Statistical Science
2017, Vol. 32, No. 1, 106–127
DOI: 10.1214/16-STS588

Forecaster's Dilemma: Extreme Events and Forecast Evaluation

Sebastian Lerch, Thordis L. Thorarinsdottir, Francesco Ravazzolo and Tilmann Gneiting

Abstract. In public discussions of the quality of forecasts, attention typically focuses on the predictive performance in cases of extreme events. However, the restriction of conventional forecast evaluation methods to subsets of extreme observations has unexpected and undesired effects, and is bound to discredit skillful forecasts when the signal-to-noise ratio in the data generating process is low. Conditioning on outcomes is incompatible with the theoretical assumptions of established forecast evaluation methods, thereby confronting forecasters with what we refer to as the forecaster's dilemma. For probabilistic forecasts, proper weighted scoring rules have been proposed as decision-theoretically justifiable alternatives for forecast evaluation with an emphasis on extreme events. Using theoretical arguments, simulation experiments and a real data study on probabilistic forecasts of U.S. inflation and gross domestic product (GDP) growth, we illustrate and discuss the forecaster's dilemma along with potential remedies.

Key words and phrases: Diebold–Mariano test, hindsight bias, likelihood ratio test, Neyman–Pearson lemma, predictive performance, probabilistic forecast, proper weighted scoring rule, rare and extreme events.

Quod male consultum cecidit feliciter, Ancus,
Arguitur sapiens, quo modo stultus erat;
Quod prudenter erat provisum, si male vortat,
Ipse Cato (populo iudice) stultus erat.[1]

John Owen, 1607

1. INTRODUCTION

Extreme events are inherent in natural or man-made systems and may pose significant societal challenges. The development of the theoretical foundations for the study of extreme events started in the middle of the last century and has received considerable interest in various applied domains, including but not limited to meteorology, climatology, hydrology, finance and economics. Topical reviews can be found in the work of

Sebastian Lerch is PostDoc, Heidelberg Institute for Theoretical Studies (HITS), and Institute for Stochastics, Karlsruhe Institute of Technology, HITS gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany (e-mail: sebastian.lerch@h-its.org). Thordis L. Thorarinsdottir is Chief Research Scientist, Norwegian Computing Center, P.O. Box 114, Blindern, 0314 Oslo, Norway (e-mail: thordis@nr.no). Francesco Ravazzolo is Associate Professor, Free University of Bozen/Bolzano, Universitätsplatz 1, 39100 Bozen-Bolzano, Italy (e-mail: francescoravazzolo@gmail.com). Tilmann Gneiting is Group Leader, Heidelberg Institute for Theoretical Studies (HITS), and Professor of Computational Statistics, Institute for Stochastics, Karlsruhe Institute of Technology, HITS gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany (e-mail: tilmann.gneiting@h-its.org).

[1] Owen (1607), 216. *Sapientia duce, comite fortuna. In Ancum.* English translation by Edith Sylla; see Bernoulli (2006):
Because what was badly advised fell out happily,
Ancus is declared wise, who just now was foolish;
Because of what was prudently prepared for, if it turns out badly,
Cato himself, in popular opinion, will be foolish.

Gumbel (1958), Embrechts, Klüppelberg and Mikosch (1997), Easterling et al. (2000), Coles (2001), Katz, Parlange and Naveau (2002), Beirlant et al. (2004) and Albeverio, Jentsch and Kantz (2006), among others. Not surprisingly, accurate predictions of extreme events are of great importance and demand. In many situations, distinct models and forecasts are available, thereby calling for a comparative assessment of their predictive performance with particular emphasis placed on extreme events.

In the public, forecast evaluation often only takes place once an extreme event has been observed, in particular, if forecasters have failed to predict an event with high economic or societal impact. Table 1 gives examples from newspapers, magazines, and broadcasting corporations that demonstrate the focus on extreme events in finance, economics, meteorology and seismology. Striking examples include the international financial crisis of 2007/08 and the L'Aquila earthquake of 2009. After the financial crisis, much attention was paid to economists who had correctly predicted the crisis, and a superior predictive ability was attributed to them. In 2011, against the protest of many scientists around the world, a group of Italian seismologists was put on trial for not warning the public of the devastating L'Aquila earthquake of 2009 that caused 309 deaths (Hall, 2011). Six scientists and a government official were found guilty of involuntary manslaughter in October 2012 and sentenced to six years of prison each. In November 2015, the scientists were acquitted by the Supreme Court in Rome, whereas the sentence of the deputy head of Italy's civil protection department, which had been reduced to two years in 2014, was upheld.

TABLE 1

Media coverage illustrating the focus on extreme events in public discussions of the quality of forecasts. A version of the table with links to the sources is provided in an online supplement (Lerch et al., 2016)

Year	**Headline**	**Source**
2008	Dr. Doom	The New York Times
2009	How did economists get it so wrong?	The New York Times
2009	He told us so	The Guardian
2010	An exclusive interview with Med Yones—The expert who predicted the financial crisis	CEO Q Magazine
2011	A seer on banks raises a furor on bonds	The New York Times
2013	Meredith Whitney redraws "map of prosperity"	USA Today
2007	Lessons learned from Great Storm	BBC
2011	Bad data failed to predict Nashville flood	NBC
2012	Bureau of Meteorology chief says super storm "just blew up on the city"	The Courier-Mail
2013	Weather Service faulted for Sandy storm surge warnings	NBC
2013	Weather Service updates criteria for hurricane warnings, after Sandy criticism	Washington Post
2015	National Weather Service head takes blame for forecast failures	NBC
2011	Italian scientists on trial over L'Aquila earthquake	CNN
2011	Scientists worry over "bizarre" trial on earthquake prediction	Scientific American
2012	L'Aquila ruling: Should scientists stop giving advice?	BBC

At first sight, the practice of selecting extreme observations, while discarding nonextreme ones, and to proceed using standard evaluation tools appears to be a natural approach. Intuitively, accurate predictions on the subset of extreme observations may suggest superior predictive ability. However, the restriction of the evaluation to subsets of the available observations has unwanted effects that may discredit even the most skillful forecast available (Denrell and Fang, 2010, Diks, Panchenko and van Dijk, 2011, Gneiting and Ranjan, 2011). In a nutshell, if forecast evaluation proceeds conditionally on a catastrophic event having been observed, always predicting calamity becomes a worthwhile strategy. Given that media attention tends to focus on extreme events, skillful forecasts are bound to fail in the public eye, and it becomes tempting to base decision-making on misguided inferential procedures. We refer to this critical issue as the *forecaster's dilemma*.[2]

To demonstrate the phenomenon, we let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with mean μ and standard deviation σ and consider the following simple experiment. Let the observation Y satisfy

$$(1.1) \quad Y|\mu \sim \mathcal{N}(\mu, \sigma^2) \quad \text{where } \mu \sim \mathcal{N}(0, 1-\sigma^2).$$

[2]Our notion of the *forecaster's dilemma* differs from a previous usage of the term in the marketing literature by Ehrman and Shugan (1995), who investigated the problem of influential forecasting in business environments. The forecaster's dilemma in influential forecasting refers to potential complications when the forecast itself might affect the future outcome, for example, by influencing which products are developed or advertised.

TABLE 2

Forecasts in the simulation study, where the observation Y satisfies (1.1) with $\sigma^2 = \frac{2}{3}$ being fixed. The mean absolute error (MAE) and mean squared error (MSE) for the point forecast X are based on a sample of size 10,000; the restricted versions rMAE and rMSE are based on the subset of observations exceeding 1.64 only. The lowest value in each column is in bold

Forecast	**Predictive distribution**	X	**MAE**	**MSE**	**rMAE**	**rMSE**
Perfect	$\mathcal{N}(\mu, \sigma^2)$	μ	**0.64**	**0.67**	1.35	2.12
Unconditional	$\mathcal{N}(0, 1)$	0	0.80	0.99	2.04	4.30
Extremist	$\mathcal{N}(\mu + \frac{5}{2}, \sigma^2)$	$\mu + \frac{5}{2}$	2.51	6.96	**1.16**	**1.61**

Table 2 introduces forecasts for Y, showing both the predictive distribution, F, and the associated point forecast, X, which we take to be the respective median or mean.[3] The perfect forecast has knowledge of μ, while the unconditional forecast is the unconditional standard normal distribution of Y. The deliberately misguided extremist forecast shows a constant bias of $\frac{5}{2}$. As expected, the perfect forecast is preferred under both the mean absolute error (MAE) and the mean squared error (MSE). However, these results change completely if we restrict attention to the largest 5% of the observations, as shown in the last two columns of the table, where the misguided extremist forecast receives the lowest mean score.

In this simple example, we have considered point forecasts only, for which there is no obvious way to abate the forecaster's dilemma by adapting existing forecast evaluation methods appropriately, such that particular emphasis can be put on extreme outcomes. Probabilistic forecasts in the form of predictive distributions provide a suitable alternative. Probabilistic forecasts have become popular over the past few decades, and in various key applications there has been a shift of paradigms from point forecasts to probabilistic forecasts, as reviewed by Tay and Wallis (2000), Timmermann (2000), Gneiting (2008) and Gneiting and Katzfuss (2014), among others. As we will see, the forecaster's dilemma is not limited to point forecasts and occurs in the case of probabilistic forecasts as well. However, in the case of probabilistic forecasts extant methods of forecast evaluation can be adapted to place emphasis on extremes in decision-theoretically coherent ways. In particular, it has been suggested that suitably weighted scoring rules allow for the comparative evaluation of probabilistic forecasts with emphasis on extreme events (Diks, Panchenko and van Dijk, 2011, Gneiting and Ranjan, 2011).

The contributions of this expository article lie in the novelty of the interpretations, rather than methodological development, and the remainder of the paper is organized as follows. In Section 2, theoretical foundations on forecast evaluation and proper scoring rules are reviewed, serving to analyze and explain the forecaster's dilemma along with potential remedies. In Section 3, this is followed up and illustrated in simulation experiments. Furthermore, we elucidate the role of the fundamental lemma of Neyman and Pearson, which suggests the superiority of tests of equal predictive performance that are based on the classical, unweighted logarithmic score. A case study on probabilistic forecasts of gross domestic product (GDP) growth and inflation for the United States is presented in Section 4. The paper closes with a discussion in Section 5.

2. FORECAST EVALUATION AND EXTREME EVENTS

We now review relevant theory that is then used to study and explain the forecaster's dilemma.

2.1 The joint distribution framework for forecast evaluation

In the following, the forecast and the observation are treated as random variables, the distributions of which are denoted by square brackets. In a seminal paper on the evaluation of point forecasts, Murphy and Winkler (1987) argued that the assessment ought to be based on the joint distribution of the forecast, X, and the observation, Y, building on both the *calibration-refinement factorization*,

$$[X, Y] = [X]\,[Y|X],$$

and the *likelihood-baserate factorization*,

$$[X, Y] = [Y]\,[X|Y].$$

[3]The predictive distributions are symmetric, so their mean and median coincide. We use X in upper case, as the point forecast may depend on μ and, therefore, is a random variable.

Gneiting and Ranjan (2013), Ehm et al. (2016) and Strähl and Ziegel (2015) extend and adapt this framework to include the case of potentially multiple probabilistic forecasts. In this setting, the probabilistic forecasts and the observation form tuples

$$(F_1, \dots, F_k, Y),$$

where the predictive distributions $F_1, \dots, F_k$ are cumulative distribution function (CDF)-valued random quantities on the outcome space of the observation, Y.

Considering the case of a single probabilistic forecast, F, the above factorizations have immediate analogues in this setting, namely, the calibration-refinement factorization

$$[F, Y] = [F]\,[Y|F] \tag{2.1}$$

and the likelihood-baserate factorization

$$[F, Y] = [Y]\,[F|Y]. \tag{2.2}$$

The components of the calibration-refinement factorization (2.1) can be linked to the sharpness and the calibration of a probabilistic forecast (Gneiting, Balabdaoui and Raftery, 2007). Sharpness refers to the concentration of the predictive distributions and is a property of the marginal distribution of the forecasts only. Calibration can be interpreted in terms of the conditional distribution of the observation, Y, given the probabilistic forecast, F.

Various notions of calibration have been proposed, with the concept of autocalibration being particularly strong. Specifically, a probabilistic forecast F is *autocalibrated* if

$$[Y|F] = F \tag{2.3}$$

almost surely (Tsyplakov, 2013). This property carries over to point forecasts, in that, given any functional T, such as the mean or expectation functional, or a quantile, autocalibration implies $\mathrm{T}([Y|F]) = \mathrm{T}(F)$. Furthermore, if the point forecast $X = \mathrm{T}(F)$ characterizes the probabilistic forecast, as is the case in Table 2, where T can be taken to be the mean or median functional, then autocalibration implies

$$\mathrm{T}([Y|X]) = \mathrm{T}([Y|F]) = \mathrm{T}(F) = X. \tag{2.4}$$

This property can be interpreted as unbiasedness of the point forecast $X = \mathrm{T}(F)$ that is induced by the predictive distribution F.

Finally, a probabilistic forecast F is *probabilistically calibrated* if the probability integral transform $F(Y)$ is uniformly distributed, with suitable technical adaptations in cases in which F may have a discrete component (Gneiting, Balabdaoui and Raftery, 2007, Gneiting and Ranjan, 2013). An autocalibrated predictive distribution is necessarily probabilistically calibrated (Gneiting and Ranjan, 2013, Strähl and Ziegel, 2015).

In contrast, the interpretation of the second component $[F|Y]$ in the likelihood-baserate factorization (2.2) is much less clear. While the conditional distribution of the forecast given the observation can be viewed as a measure of discrimination ability, it was noted by Murphy and Winkler (1987) that forecasts can be perfectly discriminatory although they are uncalibrated. Therefore, discrimination ability by itself is not informative, and forecast assessment might be misguided if one stratifies by the realized value of the observation. To demonstrate this, we return to the simpler setting of point forecasts and revisit the simulation example of equation (1.1) and Table 2, with $\sigma^2 = \frac{2}{3}$ being fixed. Figure 1 shows the perfect forecast, the deliberately misspecified extremist forecast, and the observation in this setting. The bias of the extremist forecast is readily seen when all forecast cases are taken into account. However, if we restrict attention to cases where the observation exceeds a high threshold of 2, it is not obvious whether the perfect or the extremist forecast is preferable.[4]

In this simple example, we have seen that if we stratify by the value of the realized observation, a deliberately misspecified forecast may appear appealing,

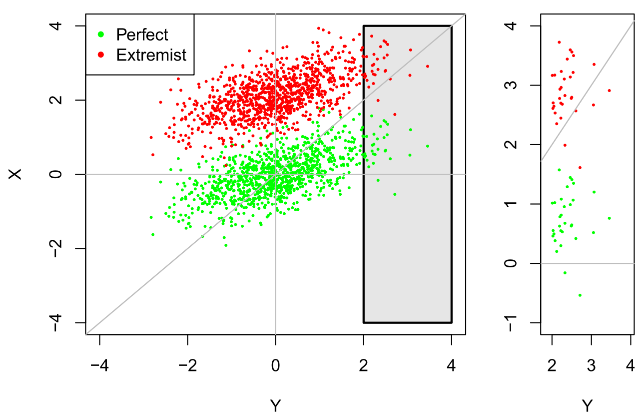

FIG. 1. *The sample illustrates the conditional distribution of the perfect forecast (green) and the extremist forecast (red, darker shade) given the observation in the setting of equation (1.1) and Table 2, where $\sigma^2 = \frac{2}{3}$. The vertical stripe, which is enlarged at right, corresponds to cases where the respective observation exceeds a threshold value of* 2.

[4]To provide analytical results, $X_{\text{perfect}}|Y = y \sim \mathcal{N}((1-\sigma^2)y, \sigma^2(1-\sigma^2))$ and $X_{\text{extr}}|Y = y \sim \mathcal{N}((1-\sigma^2)y + \frac{5}{2}, \sigma^2(1-\sigma^2))$.

while an ideal forecast may appear flawed, even though the forecasts are based on the same information set. Fortunately, unwanted effects of this type are avoided if we stratify by the value of the forecast. To see this, note that ideal predictive distributions and their induced point forecasts satisfy the autocalibration property (2.3) and, subject to conditions, the unbiasedness property (2.4), respectively.

From a bivariate extreme value theory perspective, an alternative approach to evaluating deterministic forecasts of extreme events can be described as follows. In a first step, the marginal distributions of the forecasts and the observations are compared. In case of comparing the perfect and the extremist forecasts, the difference in the respective distributions is apparent. By contrast, if point forecasts are produced by drawing random samples from the forecast distributions, the marginal distributions of the perfect and climatological forecaster are identical. In a second step, measures of asymptotic extremal dependence proposed by Coles, Heffernan and Tawn (1999) can be used to assess the closeness of the copula to perfect dependence in the upper tail. While the perfect and extremist forecaster show identical asymptotic dependence, computing such measures allows to clearly distinguish between the perfect forecaster and the climatological forecaster. Stephenson et al. (2008) use these measures of extremal dependence to construct performance measures for evaluating binary forecasts of extreme events based on contingency tables which, however, were later shown to exhibit undesirable properties (Ferro and Stephenson, 2011).

2.2 Proper scoring rules and consistent scoring functions

In the preceding section, we have introduced calibration and sharpness as key aspects of the quality of probabilistic forecasts. Proper scoring rules assess calibration and sharpness simultaneously and play key roles in the comparative evaluation and ranking of competing forecasts (Gneiting and Raftery, 2007). Specifically, let $\mathcal{F}$ denote a class of probability distributions on Ω_Y, the set of possible values of the observation Y. A *scoring rule* is a mapping $\mathrm{S} : \mathcal{F} \times \Omega_Y \longrightarrow \mathbb{R} \cup \{\infty\}$ that assigns a numerical penalty based on the predictive distribution $F \in \mathcal{F}$ and observation $y \in \Omega_Y$. We take scoring rules to be negatively oriented, that is, smaller scores indicate better predictions, and generally identify a predictive distribution with its CDF. A scoring rule is *proper* relative to the class $\mathcal{F}$ if

$$\mathbb{E}_G \mathrm{S}(G, Y) \leq \mathbb{E}_G \mathrm{S}(F, Y) \tag{2.5}$$

for all probability distributions $F, G \in \mathcal{F}$. It is *strictly proper* relative to the class $\mathcal{F}$ if the above holds with equality only if $F = G$. In what follows we assume that $\Omega_Y = \mathbb{R}$. Scoring rules provide summary measures of predictive performance, and in practical applications, competing forecasting methods are compared and ranked in terms of the mean score over the cases in a test set. Propriety is a critically important property that encourages honest and careful forecasting, as the expected score is minimized if the quoted predictive distribution agrees with the actually assumed, under which the expectation in (2.5) is computed.

The most popular proper scoring rules for real-valued quantities are the *logarithmic score* (LogS), defined as

$$\mathrm{LogS}(F, y) = -\log f(y), \tag{2.6}$$

where f denotes the density of F (Good, 1952), which applies to absolutely continuous distributions only, and the *continuous ranked probability score* (CRPS), which is defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \big(F(z) - \mathbb{1}\{y \leq z\}\big)^2 \, \mathrm{d}z \tag{2.7}$$

directly in terms of the predictive CDF (Matheson and Winkler, 1976). The CRPS can be interpreted as the integral of the proper Brier score (Brier, 1950, Gneiting and Raftery, 2007),

$$\mathrm{BS}_z(F, y) = \big(F(z) - \mathbb{1}\{y \leq z\}\big)^2, \tag{2.8}$$

for the induced probability forecast for the binary event of the observation not exceeding the threshold value z. Alternative representations of the CRPS are discussed in Gneiting and Raftery (2007) and Gneiting and Ranjan (2011).

The quality of point forecasts is typically assessed by means of a *scoring function* $\mathrm{s}(x, y)$ that assigns a numerical score based on the point forecast, x, and the respective observation, y. As in the case of proper scoring rules, competing forecasting methods are compared and ranked in terms of the mean score over the cases in a test set. Popular scoring functions include the squared error, $\mathrm{s}(x, y) = (x - y)^2$, and the absolute error, $\mathrm{s}(x, y) = |x - y|$, for which we have reported mean scores in Table 2.

To avoid misguided inferences, the scoring function and the forecasting task have to be matched carefully, either by specifying the scoring function ex ante, or by employing scoring functions that are *consistent* for a target functional T, relative to the class $\mathcal{F}$ of predictive distributions at hand, in the technical sense that

$$\mathbb{E}_F \mathrm{s}\big(\mathrm{T}(F), Y\big) \leq \mathbb{E}_F \mathrm{s}(x, Y)$$

for all $x \in \mathbb{R}$ and $F \in \mathcal{F}$ (Gneiting, 2011). For instance, the squared error scoring function is consistent for the mean or expectation functional relative to the class of the probability measures with finite first moment, and the absolute error scoring function is consistent for the median functional.

Consistent scoring functions become proper scoring rules if the point forecast is chosen to be the Bayes rule or optimal point forecast under the respective predictive distribution. In other words, if the scoring function s is consistent for the functional T, then

$$\mathrm{S}(F, y) = \mathrm{s}\big(\mathrm{T}(F), y\big)$$

defines a proper scoring rule relative to the class $\mathcal{F}$. For instance, squared error can be interpreted as a proper scoring rule provided the point forecast is the mean of the respective predictive distribution, and absolute error yields a proper scoring rule if the point forecast is the median of the predictive distribution.

2.3 Understanding the forecaster's dilemma

We are now in the position to analyze and understand the forecaster's dilemma both within the joint distribution framework and from the perspective of proper scoring rules. While there is no unique definition of extreme events in the literature, we follow common practice and take extreme events to be observations that fall into the tails of the underlying distribution. In public discussions of the quality of forecasts, attention often falls exclusively on cases with extreme observations. As we have seen, under this practice even the most skillful forecasts available are bound to fail in the public eye, particularly when the signal-to-noise ratio in the data generating process is low. In a nutshell, if forecast evaluation is restricted to cases where the observation falls into a particular region of the outcome space, forecasters are encouraged to unduly emphasize this region.

Within the joint distribution framework of Section 2.1, any stratification by, and conditioning on, the realized values of the outcome is problematic and ought to be avoided, as general theoretical guidance for the interpretation and assessment of the resulting conditional distribution $[F|Y]$ does not appear to be available. In view of the likelihood-baserate factorization (2.2) of the joint distribution of the forecast and the observation, the forecaster's dilemma arises as a consequence. Fortunately, stratification by, and conditioning on, the values of a point forecast or probabilistic forecast is unproblematic from a decision-theoretic perspective, as the autocalibration property (2.3) lends itself to practical tools and tests for calibration checks, as discussed by Gneiting, Balabdaoui and Raftery (2007), Held, Rufibach and Balabdaoui (2010) and Strähl and Ziegel (2015), among others.

From the perspective of proper scoring rules, Gneiting and Ranjan (2011) showed that a proper scoring rule S_0 is rendered improper if the product with a nonconstant weight function $w(y)$ is formed. Specifically, consider the weighted scoring rule

$$\mathrm{S}(F, y) = w(y)\mathrm{S}_0(F, y). \tag{2.9}$$

Then if Y has density g, the expected score $\mathbb{E}_g\mathrm{S}(F, Y)$ is minimized by the predictive distribution F with density

$$f(y) = \frac{w(y)g(y)}{\int w(z)g(z)\,\mathrm{d}z}, \tag{2.10}$$

which is proportional to the product of the weight function, w, and the true density, g. In other words, forecasters are encouraged to deviate from their true beliefs and misspecify their predictive densities, with multiplication by the weight function (and subsequent normalization) being an optimal strategy. Therefore, the scoring rule S in (2.9) is improper.

To connect to the forecaster's dilemma, consider the indicator weight function $w_r(y) = \mathbb{1}\{y \geq r\}$. The use of the weight function w_r does not directly correspond to restricting the evaluation set to cases where the observation exceeds or equals the threshold value r, as instead of excluding the nonextreme cases, a score of zero is assigned to them. However, when forecast methods are compared, the use of the indicator weighted scoring rule corresponds to a multiplicative scaling of the restricted score, and so the ranking of competing forecasts is the same as that obtained by restricting the evaluation set.

2.4 Tailoring proper scoring rules

The forecaster's dilemma gives rise to the question how one might apply scoring rules to probabilistic forecasts when particular emphasis is placed on extreme events, while retaining propriety. To this end, Diks, Panchenko and van Dijk (2011) and Gneiting and Ranjan (2011) consider the use of proper weighted scoring rules that emphasize specific regions of interest.

Diks, Panchenko and van Dijk (2011) propose the *conditional likelihood* (CL) score,

$$\mathrm{CL}(F, y) = -w(y)\log\left(\frac{f(y)}{\int_{-\infty}^{\infty} w(z)f(z)\,\mathrm{d}z}\right), \tag{2.11}$$

and the *censored likelihood* (CSL) score,

$$\begin{aligned}&\mathrm{CSL}(F, y)\\ &\quad = -w(y)\log f(y)\\ &\qquad - \big(1-w(y)\big)\log\left(1-\int_{-\infty}^{\infty} w(z)f(z)\,\mathrm{d}z\right).\end{aligned} \tag{2.12}$$

Here, w is a weight function such that $0 \le w(z) \le 1$ and $\int w(z)f(z)\,\mathrm{d}z > 0$ for all potential predictive distributions, where f denotes the density of F. When $w(z) \equiv 1$, both the CL and the CSL score reduce to the unweighted logarithmic score (2.6). Gneiting and Ranjan (2011) propose the *threshold-weighted continuous ranked probability score* (twCRPS), defined as

$$\begin{aligned}&\mathrm{twCRPS}(F, y)\\ &\quad = \int_{-\infty}^{\infty} w(z)\big(F(z) - \mathbb{1}\{y \le z\}\big)^2\,\mathrm{d}z,\end{aligned} \tag{2.13}$$

where, again, w is a nonnegative weight function. When $w(z) \equiv 1$, the twCRPS reduces to the unweighted CRPS (2.7). For recent applications of the twCRPS and a quantile-weighted version of the CRPS see, for example, Cooley, Davis and Naveau (2012), Lerch and Thorarinsdottir (2013) and Manzan and Zerom (2013). Zou and Yuan (2008) use the quantile-weighted version as an objective function in quantile regression.

As noted, these scoring rules are proper and can be tailored to the region of interest. When interest centers on the right tail of the distribution, we may choose $w(z) = \mathbb{1}\{z \ge r\}$ for some high threshold r. However, the indicator weight function might result in violations of the regularity conditions for the CL and CSL scoring rule, unless all predictive densities considered are strictly positive. Furthermore, predictive distributions that are identical on $[r, \infty)$, but differ on $(-\infty, r)$, cannot be distinguished. Weight functions based on Gaussian CDFs as proposed by Amisano and Giacomini (2007) and Gneiting and Ranjan (2011) provide suitable alternatives. For instance, we can set $w(z) = \Phi(z|r, \sigma^2)$ for some $\sigma > 0$, where $\Phi(\cdot|\mu, \sigma^2)$ denotes the CDF of a normal distribution with mean μ and variance σ^2. Weight functions emphasizing the left tail of the distribution can be constructed similarly, by using $w(z) = \mathbb{1}\{z \le r\}$ or $w(z) = 1 - \Phi(z|r, \sigma^2)$ for some low threshold r. In practice, the weighted integrals in (2.11), (2.12) and (2.13) may need to be approximated by discrete sums, which corresponds to the use of a discrete weight measure, rather than a weight function, as discussed by Gneiting and Ranjan (2011).

In what follows, we focus on the above proper variants of the LogS and the CRPS. However, further types of proper weighted scoring rules can be developed. Pelenis (2014) introduces the penalized weighted likelihood score and the incremental CPRS. Tödter and Ahrens (2012) and Juutilainen, Tamminen and Röning (2012) propose a logarithmic scoring rule that depends on the predictive CDF rather than the predictive density. As hinted at by Juutilainen, Tamminen and Röning (2012), page 466, this score can be generalized to a weighted version, which we call the *threshold-weighted continuous ranked logarithmic score* (twCRLS),

$$\begin{aligned}&\mathrm{twCRLS}(F, y)\\ &\quad = -\int_{\mathbb{R}} w(z)\log\big|F(z) - \mathbb{1}\{y > z\}\big|\,\mathrm{d}z.\end{aligned} \tag{2.14}$$

In analogy to the twCRPS (2.13) being a weighted integral of the Brier score in (2.8), the twCRLS (2.14) can be interpreted as a weighted integral of the discrete *logarithmic score* (LS) (Good, 1952, Gneiting and Raftery, 2007),

$$\begin{aligned}\mathrm{LS}_z(F, y) &= -\log\big|F(z) - \mathbb{1}\{y > z\}\big|\\ &= -\mathbb{1}\{y \le z\}\log F(z)\\ &\quad - \mathbb{1}\{y > z\}\log\big(1 - F(z)\big),\end{aligned} \tag{2.15}$$

for the induced probability forecast for the binary event of the observation not exceeding the threshold value z. The aforementioned weight functions and discrete approximations can be employed.

2.5 Diebold–Mariano tests

Formal statistical tests of equal predictive performance have been widely used, particularly in the economic literature. Turning now to a time series setting, we consider probabilistic forecasts F_t and G_t for an observation y_{t+k} that lies k time steps ahead. Given a proper scoring rule S, we denote the respective mean scores on a test set ranging from time $t = 1, \ldots, n$ by

$$\bar{\mathrm{S}}_n^F = \frac{1}{n}\sum_{t=1}^{n} \mathrm{S}(F_t, y_{t+k})$$

and

$$\bar{\mathrm{S}}_n^G = \frac{1}{n}\sum_{t=1}^{n} \mathrm{S}(G_t, y_{t+k}),$$

respectively. Diebold and Mariano (1995) proposed the use of the test statistic

$$t_n = \sqrt{n}\,\frac{\bar{\mathrm{S}}_n^F - \bar{\mathrm{S}}_n^G}{\hat{\sigma}_n}, \tag{2.16}$$

where $\hat{\sigma}_n^2$ is a suitable estimator of the asymptotic variance of the score difference. Under the null hypothesis of a vanishing expected score difference and standard regularity conditions, the test statistic t_n in (2.16) is asymptotically standard normal (Diebold and Mariano, 1995, Giacomini and White, 2006, Diebold, 2015). When the null hypothesis is rejected in a two-sided test, F is preferred if the test statistic t_n is negative, and G is preferred if t_n is positive.

For $j = 0, 1, \ldots,$ let $\hat{\gamma}_j$ denote the lag j sample autocovariance of the sequence $\mathrm{S}(F_1, y_{1+k}) - \mathrm{S}(G_1, y_{1+k}), \ldots, \mathrm{S}(F_n, y_{n+k}) - \mathrm{S}(G_n, y_{n+k})$ of score differences. Diebold and Mariano (1995) noted that for ideal forecasts at the k step ahead prediction horizon the respective errors are at most $(k-1)$-dependent. Motivated by this fact, Gneiting and Ranjan (2011) use the estimator

$$\hat{\sigma}_n^2 = \begin{cases} \hat{\gamma}_0 & \text{if } k = 1, \\ \hat{\gamma}_0 + 2\sum_{j=1}^{k-1} \hat{\gamma}_j & \text{if } k \geq 2 \end{cases} \tag{2.17}$$

for the asymptotic variance in the test statistic (2.16). While the at most $(k-1)$-dependence assumption might be violated in practice for various reasons, this appears to be a reasonable and practically useful choice nonetheless. Diks, Panchenko and van Dijk (2011) propose the use of the heteroskedasticity and autocorrelation consistent (HAC) estimator

$$\hat{\sigma}_n^2 = \hat{\gamma}_0 + 2\sum_{j=1}^{J}\left(1 - \frac{j}{J}\right)\hat{\gamma}_j, \tag{2.18}$$

where J is the largest integer less than or equal to $n^{1/4}$. When this latter estimator is used, larger estimates of the asymptotic variance and smaller absolute values of the test statistic (2.16) tend to be obtained, as compared to using the estimator (2.17), particularly when the sample size n is large.

3. SIMULATION STUDIES

We now present simulation studies. In Section 3.1, we mimic the experiment reported on in Table 2 for point forecasts, now illustrating the forecaster's dilemma on probabilistic forecasts. Furthermore, we consider the influence of the signal-to-noise ratio in the data generating process. Thereafter in the following sections, we investigate whether or not there is a case for the use of proper weighted scoring rules, as opposed to their unweighted counterparts, when interest focuses on extremes. As it turns out, the fundamental lemma of Neyman and Pearson (1933) provides theoretical guidance in this regard. All results in this section are based on 10,000 replications.

3.1 The influence of the signal-to-noise ratio

Let us recall that in the simulation setting of equation (1.1) the observation satisfies $Y|\mu \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu \sim \mathcal{N}(0, 1-\sigma^2)$. In Table 2, we have considered three competing point forecasts – termed the perfect, unconditional and extremist forecasts – and have noted the appearance of the forecaster's dilemma when the quality of the forecasts is assessed on cases of extreme outcomes only.

We now turn to probabilistic forecasts and study the effect of the parameter $\sigma \in (0, 1)$ that governs predictability. Small values of σ correspond to high signal-to-noise ratios, and large values of σ to small signal-to-noise ratios, respectively. Marginally, Y is standard normal for all values of σ. In the limit as $\sigma \to 0$, the perfect predictive distribution approaches the point measure in the random mean μ; as $\sigma \to 1$, it approaches the unconditional standard normal distribution. The perfect probabilistic forecast is ideal in the technical sense of Section 2.1, and thus will be preferred over any other predictive distribution (with identical information basis) by any rational user (Diebold, Gunther and Tay, 1998, Tsyplakov, 2013).

In Table 3, we report mean scores for the three probabilistic forecasts when $\sigma^2 = \frac{2}{3}$ is fixed. Under the CRPS and LogS, the perfect forecast outperforms the others, as expected, and the extremist forecast performs by far the worst. However, these results change drastically if cases with extreme observations are considered only. In analogy to the results in Table 2, the perfect forecast is discredited under the restricted scores rCRPS and rLogS, whereas the misguided extremist forecast appears to excel, thereby demonstrating

TABLE 3

Mean scores for the probabilistic forecasts in Table 2, where the observation Y satisfies (1.1) with $\sigma^2 = \frac{2}{3}$ being fixed. The CRPS and LogS are computed based on all observations, whereas the restricted versions (rCRPS and rLogS) are based on observations exceeding 1.64, the 95th percentile of the population, only. The lowest value in each column is shown in bold

Forecast	CRPS	LogS	rCRPS	rLogS
Perfect	**0.46**	**1.22**	0.96	2.30
Unconditional	0.57	1.42	1.48	3.03
Extremist	2.05	5.90	**0.79**	**1.88**

TABLE 4

Mean scores for the probabilistic forecasts in Table 2, where the observation Y satisfies (1.1) with $\sigma^2 = \frac{2}{3}$ being fixed, under the proper weighted scoring rules twCRPS, CL, and CSL. For each weight function and column, the lowest value is shown in bold

Threshold *r*	**Forecast**	**twCRPS**	**CL**	**CSL**
Indicator weight function, $w(z) = \mathbb{1}\{z \geq 1.64\}$				
1.64	Perfect	**0.018**	**<0.001**	**0.164**
	Unconditional	0.019	0.002	0.204
	Extremist	0.575	0.093	2.205
Gaussian weight function, $w(z) = \Phi(z\|1.64, 1)$				
1.64	Perfect	**0.053**	**−0.043**	**0.298**
	Unconditional	0.062	−0.028	0.345
	Extremist	0.673	0.379	1.625

the forecaster's dilemma in the setting of probabilistic forecasts. As shown in Table 4, under the proper weighted scoring rules introduced in Section 2.4 with weight functions that emphasize the right tail, the rankings under the unweighted CRPS and LogS are restored.

Next, we investigate the influence of the signal-to-noise ratio in the data generating process on the appearance and extent of the forecaster's dilemma. As noted, predictability increases with the parameter $\sigma \in (0, 1)$. Figure 2 shows the mean CRPS and LogS for the three probabilistic forecasts as a function of σ. The scores for the unconditional forecast do not depend on σ. The predictive performance of the perfect forecast decreases in σ, which is natural, as it is less beneficial to know the value of μ when σ is large. The extremist forecast yields better scores as σ increases, which can be explained by the increase in the predictive variance that allows for a better match between the probabilistic forecast and the true distribution. For the improper restricted scoring rules rCRPS and rLogS, the same general patterns can be observed in Figure 3 – the mean score increases in σ for the perfect forecast and decreases for the extremist forecast. In accordance with the forecaster's dilemma, the extremist forecast is now perceived to outperform its competitors for all sufficiently large values of σ. However, for small values of σ, when the signal in μ is strong, the rankings are the same as under the CRPS and LogS in Figure 2. This illustrates the intuitively obvious observation that the forecaster's dilemma is tied to stochastic systems with moderate to low signal-to-noise ratios, so that predictability is weak.

3.2 Power of Diebold–Mariano tests: Diks, Panchenko and van Dijk (2011) revisited

While thus far we have illustrated the forecaster's dilemma, the unweighted CRPS and LogS are well able to distinguish between the perfect forecast and its competitors. In the subsequent sections, we investigate whether there are benefits to using proper weighted scoring rules, as opposed to their unweighted versions.

To begin with, we adopt the simulation setting in Section 4 of Diks, Panchenko and van Dijk (2011). Suppose that at time $t = 1, \ldots, n$, the observations y_t are independent standard normal. We apply the two-sided Diebold–Mariano test of equal predictive performance to compare the ideal probabilistic forecast, the standard normal distribution, to a misspecified com-

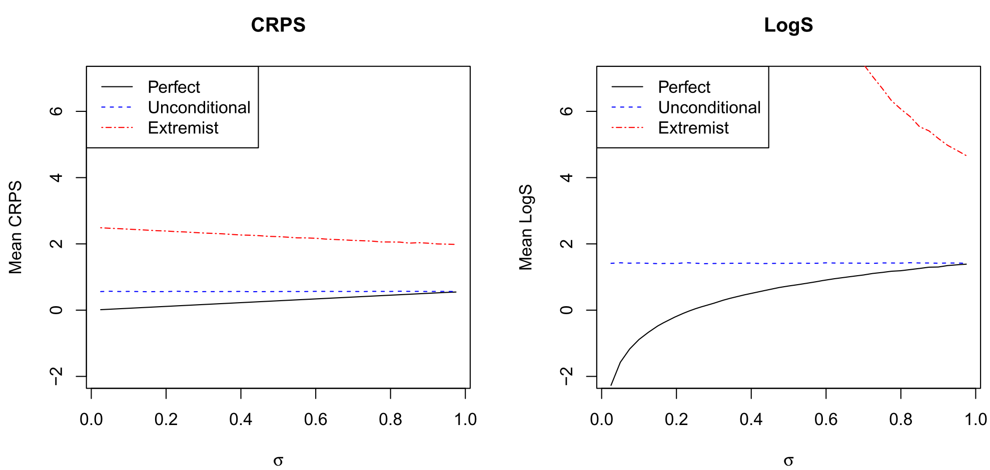

FIG. 2. *Mean CRPS and LogS for the probabilistic forecasts in the setting of equation (1.1) and Table 2 as functions of the parameter $\sigma \in (0, 1)$.*

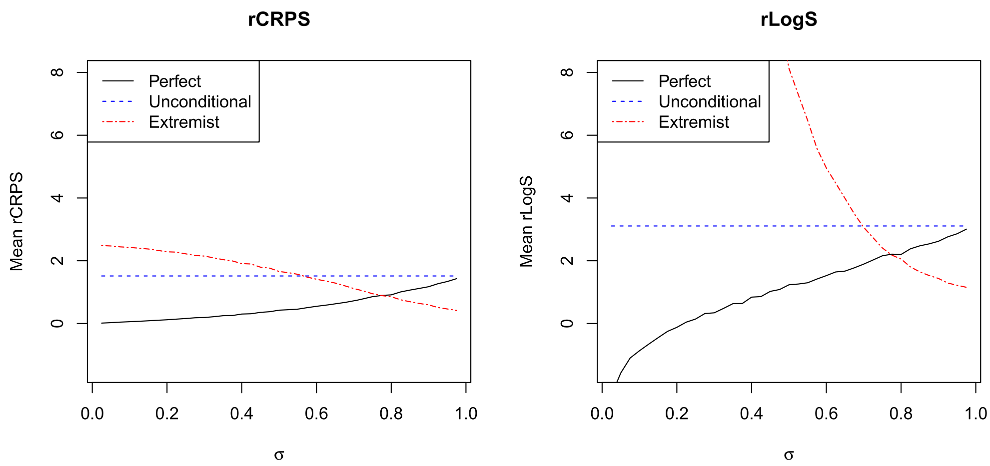

FIG. 3. *Mean of the improper restricted scoring rules rCRPS and rLogS for the probabilistic forecasts in the setting of equation* (1.1) *and Table* 2 *as functions of the parameter* $\sigma \in (0, 1)$. *The restricted mean scores are based on the subset of observations exceeding* 1.64 *only.*

petitor, a Student t distribution with five degrees of freedom, mean 0 and variance 1. Following Diks, Panchenko and van Dijk (2011), we use the nominal level 0.05, the variance estimate (2.18), and the indicator weight function $w(z) = \mathbb{1}\{z \le r\}$, and we vary the sample size, n, with the threshold value r in such a way that under the standard normal distribution the expected number, $c = 5$, of observations in the relevant region $(-\infty, r]$ remains constant.

Figure 4 shows the proportion of rejections of the null hypothesis of equal predictive performance in favor of either the standard normal or the Student t distribution, respectively, as a function of the threshold value r in the weight function. Rejections in favor of the standard normal distribution represent true power, whereas rejections in favor of the misspecified Student t distribution are misguided. The curves for the tests based on the twCRPS, CL and CSL scoring rules agree with those in the left column of Figure 5 of Diks, Panchenko and van Dijk (2011). At first sight, they might suggest that the use of the indicator weight function $w(z) = \mathbb{1}\{z \le r\}$ with emphasis on the extreme left tail, as reflected by increasingly smaller values of r, yields increased power. At second sight, we need to compare to the power curves for tests using the unweighted CRPS and LogS, based on the same sample size, n, as corresponds to the threshold r at hand. These curves suggest, perhaps surprisingly, that there may not be an advantage to using weighted scoring rules. To the contrary, the left-hand panel in Figure 4 suggests that tests based on the unweighted LogS are competitive in terms of statistical power.

3.3 The role of the Neyman–Pearson lemma

In order to understand this phenomenon, we draw a connection to a cornerstone of test theory, namely, the fundamental lemma of Neyman and Pearson (1933), following the lead of Feuerverger and Rahman (1992) and Reid and Williamson (2011). For the moment, we consider one-sided rather than two-sided tests.

In the simulation setting described by Diks, Panchenko and van Dijk (2011) and in the previous section, any test of equal predictive performance can be re-interpreted as a test of the simple null hypothesis H_0 of a standard normal population against the simple alternative H_1 of a Student t population. We write f_0 and f_1 for the associated density functions and $\mathbb{P}_0$ and $\mathbb{P}_1$ for probabilities under the respective hypotheses. By the Neyman–Pearson lemma (Lehmann and Romano, 2005, Theorem 3.2.1), under H_0 and at any level $\alpha \in (0, 1)$ the unique most powerful test of H_0 against H_1 is the likelihood ratio test. The likelihood ratio test rejects H_0 if $\prod_{t=1}^n f_1(y_t)/\prod_{t=1}^n f_0(y_t) > k$ or, equivalently, if

$$\sum_{t=1}^{n} \log f_1(y_t) - \sum_{t=1}^{n} \log f_0(y_t) > \log k, \tag{3.1}$$

where the critical value k is such that

$$\mathbb{P}_0\left(\frac{\prod_{t=1}^n f_1(y_t)}{\prod_{t=1}^n f_0(y_t)} > k\right) = \alpha.$$

Due to the optimality property of the likelihood ratio test, its power,

$$\mathbb{P}_1\left(\frac{\prod_{t=1}^n f_1(y_t)}{\prod_{t=1}^n f_0(y_t)} > k\right), \tag{3.2}$$

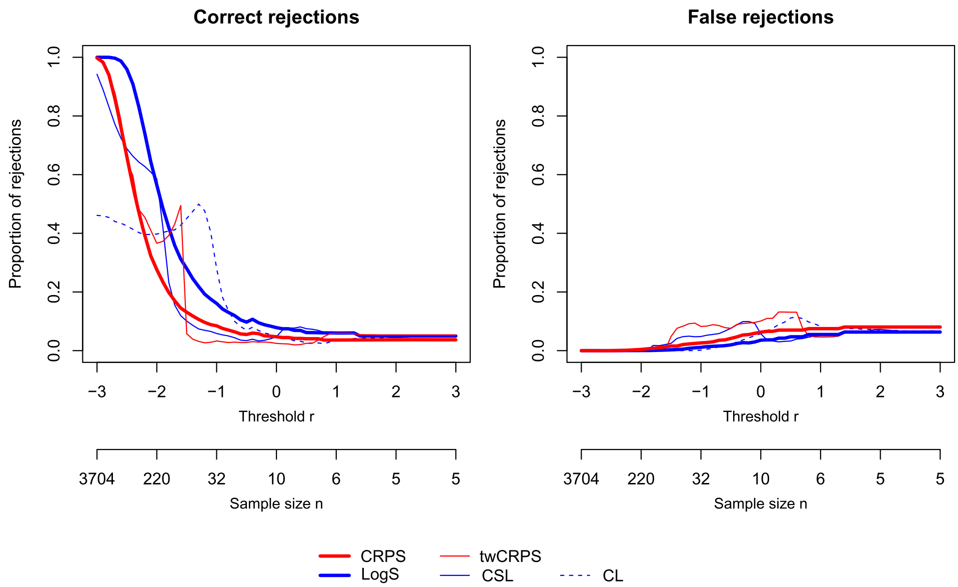

FIG. 4. *Frequency of correct rejections (in favor of the standard normal distribution, left panel) and false rejections (in favor of the Student t distribution, right panel) in two-sided Diebold–Mariano tests in the simulation setting described in Section 3.2. The panels correspond to those in the left hand column of Figure 5 in Diks, Panchenko and van Dijk (2011). The sample size n for the tests depends on the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \le r\}$ for the twCRPS, CL and CSL scoring rules such that under the standard normal distribution there are five expected observations in the relevant interval $(-\infty, r]$.*

gives a theoretical upper bound on the power of any test of H_0 versus H_1. Furthermore, the optimality result is robust, in the technical sense that minor misspecifications of either H_0 or H_1, as quantified by the Kullback–Leibler divergence, lead to minor loss of power only (Eguchi and Copas, 2006).

We now compare the likelihood ratio test to the one-sided Diebold–Mariano test based on the logarithmic score (LogS; equation (2.6)). This test uses the statistic (2.16) and rejects H_0 if

$$\sum_{t=1}^{n} \log f_1(y_t) - \sum_{t=1}^{n} \log f_0(y_t) > \sqrt{n}\hat{\sigma}_n z_{1-\alpha}, \tag{3.3}$$

where $z_{1-\alpha}$ is a standard normal quantile and $\hat{\sigma}_n^2$ is given by (2.17) or (2.18). Comparing with (3.1), we see that the one-sided Diebold–Mariano test that is based on the LogS has the same type of rejection region as the likelihood ratio test. However, the Diebold–Mariano test uses an estimated critical value, which may lead to a level less or greater than the nominal level, α, whereas the likelihood ratio test uses the (in the practice of forecasting unavailable) critical value that guarantees the desired nominal level, α.

In this light, it is not surprising that the one-sided Diebold–Mariano test based on the LogS has power close to the theoretical optimum in (3.2). We illustrate this in Figure 5, where we plot the power and size of the likelihood ratio test and one-sided Diebold–Mariano tests based on the CRPS, twCRPS, LogS, CL and CSL in the setting of the previous section. For small threshold values, the Diebold–Mariano test based on the unweighted LogS has much higher power than tests based on the weighted scores, even though it does not reach the power of the likelihood ratio test, which can be explained by the use of an estimated critical value and incorrect size properties. The theoretical upper bound on the power is violated by Diebold–Mariano tests based on the twCRPS and CL for threshold values between 0 and 1. However, the level of these tests exceeds the nominal level of $\alpha = 0.05$ with too frequent rejections of H_0. Adjusting the level of the tests to the nominal level by using simulation-based critical values instead increases the power of the tests and removes most of the nonmonotonicity of the power curves, as illustrated in the online supplement (Lerch et al., 2016). However, such adjustments are not feasible in practice.

In the setting of two-sided tests, the connection to the Neyman–Pearson lemma is less straightforward, but the general principles remain valid and provide a partial explanation of the behavior seen in Figure 4.

3.4 Power of Diebold–Mariano tests: Further experiments

In the simulation experiments just reported, Diebold–Mariano tests based on proper weighted scor-

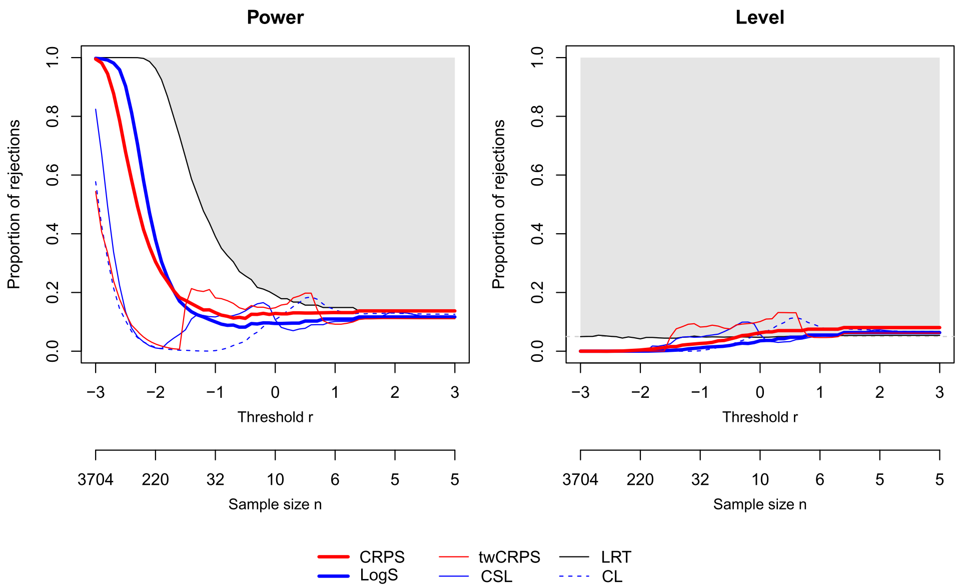

FIG. 5. *Power* (*left*) *and level* (*right*) *of the likelihood ratio test* (*LRT*) *and one-sided Diebold–Mariano tests in the simulation setting described in Section* 3.2. *The sample size* n *for the tests depends on the threshold* r *in the indicator weight function* $w(z) = \mathbb{1}\{z \le r\}$ *for the twCRPS, CL and CSL scoring rules such that under the standard normal distribution there are five expected observations in the relevant interval* $(-\infty, r]$. *In the panel for power, the shaded area above the curve for the LRT corresponds to theoretically unattainable values for a test with nominal level. In the panel for level, the dashed line indicates the nominal level.*

ing rules generally are unable to outperform tests based on traditionally used, unweighted scoring rules. Several potential reasons come to mind. As we have just seen, when the true data generating process is given by one of the competing forecast distributions, the Neyman–Pearson lemma points at the superiority of tests based on the unweighted LogS. Furthermore, in the simulation setting considered thus far, the distributions considered differ both in the center, the left tail and the right tail, and the test sample size varied with the threshold for the weight function in a peculiar way.

Therefore, we now consider a revised simulation setting, where we compare two forecast distributions neither of which corresponds to the true sampling distribution, where the forecast distributions only differ on the positive half-axis, and where the test sample size is fixed at $n = 100$. The three candidate distributions are given by Φ, a standard normal distribution with density ϕ, by a heavy-tailed distribution H with density

$$h(x) = \mathbb{1}\{x \le 0\}\phi(x) + \mathbb{1}\{x > 0\}\frac{3}{8}\left(1 + \frac{x^2}{4}\right)^{-5/2},$$

and by an equally weighted mixture F of Φ and H, with density

$$f(x) = \frac{1}{2}(\phi(x) + h(x)).$$

We perform two-sided Diebold–Mariano tests of equal predictive performance based on the CRPS, twCRPS, LogS, CL and CSL.

In Scenario A, the data are a sample from the standard normal distribution Φ, and we compare the forecasts F and H, respectively. In Scenario B, we interchange the roles of Φ and H, that is, the data are a sample from H, and we compare the forecasts F and Φ. The Neyman–Pearson lemma does not apply in this setting. However, the definition of F as a weighted mixture of the true distribution and a misspecified competitor lets us expect that F is to be preferred over the latter. Indeed, by Proposition 3 of Nau (1985), if $F = wG + (1 - w)H$ with $w \in [0, 1]$ is a convex combination of G and H, then

$$\mathbb{E}_G S(G, Y) \le \mathbb{E}_G S(F, Y) \le \mathbb{E}_G S(H, Y)$$

for any proper scoring rule S. As any utility function induces a proper scoring rule via the respective Bayes act,[5] this implies that under G any rational decision maker favors F over H (Dawid, 2007, Gneiting and Raftery, 2007).

We estimate the frequencies of rejections of the null hypothesis of equal predictive performance at level

[5] The Bayes act is the action that maximizes the ex ante expected utility (Ferguson, 1967).

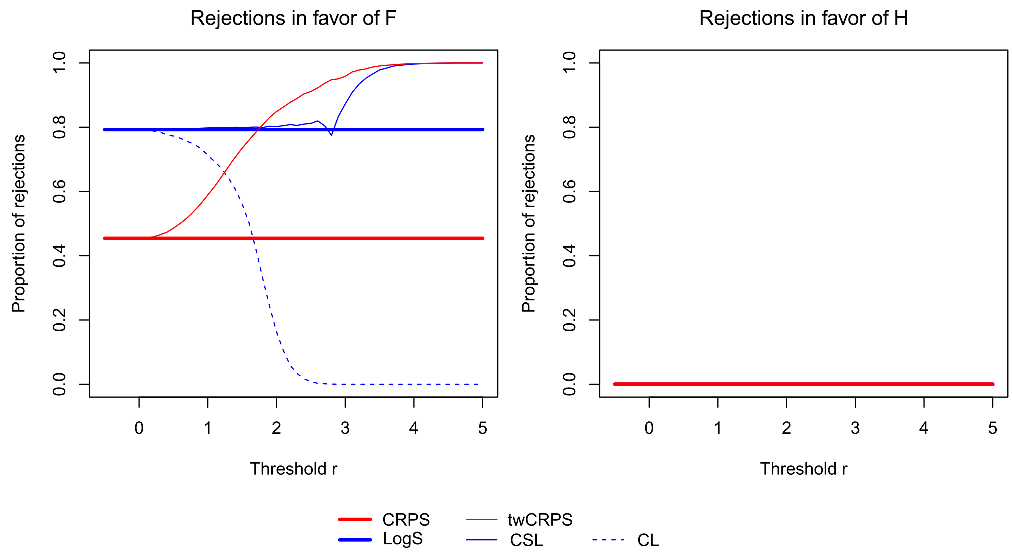

FIG. 6. *Scenario A in Section 3.4. The null hypothesis of equal predictive performance of F and H is tested under a standard normal population. The panels show the frequency of rejections in two-sided Diebold–Mariano tests in favor of either F (desired, left) or H (misguided, right). The tests under the twCRPS, CL, use the weight function $w(z) = \mathbb{1}\{z \geq r\}$, and the sample size is fixed at $n = 100$.*

$\alpha = 0.05$. The choice of the estimator for the asymptotic variance of the score difference in the Diebold–Mariano test statistic (2.16) does not have a recognizable effect in this setting, and so we show results under the estimator (2.17) with $k = 1$ only.

Figure 6 shows rejection rates under Scenario A in favor of F and H, respectively, as a function of the threshold r in the indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$ for the weighted scoring rules. The frequency of the desired rejections in favor of F increases with larger thresholds for tests based on the twCRPS and CSL, thereby suggesting an improved discrimination ability at high threshold values. Under the CL scoring rule, the rejection rate decreases rapidly for larger threshold values. This can be explained by the fact that the weight function is a multiplicative component of the CL score in (2.11). As r becomes larger and larger, none of the 100 observations in the test sample exceed the threshold, and so the mean scores under both forecasts vanish. This can also be observed in Figure 4, where, however, the effect is partially concealed by the increase of the sample size for more extreme threshold values. Interestingly, an issue very similar to that for the CL scoring rule arises in the assessment of deterministic forecasts of rare and extreme binary events, where performance measures based on contingency tables have been developed and standard measures degenerate to trivial values as events become rarer (Marzban, 1998, Stephenson et al., 2008), posing a challenge that has been addressed by Ferro and Stephenson (2011).

Figure 7 shows the respective rejection rates under Scenario B, where the sample is generated from the heavy-tailed distribution H, and the forecasts F and Φ are compared. In contrast to the previous examples the Diebold–Mariano test based on the CRPS shows a higher frequency of the desired rejections in favor of F than the test based on the LogS. However, for the tests based on proper weighted scoring rules, the frequency of the desired rejections in favor of F decays to zero with increasing threshold value, and for the tests based on the twCRPS and CSL, the frequency of the undesired rejections in favor of Φ rises for larger threshold values.

This seemingly counterintuitive observation can be explained by the tail behavior of the forecast distributions, as follows. Consider the twCRPS and CSL with the indicator weight function $w(z) = \mathbb{1}\{z \geq r\}$ and a threshold r that exceeds the maximum of the given sample. In this case, the scores do not depend on the observations, and are solely determined by the respective tail probabilities, with the lighter tailed forecast distribution receiving the better score. In a nutshell, when the emphasis lies on a low-probability region with few or no observations, the forecaster assigning smaller probability to this region will be preferred. The traditionally used unweighted scoring rules do not depend on a threshold, and thus do not suffer from this deficiency.

In comparisons of the mixture distribution F and the lighter-tailed forecast distribution Φ, this leads to a loss

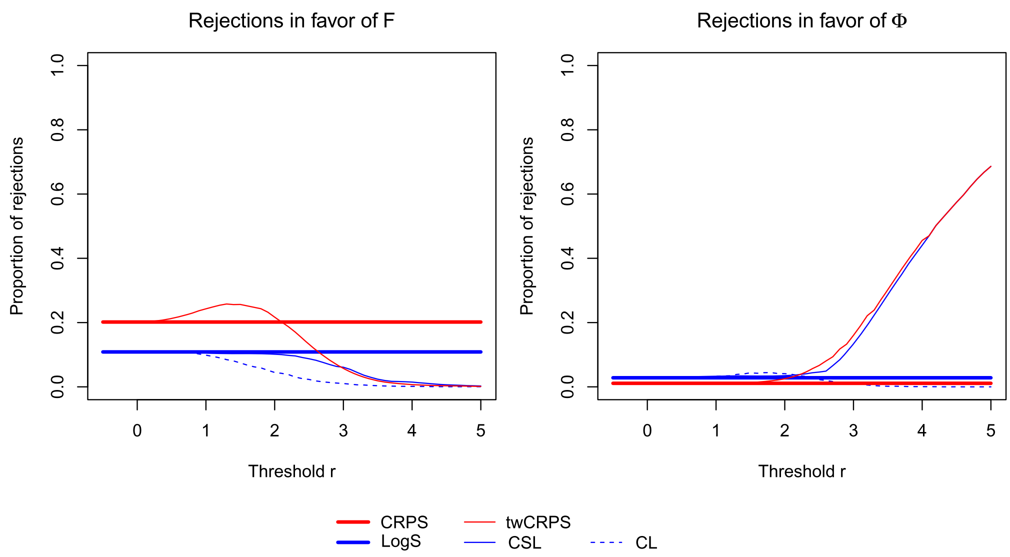

FIG. 7. *Scenario B in Section 3.4. The null hypothesis of equal predictive performance of F and Φ is tested under a Student t population. The panels show the frequency of rejections in two-sided Diebold–Mariano tests in favor of either F (desired, left) or Φ (misguided, right). The tests under the twCRPS, CL, and CSL scoring rules use the weight function $w(z) = \mathbb{1}\{z \geq r\}$, and the sample size is fixed at $n = 100$.*

of finite sample discrimination ability of the proper weighted scoring rules as the threshold r increases. This observation also suggests that any favorable finite sample behavior of the Diebold–Mariano tests based on weighted scoring rules in Scenario A might be governed by rejections due to the lighter tails of F compared to H.

In summary, even though the simulation setting at hand was specifically tailored to benefit proper weighted scoring rules, these do not consistently perform better in terms of statistical power when compared to their unweighted counterparts. Any advantages vanish at increasingly extreme threshold values in case the actually superior distribution has heavier tails.

4. CASE STUDY

Based on the work of Clark and Ravazzolo (2015), we compare probabilistic forecasting models for key macroeconomic variables for the United States, serving to demonstrate the forecaster's dilemma and the use of proper weighted scoring rules in an application setting.

4.1 Data

We consider time series of quarterly gross domestic product (GDP) growth, computed as 100 times the log difference of real GDP, and inflation in the GDP price index (henceforth *inflation*), computed as 100 times the log difference of the GDP price index, over an evaluation period from the first quarter of 1985 to the second quarter of 2011, as illustrated in Figure 8. The data are available from the Federal Reserve Bank of Philadelphia's real time dataset.[6]

For each quarter t in the evaluation period, we use the real-time data vintage t to estimate the forecasting models and construct forecasts for period t and beyond. The data vintage t includes information up to time $t-1$. The one-quarter ahead forecast is thus a current quarter (t) forecast, while the two-quarter ahead forecast is a next quarter $(t+1)$ forecast, and so forth (Clark and Ravazzolo, 2015). Here, we focus on forecast horizons of one and four quarters ahead.

As the GDP data are continually revised, it is not immediate which revision should be used as the realized observation. We follow Romer and Romer (2000) and Faust and Wright (2009) who use the second available estimates as the actual data. Specifically, suppose a forecast for quarter $t+k$ is issued based on the vintage t data ending in quarter $t-1$. The corresponding realized observation is then taken from the vintage $t+k+2$ data set. This approach may entail structural breaks in case of benchmark revisions, but is comparable to real-world forecasting situations where noisy early vintages are used to estimate predictive models (Faust and Wright, 2009).

[6]http://www.phil.frb.org/research-and-data/real-time-center/real-time-data/.

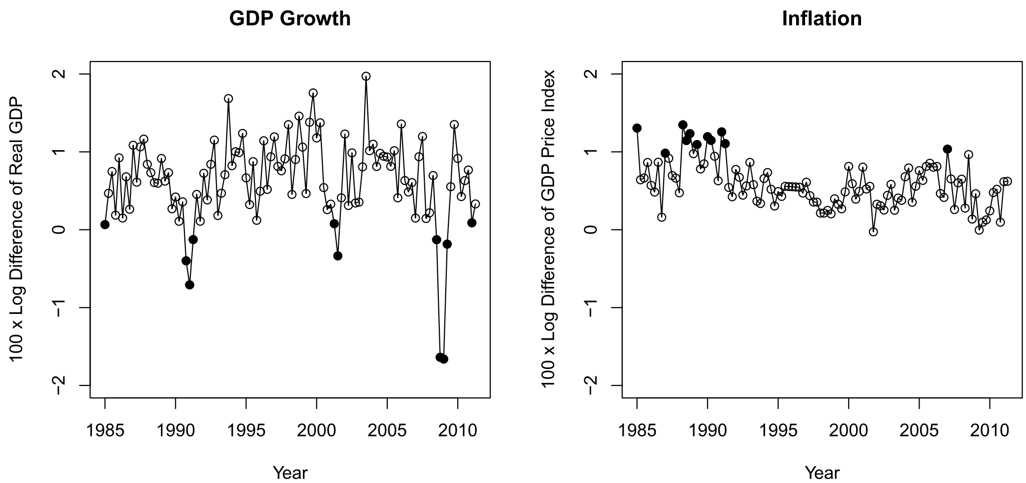

FIG. 8. *Observations of GDP growth and inflation in the U.S. from the first quarter of* 1985 *to the second quarter of* 2011. *Solid circles indicate observations considered here as extreme events.*

4.2 Forecasting models

We consider autoregressive (AR) and vector autoregressive (VAR) models, the specifications of which are given now. For further details and a discussion of alternative models, see Clark and Ravazzolo (2015).

Our baseline model is an AR(p) scheme with constant shock variance. Under this model, the conditional distribution of Y_t is given by

$$(4.1)\qquad \begin{aligned} &Y_t|\mathbf{y}_{<t}, b_0, \dots, b_p, \sigma \\ &\quad\sim \mathcal{N}\left(b_0 + \sum_{i=1}^{p} b_i y_{t-i}, \sigma^2\right), \end{aligned}$$

where $p = 2$ for GDP growth and $p = 4$ for inflation. Here, $\mathbf{y}_{<t}$ denotes the vector of the realized values of the variable Y prior to time t. We estimate the model parameters $b_0, \dots, b_p$ and σ in a Bayesian fashion using Markov chain Monte Carlo (MCMC) under a recursive estimation scheme, where the data sample $\mathbf{y}_{<t}$ is expanded as forecasting moves forward in time. The conditional predictive distribution then is the Gaussian variance-mean mixture

$$(4.2)\qquad \frac{1}{m}\sum_{j=1}^{m} \mathcal{N}\left(b_0^{(j)} + \sum_{i=1}^{p} b_i^{(j)} y_{t-i}, (\sigma^{(j)})^2\right),$$

where $m = 5000$ and $(b_0^{(1)}, \dots, b_p^{(1)}, \sigma^{(1)}), \dots, (b_0^{(m)}, \dots, b_p^{(m)}, \sigma^{(m)})$ is a sample from the posterior distribution of the model parameters. For the other forecasting models, we proceed analogously.

A more flexible approach is the Bayesian AR model with time-varying parameters and stochastic specification of the volatility (AR-TVP-SV) proposed by Cogley and Sargent (2005), which has the hierarchical structure given by

$$(4.3)\qquad \begin{aligned} &Y_t|\mathbf{y}_{<t}, b_{0,t}, \dots, b_{p,t}, \lambda_t \\ &\quad\sim \mathcal{N}\left(b_{0,t} + \sum_{i=1}^{p} b_{i,t} y_{t-i}, \lambda_t\right), \\ &b_{i,t}|b_{i,t-1}, \tau \sim \mathcal{N}(b_{i,t-1}, \tau^2), \quad i = 0, \dots, p, \\ &\log\lambda_t|\lambda_{t-1}, \sigma \sim \mathcal{N}(\log\lambda_{t-1}, \sigma^2). \end{aligned}$$

Again, we set $p = 2$ for GDP growth and $p = 4$ for inflation.

In a multivariate extension of the AR models, we consider VAR schemes where GDP growth, inflation, unemployment rate and three-month government bill rate are modeled jointly. Specifically, the conditional distribution of the four-dimensional vector $\mathbf{Y}_t$ is given by the multivariate normal distribution

$$(4.4)\qquad \begin{aligned} &\mathbf{Y}_t|\mathbf{Y}_{<t}, \mathbf{b}_0, \mathbf{B}_1, \dots, \mathbf{B}_p, \mathbf{\Sigma} \\ &\quad\sim \mathcal{N}_4\left(\mathbf{b}_0 + \sum_{i=1}^{p} \mathbf{B}_i \mathbf{y}_{t-1}, \mathbf{\Sigma}\right), \end{aligned}$$

where $\mathbf{Y}_{<t}$ denotes the data prior to time t, $\mathbf{\Sigma}$ is a 4×4 covariance matrix, $\mathbf{b}_0$ is a vector of intercepts, and $\mathbf{B}_i$ is a 4×4 matrix of lag i coefficients, where $i = 1, \dots, p$. Here, we take $p = 4$. The univariate predictive distributions for GDP growth and inflation arise as the respective margins of the multivariate posterior predictive distribution.

Finally, we consider a VAR model with time-varying parameters and stochastic specification of the volatility (VAR-TVP-SV), which is a multivariate extension of

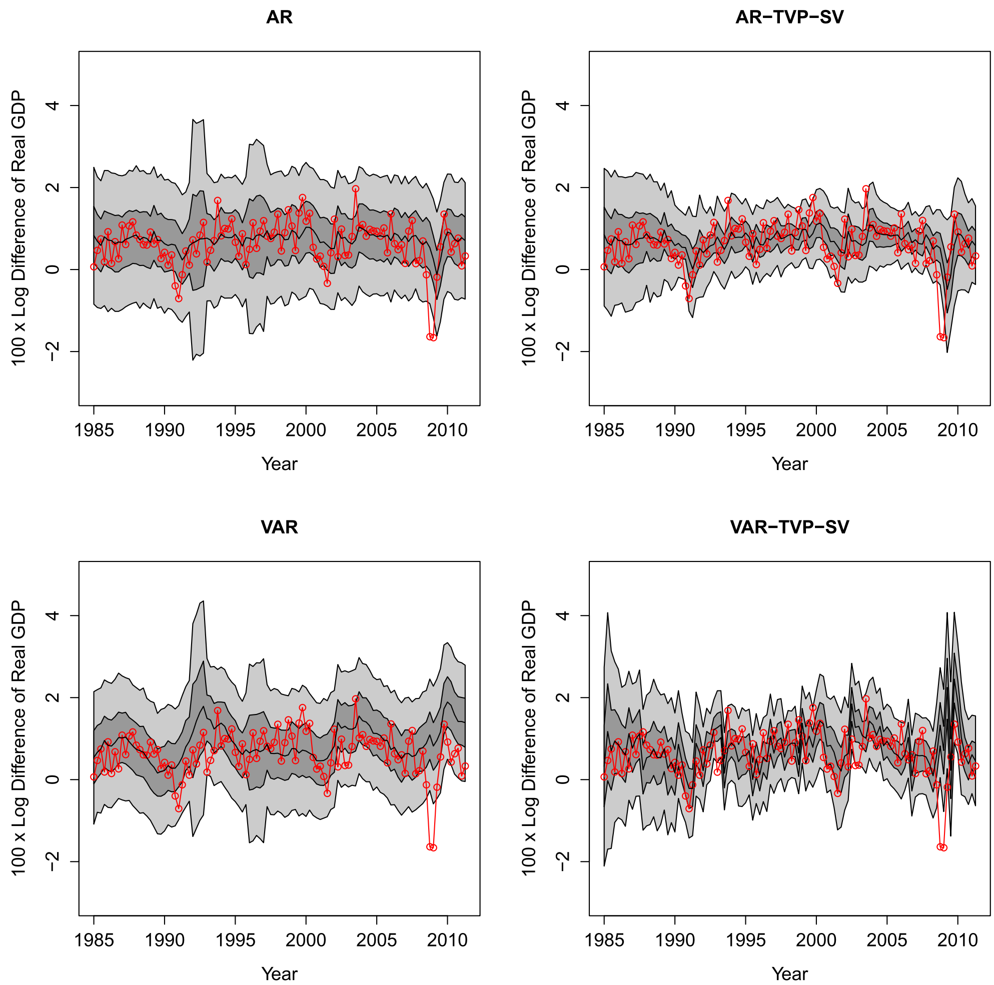

FIG. 9. *One-quarter ahead forecasts of U.S. GDP growth generated by the AR, AR-TVP-SV, VAR, and VAR-TVP-SV models. The median of the predictive distribution is shown in the black solid line, and the central* 50% *and* 90% *prediction intervals are shaded in dark and light gray, respectively. The red line shows the corresponding observations.*

the AR-TVP-SV model (Cogley and Sargent, 2005). Let $\boldsymbol{\beta}_t$ denote the vector of size $4(4p+1)$ comprising the parameters $\mathbf{b}_{0,t}$ and $\mathbf{B}_{1,t}, \ldots, \mathbf{B}_{p,t}$ at time t, set $\boldsymbol{\Lambda}_t = \operatorname{diag}(\lambda_{1,t}, \ldots, \lambda_{4,t})$ and let $\mathbf{A}$ be a lower triangular matrix with ones on the diagonal and nonzero random coefficients below the diagonal. The VAR-TVP-SV model takes the hierarchical form

$$\begin{aligned}
&\mathbf{Y}_t | \mathbf{Y}_{<t}, \boldsymbol{\beta}_t, \boldsymbol{\Lambda}_t, \mathbf{A} \\
&\quad \sim \mathcal{N}_4\left(\mathbf{b}_{0,t} + \sum_{i=1}^{p} \mathbf{B}_{i,t}\mathbf{y}_{t-1}, \mathbf{A}^{-1}\boldsymbol{\Lambda}_t(\mathbf{A}^{-1})^{\top}\right), \\
&\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1}, \mathbf{Q} \sim \mathcal{N}_{4(4p+1)}(\boldsymbol{\beta}_{t-1}, \mathbf{Q}), \\
&\log \lambda_{i,t} | \lambda_{i,t-1}, \sigma_i \sim \mathcal{N}(\log \lambda_{i,t-1}, \sigma_i^2), \\
&\quad i = 1, \ldots, 4.
\end{aligned} \tag{4.5}$$

We set $p = 2$ and refer to Clark and Ravazzolo (2015) for further details of the notation, the model, and its estimation.

Figure 9 shows one-quarter ahead forecasts of GDP growth over the evaluation period. The baseline models with constant volatility generally exhibit wider prediction intervals, while the TVP-SV models show more pronounced fluctuations both in the median forecast and the associated uncertainty. In 1992 and 1996, the Bureau of Economic Analysis performed benchmark data revisions, which causes the prediction uncertainty of the baseline models to increase substantially. The more flexible TVP-SV models seem less sensitive to these revisions.

4.3 Results

To compare the predictive performance of the four forecasting models, Table 5 shows the mean CRPS

TABLE 5
Mean CRPS and mean LogS for probabilistic forecasts of GDP growth and inflation in the U.S. at prediction horizons of $k = 1$ and $k = 4$ quarters, respectively, for the first quarter of 1985 to the second quarter of 2011. For each variable and column, the lowest value is in bold

	CRPS		**LogS**	
	$k = 1$	$k = 4$	$k = 1$	$k = 4$
GDP growth				
AR	0.330	0.359	1.044	1.120
AR-TVP-SV	**0.292**	**0.329**	**0.833**	**1.019**
VAR	0.385	0.402	1.118	1.163
VAR-TVP-SV	0.359	0.420	0.997	1.257
Inflation				
AR	0.167	0.187	0.224	0.374
AR-TVP-SV	**0.143**	**0.156**	**0.047**	**0.175**
VAR	0.170	0.198	0.235	0.428
VAR-TVP-SV	0.162	0.201	0.179	0.552

and LogS over the evaluation period. For the LogS, we follow extant practice in the economic literature and employ the quadratic approximation proposed by Adolfson, Lindé and Villani (2007). Specifically, we find the mean, $\hat{\mu}_F$ and variance, $\hat{\sigma}_F^2$, of a sample $\hat{x}_1, \ldots, \hat{x}_m$, where $\hat{x}_i$ is a random number drawn from the ith mixture component of the posterior predictive distribution (4.2), and compute the logarithmic score under the assumption of a normal predictive distribution with mean $\hat{\mu}_F$ and variance $\hat{\sigma}_F^2$.[7] To compute the CRPS and the threshold-weighted CRPS, we use the numerical methods proposed by Gneiting and Ranjan (2011).

The relative predictive performance of the forecasting models is consistent across the two variables and the two proper scoring rules. The AR-TVP-SV model has the best predictive performance and outperforms the baseline AR model. The p-values for the respective two-sided Diebold–Mariano tests range from 0.00 to 0.06, except for the LogS for GDP growth at a prediction horizon of $k = 4$ quarters, where the p-value is 0.37. However, the VAR models fail to outperform the simpler AR models. As we do not impose sparsity constraints on the parameters of the VAR models, this is likely due to overly complex forecasting models and overfitting, in line with results of Holzmann and Eulert (2014) and Clark and Ravazzolo (2015) in related economic and financial case studies.

To relate to the forecaster's dilemma, we restrict attention to extremes events. For GDP growth, we consider quarters with observed growth less than $r = 0.1$ only. For inflation, we restrict attention to high values in excess of $r = 0.98$. In either case, this corresponds to using about 10% of the observations. Table 6 shows the results of restricting the computation of the mean CRPS and the mean LogS to these observations only. For both GDP growth and inflation, the baseline AR model is considered best, and the AR-TVP-SV model appears to perform poorly. These restricted scores thus result in substantially different rankings than the proper scoring rules in Table 5, thereby illustrating the forecaster's dilemma. Strikingly, under the restricted assessment all four models seem less skillful at predicting inflation in the current quarter than four quarters ahead. This is a counterintuitive result that illustrates the dangers of conditioning on outcomes and should

TABLE 6
Mean restricted CRPS (rCRPS) and restricted LogS (rLogS) for probabilistic forecasts of GDP growth and inflation in the U.S. at prediction horizons of $k = 1$ and $k = 4$ quarters, respectively, for the first quarter of 1985 to the second quarter of 2011. The means are computed on instances when the observation is smaller than 0.10 (GDP) or larger than 0.98 (inflation) only. For each variable and column, the lowest value is shown in bold

	rCRPS		**rLogS**	
	$k = 1$	$k = 4$	$k = 1$	$k = 4$
GDP growth				
AR	**0.654**	**0.870**	**1.626**	**2.010**
AR-TVP-SV	0.659	0.970	2.016	3.323
VAR	0.827	0.924	2.072	2.270
VAR-TVP-SV	0.798	0.978	2.031	2.409
Inflation				
AR	0.214	0.157	0.484	**0.296**
AR-TVP-SV	0.236	0.179	0.619	0.327
VAR	**0.203**	**0.147**	**0.424**	0.317
VAR-TVP-SV	0.302	0.247	0.950	0.849

[7]We believe that there are more efficient and more theoretically principled ways of approximating the LogS in Bayesian settings. However, these considerations are beyond the scope of the paper, and we leave them to future work. Here, we use the quadratic approximation based on a sample. This very nearly corresponds to replacing the LogS by the proper Dawid–Sebastiani score (Dawid and Sebastiani, 1999, DSS; Gneiting and Raftery, 2007), which for a predictive distribution F with mean μ_F and finite variance σ_F^2 is given by

$$\mathrm{DSS}(F, y) = 2 \log \sigma_F + \frac{(y - \mu_F)^2}{\sigma_F^2}.$$

The quadratic approximation is infeasible for the CL and CSL scoring rules, as it then leads to improper scoring rules; see Appendix.

TABLE 7

Mean threshold-weighted CRPS for probabilistic forecasts of GDP growth and inflation in the U.S. at prediction horizons of $k = 1$ and $k = 4$ quarters, respectively, under distinct weight functions, for the first quarter of 1985 to the second quarter of 2011. For each variable and column, the lowest value is shown in bold

	twCRPS			
	$\boldsymbol{k = 1}$	$\boldsymbol{k = 4}$	$\boldsymbol{k = 1}$	$\boldsymbol{k = 4}$
GDP growth	$w_{\mathrm{I}}(z) = \mathbb{1}\{z \le 0.1\}$		$w_{\mathrm{G}}(z) = 1 - \Phi(z\|0.1, 1)$	
AR	0.062	0.068	0.111	0.120
AR-TVP-SV	**0.052**	**0.062**	**0.101**	**0.115**
VAR	0.062	**0.062**	0.119	0.119
VAR-TVP-SV	0.059	0.080	0.115	0.135
Inflation	$w_{\mathrm{I}}(z) = \mathbb{1}\{z \ge 0.98\}$		$w_{\mathrm{G}}(z) = \Phi(z\|0.98, 1)$	
AR	0.026	0.032	0.027	0.031
AR-TVP-SV	**0.018**	**0.018**	**0.021**	**0.022**
VAR	0.026	0.033	0.025	0.031
VAR-TVP-SV	0.022	0.037	0.024	0.034

be viewed as a further manifestation of the forecaster's dilemma.

In Table 7, we show results for the proper twCRPS under weight functions that emphasize the respective region of interest. For both variables, this yields rankings that are similar to those in Table 5. However, the p-values for binary comparisons with two-sided Diebold–Mariano tests generally are larger than those under the unweighted CRPS. The AR-TVP-SV model is predominantly the best, and the current quarter forecasts are deemed more skillful than those four quarters ahead. To summarize, our case study suggests that modeling volatility with time-varying parameters improves predictive performance, and that univariate models outperform multivariate models, at least in the absence of sparsity constraints. These findings also hold when interest centers on events in the tails of the distributions, and proper weighted scoring rules are used for forecast evaluation. The model rankings and relative score differences are largely consistent when the threshold in the weight functions is varied, as illustrated in the online supplement (Lerch et al., 2016).

5. DISCUSSION

We have studied the dilemma that occurs when forecast evaluation is restricted to cases with extreme observations, a procedure that appears to be common practice in public discussions of forecast quality. As we have seen, under this practice even the most skillful forecasts available are bound to be discredited when the signal-to-noise ratio in the data generating process is low. Key examples might include macroeconomic and seismological predictions. Notably, in operational earthquake forecasting predicted event probabilities are low, but high probability gains are achieved by state of the art forecasting methods (Jordan et al., 2011). In such settings, it is important for forecasters, decision makers, journalists and the general public to be aware of the forecaster's dilemma. Otherwise, charlatans might be given undue attention and recognition, and critical societal decisions could be based on misguided predictions. The forecaster's dilemma is closely connected to the concept of hindsight bias in psychology (Kahneman, 2012), and can be interpreted as an extreme form thereof.

We have offered two complementary explanations of the forecaster's dilemma. From the joint distribution perspective of Section 2.1 stratifying by, and conditioning on, the realized value of the outcome is problematic in forecast evaluation, as theoretical guidance for the interpretation and assessment of the resulting conditional distributions is unavailable. In contrast stratifying by, and conditioning on, the forecast is unproblematic. From the perspective of proper scoring rules in Section 2.3, restricting the outcome space corresponds to the multiplication of the scoring rule by an indicator weight function, which renders any proper score improper, with an explicit hedging strategy being available.

Arguably the only remedy is to consider all available cases when evaluating predictive performance. Proper weighted scoring rules emphasize specific regions of interest and facilitate interpretation (Haiden, Magnusson and Richardson, 2014). By identifying which of several competing forecast models perform best for regions of interest, they may further prove useful for combining forecasts; see Gneiting and Ranjan (2013) for a recent review of combination methods for predictive distributions, and Lerch and Thorarinsdottir (2013) for a related approach in probabilistic weather forecasting. Interestingly, however, the Neyman–Pearson lemma and our simulation studies suggest that in general the benefits of using proper weighted scoring rules in terms of power are rather limited, as compared to using standard, unweighted scoring rules. Any potential advantages vanish under weight functions with increasingly extreme threshold values, where the finite sample behavior of Diebold–Mariano tests depends on the tail properties of the forecast distributions only.

When evaluating probabilistic forecasts with emphasis on extremes, one could also consider functionals of the predictive distributions, such as the induced probability forecasts for binary tail events, as utilized in a recent comparative study by Williams, Ferro and Kwasniok (2014). Another option is to consider the induced quantile forecasts, or related point summaries of the (tails of the) predictive distributions, at low or high levels, say $\alpha = 0.975$ or $\alpha = 0.99$, as is common practice in financial risk management, both for regulatory purposes and internally at financial institutions (McNeil, Frey and Embrechts, 2015). In this context, Holzmann and Eulert (2014) studied the power of Diebold–Mariano tests for quantile forecasts at extreme levels, and Fissler, Ziegel and Gneiting (2016) raise the option of comparative backtests of Diebold–Mariano-type in banking regulation. Ehm et al. (2016) propose decision-theoretically principled, novel ways of evaluating quantile and expectile forecasts.

Variants of the forecaster's dilemma have been discussed in various strands of literature. Centuries ago, Bernoulli (1713) argued that even the most foolish prediction might attract praise when a rare event happens to materialize, referring to lyrics by Owen (1607) that are quoted in the front matter of our paper.

Tetlock (2005) investigated the quality of probability forecasts made by human experts for U.S. and world events. He observed that while forecast quality is largely independent of an expert's political views, it is strongly influenced by how a forecaster thinks. Forecasters who "know one big thing" tend to state overly extreme predictions and, therefore, tend to be outperformed by forecasters who "know many little things". Furthermore, Tetlock (2005) found an inverse relationship between the media attention received by the experts and the accuracy of their predictions, and offered psychological explanations for the attractiveness of extreme predictions for both forecasters and forecast consumers. Media attention might thus not only be centered around extreme events, but also around less skillful forecasters with a tendency towards misguided predictions.

Denrell and Fang (2010) reported similar observations in the context of managers and entrepreneurs predicting the success of a new product. In a study of the Wall Street Journal Survey of Economic Forecasts, they found a negative association between the predictive performance on a subset of cases with extreme observations and measures of general predictive performance based on all cases, and argued that accurately predicting a rare and extreme event actually is a sign of poor judgment. Their discussion was limited to point forecasts, and the suggested solution was to take into account all available observations, much in line with the findings and recommendations in our paper.

APPENDIX: IMPROPRIETY OF QUADRATIC APPROXIMATIONS OF WEIGHTED LOGARITHMIC SCORES

Let F be a predictive distribution with mean μ_F and standard deviation σ_F. As regards the conditional likelihood (CL) score (2.11), the quadratic approximation is given by

$$\mathrm{CL}^q(F, y) = -w(y) \log\left(\frac{\phi(y|F)}{\int w(x)\phi(x|F)\,\mathrm{d}x}\right),$$

where $\phi(\cdot|F)$ denotes a normal density with mean μ_F and standard deviation σ_F, respectively. Let

$$c_F = \int w(x)\phi(x|F)\,\mathrm{d}x,$$

$$c_G = \int w(x)\phi(x|G)\,\mathrm{d}x,$$

$$c_g = \int w(x)g(x)\,\mathrm{d}x,$$

and recall that the Kullback–Leibler divergence between two probability densities u and v is given by

$$K(u, v) = \int u(x) \log\left(\frac{u(x)}{v(x)}\right) \mathrm{d}x.$$

Assuming that CL^q is proper, it is true that

$$\begin{aligned}
&\mathbb{E}_G\bigl(\mathrm{CL}^q(F, Y) - \mathrm{CL}^q(G, Y)\bigr) \\
&\quad = c_g\left[K\left(\frac{w(y)g(y)}{c_g}, \frac{w(y)\phi(y|F)}{c_F}\right)\right. \\
&\qquad \left. - K\left(\frac{w(y)g(y)}{c_g}, \frac{w(y)\phi(y|G)}{c_G}\right)\right]
\end{aligned}$$

is nonnegative. Let G be uniform on $[-\sqrt{3}, \sqrt{3}]$ so that $\mu_G = 0$ and $\sigma_G = 1$, and let $w(y) = \mathbb{1}\{y \geq 1\}$. Denoting the cumulative distribution function of the standard normal distribution by Φ, we find that

$$\begin{aligned}
&K\left(\frac{w(y)g(y)}{c_g}, \frac{w(y)\phi(y|F)}{c_F}\right) \\
&\quad - K\left(\frac{w(y)g(y)}{c_g}, \frac{w(y)\phi(y|G)}{c_G}\right) \\
&= \log\left(\sigma_F \frac{1 - \Phi((1 - \mu_F)/\sigma_F)}{1 - \Phi(1)}\right) \\
&\quad + \frac{3(\sqrt{3} - 1)\mu_F^2 - 6\mu_F + (3\sqrt{3} - 1)(1 - \sigma_F^2)}{6(\sqrt{3} - 1)\sigma_F^2},
\end{aligned}$$

which is strictly negative in a neighborhood of $\mu_F = 1.314$ and $\sigma_F = 0.252$, for the desired contradiction. Therefore, CL^q is not a proper scoring rule.

As regards the censored likelihood (CSL) score (2.12), the quadratic approximation is

$$\begin{aligned}\mathrm{CSL}^q(F, y) \\ &= -w(y)\log\bigl(\phi(y|F)\bigr) \\ &\quad - \bigl(1-w(y)\bigr)\log\Bigl(1-\int w(z)\phi(z|F)\,\mathrm{d}z\Bigr).\end{aligned}$$

Under the same choice of w, F, and G as before, we find that

$$\begin{aligned}&\mathbb{E}_G\bigl(\mathrm{CSL}^q(F, Y) - \mathrm{CSL}^q(G, Y)\bigr) \\ &= \frac{\sqrt{3}-1}{2\sqrt{3}}\log\sigma_F \\ &\quad - \frac{\sqrt{3}+1}{2\sqrt{3}}\log\Bigl(\frac{\Phi((1-\mu_F)/\sigma_F)}{\Phi(1)}\Bigr) \\ &\quad + \frac{3(\sqrt{3}-1)\mu_F^2 - 6\mu_F + (3\sqrt{3}-1)(1-\sigma_F^2)}{12\sqrt{3}\sigma_F^2},\end{aligned}$$

which is strictly negative in a neighborhood of $\mu_F = 0.540$ and $\sigma_F = 0.589$. Therefore, CSL^q is not a proper scoring rule.

ACKNOWLEDGEMENTS

The support of the Volkswagen Foundation through the project 'Mesoscale Weather Extremes – Theory, Spatial Modeling and Prediction (WEX-MOP)' is gratefully acknowledged. Sebastian Lerch also acknowledges support by the Deutsche Forschungsgemeinschaft through Research Training Group 1953, and Tilmann Gneiting and Sebastian Lerch are grateful for support by the Klaus Tschira Foundation. The initial impetus for this work stems from a meeting with Jeff Baars, Cliff Mass and Adrian Raftery at the University of Washington, where Jeff Baars presented a striking meteorological example of what we here call the forecaster's dilemma. We are grateful to our colleagues for the inspiration. We thank Norbert Henze for insightful comments on initial versions of our simulation studies, and Alexander Jordan for suggesting the simulation setting in Section 3.4. We also are grateful to Richard Chandler for pointing us to the Neyman–Pearson connection and the paper by Feuerverger and Rahman (1992). We further thank the Associate Editor and a referee for constructive comments.

SUPPLEMENTARY MATERIAL

Additional figures and tables (DOI: 10.1214/16-STS588SUPP; .pdf). We provide further figures for Sections 3.3 and 4.3 and a version of Table 1 with direct links to the original media sources.

REFERENCES

ADOLFSON, M., LINDÉ, J. and VILLANI, M. (2007). Forecasting performance of an open economy DSGE model. *Econometric Rev.* **26** 289–328. MR2364364

ALBEVERIO, S., JENTSCH, V. and KANTZ, H., eds. (2006). *Extreme Events in Nature and Society*. Springer, Berlin.

AMISANO, G. and GIACOMINI, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *J. Bus. Econom. Statist.* **25** 177–190. MR2367773

BEIRLANT, J., GOEGEBEUR, Y., TEUGELS, J. and SEGERS, J. (2004). *Statistics of Extremes: Theory and Applications*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2108013

BERNOULLI, J. (1713). *Ars Conjectandi*. Impensis Thurnisiorum, Basileae. Reproduction of original from Sterling Memorial Library, Yale University. Online edition of Gale Digital Collections: The Making of the Modern World: Part I: The Goldsmiths'-Kress Collection, 1450–1850. Available at http://nbn-resolving.de/urn:nbn:de:gbv:3:1-146753.

BERNOULLI, J. (2006). *The Art of Conjecturing: Together with "Letter to a friend on sets in court tennis"*, *Translated from the Latin and with an introduction and notes by Edith Dudley Sylla*. Johns Hopkins Univ. Press, Baltimore, MD. MR2195221

BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78** 1–3.

CLARK, T. E. and RAVAZZOLO, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *J. Appl. Econometrics* **30** 551–575. MR3358636

COGLEY, T. S. M. and SARGENT, T. J. (2005). Drifts and volatilities: Monetary policies and outcomes in the post-World War II U.S. *Rev. Econ. Dyn.* **8** 262–302.

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. *Springer Series in Statistics*. Springer, London. MR1932132

COLES, S., HEFFERNAN, J. and TAWN, J. (1999). Dependence measures for extreme value analyses. *Extremes* **2** 339–365.

COOLEY, D., DAVIS, R. A. and NAVEAU, P. (2012). Approximating the conditional density given large observed values via a multivariate extremes framework, with application to environmental data. *Ann. Appl. Stat.* **6** 1406–1429. MR3058669

DAWID, A. P. (2007). The geometry of proper scoring rules. *Ann. Inst. Statist. Math.* **59** 77–93. MR2396033

DAWID, A. P. and SEBASTIANI, P. (1999). Coherent dispersion criteria for optimal experimental design. *Ann. Statist.* **27** 65–81. MR1701101

DENRELL, J. and FANG, C. (2010). Predicting the next big thing: Success as a signal of poor judgment. *Manage. Sci.* **56** 1653–1667.

DIEBOLD, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *J. Bus. Econom. Statist.* **33** 1–9. MR3303732

DIEBOLD, F. X., GUNTHER, T. A. and TAY, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *Internat. Econom. Rev.* **39** 863–883.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.

DIKS, C., PANCHENKO, V. and VAN DIJK, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *J. Econometrics* **163** 215–230. MR2812867

EASTERLING, D. R., MEEHL, G. A., PARMESAN, C., CHANGNON, S. A., KARL, T. R. and MEARNS, L. O. (2000). Climate extremes: Observations, modeling, and impacts. *Science* **289** 2068–2074.

EGUCHI, S. and COPAS, J. (2006). Interpreting Kullback–Leibler divergence with the Neyman–Pearson lemma. *J. Multivariate Anal.* **97** 2034–2040. MR2301274

EHM, W., GNEITING, T., JORDAN, A. and KRÜGER, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 505–562. MR3506792

EHRMAN, C. M. and SHUGAN, S. M. (1995). The forecaster's dilemma. *Mark. Sci.* **14** 123–147.

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events*: *For Insurance and Finance*. *Applications of Mathematics* **33**. Springer, Berlin. MR1458613

FAUST, J. and WRIGHT, J. H. (2009). Comparing Greenbook and reduced form forecasts using a large realtime dataset. *J. Bus. Econom. Statist.* **27** 468–479. MR2572034

FERGUSON, T. S. (1967). *Mathematical Statistics*: *A Decision Theoretic Approach*. *Probability and Mathematical Statistics*, Vol. **1**. Academic Press, New York–London. MR0215390

FERRO, C. A. T. and STEPHENSON, D. B. (2011). Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather Forecast.* **26** 699–713.

FEUERVERGER, A. and RAHMAN, S. (1992). Some aspects of probability forecasting. *Comm. Statist. Theory Methods* **21** 1615–1632. MR1173318

FISSLER, T., ZIEGEL, J. F. and GNEITING, T. (2016). Expected shortfall is jointly elicitable with value-at-risk: Implications for backtesting. *Risk* 58–61.

GIACOMINI, R. and WHITE, H. (2006). Tests of conditional predictive ability. *Econometrica* **74** 1545–1578. MR2268409

GNEITING, T. (2008). Editorial: Probabilistic forecasting. *J. Roy. Statist. Soc. Ser. A* **171** 319–321. MR2427336

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. MR2325275

GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application* **1** 125–151.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422. MR2848512

GNEITING, T. and RANJAN, R. (2013). Combining predictive distributions. *Electron. J. Stat.* **7** 1747–1782. MR3080409

GOOD, I. J. (1952). Rational decisions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **14** 107–114. MR0077033

GUMBEL, E. J. (1958). *Statistics of Extremes*. Columbia Univ. Press, New York. MR0096342

HAIDEN, T., MAGNUSSON, L. and RICHARDSON, D. (2014). Statistical evaluation of ECMWF extreme wind forecasts. *ECMWF Newsletter* **139** 29–33.

HALL, S. S. (2011). Scientists on trial: At fault? *Nature* **477** 264–269.

HELD, L., RUFIBACH, K. and BALABDAOUI, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics* **66** 1295–1305. MR2758518

HOLZMANN, H. and EULERT, M. (2014). The role of the information set for forecasting—With applications to risk management. *Ann. Appl. Stat.* **8** 595–621. MR3192004

JORDAN, T., CHEN, Y.-T., GASPARINI, P., MADARIAGA, R., MAIN, I., MARZOCCHI, W., PAPADOPOULOS, G., YAMAOKA, K. and ZSCHAU, J. (2011). Operational earthquake forecasting: State of knowledge and guidelines for implementation *Ann. Geophys.* **54** 315–391.

JUUTILAINEN, I., TAMMINEN, S. and RÖNING, J. (2012). Exceedance probability score: A novel measure for comparing probabilistic predictions. *J. Stat. Theory Pract.* **6** 452–467. MR3196559

KAHNEMAN, D. (2012). *Thinking*, *Fast and Slow*. Penguin Books, London.

KATZ, R. W., PARLANGE, M. B. and NAVEAU, P. (2002). Statistics of extremes in hydrology. *Adv. Water Resour.* **25** 1287–1304.

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

LERCH, S. and THORARINSDOTTIR, T. L. (2013). Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus*, *Ser. A Dyn. Meteorol. Oceanogr.* **65** 21206.

LERCH, S., THORARINSDOTTIR, T. L., RAVAZZOLO, F. and GNEITING, T. (2016). Supplement to "Forecaster's dilemma: Extreme events and forecast evaluation".

MANZAN, S. and ZEROM, D. (2013). Are macroeconomic variables useful for forecasting the distribution of US inflation? *Int. J. Forecast.* **29** 469–478.

MARZBAN, C. (1998). Scalar measures of performance in rare-event situations. *Weather Forecast.* **13** 753–763.

MATHESON, J. E. and WINKLER, R. L. (1976). Scoring rules for continuous probability distributions. *Manage. Sci.* **22** 1087–1096.

MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2015). *Quantitative Risk Management*: *Concepts*, *Techniques and Tools*, Revised ed. *Princeton Series in Finance*. Princeton Univ. Press, Princeton, NJ. MR3445371

MURPHY, A. H. and WINKLER, R. L. (1987). A general framework for forecast verification. *Mon. Weather Rev.* **115** 1330–1338.

NAU, R. F. (1985). Should scoring rules be 'effective'? *Manage. Sci.* **31** 527–535.

NEYMAN, J. and PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **231** 289–337.

OWEN, J. (1607). *Epigrammatum*, *Book IV*. Hypertext critical edition by Dana F. Sutton, The Univ. California, Irvine (1999). Available at http://www.philological.bham.ac.uk/owen/.

PELENIS, J. (2014). Weighted scoring rules for comparison of density forecasts on subsets of interest. Preprint. Available at http://elaine.ihs.ac.at/~pelenis/JPelenis_wsr.pdf.

REID, M. D. and WILLIAMSON, R. C. (2011). Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.* **12** 731–817. MR2786911

ROMER, C. D. and ROMER, D. H. (2000). Federal Reserve information and the behavior of interest rates. *Am. Econ. Rev.* **90** 429–457.

STEPHENSON, D. B., CASATI, B., FERRO, C. A. T. and WILSON, C. A. (2008). The extreme dependency score: A nonvanishing measure for forecasts of rare events. *Meteorol. Appl.* **15** 41–50.

STRÄHL, C. and ZIEGEL, J. F. (2015). Cross-calibration of probabilistic forecasts. Preprint. Available at http://arxiv.org/abs/1505.05314.

TAY, A. S. and WALLIS, K. F. (2000). Density forecasting: A survey. *J. Forecast.* **19** 124–143.

TETLOCK, P. E. (2005). *Expert Political Judgment*: *How Good Is It? How Can We Know?* Princeton Univ. Press, Princeton.

TIMMERMANN, A. (2000). Density forecasting in economics and finance. *J. Forecast.* **19** 231–234.

TÖDTER, J. and AHRENS, B. (2012). Generalization of the ignorance score: Continuous ranked version and its decomposition. *Mon. Weather Rev.* **140** 2005–2017.

TSYPLAKOV, A. (2013). Evaluation of Probabilistic Forecasts: Proper Scoring Rules and Moments. Available at SSRN: http://ssrn.com/abstract=2236605.

WILLIAMS, R. M., FERRO, C. A. T. and KWASNIOK, F. (2014). A comparison of ensemble post-processing methods for extreme events. *Q. J. R. Meteorol. Soc.* **140** 1112–1120.

ZOU, H. and YUAN, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36** 1108–1126. MR2418651