*Biometrika* (2022), **00**, 0, *pp.* 1–17
*Printed in Great Britain*

doi: 10.1093/biomet/asab047

# Valid sequential inference on probability forecast performance

BY ALEXANDER HENZI and JOHANNA F. ZIEGEL

*Institute of Mathematical Statistics and Actuarial Science, University of Bern,*
*Alpeneggstrasse 22, 3012 Bern, Switzerland*

alexander.henzi@stat.unibe.ch johanna.ziegel@stat.unibe.ch

## SUMMARY

Probability forecasts for binary events play a central role in many applications. Their quality is commonly assessed with proper scoring rules, which assign forecasts numerical scores such that a correct forecast achieves a minimal expected score. In this paper, we construct e-values for testing the statistical significance of score differences of competing forecasts in sequential settings. E-values have been proposed as an alternative to $p$-values for hypothesis testing, and they can easily be transformed into conservative $p$-values by taking the multiplicative inverse. The e-values proposed in this article are valid in finite samples without any assumptions on the data-generating processes. They also allow optional stopping, so a forecast user may decide to interrupt evaluation, taking into account the available data at any time, and still draw statistically valid inference, which is generally not true for classical $p$-value-based tests. In a case study on post-processing of precipitation forecasts, state-of-the-art forecast dominance tests and e-values lead to the same conclusions.

*Some key words*: Consistent scoring function; E-value; Forecast dominance; Optional stopping; Probability forecast; Proper scoring rule; Sequential inference.

## 1. INTRODUCTION

Consider a forecast user who compares probability predictions $p_t, q_t \in [0, 1]$, $t \in \mathbb{N}$, for a binary event $Y_{t+h} \in \{0, 1\}$, where $h \geqslant 1$ is the time lag between the forecasts and the observations. At time $t$, the forecasts $p_t$ and $q_t$, as well as any predictions and observations before $t$, are known. This setting encompasses many practical situations, such as probability-of-precipitation forecasts $h$ days ahead or predictions of negative economic growth in the next quarter. The forecast user wants to draw conclusions about the relative performance of $p_t$ and $q_t$, that is, to identify the better of the two forecasts.

Probability forecasts for binary events are arguably the simplest and best-understood type of probabilistic forecasts; see Winkler (1996) for an earlier overview and more recent reviews in Gneiting & Raftery (2007), Ranjan & Gneiting (2010) and Lai et al. (2011). The key requirements for probability forecasts are calibration, meaning that events with a predicted probability of $p$ should occur at a frequency of $p$, and sharpness, which requires the forecast probabilities to be as informative as possible, i.e., close to 0 or 1. These properties are simultaneously assessed with proper scoring rules (Gneiting & Raftery, 2007), which coincide with consistent scoring functions for the mean (Gneiting, 2011) in the case of probability forecasts, and will be simply referred to as scoring functions in this article. A scoring function $S = S(p, y)$ maps a forecast probability $p$ and an observation $y$ to a numerical score, with smaller scores indicating a better forecast. More precisely, $S$ satisfies

$$\mathbb{E}_\pi\{S(\pi, Y)\} \leqslant \mathbb{E}_\pi\{S(p, Y)\} \tag{1}$$

for all $p, \pi \in [0, 1]$, where $\mathbb{E}_\pi(\cdot)$ denotes the expected value under the assumption that $Y = 1$ with probability $\pi$. That is, the true event probability attains a minimal expected score, and $S$ is strictly consistent if equality in (1) holds only for $p = \pi$. Well-known examples are the Brier score $(y - p)^2$ and the logarithmic score $-\log(|1 - y - p|)$.

To compare the predictions $p_t$ and $q_t$, the forecast user would therefore collect a sample $y_{t+h}, p_t, q_t$, $t = 1, \ldots, T$, and compute the empirical score difference $(1/T)\sum_{t=1}^{T}\{S(p_t, y_{t+h}) - S(q_t, y_{t+h})\}$. To take into account the sampling uncertainty, such score differences are accompanied by $p$-values indicating whether the mean score differs significantly from zero. If the observations are not independent, as is usual in sequential settings, a number of asymptotic tests are available for computing $p$-values, prominent ones being the Diebold–Mariano test (Diebold & Mariano, 1995) and the test of conditional predictive ability proposed by Giacomini and White (Giacomini & White, 2006). Further examples are the martingale-based approaches of Seillier-Moiseiwitsch & Dawid (1993) and Lai et al. (2011), and more recent tests of forecast dominance (Ehm & Krüger, 2018; Yen & Yen, 2021).

In this article, we expand the tools for drawing inference on probability forecast performance by using e-values. E-values, where the ‘e’ refers to expectation, have been introduced as an alternative to $p$-values for testing. The term e-value was first used in the literature by Vovk & Wang (2021), but the concept also appears in Shafer (2021), under the name ‘betting score’, and in Grünwald et al. (2020); see also the series of working papers at `http://alrw.net/e/`. In brief, an e-value is a random variable $E \geqslant 0$, satisfying $\mathbb{E}(E) \leqslant 1$ under a given null hypothesis. By Markov’s inequality, this implies that $\mathrm{pr}(E > 1/\alpha) \leqslant \alpha$ for any $\alpha \in (0, 1)$, i.e., large realizations of an e-value can be taken as evidence against the null hypothesis, and the value $1/E$ is a conservative $p$-value. A main motivation for using e-values instead of $p$-values, explained in more detail in Shafer (2021), Grünwald et al. (2020) and Wang & Ramdas (2020), is their simple behaviour under combinations. The arithmetic average of e-values is again an e-value, and so is the product of independent or sequential e-values. E-values also have advantages over $p$-values with respect to false discovery rate control (Wang & Ramdas, 2020), which may be beneficial for the comparison of forecasts over many locations, such as over a fine latitude-longitude grid around the globe. The central property for this work is that e-values are valid under optional stopping and continuation; that is, the collection of data for computing an e-value may be stopped or continued based on seeing the past observations and e-values. It is well known that $p$-values in general do not have these properties.

Our main contribution is the result that for any scoring rule $S$ and forecasts $p$ and $q$ for $Y \in \{0, 1\}$, there exists an e-value which satisfies $\mathbb{E}_\pi(E) \leqslant 1$ if and only if $\mathbb{E}_\pi\{S(p, Y) - S(q, Y)\} \leqslant 0$. This e-value allows one to draw inference on the relative performance of the forecasts $p$ and $q$ with respect to $S$ based on only a single observation. In a sequential setting, e-values from different time-points can be merged by products into a nonnegative supermartingale or test-martingale, which are analysed in detail by Ramdas et al. (2020). This gives a statistical test of forecast dominance that is valid in finite samples without any further assumptions on the data-generating process. Moreover, the constructed e-values are valid under optional stopping, so a forecast user may decide to continue or stop forecast comparison based on only a part of the data. These advantages are inherent to any e-value, but we believe that they make e-values a particularly attractive tool in sequential forecast evaluation. The tests mentioned above for comparing probability forecasts are all only asymptotically valid, and the underlying assumptions are often difficult or impossible to verify. In the case of tests with asymptotic normality, the selection of the variance estimator for the test statistic may have a dramatic impact on the test validity; see, for example, Lazarus et al. (2018, Table 1). More serious is the problem of optional stopping. In a simple but realistic simulation example, we demonstrate that commonly used tests for forecast superiority at the level of 0.05 may yield rejection rates of up to 0.15 under optional stopping, grossly misleading and invalidating statistical inference. Although statisticians and practitioners should know that the sample size for classical tests must be determined in advance, we believe that optional stopping is quite common in forecast evaluation, where data arrive sequentially and it might be tempting to stop, or continue, an expensive or time-consuming experiment upon seeing enough, or just not enough, evidence against a hypothesis. Moreover, also in the analysis of past datasets, optional continuation may occur implicitly, in that methods are often first evaluated on a smaller, manageable part of the data and the analysis is continued if the results are promising. Last, but not least, even to a statistician fully aware of the problem of optional stopping, it may be desirable to have a tool that allows the stopping of an evaluation when enough evidence is collected, without having to bother about the implications for inference.

The advantages of e-values for forecast comparison relative to the currently available methods come at a price, namely lower power. This is well known, not only for e-values, and is a general phenomenon when tools for anytime-valid inference are compared with methods for inference with a fixed sample size; see, for example, Fig. 1 in Waudby-Smith & Ramdas (2021), which displays the widths of time-uniform and fixed-time confidence intervals for a mean. However, in the case study in this article, $p$-values from classical tests and e-values lead to qualitatively the same results.

## 2. Preliminaries

### 2.1. *Scoring functions for probabilities*

Throughout the article, $\mathbb{E}_{\mathbb{Q}}(\cdot)$ denotes the expected value of the quantity in parentheses under the probability distribution $\mathbb{Q}$. If the measure $\mathbb{Q}$ is the probability $\pi \in [0,1]$ of a binary event, we simply write $\mathbb{E}_{\pi}(\cdot)$.

When comparing probability forecasts with scoring functions, the choice of the scoring function plays a crucial role. While (1) guarantees that the true event probability always achieves a minimal expected score, different scoring functions may yield different rankings when misspecified forecasts are compared (Patton, 2020). This problem can be avoided by basing forecast comparisons on several or all scoring rules simultaneously. For probabilities of binary events, under mild regularity conditions stated in Gneiting et al. (2007, Theorem 2.3), all consistent scoring functions are of the form

$$S(p,y) = \int_{(0,1)} S_\theta(p,y)\,\mathrm{d}\nu(\theta), \tag{2}$$

where $\nu$ is a locally finite Borel measure on $(0,1)$ and

$$S_\theta(p,y) = (\theta - y)\{\mathbb{1}(p > \theta) - \mathbb{1}(y > \theta)\} = \begin{cases} \theta, & y = 0,\ p > \theta, \\ 1-\theta, & y = 1,\ p \leqslant \theta, \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

In (3), $\mathbb{1}$ denotes the indicator function. This representation originally dates back to Schervish (1989); see also Ehm et al. (2016). The scoring function $S$ is strictly consistent if and only if $\nu$ assigns positive mass to all nondegenerate intervals in $(0,1)$.

### 2.2. *Forecast dominance and hypotheses*

Let $(\Omega, \mathcal{F}, \mathbb{Q})$ be a probability space with a filtration $\mathcal{F}_t$, $t \in \mathbb{N}$. We assume that the competing forecasts $p_t$ and $q_t$ and the observation $Y_t$, constitute a random vector $(Y_t, p_t, q_t)$ adapted to $\mathcal{F}_t$, and that $(p_t, q_t)$ are forecasts for $Y_{t+h}$ for some integer lag $h \geqslant 1$. The measure $\mathbb{Q}$ describes the joint dynamics of the forecasts and the observations.

When comparing forecasts using a given scoring function $S$, the quantity of interest is often not the unconditional expected score difference $\mathbb{E}_{\mathbb{Q}}\{S(p_t, Y_{t+h}) - S(q_t, Y_{t+h})\}$, which describes the average relative performance of $p_t$ and $q_t$. More interesting is the question of whether, given the information $\mathcal{F}_t$ at the time of forecasting, the conditional event probability is closer to $p_t$ than to $q_t$, i.e., $\mathbb{E}_{\mathbb{Q}}\{S(p_t, Y_{t+h}) - S(q_t, Y_{t+h}) \mid \mathcal{F}_t\} \leqslant 0$. This notion of forecast dominance is called conditional forecast dominance and was introduced by Giacomini & White (2006).

The definition of forecast dominance used here does not require knowledge about the processes generating $(Y_t, p_t, q_t)$, which are often unknown or not well enough understood to formulate a suitable stochastic model. The relative performance of the forecasts $p_t$ and $q_t$ is governed by the underlying distribution $\mathbb{Q}$, and hypotheses about forecast dominance are hypotheses about the data-generating process. Denoting by $\mathcal{P}$ the set of probability measures on $(\Omega, \mathcal{F})$, we will construct tests for the following hypotheses:

$$\mathcal{H}_{S;c} = \left[\mathbb{P} \in \mathcal{P} : c_t\, \mathbb{E}_{\mathbb{P}}\{S(p_t, Y_{t+h}) - S(q_t, Y_{t+h}) \mid \mathcal{F}_t\} \leqslant 0 \text{ a.s.}, \ t \in \mathbb{N}\right], \tag{4}$$

$$\mathcal{H}_c = \left[\mathbb{P} \in \mathcal{P} : \sup_{\theta \in [0,1]} c_t\, \mathbb{E}_{\mathbb{P}}\{S_\theta(p_t, Y_{t+h}) - S_\theta(q_t, Y_{t+h}) \mid \mathcal{F}_t\} \leqslant 0 \text{ a.s.}, \ t \in \mathbb{N}\right], \tag{5}$$

where a.s. stands for almost surely. Here, $(c_t)_{t\in\mathbb{N}}$ is a sequence of $\mathcal{F}_t$-measurable random variables $c_t \in \{0, 1\}$. If $c_t = 1$ for all $t$, we write $\mathcal{H}_{S;c} = \mathcal{H}_S$ and $\mathcal{H}_c = \mathcal{H}$. In this case, hypothesis (4) states that at all times $t$, forecast $p_t$ is at least as good as forecast $q_t$ under the scoring rule $S$, given the information available at the time of forecasting. Hypothesis (5) is stronger and states that $p_t$ is preferred over $q_t$ under all elementary scores (3), and it corresponds to what is denoted by $H^S_-$ in Ehm & Krüger (2018, (2.5)). Recently, hypotheses of the type $\mathcal{H}$ or $\mathcal{H}_S$ have been called into question by Zhu & Timmermann (2020), who demonstrate that the null hypothesis of equal conditional predictive accuracy is basically never satisfied in realistic settings. Their criticism does not directly apply to one-sided hypotheses, but we emphasize that the null hypotheses $\mathcal{H}_S$ and $\mathcal{H}$ are rather strong in that they require conditional dominance at all time-points. Tests for these hypotheses are therefore most suitable for comparing a new method with an established benchmark or a state-of-the-art method, where rejecting the null means that the new method outperforms the benchmark at least in some situations, a minimal requirement.

The classical example for a situation with $\mathbb{P} \in \mathcal{H}$ is $p_t = \mathbb{P}(Y_{t+h} = 1 \mid \mathcal{F}_t)$, i.e., $p_t$ is the ideal forecast in the sense of Gneiting & Ranjan (2013). For the hypotheses $\mathcal{H}_S$, one may easily construct situations with dominance relations also among noncalibrated forecasts; see the simulation examples in § 4.

In many practical situations, it cannot be expected that a certain forecast method will always outperform another one, and forecast users want to know under what conditions a particular forecast should be preferred. Choosing the sequence $(c_t)_{t\in\mathbb{N}}$ such that $c_t = 1$ if the condition holds and $c_t = 0$ otherwise, allows us to formalize this question. Here the variables $c_t$ must be $\mathcal{F}_t$-measurable, i.e., known at the time of forecasting. In practice this is not a severe limitation, since the information that one forecast is more accurate than another under a given condition is useful only if this condition is known at the time of forecasting, and not ex post. But also from a theoretical point of view, forecast evaluation should only be conditioned on the forecasts themselves, and not on the observations or on information not available at the time of forecasting; see Lerch et al. (2017) for a detailed analysis of this issue in the case of extreme events.

## 3. E-values for testing forecast dominance

### 3.1. *One-period setting*

We first construct e-values for the comparison of probability forecasts in a one-period setting, where $Y = 1$ with probability $\pi$ and the forecasts $p$ and $q$ are assumed to be fixed numbers in $(0, 1)$. These e-values give an absolute and valid interpretation of predictive performance with only a single observation, e.g., for a single time-point in the sequential setting of § 2.2 or in binary classification problems with independent forecast-observation pairs, where the competing forecasts are based on covariates and $\pi$ is the probability of $Y = 1$ conditional on the covariate values. The null hypotheses that $p$ is a better forecast than $q$ with respect to a given score $S$ or with respect to all scoring functions simultaneously correspond to

$$H_S = \big[\pi \in [0, 1] : \mathbb{E}_\pi\{S(p, Y) - S(q, Y)\} \leqslant 0\big],$$

$$H = \left[\pi \in [0, 1] : \sup_{\theta\in[0,1]} \mathbb{E}_\pi\{S_\theta(p, Y) - S_\theta(q, Y)\} \leqslant 0\right].$$

For $p < q$, a direct computation shows that $H_S$ is the interval $[0, \kappa_\nu\{[p, q)\}]$ with

$$\kappa_\nu\{[a, b)\} = \frac{\int_{[a,b)} \theta \, d\nu(\theta)}{\nu\{[a, b)\}} \quad (0 < a < b < 1).$$

The stronger null hypothesis $H$ is the intersection of these intervals for all mixing measures $\nu$, that is, $[0, p]$. In the case of $q > p$, the intervals $H_S$ and $H$ take the form $[\kappa_\nu\{[q, p)\}, 1]$ and $[p, 1]$, respectively. Table 1 gives the boundary $\kappa_\nu\{[p, q)\}$ for commonly used scoring functions.

For a set $\mathcal{P}$ of probability measures and disjoint $H, H' \subset \mathcal{P}$, we say that an e-value $E$ has null hypothesis $H$ and alternative $H'$ if $\mathbb{E}_\mathbb{P}(E) \leqslant 1$ for all $\mathbb{P} \in H$ and $\mathbb{E}_\mathbb{Q}(E) > 1$ for all $\mathbb{Q} \in H'$. The following theorem characterizes e-values for testing $H_S$.

Theorem 1. *Let $S$ be a consistent scoring function and let $p, q \in (0, 1)$ with $p \nmid q$. Assume that the mixing measure $\nu$ of $S$ satisfies $\nu\{[\min(p, q), \max(p, q))\} > 0$. Then a function $E = E(y)$ is an e-value with null hypothesis $H_S$ and alternative $[0, 1] \setminus H_S$ if and only if for some $\lambda \in (0, 1]$,*

$$E(y) = E_{p,q;\lambda}(y) = 1 + \lambda \frac{S(p, y) - S(q, y)}{|S(p, \mathbb{1}\{p > q\}) - S(q, \mathbb{1}\{p > q\})|}. \tag{6}$$

Theorem 1 gives a family of e-values for testing forecast dominance with a given score $S$, and in a next step we tune the parameter $\lambda$ in (6) such that the corresponding e-value has maximal power against a given alternative. The notion of power for e-values differs from the classical power for $p$-values, and it is motivated in detail by Shafer (2021) and Grünwald et al. (2020). An e-value can be interpreted as a bet against the null hypothesis, and a product $\prod_{t=1}^T E_t$ of e-values represents the accumulated capital at time $T$ if the initial capital is 1 and all money is invested in the bet at each step. Maximizing the gains is equivalent to maximizing the growth rate $(1/T) \log \prod_{t=1}^T E_t = (1/T) \sum_{t=1}^T \log(E_t)$, a strategy sometimes called Kelly betting after Kelly Jr (1956). If an e-value maximizes $\mathbb{E}_\mathbb{P}\{\log(E)\}$ under a measure $\mathbb{P}$ representing an alternative

Table 1. *Commonly used scoring rules and the corresponding denominators in the GROW e-values under the assumption $p < q$. The case of $p > q$ is obtained by interchanging the roles of $p$ and $q$. The mixing measure $\nu$ is given in the form of its Lebesgue density $h(\theta)$, $\theta \in (0,1)$. For the spherical score, $\|p\| = (2p^2 - 2p + 1)^{1/2}$ denotes the Euclidean norm of the vector $(p, 1-p)$*

| Score | $S(p,y)$ | Mixing density $\nu$ | $\kappa_\nu\{[p,q)\}$ |
|---|---|---|---|
| Brier | $(p-y)^2$ | 2 | $(p+q)/2$ |
| Logarithmic | $-\log(\|1-y-p\|)$ | $\theta^{-1}(1-\theta)^{-1}$ | $\log\left(\frac{1-p}{1-q}\right)/\log\left\{\frac{q(1-p)}{p(1-q)}\right\}$ |
| Spherical | $1 - \|1-y-p\|/\|p\|$ | $(2\theta^2 - 2\theta + 1)^{-3/2}$ | $\frac{(q-1)\|p\|-(p-1)\|q\|}{(2q-1)\|p\|-(2p-1)\|q\|}$ |

hypothesis, it is said to be growth-rate-optimal, abbreviated GROW (Grünwald et al., 2020). One such alternative could be that $Y = 1$ with probability $q$, but one can maximize the power under any other alternative $\pi_1 \notin H_S$.

Theorem 2. *Under the assumptions of Theorem 1, for any $\pi_1 \notin H_S$, $\mathbb{E}_{\pi_1}\{\log(E_{p,q;\lambda})\}$ is maximal in $\lambda$ if and only if*

$$\lambda = \begin{cases} (1-\pi_1) + \pi_1 \dfrac{S(p,1) - S(q,1)}{S(p,0) - S(q,0)}, & p > q, \\[2ex] \pi_1 + (1-\pi_1)\dfrac{S(p,0) - S(q,0)}{S(p,1) - S(q,1)}, & p < q. \end{cases}$$

*The corresponding e-value equals*

$$E_{p,q}^{\pi_1}(y) = \begin{cases} \dfrac{1-\pi_1}{1-\kappa_\nu\{[\min(p,q), \max(p,q))\}}, & y = 0, \\[2ex] \dfrac{\pi_1}{\kappa_\nu\{[\min(p,q), \max(p,q))\}}, & y = 1. \end{cases}$$

Theorem 2 shows that the GROW e-values for the comparison of probability forecasts take the form of likelihood ratios with the alternative probability in the numerator and the integral of the mixing measure $\nu$ over the interval $[\min(p,q), \max(p,q)$, suitably normalized, in the denominator. It is possible to obtain this result directly by applying Theorem 1 of Grünwald et al. (2020), since $\kappa_\nu\{[\min(p,q), \max(p,q))\}$ is the boundary of the null hypothesis $H_S$. We have chosen to take the indirect but more instructive approach via Theorem 1, because to the best of our knowledge this is the first application of e-values to forecast comparison, and similar approaches might be used to construct e-values for score differences in more general settings than the evaluation of binary event forecasts. In fact, Waudby-Smith & Ramdas (2021, Proposition 2) contains a similar representation of e-values to that in (6) for testing hypotheses about a constant mean.

For the test of the null hypothesis $H$, applying Theorem 1 of Grünwald et al. (2020) shows that the GROW e-value is the likelihood ratio.

Theorem 3. *Let $p, q \in (0,1)$. Then the GROW e-value with null hypothesis $H$ and alternative hypothesis that $Y = 1$ with probability $\pi_1 \notin H$ is*

$$E_{p,q}^{\pi_1 *}(y) = \begin{cases} (1-\pi_1)/(1-p), & y = 0, \\ \pi_1/p, & y = 1. \end{cases}$$

In testing with e-values, the GROW e-value for testing the point null hypothesis $\{p\}$ against the alternative $\pi_1$ is exactly the likelihood ratio, and Theorem 3 states that this is equivalent to testing forecast dominance with respect to all scoring functions. Dominance with respect to all scoring functions is a very strong requirement on $p$, since the null hypothesis is false as soon as the true probability $\pi$ is on the same side of $p$ as $q$, that is, in $(p, 1]$ for $p < q$ or in $[0, p)$ for $q < p$, and the choice of $\pi_1$ is restricted to these sets. Unlike the e-values $E_{p,q}^{\pi_1}$, $E_{p,q}^{\pi_1 *}$ does not depend directly on $q$, but rather indirectly via the admissible values for $\pi_1$.

### 3.2. *Sequential inference*

We now turn to the sequential model with observations $Y_t$ and forecasts $p_t$ and $q_t$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ with a filtration $\mathcal{F}_t$, $t \in \mathbb{N}$. In the $h = 1$ case, for any $\mathbb{Q} \in \mathcal{H}_{S;c}$ and any adapted sequence $\lambda_t \in [0, 1]$, $t \in \mathbb{N}$, with $E_{p_t,q_t;\lambda_t}$ as defined in (6),

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\left\{\prod_{t=1}^{T} E_{p_t,q_t;\lambda_t}(Y_{t+1})\right\} &= \mathbb{E}_{\mathbb{Q}}\left[\mathbb{E}_{\mathbb{Q}}\left\{\prod_{t=1}^{T} E_{p_t,q_t;\lambda_t}(Y_{t+1}) \,\middle|\, \mathcal{F}_T\right\}\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[\prod_{t=1}^{T-1} E_{p_t,q_t;\lambda_t}(Y_{t+1})\mathbb{E}_{\mathbb{Q}}\{E_{p_T,q_T;\lambda_T}(Y_{T+1}) \mid \mathcal{F}_T\}\right].\end{aligned}$$

If $c_t = 0$, then there is no hypothesis about $p_t$ and $q_t$. For these cases, the definition in (6) may be extended to $\lambda = 0$, so that $E_{p_t,q_t;0} \equiv 1$ if $c_t = 0$. Then, if $\lambda_T = 0$ when $c_T = 0$,

$$\mathbb{E}_{\mathbb{Q}}\{E_{p_T,q_T;\lambda_T}(Y_{T+1}) \mid \mathcal{F}_T\} = (1 - c_T) + c_T \mathbb{E}_{\mathbb{Q}}\{E_{p_T,q_T;\lambda_T}(Y_{T+1}) \mid \mathcal{F}_T\} \leqslant 1$$

almost surely for $\mathbb{Q} \in \mathcal{H}_{S;c}$, so

$$\mathbb{E}_{\mathbb{Q}}\left\{\prod_{t=1}^{T} E_{p_t,q_t;\lambda_t}(Y_{t+1})\right\} \leqslant \mathbb{E}_{\mathbb{Q}}\left\{\prod_{t=1}^{T-1} E_{p_t,q_t;\lambda_t}(Y_{t+1})\right\}.$$

Iterating this argument shows that the product $\prod_{t=1}^{T} E_{p_t,q_t;\lambda_t}(Y_{t+1})$ is an e-value for $\mathcal{H}_{S;c}$; more precisely, the process $\prod_{j=1}^{t} E_{p_j,q_j;\lambda_j}(Y_{j+1})$, $t = 1, 2, 3, \ldots$, is a nonnegative supermartingale with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}}$. For a general lag $h$, sequential conditioning at time steps of 1 is not possible, and one option is to average the products of all e-values with a time difference of $h$, in the spirit of the U-statistics merging functions suggested by Vovk & Wang (2021). We summarize this in the following proposition.

PROPOSITION 1. *Let* $(Y_t, c_t, p_t, q_t, \lambda_t) \in \{0, 1\}^2 \times (0, 1)^2 \times [0, 1]$ *be defined on a measurable space* $(\Omega, \mathcal{F})$ *and adapted to the filtration* $\mathcal{F}_t$ $(t \in \mathbb{N})$*, and assume that* $\lambda_t = 0$ *if* $c_t = 0$*. Further, let S be a strictly consistent scoring function. Then for all* $T \geqslant h + 1$,

$$e_T = \frac{1}{h}\sum_{k=1}^{h}\prod_{l\in I_k} E_{p_l,q_l;\lambda_l}(Y_{l+h})$$

*with* $I_k = \{k + hs : s = 0, \ldots, \lfloor (T - k)/h \rfloor - 1\}$ *are* $\mathcal{F}_T$*-measurable and are e-values under* $\mathcal{H}_{S;c}$.

Proposition 1 is an analogous result to Theorem 1 in the sense that it only characterizes possible e-values for testing forecast dominance, but the parameters $\lambda_t$ could be any adapted sequence $(\lambda_t)_{t\in\mathbb{N}} \subset [0,1]$. E-values for dominance testing under the conditions $(c_t)_{t\in\mathbb{N}}$ are obtained by setting all e-values for which the condition is not satisfied to 1. The forecast user may, and in fact has to, tune the $(\lambda_t)_{t\in\mathbb{N}}$ in order to attain good power against a given alternative. Recall that at any $t$, $\lambda_t$ may be a function of all the forecasts and observations before time $t$. Instead of the parameters $\lambda_t$, it is usually more intuitive to think of an alternative probability $\eta_t$ for the event $Y_{t+h} = 1$ and then directly use the GROW e-values $E^{\eta_t}_{p_t,q_t}$ constructed in Theorem 2. In that respect, testing forecast dominance with e-values differs from $p$-value-based tests of a zero score difference, which do not require the user to explicitly specify an alternative hypothesis. In the applications in § 4 and § 5, we will give guidance on the selection of alternative hypotheses and show that reasonable power can be attained with simple heuristic methods.

As a side remark, choosing an alternative hypothesis for e-values in sequential forecast dominance testing is similar to the conditional predictive ability tests of Giacomini & White (2006), where $\mathcal{F}_t$-measurable test functions are used to weight score differences and improve power. Whereas selection of the test functions in the Giacomini–White test is delicate, because they may have an impact on the variance estimates and the finite-sample validity of the tests, e-values remain valid under any choice of adapted weights $(\lambda_t)_{t\in\mathbb{N}}$.

Our final theoretical result states that the e-values $e_T$ constructed above are also valid when $T$ is replaced by a stopping time $\tau$. This is a consequence of the fact that $(e_t)_{t\geqslant h+1}$ is a nonnegative supermartingale (see Ramdas et al., 2020, § 3).

PROPOSITION 2. *Let $\tau \in \mathbb{N}$ be a stopping time. Then under the assumptions of Proposition 1,*

$$\mathbb{E}_{\mathbb{Q}}(e_\tau) \leqslant 1, \quad \mathbb{Q} \in \mathcal{H}_S.$$

To understand validity under optional stopping intuitively, recall that at time $t$ the forecast user has to determine the parameter $\lambda_t$ in the e-value $E_{p_t,q_t;\lambda_t}(Y_{t+h})$. Optional stopping at $t_0$ corresponds to setting $\lambda_t \equiv 0$, or equivalently $E_{p_t,q_t;\lambda_t}(Y_{t+h}) \equiv 1$, for $t \geqslant t_0$, i.e., ignoring all observations starting from time $t_0 + h$. In the case of forecast lag 1, this allows the forecast user to stop evaluation at any time, since $\lambda_t$ in $E_{p_t,q_t;\lambda_t}(Y_{t+1})$ is defined at the same time as $Y_t$ is observed. However, when $h > 1$, the coefficients $\lambda_t$ in $E_{p_t,q_t;\lambda_t}(Y_{t+h})$ for $t = t_0 - h + 1, \ldots, t_0 - 1$ have already been determined in the past and may not be set to zero at $t_0$, since they must be $(\mathcal{F}_t)_{t\in\mathbb{N}}$-adapted. This implies that the stopped e-value depends on the unknown, future observations $Y_{t_0+1}, \ldots, Y_{t_0+h-1}$ and so is not deterministic at time $t_0$.

In the case $h = 1$, optional stopping is a powerful strategy when the goal is to assess forecast superiority at a significance level $\alpha \in (0, 1)$, because the stopping time

$$\tau_\alpha = \min\{T, \inf(t \geqslant 2 : e_t \geqslant 1/\alpha)\}$$

allows us to reject the null hypothesis as soon as the sequential e-value $e_t$ exceeds $1/\alpha$. If $h > 1$, one may similarly define

$$\tau_{\alpha,h} = \min\Big(T, \inf\big[t \geqslant h+1 : e_t \geqslant \max_{j=t-h+1,\ldots,t-1} E_{p_j,q_j;\lambda_j}\{\mathbb{1}(p_j > q_j)\}^{-1}/\alpha\big]\Big),$$

which guarantees that when stopping at $t_0$, the level $1/\alpha$ is exceeded no matter what values $Y_{t_0+1}, \ldots, Y_{t_0+h-1}$ take; see the Supplementary Material. Instead of specifying a significance level $\alpha$ in advance, one may as well transform the sequence $(e_t)_{t\in\mathbb{N}}$ into so-called anytime-valid

$p$-values $p_{t_0} = \min\{1, \inf_{s=1,\ldots,t_0} 1/e_s\}$, which are valid simultaneously for all $t_0 \geqslant h+1$ (see Ramdas et al., 2020, § 3.1).

## 4. Simulation examples

### 4.1. *Basic properties*

For the simulation examples in this subsection and the next, we transform e-values $E$ into $p$-values by taking their inverse $1/E$, so that direct comparisons with $p$-values are possible. Further variations of these simulation examples are presented in the Supplementary Material. An `R` package for the proposed methods and replication material for all results in this article are available at `https://github.com/AlexanderHenzi/eprob`.

In the first example, for varying $\mu \in (0,1)$, we simulate independent forecasts $p_t, q_t \sim \text{Unif}\,(0,1)$, define $\pi_t = \mu q_t + (1-\mu)p_t$, and generate independent Bernoulli observations $Y_{t+1}$ with mean $\pi_t$ conditional on $p_t$ and $q_t$. This represents a situation where forecasters only have access to partial information and both forecasts are not calibrated, i.e., $\mathbb{P}(Y_{t+1} = 1 \mid p_t) \neq p_t$ and $\mathbb{P}(Y_{t+1} = 1 \mid q_t) \neq q_t$. We choose $S$ to be the Brier score, so that $p_t$ outperforms $q_t$ if and only if $\pi_t \in [0, (p_t+q_t)/2]$ if $p_t < q_t$ or $\pi_t \in [(p_t+q_t)/2, 1]$ if $p_t > q_t$, i.e., if and only if $\mu \leqslant 0.5$. When $\mu > 0.5$, the GROW e-value is obtained by choosing $\pi_t$ as the alternative hypothesis probability, but in practice $\pi_t$ is not known. The forecast user might assume that the true probability of $Y_{t+1} = 1$ lies somewhere between $(p_t+q_t)/2$ and $q_t$, and choose a convex mixture $\eta_t(\xi) = \xi(p_t+q_t)/2 + (1-\xi)q_t$ with some $\xi \in (0,1)$ as an alternative. Proposition 1 implies that for $k \in \mathbb{N}$ and $\xi_1, \ldots, \xi_k \in (0,1)$,

$$e_{t;\xi_j} = \prod_{i=1}^{t} E_{p_i,q_i}^{\eta_i(\xi_j)}(Y_{i+1}), \quad e_t = \frac{1}{k}\sum_{j=1}^{k} e_{t;\xi_j}$$

are e-values under $\mathcal{H}_S$. In Fig. 1, we compare the rejection rates at the 5% level, corresponding to e-values greater than or equal to 20, when the $\xi_j$ are $k$ equally spaced weights in $(0,1)$ for $k=1$ and $k=5$, i.e., $\xi_1 = 0.5$ if $k=1$ and $\xi_l = l/6$ for $l = 1,\ldots,5$ in the case of $k=5$. We computed both the unstopped e-value $e_T$ and the stopped variant $e_{\tau_{0.05}}$, and the e-values under alternatives $\eta_t = \pi_t$ and $\eta_t = q_t$. The rejection rates are compared with those of one-sided $t$-tests of the null hypothesis that the mean Brier score difference is nonpositive. Additionally, we report the rejection rates when the $p$-value is used for optional stopping at given time-points upon seeing a significant difference.

Our simulations illustrate the known fact that classical statistical tests are not valid under stopping. At the boundary of the null hypothesis, the rejection rate of the $t$-test amounts to 0.12 for $T = 600$ and optional stops at times 150, 300 and 450; given the number of optional stops, this phenomenon occurs independently of the sample size. As for the e-values, stopping, i.e., $e_{\tau_{0.05}}$, is always a more powerful but valid strategy compared to the e-value $e_T$. While the heuristic alternatives achieve a power close to that under the correct alternative hypothesis, the misspecified hypothesis $\eta_t = q_t$ is clearly weaker. Interestingly, the correct alternative $\eta_t = \pi_t$ has lower power than the heuristic alternatives close to the boundary of the null hypothesis. This is not an error: specifying $\eta_t = \pi_t$ yields the optimal growth rate for the e-value, but this does not necessarily mean that it gives optimal power for the stopped e-value at the threshold $1/\alpha = 20$ in finite samples. The $t$-test generally achieves higher power than the e-values, which is to be expected given the absence of assumptions on the data-generating process and the validity under optional stopping for the e-values. See also Waudby-Smith & Ramdas (2021).

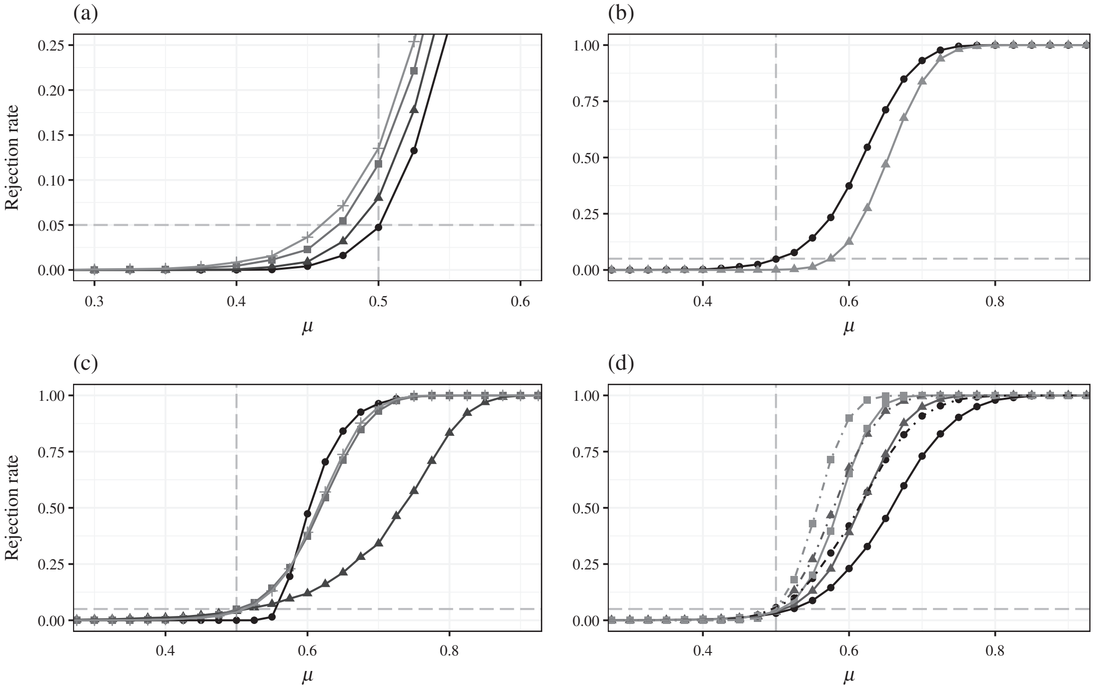

Fig. 1. Rejection rates of e-values and Student's $t$-test for the hypothesis that $p_t$ dominates $q_t$ with respect to the Brier score in the simulation of § 4.1. The sample size is $T = 600$ for panels (a)–(c) and the significance level is $\alpha = 0.05$ for all panels. (a) Rejection rate of $t$-test under optional stopping at one (triangles), three (squares) and five (crosses) equispaced time-points between 1 and $T = 600$, and without optional stopping (dots). (b) Rejection rates of stopped (dots) and unstopped (triangles) e-values with $k = 1$. (c) Rejection rates of e-values with different alternative hypotheses: $q_t$ (triangles), $\pi_t$ (dots), $k = 1$ (crosses) and $k = 5$ (squares). (d) Rejection rates of e-values with $k = 5$ (solid lines) and $t$-test without stopping (dot-dashed lines) for sample sizes $T = 300$ (dots), 600 (triangles) and 1200 (squares).

### 4.2. *Time series example*

We simulate $Z_t$ from a moving-average process $Z_t = \epsilon_t + \theta \sum_{j=1}^{4} \epsilon_{t-j}$ and define

$$Y_t = \mathbb{1}\{Z_t > 0\}, \quad \pi_{t;h} = \mathbb{P}(Z_t > 0 \mid Z_{t-j}, j = h, \ldots, 4) \quad (h = 1, \ldots, 4). \tag{7}$$

The probability $\pi_{t;h}$ corresponds to the ideal forecast at lag $h$. We compare $q_{t;h} = \pi_{t;h}$ and $p_{t;h} = \pi_{t;h+1}$ for lags $h = 1, 2, 3$, so that $q_{t;h}$ always outperforms $p_{t;h}$. As the parameter $\theta$ decreases, serial dependence decreases and the forecasting skills of $p_{t;h}$ and $q_{t;h}$ become similar. The alternative hypothesis for the e-values is the correct alternative $\eta_{t;h} = q_{t;h}$, so that the effect of a higher lag can be analysed in isolation from the question of how to choose the alternative hypothesis. Rejection rates are compared with the Diebold–Mariano test at the 5% level.

Figure 2 shows the dependence of the rejection rates on the parameter $\theta$ for different sample sizes $T$. The e-values use the stopping time $\tau_{0.05}$ for lag 1 and the stopping time $\tau_{0.05;h}$ for lags $h = 2$ and $h = 3$. As in the previous simulations, the power of the e-values is below that of the $p$-values for the lag-1 forecasts, where the Diebold–Mariano test essentially coincides with the $t$-test. For lags 2 and 3 this difference increases, since the combination method for e-values becomes less powerful. With increasing lag, the rejection rates of both methods decrease, but the difference to lag 1 is smaller for the Diebold–Mariano test than for the e-value. In this example, the Diebold–Mariano test is valid because the forecasts are ideal and the data-generating process is stationary. For the e-values, validity is guaranteed without such assumptions, which may be a great advantage in applications.

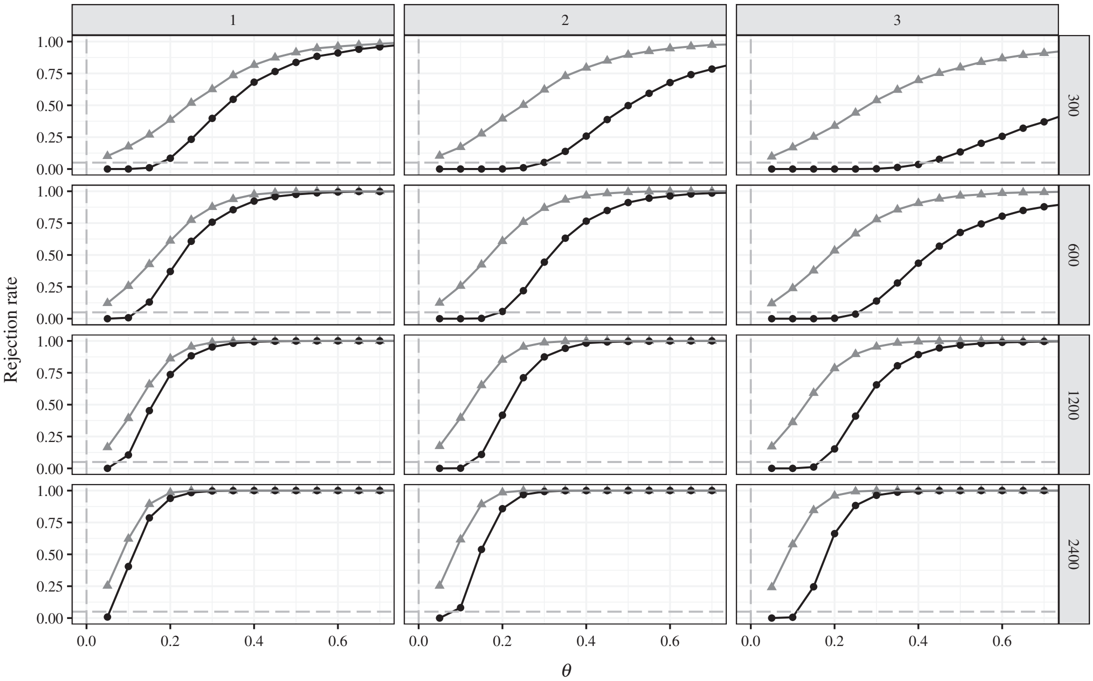

Fig. 2. Rejection rates of e-values (dots) and the Diebold–Mariano test (triangles) in the example (7) at the 5% level for different sample sizes $T$ (rows) and lags $h$ (columns).

## 5. Case study

### 5.1. *Data and methods*

Henzi et al. (2021) compared post-processing methods for precipitation forecasts with lags of one to five days at Brussels, Frankfurt, London Heathrow and Zurich airports. In their case study, probability-of-precipitation, or PoP, forecasts were evaluated with the Brier score, but no tests for significance of score differences were performed. We demonstrate here how to apply e-values to probability forecasts, and we will compare the results with state-of-the-art forecast dominance tests.

A detailed description of the dataset and methods is given in Henzi et al. (2021, § 5), so here we only summarize the key information. The dataset covers the period from 6 January 2007 to 1 January 2017, and upon accounting for missing values, the numbers of available observations are 3406 for Brussels, 3617 for Frankfurt, 2256 for London and 3241 for Zurich. Post-processing is applied to the ensemble forecasts of the European Centre for Medium-Range Weather Forecasts (Molteni et al., 1996; Buizza et al., 2005), which are issued on a latitude-longitude grid and consist of a high-resolution forecast, 50 perturbed ensemble forecasts at a lower resolution, and the control run for the perturbed forecasts. In simple terms, ensemble forecasts account for uncertainty by running a numerical weather prediction model several times, each time under slightly perturbed initial conditions; each run of the model yields a different forecast, and these forecasts together form a so-called ensemble (Leutbecher & Palmer, 2008). Ensemble forecasts are usually subject to biases and dispersion errors, which can be corrected by estimating the conditional distribution of the weather variable, given the numerical weather prediction ensemble. This statistical procedure is known as post-processing of ensemble forecasts (Vannitsem et al., 2018).

Henzi et al. (2021) proposed isotonic distributional regression, IDR, as a benchmark for such post-processing methods. IDR estimates conditional distributions nonparametrically and without any tuning parameters. The method is not specifically tailored to forecasting precipitation, and one would expect that a parametric model designed for this purpose will give more precise forecasts. One such method is heteroscedastic censored logistic regression, HCLR (Messner et al., 2014), which assumes that the square root of the precipitation follows a logistic distribution censored at zero. The implementation is as in Henzi et al. (2021). While the covariates in IDR are only the high-resolution forecast, the control forecast and the ensemble mean, the HCLR model additionally includes a scale parameter depending on the ensemble standard deviation.

In contrast to the study in Henzi et al. (2021), which uses an expanding window for the post-processing, we estimate both post-processed forecasts on half of the data for each airport for simplicity, and keep the remaining half for validation.

### 5.2. *Hypothesis tests*

We illustrate the usage of e-values in the following hypothesis tests. Firstly, we try to reject the null hypothesis that IDR PoP forecasts are better than HCLR PoP forecasts with respect to the Brier score. Secondly, we modify HCLR by dropping the scale parameter. It is expected that this variant, denoted by $\text{HCLR}_{-}$, will be outperformed by the original version of HCLR and also by IDR, since both IDR and $\text{HCLR}_{-}$ assume a monotone relationship between the covariates and the PoP, but the nonparametric IDR can estimate a broader class of functions. Finally, we further investigate the effect of the scale parameter on HCLR predictions for high precipitation. Suppose a weather forecaster issues a warning if the probability that the precipitation exceeds a high threshold is greater than 50%. As thresholds, we chose the empirical 90% quantile of precipitation in the training data for each airport. Intuitively, the HCLR model should yield more accurate warnings than $\text{HCLR}_{-}$, because it includes the ensemble standard deviation as an uncertainty measure.

The first and second sets of hypotheses are tested with the Brier score and the corresponding e-values. As an alternative probability, we take the convex mixtures $\eta_t = 0.25p_t + 0.75q_t$, which were explored in § 4, denoting by $p_t$ the forecasting method that is expected to have a better performance than $q_t$ under the null hypothesis. The hypothesis about the extreme precipitation warnings is a conditional comparison with the conditions $c_t = \mathbb{1}\{\max(p_t, q_t) \geqslant 0.5\}$. For this hypothesis, instead of dominance with respect to the Brier score, we test the stronger hypothesis of forecast dominance with respect to all scoring rules. The rationale is that the forecast dominance hypothesis should be easily rejected if the HCLR model truly issues the better tail forecasts; and, on the other hand, failing to reject may indicate either that, even with data of 10 years it is not possible to clearly discriminate the quality of such warnings, or that the ensemble standard deviation does not bring a benefit. For this hypothesis we define $\eta_t = q_t$, assuming that the conditional event probabilities should be much closer to those issued by HCLR than by $\text{HCLR}_{-}$. No optional stopping is applied in all e-values.

For comparison, we also compute $p$-values for the significance of score differences. The first two hypotheses are tested with one-sided Diebold-Mariano tests (Diebold & Mariano, 1995; see also Giacomini & White, 2006). To estimate the variance of the test statistics, we use the heteroscedasticity and autocorrelation consistent estimator with Bartlett weights; see Lerch et al. (2017, equation 2.18). For testing dominance of the tail probability forecasts, the test of Yen & Yen (2021) would allow arbitrary forecast lags, but it assumes strict stationarity. Since the sequence $c_t$ selects only particular instances, with possibly strongly varying time gaps in between, stationarity is highly questionable. We therefore apply the dominance test of Ehm & Krüger (2018), which is valid under weaker assumptions, but is limited to lag 1. Strictly speaking, both the Diebold-Mariano test and the forecast dominance test are valid under larger null hypotheses than the

Table 2. *Brier scores for different PoP forecasting methods, along with e-values and p-values for testing significance of score differences. The columns under HCLR/IDR show e-values and p-values for tests of the null hypothesis that IDR PoP forecasts achieve a lower Brier score than HCLR forecasts, with analogous interpretations for the other forecast pairs*

| | | Average Brier score | | | HCLR/IDR | | IDR/HCLR$_-$ | | HCLR/HCLR$_-$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Lag | IDR | HCLR | HCLR$_-$ | $E$ | $p$ | $E$ | $p$ | $E$ | $p$ |
| BRU | 1 | 0.107 | 0.117 | 0.118 | 0 | 0.9998 | $>100$ | $<10^{-4}$ | $>100$ | 0.0702 |
| | 2 | 0.119 | 0.123 | 0.125 | 0.01 | 0.9471 | $>100$ | 0.0101 | 13.602 | 0.0294 |
| | 3 | 0.134 | 0.133 | 0.136 | 0.425 | 0.4405 | $>100$ | 0.1916 | 15.185 | 0.0019 |
| | 4 | 0.152 | 0.145 | 0.148 | 4.804 | 0.0138 | 1.943 | 0.9358 | 5.165 | 0.0074 |
| | 5 | 0.171 | 0.161 | 0.164 | 16.969 | 0.0002 | 0.415 | 0.9965 | 3.436 | 0.0003 |
| FRA | 1 | 0.109 | 0.111 | 0.114 | 0 | 0.7784 | $>100$ | 0.0213 | $>100$ | $<10^{-4}$ |
| | 2 | 0.114 | 0.119 | 0.122 | 0.054 | 0.9643 | $>100$ | 0.0002 | $>100$ | 0.0004 |
| | 3 | 0.123 | 0.127 | 0.132 | 0.078 | 0.9352 | $>100$ | 0.0001 | 26.569 | $<10^{-4}$ |
| | 4 | 0.147 | 0.144 | 0.147 | 2.291 | 0.0966 | 9.618 | 0.5245 | 5.54 | 0.0001 |
| | 5 | 0.166 | 0.161 | 0.163 | 1.526 | 0.0305 | 2.362 | 0.8871 | 3.227 | 0.0051 |
| LHR | 1 | 0.135 | 0.138 | 0.139 | 0.029 | 0.8136 | 14.979 | 0.1314 | 2.845 | 0.3721 |
| | 2 | 0.138 | 0.143 | 0.143 | 0.188 | 0.9189 | $>100$ | 0.0509 | 2.868 | 0.4369 |
| | 3 | 0.152 | 0.154 | 0.155 | 0.734 | 0.7549 | 40.905 | 0.1394 | 2.488 | 0.3400 |
| | 4 | 0.169 | 0.167 | 0.169 | 1.429 | 0.2455 | 1.7 | 0.5442 | 1.744 | 0.0785 |
| | 5 | 0.186 | 0.181 | 0.182 | 1.577 | 0.0753 | 0.379 | 0.9288 | 1.118 | 0.3216 |
| ZRH | 1 | 0.104 | 0.108 | 0.110 | 0.003 | 0.9306 | $>100$ | 0.0055 | 61.747 | 0.0003 |
| | 2 | 0.110 | 0.112 | 0.114 | 0.116 | 0.7219 | 36.891 | 0.0304 | 10.276 | 0.0001 |
| | 3 | 0.121 | 0.118 | 0.121 | 1.516 | 0.0892 | 31.924 | 0.4410 | 5.098 | 0.0001 |
| | 4 | 0.138 | 0.132 | 0.134 | 4.069 | 0.0027 | 1.276 | 0.9588 | 2.771 | 0.0015 |
| | 5 | 0.165 | 0.156 | 0.159 | 15.151 | $<10^{-4}$ | 0.842 | 0.9978 | 2.383 | 0.0002 |

IDR, isotonic distributional regression; HCLR, heteroscedastic censored logistic regression; HCLR$_-$, heteroscedastic censored logistic regression without the scale parameter; BRU, Brussels; FRA, Frankfurt; LHR, London Heathrow; ZRH, Zurich; E, e-values; $p$, $p$-values.

e-values, as they only require the average score difference between $p_t$ and $q_t$ to be nonpositive, whereas the null hypothesis for the e-values asks for conditional superiority at each time-point. A comparison is nevertheless interesting, since these two tests represent commonly used methods for testing the significance of score differences.

Tables 2 and 3 show the e-values and one-sided $p$-values for the hypotheses described above, computed separately for each airport and forecast lag. The e-values are not transformed to $p$-values here. For interpretation, Vovk & Wang (2020, § 3) suggested a discrete scale such that e-values in $(0, 1]$, $(1, 3.16]$, $(3.16, 10]$, $(10, 31.6]$, $(31.6, 100]$ and $(100, \infty)$ represent no, poor, substantial, strong, very strong and decisive evidence against the null hypothesis, respectively. E-values greater than 100 are not displayed to improve readability, but an untruncated version of Table 2 is included in the Supplementary Material so that it is possible to update the e-values with more recent data. For all hypotheses, the $p$-values and e-values largely lead to the same conclusions. HCLR does not outperform IDR for PoP forecasts at lags 1–3, but for the Brussels and Zurich airports there is substantial to strong evidence that it achieves lower Brier scores at lags 4 and 5. HCLR$_-$ is clearly outperformed by the more complex variant with the ensemble-dependent scale parameter at short lags; also, for the longer lead times there is some evidence that including the scale parameter improves the forecasts, except for London airport. As for the difference between IDR and HCLR$_-$, both the e-values and the $p$-values suggest that IDR yields the better forecasts at lags 1–3, but at lags 4 and 5 there are no rejections of the null hypothesis. Figure 3 shows how

Table 3. *Sample sizes, e-values and p-values for the comparison of tail probability forecasts; the sample size is the number of observations for which the condition* $\min(p_t, q_t) \geqslant 0.5$ *holds*

| | Brussels | | Frankfurt | | London | | Zurich | |
|---|---|---|---|---|---|---|---|---|
| Lag | *n* | *E* (*p*) | *n* | *E* (*p*) | *n* | *E* (*p*) | *n* | *E* (*p*) |
| 1 | 116 | >100 (0.050) | 79 | 0.175 (0.814) | 72 | 0.45 (0.724) | 92 | 0.047 (0.892) |
| 2 | 88 | 23.409 | 87 | 3.327 | 69 | 1.332 | 99 | 2.961 |
| 3 | 68 | 10.704 | 62 | 3.542 | 60 | 1.429 | 75 | 0.567 |
| 4 | 49 | 2.338 | 53 | 1.166 | 39 | 0.868 | 52 | 0.773 |
| 5 | 28 | 1.029 | 26 | 1.033 | 30 | 1.077 | 36 | 1.073 |

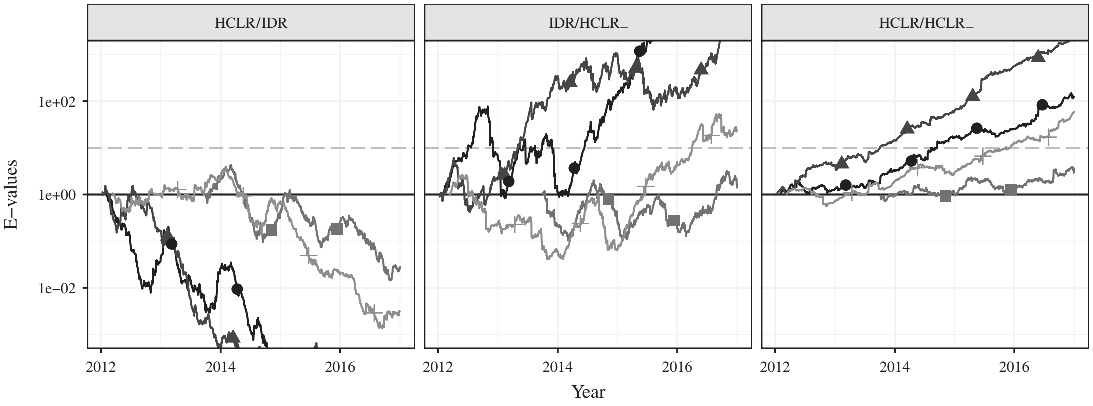

Fig. 3. E-values for the hypotheses tests at lag 1 for Brussels (dots), Frankfurt (triangles), London (squares) and Zurich (crosses). The abbreviations of the hypotheses are as in Table 2.

the cumulative products of the e-values for the hypothesis tests at lag 1 evolve over time. If the goal was to accumulate strong evidence against the hypotheses, say exceeding the level 10, then the hypothesis that IDR outperforms HCLR$_-$ could already be rejected with only 9% or 27% of the data for Brussels and Frankfurt airport, respectively, which is where the corresponding lines first cross the level 10. For Zurich airport, rejection happens at 85% of the total sample size.

Interestingly, in the comparison of HCLR and HCLR$_-$ for Brussels at lag 1, the *p*-value is nonsignificant, at 0.07, but the e-value gives decisive evidence, being greater than 100. We attribute this to the different null hypotheses of the tests. The mean difference in Brier score is only 0.001 with an estimated standard deviation of 0.03, giving little evidence against the null hypothesis of the Diebold-Mariano test. However, the null hypothesis for the e-value is smaller, requiring that HCLR$_-$ outperform HCLR at all time-points. Even if the score differences are only small, evidence eventually accumulates over the whole time period; see the rightmost panel of Fig. 3. The fact that the e-values in the HCLR/HCLR$_-$ comparison decrease with the forecast lag is an effect of the less powerful merging method for e-values with higher lag.

In the comparisons of extreme precipitation warnings, the *p*-value gives some evidence against the null hypothesis for Brussels airport, and the corresponding e-value is decisive, $E = 3703$. For the other lag-1 forecasts, both *p*-values and e-values do not indicate that including the ensemble standard deviation brings a benefit. As for the higher lags, for London and Zurich airports there is no evidence that HCLR outperforms HCLR$_-$, and for Brussels and Frankfurt airports there is evidence only at lags 2 and 3. Overall, the evidence in favour of the HCLR model for issuing extreme precipitation warnings as compared to HCLR$_-$ is surprisingly weak.

## Acknowledgement

The authors acknowledge financial support from the Swiss National Science Foundation. They are grateful to Ruodu Wang for introducing them to e-values for hypothesis testing, and to Tobias Fissler and Aaditya Ramdas for helpful comments.

## Supplementary material

Supplementary Material available at *Biometrika* online includes extensions of the simulation examples, a proof of the validity of the proposed stopping rule for lags $h > 1$, and a version of Table 2 without truncation of the e-values.

## Appendix

*Proof of Theorem* 1. If $E(y)$ is of the stated form, then $E(y) \geqslant E\{\mathbb{1}(p > q)\} = 1 - \lambda \geqslant 0$, and one can easily verify that $E$ has the given null hypothesis. Assume that $p < q$; the case of $p > q$ is analogous. Define $d_{p,q}(y) = S(p, y) - S(q, y)$ and, for $\pi \in [0, 1]$,

$$f(\pi) = \mathbb{E}_{\pi}\{d_{p,q}(Y)\} = (1 - \pi)d_{p,q}(0) + \pi d_{p,q}(1).$$

The elementary score representation (2) and $\nu\{[p, q)\} > 0$ imply that $d_{p,q}(0) < 0 < d_{p,q}(1)$, so $f(\pi)$ is strictly increasing in $\pi$ and equal to zero for some $\pi_0 \in (0, 1)$. Let $E = E(y)$ be an e-value under $H_S$ with alternative $H_S^{\mathrm{c}}$, i.e., $E(y) \geqslant 0$ and

$$\mathbb{E}_{\pi}\{E(Y)\} = (1 - \pi)E(0) + \pi E(1) \leqslant 1 \iff f(\pi) \leqslant 0. \tag{A1}$$

Condition (A1) implies that $\mathbb{E}_{\pi}\{E(Y)\} = 1$ if and only if $f(\pi) = 0$, which yields

$$\frac{d_{p,q}(0)}{d_{p,q}(1) - d_{p,q}(0)} = \frac{E(0) - 1}{E(1) - E(0)}. \tag{A2}$$

Rearranging this equation gives $E(1) = 1 - \{1 - E(0)\}d_{p,q}(1)/d_{p,q}(0)$. It follows from (A1) and (A2) that $E(0) \in (0, 1)$, so with $\lambda = 1 - E(0)$ we obtain $E(y) = 1 + \lambda d_{p,q}(y)/|d_{p,q}(0)|$. Similar arguments for the case $p > q$ show that in general,

$$E(y) = 1 + \lambda \frac{d_{p,q}(y)}{|d_{p,q}\{\mathbb{1}(p > q)\}|}.$$

□

*Proof of Theorem* 2. All e-values for the given null hypothesis are of the form (6). To find the GROW e-value under the alternative that $Y = 1$ with probability $\pi_1$, we have to maximize

$$\mathbb{E}_{\pi_1}[\log\{E_{p,q;\lambda}(Y)\}] = (1 - \pi_1)\log\left[1 - \lambda\frac{d_{p,q}(0)}{d_{p,q}\{\mathbb{1}(p > q)\}}\right] + \pi_1 \log\left[1 - \lambda\frac{d_{p,q}(1)}{d_{p,q}\{\mathbb{1}(p > q)\}}\right],$$

where again $d_{p,q}(y) = S(p, y) - S(q, y)$. Let $p < q$; the $p > q$ case is analogous. Under this assumption $d_{p,q}(0) < 0 < d_{p,q}(1)$, and $g(\lambda) = \mathbb{E}_{\pi_1}[\log\{E_{p,q;\lambda}(Y)\}]$ is continuous in $\lambda$ with $g(0) = 0$ and $\lim_{\lambda \to 1} g(\lambda) = -\infty$, so a maximum is attained at some $\lambda \in [0, 1)$. Define $h = d_{p,q}(1)/d_{p,q}(0) < 0$, so that

$$g(\lambda) = (1 - \pi_1)\log(1 - \lambda) + \pi_1 \log(1 - \lambda h), \quad g'(\lambda) = -\frac{1 - \pi_1}{1 - \lambda} - \pi_1\frac{h}{1 - \lambda h}$$

and $g'(\lambda_0) = 0$ is equivalent to $\lambda_0 = \pi_1 + (1 - \pi_1)/h$. By the definition of $H_S$, $\pi_1 \notin H_S$ holds if and only if $\mathbb{E}_{\pi_1}\{d_{p,q}(Y)\} > 0$, which is equivalent to $\pi_1 + (1 - \pi_1)/h > 0$, so indeed $\lambda_0 > 0$ for all $\pi_1 \notin H_S$, and

$$E_{p,q;\lambda_0}(0) = 1 - \lambda_0 = (1 - \pi_1)\left(1 - \frac{1}{h}\right) = (1 - \pi_1)\frac{d_{p,q}(1) - d_{p,q}(0)}{d_{p,q}(1)},$$
$$E_{p,q;\lambda_0}(1) = 1 - \lambda_0\,\frac{d_{p,q}(1)}{d_{p,q}(0)} = \pi_1\frac{d_{p,q}(0) - d_{p,q}(1)}{d_{p,q}(0)}.$$

With $d_{p,q}(y) = \int \mathbb{1}\{p \leqslant \theta < q\}(\theta - y)\,\mathrm{d}\nu(\theta)$, it now follows that

$$\frac{d_{p,q}(1) - d_{p,q}(0)}{d_{p,q}(1)} = \frac{-\nu\{[p,q)\}}{-\nu\{[p,q)\} + \int_{[p,q)} \theta\,\mathrm{d}\nu(\theta)} = \frac{1}{1 - \kappa_\nu\{[p,q)\}}$$

and $1 - h = \{d_{p,q}(0) - d_{p,q}(1)\}/d_{p,q}(0) = \kappa_\nu\{[p,q)\}^{-1} > \pi_1^{-1}$, which gives the desired result. □

*Proof of Theorem* 3. A direct computation shows that $H = [0, p]$ if $p < q$ and $H = [p, 1]$ if $p > q$, and that $\mathbb{E}_\pi\{E_{p,q}^{\pi_1 *}(Y)\} \leqslant 1$ for all $\pi \in H$ and $\mathbb{E}_\pi\{E_{p,q}^{\pi_1 *}(Y)\} > 1$ for $\pi \notin H$. The result then follows by Theorem 1 of Grünwald et al. (2020), with $W_1$ being the Dirac measure of the point $\{\pi_1\}$. □

*Proof of Proposition* 1. This follows as in the $h = 1$ case with sequential conditioning on $\mathcal{F}_{k+hl}$, $l = 1, \ldots, \lfloor (T - k)/h \rfloor$, for each of the $h$ products $\prod_{l \in I_k} E_{p_l,q_l;\lambda_l}(Y_{l+h})$. □

## REFERENCES

BUIZZA, R., HOUTEKAMER, P. L., PELLERIN, G., TOTH, Z., ZHU, Y. & WEI, M. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.* **133**, 1076–97.

DIEBOLD, F. X. & MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econ. Statist.* **13**, 253–63.

EHM, W., GNEITING, T., JORDAN, A. & KRÜGER, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *J. R. Statist. Soc. B.* **78**, 505–62.

EHM, W. & KRÜGER, F. (2018). Forecast dominance testing via sign randomization. *Electron. J. Statist.* **12**, 3758–93.

GIACOMINI, R. & WHITE, H. (2006). Tests of conditional predictive ability. *Econometrica* **74**, 1545–78.

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Am. Statist. Assoc.* **106**, 746–62.

GNEITING, T., BALABDAOUI, F. & RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Statist. Soc. B.* **69**, 243–68.

GNEITING, T. & RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Assoc.* **102**, 359–78.

GNEITING, T. & RANJAN, R. (2013). Combining predictive distributions. *Electron. J. Statist.* **7**, 1747–82.

GRÜNWALD, P., DE HEIDE, R. & KOOLEN, W. M. (2020). Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*. IEEE.

HENZI, A., ZIEGEL, J. F. & GNEITING, T. (2021). Isotonic distributional regression. *J. R. Statist. Soc. B* **83**, 963–93.

KELLY JR, J. L. (1956). A new interpretation of information rate. *Bell System Tech. J.* **35**, 917–26.

LAI, T. Z., GROSS, S. T. & SHEN, D. B. (2011). Evaluating probability forecasts. *Ann. Statist.* **39**, 2356–82.

LAZARUS, E., LEWIS, D. J., STOCK, J. H. & WATSON, M. W. (2018). HAR inference: Recommendations for practice. *J. Bus. Econ. Statist.* **36**, 541–59.

LERCH, S., THORARINSDOTTIR, T. L., RAVAZZOLO, F. & GNEITING, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statist. Sci.* **32**, 106–7.

LEUTBECHER, M. & PALMER, T. N. (2008). Ensemble forecasting. *J. Comp. Phys.* **227**, 3515–39.

MESSNER, J. W., MAYR, G. J., WILKS, D. S. & ZEILEIS, A. (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Weather Rev.* **142**, 3003–14.

MOLTENI, F., BUIZZA, R., PALMER, T. N. & PETROLIAGIS, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.

PATTON, A. J. (2020). Comparing possibly misspecified forecasts. *J. Bus. Econ. Statist.* **38**, 796–809.

RAMDAS, A., RUF, J., LARSSON, M. & KOOLEN, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*.

RANJAN, R. & GNEITING, T. (2010). Combining probability forecasts. *J. R. Statist. Soc. B.* **72**, 71–91.

Schervish, M. J. (1989). A general method for comparing probability assessors. *Ann. Statist.* **17**, 1856–79.

Seillier-Moiseiwitsch, F. & Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. *J. Am. Statist. Assoc.* **88**, 355–59.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *J. R. Statist. Soc. A* **184**, 407–31.

Vannitsem, S., Wilks, D. S. & Messner, J., eds. (2018). *Statistical Postprocessing of Ensemble Forecasts*. Amsterdam: Elsevier.

Vovk, V. & Wang, R. (2020). True and false discoveries with independent e-values. *arXiv:2003.00593*.

Vovk, V. & Wang, R. (2021). E-values: Calibration, combination, and applications. *Ann. Statist.* **49**, 1739–54.

Wang, R. & Ramdas, A. (2020). False discovery rate control with e-values. *arXiv:2009.02824v2*.

Waudby-Smith, I. & Ramdas, A. (2021). Estimating means of bounded random variables by betting. *arXiv:2010.09686v4*.

Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test* **5**, 1–60.

Yen, Y. & Yen, T. (2021). Testing forecast accuracy of expectiles and quantiles with the extremal consistent loss functions. *Int. J. Forecast.* **37**, 733–58.

Zhu, Y. & Timmermann, A. (2020). Can two forecasts have the same conditional expected accuracy? *arXiv:2006.03238*.

[*Received on* 15 *March* 2021. *Editorial decision on* 6 *September* 2021]