*(*) = 0$. The extension of the original space ensures that $\mathcal{G}_w^\flat \subseteq \mathcal{G}^*$. For clarity, we henceforth omit the superscripts '$*$'. Under the censored measure, the censoring event $*$ carries probability $\mathrm{F}_w^\flat(*) = \bar{F}_w$. In contrast to indicator functions, however, the probability $\mathrm{F}_w^\flat(*)$ is generally different from $\mathrm{F}(A_w^c)$, as probabilities on A_w are also weighted by $w(y)$. Hence, the event $*$ is not directly equivalent to A_w^c in the original outcome space but should be regarded as an abstract event, interchangeable with 'NaN'.

The censored distribution admits the $(\mu + \delta_*)$-density $f_w^\flat(s) = w(s)f(s)\mathbb{1}_{s\neq *} + \bar{F}_w \mathbb{1}_{s=*}$, $s \in \mathcal{Y}_w^\flat$, provided that $\mathrm{F} \ll \mu$; see Appendix B.1 for details. For indicator functions, the censored distribution recovers the minimal localization $\mathrm{F}_A^\flat \equiv \mathrm{F}_{|\mathcal{G}_A^\flat}$. In this case, the density simplifies to $f_A^\flat(y) = f(y)\mathbb{1}_A(y) + \bar{F}_A \mathbb{1}_{A^c}(y)$, similar to Borowska et al. (2020). To ease notation, we adopt the subscript A instead of $\mathbb{1}_A$ when referencing indicator functions.

An important difference from the conditional random variable $Y_w^\sharp : (\mathcal{Y}, \mathcal{G}, \mathrm{F}) \to (\mathcal{Y}_w^\sharp, \mathcal{G}_w^\sharp)$, proposed by Holzmann and Klar (2017a), with distribution $\mathrm{dF}_w^\sharp = \frac{1}{1-\bar{F}_w}\mathrm{dF}_w$, is that $\mathrm{F}_A^\sharp(A^c) \neq \mathrm{F}(A^c) = \mathrm{F}_A^\flat(A^c)$ unless $\mathcal{Y} \subseteq A$. In other words, conditioning does not preserve all probabilities of interest and does accordingly not reduce to the minimal localization of F to A. The symbols 'sharp' ($\sharp$) and 'flat' ($\flat$) 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. Holzmann and Klar (2017a) remedy the lack of strict propriety of $S_w^\sharp(\mathrm{F}, y) := w(y)S(\mathrm{F}_w, y)$ 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. A further discussion is deferred to Section 3.4.

3.1 Censored scoring

Ideally, the censored scoring rule would be given by the identity $S_A^\flat(\mathrm{F}, y) = S(\mathrm{F}_A^\flat, y_A^\flat)$, as this would fully respect the forecaster's specific choice of the unweighted scoring rule S. The censored scoring rule given by Definition 5 below indeed reduces to this definition for the indicator weight function $w(y) = \mathbb{1}_A(y)$. The censored scoring rule is also attractive for general weight functions, for which an explicit transformation in terms of the unweighted random variable, $Y \mapsto Y_w^\flat$, is not always available. However, a randomization perspective developed in Appendix E, yields a similar identity for general weight functions.

Definition 5 (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 unweighted scoring rule. Then, the corresponding censored scoring rule is given by the map $S_\cdot^\flat : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \to \bar{\mathbb{R}}$,*

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

where the censored distribution $\mathrm{F}_w^\flat$ is defined in Equation (2)*.*

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

Theorem 1. *If the unweighted scoring rule S is strictly proper relative to $\mathcal{P}^\flat$, the censored scoring rule $S^\flat$ in Definition 5 is strictly locally proper relative to $(\mathcal{P}, \mathcal{W})$. Moreover, its associated score divergence $\mathbb{D}_{S_w^\flat}$ 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 imposed in Theorem 1 ensures that the unweighted scoring rule is well-defined with respect to mixed continuous-discrete distributions on measurable spaces extended by '$*$'. In Example 3, we verify this explicitly for the PseudoSpherical scoring rule.

For the Logarithmic scoring rule, the score divergence of $\mathrm{LogS}_w^\flat$ coincides with the one by Diks et al. (2011). As shown by Ehm and Gneiting (2012), the Logarithmic scoring rule is a *local scoring rule* in the sense that it only depends on the observation through the scoring rule, a unique property of this scoring rule. Consequently, $\mathrm{LogS}(f_w^\flat, *) = \mathrm{LogS}(\bar{F}_w, *)$ and hence $\mathrm{LogS}_w^\flat(f, *)$ depends only on $\bar{F}_w$, a sufficient condition for property (i) in Definition 2, ensuring the score divergence to be localizing. However, as illustrated by Example 3, this condition is not necessary; $S(f_w^\flat, *)$ does not need to depend solely on $\bar{F}_w$.

Example 3 (Censored PseudoSpherical score)**.** *Consider a class of μ-densities $\mathcal{P}_\alpha$ on $(\mathcal{Y}, \mathcal{G}, \mu)$ with finite L^α-norm, i.e. $\|f\|_\alpha := \left(\int_\mathcal{Y} f^\alpha \mathrm{d}\mu\right)^{1/\alpha} < \infty, \forall f \in \mathcal{P}_\alpha$. As advocated by Gneiting and Raftery (2007), the PseudoSpherical family,* $\mathrm{PsSphS}_\alpha(f, y) = f(y)^{\alpha-1}/\|f\|_\alpha^{\alpha-1}$, *$\alpha > 1$, is strictly proper relative to $\mathcal{P}_\alpha$. Hence, one can easily verify its strict propriety with respect to $\mathcal{P}_\alpha^\flat$ as required for Theorem 1, since $\|f_w^\flat\|_\alpha^\alpha \leq 1 + \|f\|_\alpha^\alpha < \infty$, $\forall f \in \mathcal{P}_\alpha$, $\forall w \in \mathcal{W}$, meaning that $S_w^\flat(f, y)$ is strictly proper relative to $(\mathcal{P}, \mathcal{W})$. Notably, while $S_w^\flat(f, *) = \bar{F}_w^{\alpha-1}/(\|wf\|_\alpha + \bar{F}_w^\alpha)^{(\alpha-1)/\alpha}$ does not depend solely on $\bar{F}_w$, it holds that $S_w^\flat(f, *) = S_w^\flat(g, *)$, if $f = g$ a.s. on A_w.*

In Theorem 1, we assume that the scoring rule S (and thus $\mathbb{D}_S$) is well-defined relative to the class of censored measures $\mathcal{P}^\flat$ that have a point mass at $*$. However, distance-sensitive scoring rules such as kernel scores are generally incompatible with such measures because the distance to $*$ is undefined. This incompatibility also precludes minimal localization, which requires $* \equiv A^c \notin \mathcal{Y}$. Nonetheless, Section 3.2 shows that Theorem 1 continues to hold for any $y_0 \in \mathcal{Y}$ at which all distributions in $\mathcal{P}$ have zero mass, or to which all weight functions in $\mathcal{W}$ assign weight zero (both conditions hold trivially by construction at $*$). In practice, one can thus generally replace $*$ with y_0 without affecting theoretical results, as also suggested by Allen et al. (2023), whose threshold weighted kernel scores coincide with

censored kernel scores in particular cases; see Section 3.4. However, they also argue that some choices for y_0 are more natural than others and choose points at the boundary of their tail-based regions of interest.

In Appendix C.1, we formalize this idea by showing that for $A = (-\infty, r)$, with $r \in \mathbb{R}$, choosing $y_0 = r$ yields a generalized censored measure $\mathrm{dF}^{\flat}_{A,r} := \mathrm{dF}_w + \bar{F}_w \mathrm{d}\delta_r$ which coincides with F on the minimal σ-algebra $\mathcal{B}^{\flat}_A = \sigma(\{(-\infty, y] \cap (-\infty, r) : y \in \mathbb{R}\})$, while being dominated by the Lebesgue measure on $\mathbb{R}$, in contrast to the censored measure in Equation (2). If $A = (-\infty, r_1) \cup (r_2, \infty)$, we cannot find an equivalent Lebesgue dominated generalized censored measure that coincides with the minimal localization of F to A. This would require splitting $\bar{F}_w = \mathrm{F}\big((-\infty, r_1) \cup (r_2, \infty)\big)$ into the probabilities $\mathrm{F}\big((-\infty, r_1)\big)$ and $\mathrm{F}\big((r_2, \infty)\big)$, conflicting the localizing property of an associated weighted scoring rule; see Appendix C.2. In response, Section 3.2 introduces the flexibility to distribute $\bar{F}_w$ over r_1 and r_2 independent from F, using $\mathrm{dF}^{\flat}_{A,r_1,r_2} := \mathrm{dF}_w + \bar{F}_w \mathrm{d}\big(\gamma\delta_{r_1} + (1-\gamma)\delta_{r_2}\big)$, where $\gamma \in [0, 1]$.

3.2 Generalized censored scoring

This section formalizes the more flexible censoring framework introduced above. In general, suppose that a weight function introduces k pivotal points $r_1, \ldots, r_k \in \mathcal{Y}$, then a natural generalization of the censored distribution in Equation (2) reads

$$\mathrm{dF}^{\flat}_{w,\boldsymbol{\gamma}} := \mathrm{dF}_w + \bar{F}_w \sum_{i=1}^{k} \gamma_i \mathrm{d}\delta_{r_i}, \qquad \boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_k)' \in \Delta(k), \tag{3}$$

where $\Delta(k)$ denotes the unit simplex and $r_i \in \mathcal{Y}, \forall i$. Section 3.5 provides guidance on choosing (r_i, γ_i). Definition 6 formalizes the adaptation of the censored scoring rule to generalized censored measures for which $\mathrm{dH} = \sum_{i=1}^{k} \gamma_i \mathrm{d}\delta_{r_i}$ serves as the specific choice reducing the generalized censored distribution to Equation (3). 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 $S : \mathcal{P}^\flat \times \mathcal{Y} \to \bar{\mathbb{R}}$ denote an unweighted scoring rule and $\mathcal{H} \subseteq \mathcal{P}$ a class of nuisance distributions. The associated generalized censored scoring rule is given by the map $S^\flat_{\cdot,\cdot} : \mathcal{P} \times \mathcal{Y} \times \mathcal{W} \times \mathcal{H} \to \bar{\mathbb{R}}$,*

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

where $\mathrm{F}^\flat_{w,\mathrm{H}}$ is referred to as the generalized censored distribution of F and $\mathrm{H} \in \mathcal{H}$ denotes the distribution of the random variable Q.

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

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

A corollary to Lemma A2 in the proof of Theorem 2 is the mathematical identity

$$\mathbb{D}_{S^\flat_{w,\mathrm{H}}}(\mathrm{P}\|\mathrm{F}) = \mathbb{D}_S(\mathrm{P}^\flat_{w,\mathrm{H}}\|\mathrm{F}^\flat_{w,\mathrm{H}}), \tag{4}$$

which almost trivially is a localized divergence (Definition 3). The key subtlety is ensuring that $\mathbb{D}_{S^\flat_{w,\mathrm{H}}}$ is zero if and only if $\mathrm{P} \overset{A_w}{=} \mathrm{F}$. Since $\mathbb{D}_S$ is a divergence, $\mathrm{P} \overset{A_w}{=} \mathrm{F}$ should be equivalent to $\mathrm{P}^\flat_{w,\mathrm{H}} = \mathrm{F}^\flat_{w,\mathrm{H}}$, for fixed w, H satisfying Assumption 1. Equivalence holds as (i) w and H do not depend on P and F, so the transformation is identical for both; and (ii) it

is recoverable, as the existence of $\tilde{E} \in \mathcal{G}$ such that $\mathrm{F}_w(\tilde{E}) = 0$ and $\mathrm{H}(\tilde{E}) > 0$ ensures that $\bar{P}_w = \bar{F}_w$ if $\mathrm{P}^\flat_{w,\mathrm{H}} = \mathrm{F}^\flat_{w,\mathrm{H}}$, so $\mathrm{P} \overset{A_w}{=} \mathrm{F}$ if $\mathrm{P}^\flat_{w,\mathrm{H}} = \mathrm{F}^\flat_{w,\mathrm{H}}$. The 'only if' is trivial, i.e., $\mathrm{P}^\flat_{w,\mathrm{H}} = \mathrm{F}^\flat_{w,\mathrm{H}}$ if $\mathrm{P} \overset{A_w}{=} \mathrm{F}$. In the setting of Example 1, Assumption 1 excludes the possibility that Jim's outcome of interest (the 'silver ball' with $w(\{s\}) = 1$) is also the censoring outcome. If it were, all censored measures would degenerate to $\delta_{\{s\}}$, leaving $\bar{F}_w$ unidentifiable.

A rich class of scoring rules obtained by the generalization of the censored scoring rule is that of kernel scores (Gneiting and Raftery 2007). 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. *Consider a class of distributions $\mathcal{P}_r$ on some measurable space $(\mathcal{Y}, \mathcal{G})$, such that $\mathrm{F}(r) = 0, \forall \mathrm{F} \in \mathcal{P}_r$, where $r \in \mathcal{Y}$, including all continuous distributions on $\mathcal{Y}$. Consider the kernel score family $S_\rho(\mathrm{F}, y) = \frac{1}{2}\mathbb{E}_\mathrm{F}\rho(X, X') - \mathbb{E}_\mathrm{F}\rho(X, y) + \frac{1}{2}\rho(y, y)$ with divergence $\mathbb{D}_{S_\rho}(\mathrm{P}\|\mathrm{F}) = \mathbb{E}_{\mathrm{P},\mathrm{F}}\rho(X, Y) - \frac{1}{2}\mathbb{E}_\mathrm{P}\rho(X, X') - \frac{1}{2}\mathbb{E}_\mathrm{F}\rho(Y, Y')$. The associated generalized censored scoring rule for $\mathrm{H} = \delta_r$ reads $S^\flat_{\rho,w}(\mathrm{F}, y) = \frac{1}{2}\mathbb{E}_{\mathrm{F}^\flat_w}\rho(X, X') - w(y)\left(\mathbb{E}_{\mathrm{F}^\flat_w}\rho(X, y) - \frac{1}{2}\rho(y, y)\right) - \left(1 - w(y)\right)\left(\mathbb{E}_{\mathrm{F}^\flat_w}\rho(X, r) - \frac{1}{2}\rho(r, r)\right)$. Assumption 1 is clearly satisfied for all weight functions and distributions $\mathrm{F} \in \mathcal{P}_r$. Therefore, the score divergence $\mathbb{D}_{S^\flat_{\rho,w}}(\mathrm{P}\|\mathrm{F}) = \mathbb{E}_{\mathrm{P}^\flat_w,\mathrm{F}^\flat_w}\rho(X, Y) - \frac{1}{2}\mathbb{E}_{\mathrm{P}^\flat_w}\rho(X, X') - \frac{1}{2}\mathbb{E}_{\mathrm{F}^\flat_w}\rho(Y, Y')$ is a localized divergence if S_ρ is strictly proper relative to $\mathcal{P}^\flat_r$, which follows from the conditions under which S_ρ is strictly proper with respect to $\mathcal{P}_r$. Further details are given in Appendix D.4.*

3.3 Localized Neyman Pearson

Anticipating the applications in the next section, we now consider an explicit time-series context. Specifically, we consider a stochastic process $\{Y_t : \Omega \to \mathcal{Y}\}_{t=1}^T$ 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 $\{\mathcal{F}_t\}_{t=1}^T$, where $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$ 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.

The aim is to derive a uniformly most powerful (UMP) test for the hypotheses

$$\mathbb{H}_0 : p_t \mathbb{1}_{A_t} = f_{0t} \mathbb{1}_{A_t}, \ \forall t \qquad \text{vs.} \qquad \mathbb{H}_1 : p_t \mathbb{1}_{A_t} = f_{1t} \mathbb{1}_{A_t}, \quad \forall t, \tag{5}$$

with $f_{jt}, \ j \in \{0, 1\}$, fixed. Although the predictive densities under the null and alternative are known, the test concerns a multiple versus multiple hypothesis test due to the lacking specification of the density outside the regions of interest A_t. In other words, the densities $\big[f_{0t}\mathbb{1}_{A_t} + \big(\mathrm{F}_{0t}(A_t^c)/\mathrm{H}_t(A_t^c)\big) h_t \mathbb{1}_{A_t^c}\big]_{A_t}^\flat$ and $[f_{0t}]_{A_t}^\flat$ coincide, assuming $\mathrm{H}_t(A_t^c) > 0$. Here, $[\cdot]_w^\flat$ refers to censoring a distribution (function) or density according to Equation (2). Similarly, we use $[\cdot]_w^\sharp$ for conditioning. Theorem 3 reveals that this setting admits a UMP test, reducing to the Neyman and Pearson (1933) lemma when $A_t = \mathcal{Y}, \ \forall t$.

Theorem 3 (Localized Neyman Pearson)**.** *For any given $\alpha \in (0, 1)$, the UMP test of size α for testing problem* (5) *reads*

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

where $\phi_A^\flat : \mathcal{Y}^T \to [0, 1]$ denotes a test function specifying the rejection probability, 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 $\alpha = [\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) > c\big) + \gamma [\mathrm{F}_0]_A^\flat\big(\lambda(\mathbf{y}) = c\big)$.

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 the conditional operator does not bear a UMP test.

Corollary 1. *An alternative formulation of the UMP test for testing problem* (5) *is given by the test defined in Theorem 3 with* $\lambda(\mathbf{y})$ *replaced by* $\tilde{\lambda}(\mathbf{y}) := \sum_{t=0}^{T-1} \big(LogS^{\flat}_{A_t}(f_{1t}, y_{t+1}) - LogS^{\flat}_{A_t}(f_{0t}, y_{t+1})\big)$*, i.e., in terms of the CSL.*

Corollary 2. *For testing problem* (5)*, the test* $\phi^{\sharp}_A$*, which is defined as* $\phi^{\flat}_A$ *upon replacing* $\flat$ *by* $\sharp$*, is not UMP.*

3.4 Related weighted scoring rules

First, we compare our procedure with Holzmann and Klar (2017a), who base their approach on the conditional distribution $\mathrm{F}^{\sharp}_w$. Due to the normalization with respect to A_w, however, $\mathrm{P}^{\sharp}_w = \mathrm{F}^{\sharp}_w$ does *not* hold if and only if $\mathrm{P} \overset{A_w}{=} \mathrm{F}$. Consequently, the score divergence $\mathbb{D}_{S^{\sharp}_w}(\mathrm{P}\|\mathrm{F}) = (1-\bar{P}_w)\mathbb{D}_S(\mathrm{P}^{\sharp}_w\|\mathrm{F}^{\sharp}_w)$ fails to satisfy condition (ii) of localized divergences in Definition 2, unless $\bar{P}_w = \bar{F}_w$. To resolve this, they add an auxiliary scoring rule s, enforcing the score divergence to be zero if and only $\mathrm{P} \overset{A_w}{=} \mathrm{F}$. Example 5 describes their composite scoring rule, for which the score divergence is a local divergence.

Example 5. *Holzmann and Klar (2017a) propose a class of weighted scoring rules defined as* $\tilde{S}_{w,s}(\mathrm{F}, y) := S^{\sharp}_w(\mathrm{F}, y) + w(y)s(b_{1-\bar{F}_w}, 1) + \big(1-w(y)\big)s(b_{1-\bar{F}_w}, 0)$*, 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* $\mathbb{D}_{S_{(w,s)}}(\mathrm{P}\|\mathrm{F}) = (1-\bar{P}_w)\mathbb{D}_S(\mathrm{P}^{\sharp}_w\|\mathrm{F}^{\sharp}_w) + \mathbb{D}_s(b_{1-\bar{P}_w}\|b_{1-\bar{F}_w})$ *a local divergence. It is generally not a localized divergence due to the dependence on the auxiliary divergence. Indeed, this dependence can become substantial, including the extreme case that the composite divergence becomes independent of the original divergence* $\mathbb{D}_S$ *if* P *and* F *are proportional on* A_w*.*

Holzmann and Klar (2017a) promote two specific choices for s: $\mathrm{Logs}(b_{1-\bar{F}_w}, z) = z\log(1-\bar{F}_w) + (1-z)\log\bar{F}_w$ and $\bar{s}(b_{1-\bar{F}_w}, z) := z(\log(1-\bar{F}_w)+1) - (1-\bar{F}_w)$. The combination

$S = \text{LogS}$ and $s = \text{Logs}$ recovers $\text{LogS}^\flat_w$ and hence a localized divergence. In contrast, setting $S = \text{LogS}$ and $s = \bar{s}$ leads to the weighted likelihood score by Pelenis (2014), for which the score divergence $\mathbb{D}_{\text{pwl}_w}(\text{P}\|\text{F}) = \mathbb{D}_{\text{LogS}}(\text{P}^\flat_w\|\text{F}^\flat_w) - \bar{P}_w \log \frac{\bar{P}_w}{\bar{F}_w} + (1-\bar{P}_w)\log\frac{1-\bar{P}_w}{1-\bar{F}_w} + (\bar{P}_w - \bar{F}_w)$ is not a localized divergence of $\mathbb{D}_{\text{LogS}}$. The flexibility of the auxiliary scoring rule s allows for weighted scoring rules beyond the scope of (generalized) censored scoring rules. An example is the combination $S = \text{QS}$ and $s = \text{Logs}$, with local divergence $\mathbb{D}_{\text{QS}_{(w,\text{Logs})}}(p\|f) = (1-\bar{P}_w)\|p^\sharp_w - f^\sharp_w\|_2^2 + \mathbb{D}_{\text{Logs}}(b_{1-\bar{P}_w}\|b_{1-\bar{F}_w})$. A key distinction from the censored divergence $\mathbb{D}_{\text{QS}^\flat_{w,\text{H}}}(p\|f) = \|p^\flat_{w,\text{H}} - f^\flat_{w,\text{H}}\|_2^2$ is the role of the KL divergence, which may dominate the L^2-norm, particularly when $\bar{F}_w$ is large. Censoring is also not nested within the framework of Holzmann and Klar (2017a). Unlike censoring (see Example 3), the composite scoring rule $\tilde{S}_{w,s}(\text{F}, y)$ in Example 5 enforces $\tilde{S}_{w,s}(\text{F}, y)$ to be only a function of $\bar{F}_w$ if $y \in A^c_w$.

The second comparison concerns the threshold weighted kernel score introduced by Allen et al. (2023), which generalizes the twCRPS by introducing the kernel $\rho(v(x), v(x'))$ based on a measurable chaining function $v : \mathcal{Y} \mapsto \mathcal{Y}$. For indicator weight functions $w(y) = \mathbb{1}_A(y)$, the $\text{twS}_\rho(\text{F}, y; v, y_0)$ based on the identity chaining function on A reduces to the censored kernel $S^\flat_{\rho,r}$ score given in Example 4 for $r = y_0$. Given that $S^\flat_{\rho,r}$ is strictly locally proper with respect to $\mathcal{P}_r$ for all r, we concur with Allen et al. (2023) that the specific choice of y_0 is irrelevant for their theoretical results. However, as illustrated by Example C.1 and emphasized by Allen et al. (2023), taking y_0 at the boundary of the region of interest is more natural; and always feasible due to their focus on high impact events.

The injectivity condition on the chaining function for strict propriety is, at best, less evidently satisfied for general weight functions. While Allen et al. (2023) provide an example of a chaining function that is convex along each dimension for high-impact weight functions, weight functions associated with non-high-impact interest do not readily translate

into chaining functions that ensure the strict local propriety of twS_ρ. Similarly, choosing two pivotal points as in Example C.2 with a center indicator is incompatible with the condition $\rho(v(z), v(\cdot)) = \rho(v(z'), v(\cdot))$ required for the strict local propriety of twS_ρ.

Third and finally, we compare with Mitchell and Weale (2023), who, for real-valued, unimodal densities with the center as the region of interest, also consider censored density forecasts based on LogS. However, unlike our setting and Diks et al. (2011), they do not evaluate multiple candidate densities. As illustrated by Example 6, the adapted versions are not suitable for evaluating multiple density forecasts because the introduced dependence of the region of interest on the candidate distribution renders the scoring rule improper.

Example 6. *Mitchell and Weale (2023) consider the alternative censored likelihood score* $\text{Log}_\alpha^{\text{MW}}(f, y) := \log f(y)\mathbb{1}_{A(F;\alpha)}(y) + \log(\alpha)\mathbb{1}_{A(F;\alpha)^c}(y)$, *where* $A(F;\alpha) := [F^{-1}(\alpha/2), F^{-1}(1-\alpha/2)]$. *An important difference with the censored likelihood score* $\log f_A^{\text{b}}(y)$ *is the dependence of the region of interest on* f, *by which* $\text{Log}_\alpha^{\text{MW}}(f, y)$ *is improper. For instance, consider Gaussian densities* p *and* f *with mean zero and variances* $\sigma_f^2 = 1/2$ *and* $\sigma_p^2 = 1$, *respectively. Then,* $\mathbb{E}_p\text{LogS}_\alpha^{\text{MW}}(p, Y) - \mathbb{E}_p\text{LogS}_\alpha^{\text{MW}}(f, Y) < 0$, $\forall \alpha > 0.052$.

3.5 Practical guidance

The censoring procedure proposed in this paper is sufficiently general to enable researchers to construct censored analogs of their preferred scoring rules for practically any chosen family of weight functions. In particular, for semi-local scoring rules Definition 5 provides an analytical 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 Logarithmic (LogS), Power (PowS_α) and PseudoSpeherical (PsSphS_α) families of scoring rules, as well as the corresponding localized divergences derived from the identity

$\mathbb{D}_{S_w^\flat}(\mathrm{P}\|\mathrm{F}) = \mathbb{D}_S(\mathrm{P}_w^\flat\|\mathrm{F}_w^\flat)$. For the unweighted rules, we rely on the results advocated by Gneiting and Raftery (2007) and Ovcharov (2018).

Finiteness of $\mathbb{D}_S$ when applied to censored measures, as desired for a localized divergence, is generally easily guaranteed; see Theorem 1 and Example 3. For comparison, the conditional rules are also included, as well as the main characteristics of the original rules and the generalized censored scoring rules. The latter reveals insensitivity to the reference distribution because the censored scoring rules are independent of this choice as long as the reference distribution is normalized to $\|h\|_\alpha = 1$.

For distance-sensitive scoring rules, such as the kernel scores introduced in Example 4, we use the generalized censoring approach outlined in Definition 6 based on the censored measure in Equation (3). The localized versions of distance-sensitive scoring rules thereby depend on the choice of the pivotal points and associated weights. In agreement with Allen et al. (2023), the theoretical properties of censored distance-sensitive scoring rules typically remain unaffected, but as illustrated by Examples C.1 and C.2, weight functions often suggest natural choices for pivotal points, and it is these points we recommend incorporating into the censored measure. Specifically, for the real-valued weight functions $I_{\mathrm{L}}(y;r) := \mathbb{1}_{(-\infty,r)}(y)$, $I_{\mathrm{R}}(y;r) := \mathbb{1}_{(-\infty,r)}(y)$, $\Lambda_{a,\mathrm{L}}(y;r) := \frac{1}{1+\exp(a(y-r))}$, $a > 0$, $r \in \mathbb{R}$ is pivotal. Similarly, for the weight functions on $\mathbf{y} \in \mathbb{R}^2$ given by $I^2_{\mathrm{L}}(\mathbf{y};\mathbf{r}) := I_{\mathrm{L}}(y_1;r_1) \times I_{\mathrm{L}}(y_2;r_2)$, $I^2_{\mathrm{R}}(\mathbf{y};\mathbf{r}) := I_{\mathrm{R}}(y_1;r_1) \times I_{\mathrm{R}}(y_2;r_2)$, $\Lambda^2_{a,\mathrm{L}}(\mathbf{y};\mathbf{r}) := \Lambda_{a,\mathrm{L}}(y_1;r_1) \times \Lambda_{a,\mathrm{L}}(y_2;r_2)$ and $I^2_{\mathrm{LS}}(\mathbf{y};\mathbf{r}) := \mathbb{1}_{(-\infty,r)}(y_1+y_2)$ the use of $\mathbf{r} \in \mathbb{R}^2$, or $r \in \mathbb{R}$ in the last case, is considered natural. For the center indicator, $I_{\mathrm{C}}(y;r,\ell) := \mathbb{1}_{(r-\ell,r+\ell)}(y)$, there are two pivotal points $r_{1,2} = r \pm \ell$. In Section 4, we use the proportion of data smaller than r_1 as estimate for the weight γ.

Furthermore, as clarified in Section 3.2, Assumption 1 is readily satisfied by the censored distribution in Equation (3). For instance, if the underlying distributions are continuous,

Table 1: Examples of semi-local scoring rules.

Name	Logarithmic	Power family	PseudoSpherical family
		unweighted	
$S(f,y)$	$\mathrm{LogS}(f,y)=\log f(y)$	$\mathrm{PowS}_\alpha(f,y)=\alpha f(y)^{\alpha-1}-(\alpha-1)\|f\|_\alpha^\alpha,\quad \alpha>1$	$\mathrm{PsSphS}_\alpha(f,y)=\frac{f(y)^{\alpha-1}}{\|f\|_\alpha^{\alpha-1}},\quad \alpha>1$
Special cases	–	$\mathrm{QS}(f,y)=\mathrm{PowS}_2(f,y)$	$\mathrm{SphS}(f,y)=\mathrm{PsSphS}_2(f,y)$
		$\mathrm{LogS}(f,y)=\lim_{\alpha\downarrow 1}\mathrm{PowS}_\alpha(f,y)$	$\mathrm{LogS}(f,y)=\lim_{\alpha\downarrow 1}\mathrm{PsSphS}_\alpha(f,y)$
$\mathbb{D}_S(p\|f)$	$\mathrm{KL}(p\|f)=\mathbb{E}_p\log\left(\frac{p}{f}\right)$	$\|p\|_\alpha^\alpha-\alpha\int f^{\alpha-1}(p-f)\mathrm{d}\mu-\|f\|_\alpha^\alpha$	$\|p\|_\alpha-\frac{\int pf^{\alpha-1}\mathrm{d}\mu}{\|f\|_\alpha^{\alpha-1}}$
$\alpha=2$	–	$\|p-f\|_2^2$	$\|p\|_2\,(1-C(p,f))$
SP class	$\mathcal{P}_{\alpha=1}$	$\mathcal{P}_\alpha$	$\mathcal{P}_\alpha$
$\zeta(t)$	$t\log t$	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	
$S_w^\sharp(f,y)$	$w(y)\log\left(\frac{f(y)}{1-F_w}\right)$	$w(y)\left(\alpha\left(\frac{f_w(y)}{1-F_w}\right)^{\alpha-1}-(\alpha-1)\left\|\frac{f_w(y)}{1-F_w}\right\|_\alpha^\alpha\right)$	$w(y)\frac{f_w(y)^{\alpha-1}}{\|f_w\|_\alpha^{\alpha-1}}$
$S_w^\flat(f,y)$	$w(y)\log f(y)+\big(1-w(y)\big)\log\bar F_w$	$w(y)\alpha f_w(y)^{\alpha-1}+\big(1-w(y)\big)\alpha\bar F_w^{\alpha-1}$ $-(\alpha-1)\left(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\right)$	$\frac{w(y)f_w(y)^{\alpha-1}+(1-w(y))\bar F_w^{\alpha-1}}{\left(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$
$S_{w,h}^\flat(f,y)$	$w(y)\log f(y)+\big(1-w(y)\big)\log\bar F_w$	$w(y)\alpha f_w(y)^{\alpha-1}+\big(1-w(y)\big)\alpha\bar F_w^{\alpha-1}\|h\|_\alpha^\alpha$ $-(\alpha-1)\left(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\|h\|_\alpha^\alpha\right)$	$\frac{w(y)f_w(y)^{\alpha-1}+\big(1-w(y)\big)\bar F_w^{\alpha-1}\|h\|_\alpha^\alpha}{\left(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\|h\|_\alpha^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$
$\mathbb{D}_{S_w^\flat}(p\|f)$	$\int\log\left(\frac{p_w}{f_w}\right)p_w\mathrm{d}\mu+\log\left(\frac{\bar P_w}{\bar F_w}\right)\bar P_w$	$\|p_w\|_\alpha^\alpha+\bar P_w^\alpha-\int p_w f_w^{\alpha-1}\mathrm{d}\mu-\bar P_w\bar F_w^{\alpha-1}$ $-(\alpha-1)\big(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\big)$	$\big(\|p_w\|_\alpha^\alpha+\bar P_w^\alpha\big)^{\frac{1}{\alpha}}-\frac{\int p_w f_w^{\alpha-1}\mathrm{d}\mu+\bar P_w\bar F_w^{\alpha-1}}{\left(\|f_w\|_\alpha^\alpha+\bar F_w^\alpha\right)^{\frac{\alpha-1}{\alpha}}}$

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 $\|p\|_\alpha=(\int_{\mathcal{Y}}p^\alpha\mathrm{d}\mu)^{1/\alpha}$. The common limiting case of PowS_α and PsSphS_α remains to hold for conditioning and censoring, i.e., $\lim_{\alpha\downarrow1}\frac{1}{\alpha-1}\mathrm{PsSphS}_{\alpha,w}^x(f,y)=\lim_{\alpha\downarrow1}\frac{1}{\alpha-1}\mathrm{PowS}_{\alpha,w}^x(f,y)=\mathrm{LogS}^x(f,y)$, $x\in\{\sharp,\flat\}$. $\mathbb{D}_S(p\|f)$ denotes the score divergence of f from p and $C(p,f)=\int pf\mathrm{d}\mu/\sqrt{\int p^2\mathrm{d}\mu\int f^2\mathrm{d}\mu}$, 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. $\mathcal{P}_\alpha$ 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 $S_\zeta:\mathcal{P}(\mathcal{Y},\mathcal{G})\times\mathcal{Y}\to\mathbb{R}$ of the form $S_\zeta(p,y)=\zeta'(p(y))-\int_{\mathcal{Y}}\zeta'(p(y))p(y)-\zeta(p(y))\mu(\mathrm{d}y)$. $S(\tilde f,\tilde y)$ denotes the score of the real-valued random variable $\tilde Y=bY+a$, where $a\in\mathbb{R}$ and $b\in\mathbb{R}\backslash\{0\}$, with density $\tilde f(\tilde y)=\frac{1}{|b|}f\big(\frac{\tilde y-a}{b}\big)$. 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 $S_{w,h}^\flat$ departs from a density h of which the support is a subset of $\{w=0\}\subseteq\mathcal{Y}$. The weight function is restricted accordingly. Appendix D details the derivations of the results presented in this table.

any value of r is valid; if the distribution is discrete or continuous-discrete, any r at which the distributions under consideration exhibit no point mass may be chosen.

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 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 to the forecast method itself. Let $\hat{f}_t$ and $\hat{g}_t$ denote density forecasts resulting from competing methods, each estimated with a rolling window of length m. Following Diks et al. (2011) and Holzmann and Klar (2017b), we test the null hypothesis of equal predictive ability, $\mathbb{H}_0 : \mathbb{E}_{p_t} S_w(\hat{f}_t, Y_{t+1}) = \mathbb{E}_{p_t} S_w(\hat{g}_t, Y_{t+1})$, by means of the test statistic $t_{m,n} := \frac{1}{n} \sum_{t=m}^{T-1} \big(S_w(\hat{f}_t, Y_{t+1}) - S_w(\hat{g}_t, Y_{t+1})\big) / \sqrt{\hat{\sigma}^2_{m,n}/n}$, where $n = T - m$ is the number of observations used for evaluation and $\hat{\sigma}^2_{m,n}$ is a heteroskedasticity and autocorrelation-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. Because this null differs from that in (5), we currently lack 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 F.

In addition to conditional scoring rules as such, we also evaluate the proposal of Holzmann and Klar (2017a) to augment the conditional scoring rule with the auxiliary rule

sbar or slog, see Section 3.4. This procedure yields a composite scoring rule, for which the logarithmic component can become dominant. Consequently, it becomes difficult to attribute higher power of the test to either the localization method or the advantageous properties of the Logarithmic score. In Appendix **??**, we analyze the contribution of the auxiliary rules sbar and slog to standardized $\mathbb{D}_{S_w}(p_t\|\hat{g}_t)$.

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 TR:$= \max_{i,j\in\mathcal{M}_k} |t_{m,n}^{(i,j)}|$ statistic, where $t_{m,n}^{(i,j)}$ corresponds to the pairwise $t_{m,n}$-statistic between forecast methods i and j introduced above. Table H.1.b in Appendix H.1 includes results for the other statistic proposed by Hansen et al. (2011). 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 **??**, using a block bootstrap with $B = 10,000$ replications and block length $b = 5$, unless stated otherwise. Our results are robust to variations in these parameters. Model confidence sets obtained with censored and conditional scoring rules are denoted $\text{MCS}^{\flat}$ and $\text{MCS}^{\sharp}$, respectively.

In each application below, the unweighted scoring rules are given by LogS, QS, SphS and S_{ρ_1}, with kernel $\rho_1(\mathbf{x}_1, \mathbf{x}_2) := \|\mathbf{x}_1 - \mathbf{x}_2\|$, 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. If $\text{CRPS}^{\flat}$ and twCRPS do not coincide, twCRPS is included as

reference. Hence, we consider 12 or 13 weighted scoring rules per application. There are $|\mathcal{M}_0| = 6$ candidate forecast methods, with specifications differing by application. For reproducibility, Appendix G includes details on the specification of the individual methods. Moreover, Appendix H discusses the MCS p-values per individual scoring rule and weight function underlying the summary results presented in this section.

4.1 Financial risk management

Evaluating 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 (VaR) and Expected Shortfall (ES). To achieve the associated focus on the left tail of the density forecast, we use the indicator weight function $I_{\mathrm{L}}(y_t; \hat{r}_t^q)$, where $\hat{r}_t^q$ denotes the q-th empirical quantile of y_t, based on the same fixed rolling window of length $m = 1000$ as for the estimation of the forecast methods. We evaluate the resulting scoring rules for 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.

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 $h = 1$ and 5 days.

Table 2 reveals stark differences in the cardinality of $\mathrm{MCS}^\flat$ and $\mathrm{MCS}^\sharp$, particularly at

$h = 1$. In case no correction is applied to the conditional scoring rules, $\text{MCS}^{\flat}$ is strictly smaller than $\text{MCS}^{\sharp}$ in more than 70% of all replications, while $\text{MCS}^{\sharp}$ contains more than twice the number of methods compared to $\text{MCS}^{\flat}$ on average. These results moderate somewhat when the conditional scoring rules are appended with a Holzmann-Klar correction term. Nevertheless, $\text{MCS}^{\flat}$ remains strictly smaller than $\text{MCS}^{\sharp}$ in close to 40% of the cases, with an average difference in cardinality of 20%. For $h = 5$, the differences become smaller but remain in favor of censoring, averaging between 10 and 30%.

We supplement our statistical forecast assessment with one and five step ahead VaR and ES calculations, while recognizing that these single quantile or conditional moment measures do not fully capture a forecast distribution's tail. As shown in Appendix G.4, the censored MCS, while often being smaller, contains well-fitted (VaR, ES) pairs more than twice as often as its conditional analogs.

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). 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 $I_{\text{L}}^2(\mathbf{y}_t; \tilde{\mathbf{r}}_t^q)$ and $\Lambda_{a,\text{L}}^2(\mathbf{y}_t; \tilde{\mathbf{r}}_t^q)$, with $a = 2$, while having verified stability of results for $a \in \{1, 3\}$. 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.

Similar findings emerge for the bivariate density forecasts for Energy and Financial sector returns, albeit with some notable differences. 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 78% and 94%. Between 33% and 44% of cases $\text{MCS}^{\flat}$ even is strictly smaller than $\text{MCS}^{\sharp}$. Second, the former percentages are hardly affected by adding the Holzmann-Klar correction terms to the conditional scoring rule, while the latter decline quite substantially. For example, using the slog correction term for the scoring rules focused with weight function $I^2_{\text{L}}(\mathbf{y}_t; \tilde{\mathbf{r}}^q_t)$, it drops from 44% to 17% for $h = 1$. Hence, the correction terms result in equally large MCSs for the censored and conditional scoring rules more frequently, where closer inspection reveals that their compositions almost always are identical as well. Third, the results for the logistic weight function $\Lambda^2_{a,\text{L}}(\mathbf{y}_t; \tilde{\mathbf{r}}^q_t)$ do not differ much from those for the indicator weight function $I^2_{\text{L}}(\mathbf{y}_t; \tilde{\mathbf{r}}^q_t)$, which is recovered for $a \to \infty$. However, the discrepancies in cardinality are somewhat moderated, particularly for $h = 5$, 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, the results for $h = 1$ and 5 are quite similar; hence, 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 $h = 5$.

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[1] and European Central Bank[2] target an annual inflation rate of 2%, we focus on the central range $A_\ell = (2 - \ell, 2 + \ell)$, where $\ell > 0$, by using the weight function $I_{\text{C}}(y_t; 2, \ell)$. To address policymakers' concerns for deviations beyond A_ℓ, termed 'Inflation at Risk' (Lopez-Salido and Loria 2020), we additionally consider its complement

[1]Source: `https://federalreserve.gov/monetarypolicy/files/fomc_longerrungoals.pdf`

[2]Source: `https://ecb.europa.eu/mopo/implement/app/html/index.en.html`

Table 2: Changes in MCS cardinality between censored and conditional scoring rules

			no correction			sbar			slog		
Sec.	w_t	h	$\leq$	$<$	$\sharp/\flat$	$\leq$	$<$	$\sharp/\flat$	$\leq$	$<$	$\sharp/\flat$
4.1	$I_{\mathrm{L}}(y_t;\hat{r}_t^q)$	1	96%	71%	2.28	67%	38%	1.18	71%	38%	1.20
		5	62%	29%	1.29	54%	25%	1.08	62%	25%	1.10
	$I_{\mathrm{L}}^2(\mathbf{y}_t;\tilde{\mathbf{r}}_t^q)$	1	78%	44%	1.38	78%	28%	1.16	72%	17%	1.05
		5	94%	33%	1.31	94%	28%	1.37	100%	28%	1.38
	$\Lambda_{2,\mathrm{L}}^2(\mathbf{y}_t;\tilde{\mathbf{r}}_t^q)$	1	83%	44%	1.20	83%	28%	1.07	83%	28%	1.07
		5	78%	44%	1.28	94%	33%	1.28	100%	28%	1.28
4.2	$I_{\mathrm{C}}(y_t;2,\ell_1)$	6	100%	92%	2.00	83%	50%	1.13	67%	33%	1.17
		24	92%	75%	2.86	42%	33%	1.35	67%	8%	0.88
	$I_{\mathrm{C}}^c(y_t;2,\ell_1)$	6	100%	83%	2.91	100%	83%	2.66	100%	75%	2.01
		24	100%	67%	2.33	75%	33%	1.22	83%	17%	1.23
4.3	$I_{\mathrm{R}}(y_t;\hat{r}_t^q)$	1	83%	58%	1.94	92%	67%	1.90	79%	25%	1.21
		3	79%	46%	1.56	88%	42%	1.44	79%	4%	0.94
	$I_{\mathrm{C}}(y_t;18,\ell_2)$	1	92%	42%	1.54	92%	8%	1.04	75%	8%	0.96
		3	100%	58%	1.58	100%	0%	1.00	100%	0%	1.00
Total average			89%	55%	1.83	81%	35%	1.36	81%	23%	1.18

NOTE: This table presents changes in cardinality of the MCS in absolute and relative terms, at confidence level 0.90, across different forecast horizons h, corresponding to the forecasting applications in risk management (Sec. 4.1), inflation (Sec. 4.2) and temperature (Sec. 4.4). 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 $\mathrm{MCS}^{\flat}$ contains (strictly) fewer forecast methods than $\mathrm{MCS}^{\sharp}$ and the column labeled $\sharp/\flat$ reports the ratio $|\mathrm{MCS}^{\sharp}|/|\mathrm{MCS}^{\flat}|$. Each result represents an average over a set of scoring rules $S \in \{\mathrm{LogS}, \mathrm{QS}, \mathrm{SphS}, \mathrm{CRPS}/S_{\rho_1}\}$ and quantile levels $q \in \{0.01, 0.05, 0.10, 0.15, 0.20, 0.25\}$ or levels $\ell_1 \in \{1, 1.5, 2\}$ and $\ell_2 \in \{1, 2, 4\}$. The empirical q-th quantiles $\hat{r}_t^q$ of y_t are based on the forecast method estimation window ($m = 1,000$), 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 H.

$I_{\mathrm{C}}^c(y_t;2,\ell) := 1 - I_{\mathrm{C}}(y_t;2,\ell)$. For the CRPS, we adopt two pivotal points for $I_{\mathrm{C}}(y_t;2,\ell)$,while using $r = 2$ 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 h-month inflation rates $y_{t+h}^h = (1200/h)\log\left(P_{t+h}/P_t\right)$, where P_t denotes the U.S. consumer price index (CPI) in month t, for horizons $h = 6$ 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 timeframe.

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 $y^h_{t+h} = \mu^h_{j,t+h}(\mathbf{x}_t) + u^h_{t+h}$, where we consider the following subset of methods listed by Medeiros et al. (2021) for the conditional mean $\mu^h_{j,t+h}$: Random Walk, Auto-Regressive model (AR), Bagging, Complete Subset Regression (CSR), Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest. The error u^h_{t+h} 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 $\text{MCS}^{\flat}$ are almost always (weakly) smaller than those of $\text{MCS}^{\sharp}$. The relative increase in set cardinality when opting for conditioning over censoring is substantial at at least 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, albeit 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 $I_{\mathrm{R}}(y_t; \hat{r}_t^q)$ and temperatures near the optimal temperature for tuber growth, approximately 18 degrees Celsius (Struik 2007, Section 18.5.5), using $I_{\mathrm{C}}(y_t; 18, \ell)$. 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 the first $m = 2922$ days (or 8 years) serving as the initial estimation window. 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 weight function $I_{\mathrm{R}}(y_t; \hat{r}_t^q)$ to focus on the right tail, corresponding to 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 $h = 1$ day-ahead forecasts; and the differences diminish at the longer forecast horizon $h = 3$ 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 $I_{\mathrm{C}}(y_t; 18, \ell)$, we find that there are no instances where conditioning leads to a smaller MCS for $h = 3$ and almost no such cases for $h = 1$, similar to the inflation forecasts focused on the central range around the 2% target. Relative to inflation, there is a notable increase in cases in which the MCSs possess identical cardinality, which is also reflected in the smaller ratios $|\mathrm{MCS}^\sharp|/|\mathrm{MCS}^\flat|$.

5 CONCLUSION

In many applications, forecasters are particularly interested in specific areas of the outcome space. Addressing this, we propose 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 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 solving its localization bias for 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 one for all outcomes. By 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 stylized scenario in which the candidates' tails are close to proportional.

To analyze real performance, we use the size of the Model Confidence Set (MCS) as a proxy for 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 focused forecast assessments of S&P500 and temperature data, a comparable pattern emerges, corroborating the enhanced power of censoring.

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.

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.

Avalaible 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.