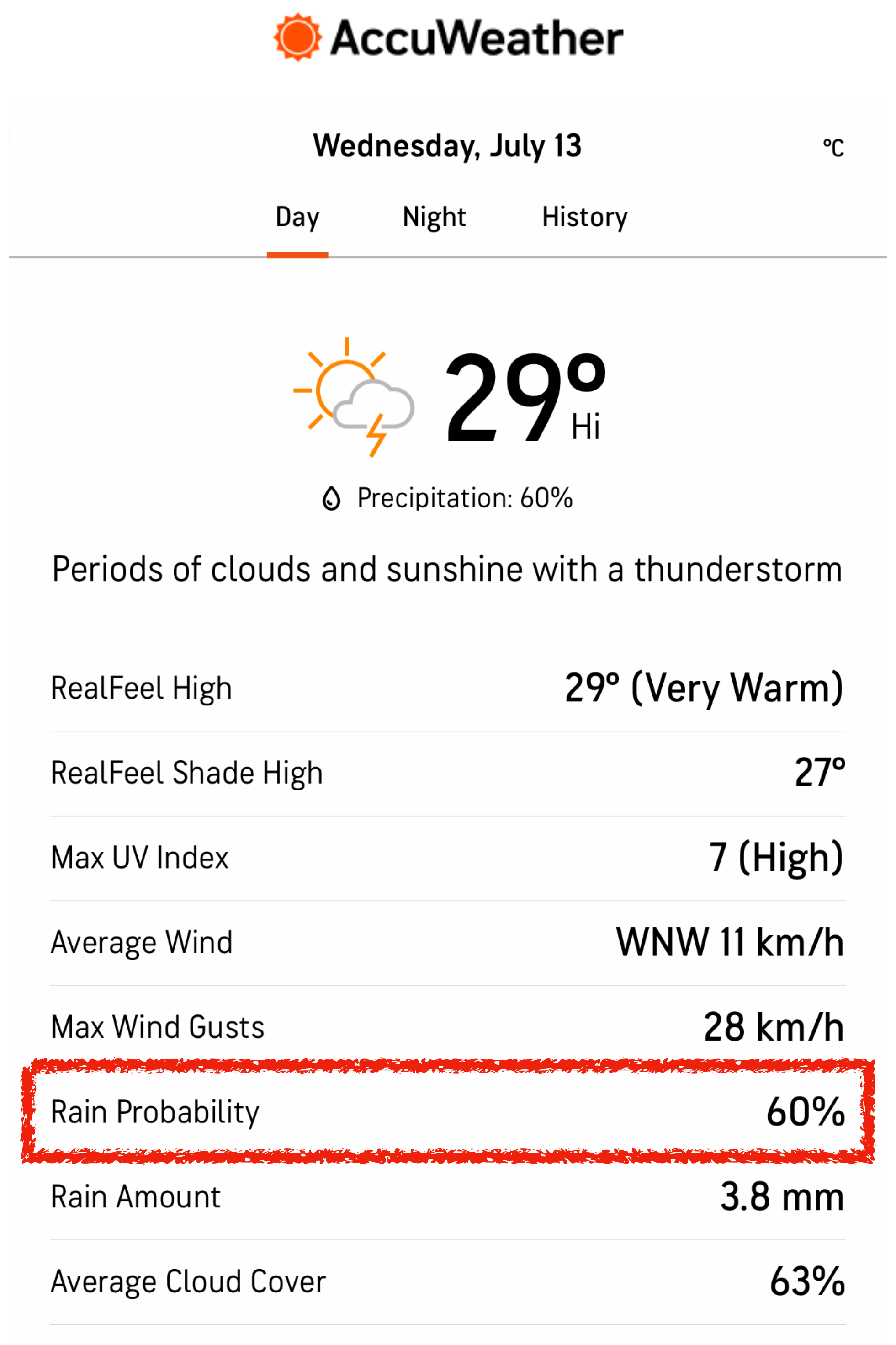
 

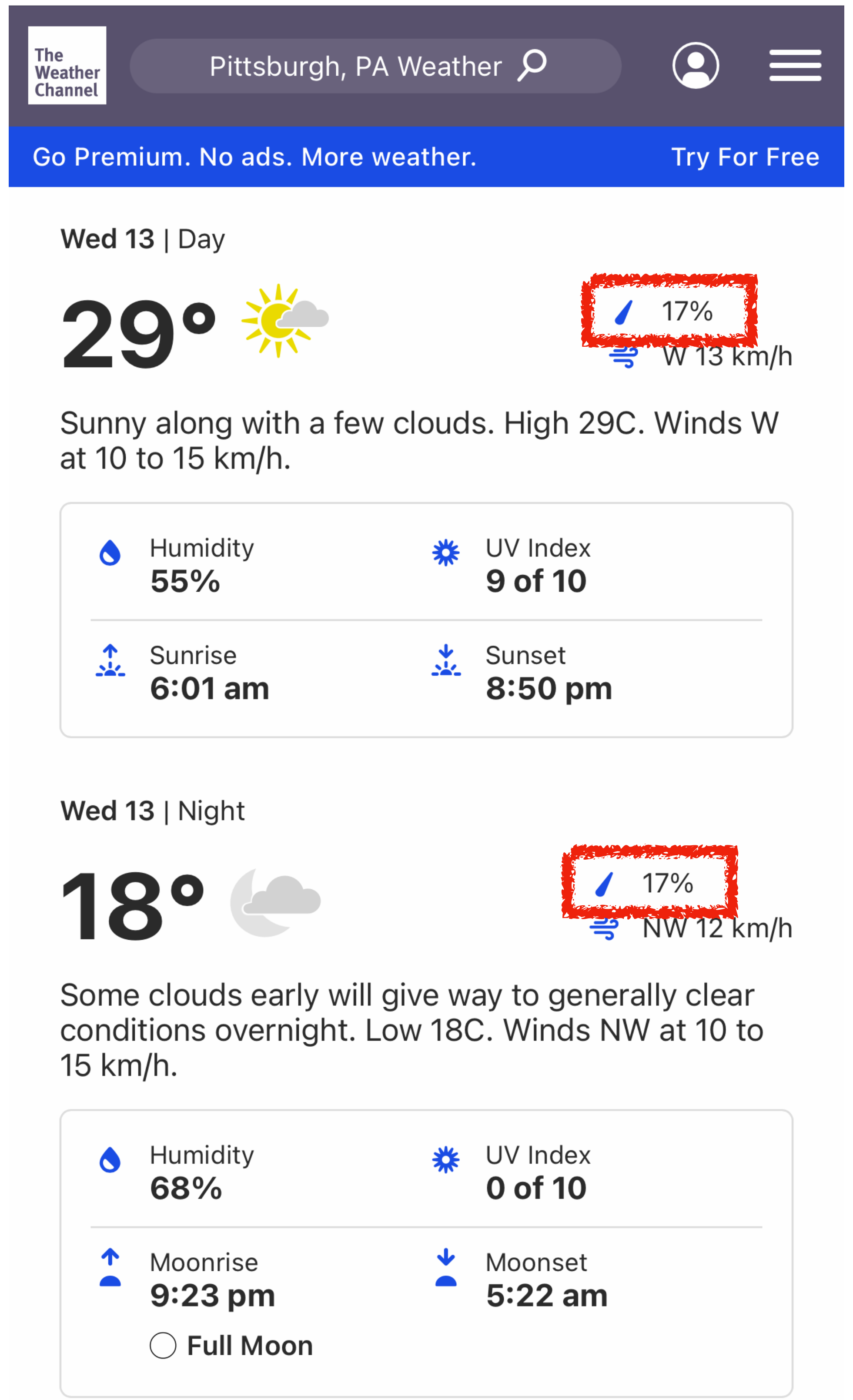
Comparing Sequential Forecasters

Yo Joong "YJ" Choe & Aaditya Ramdas
Dept. of Statistics & Data Science; Machine Learning Dept.
Carnegie Mellon University

Paper: https://arxiv.org/abs/2110.00115
Code: https://github.com/yjchoe/ComparingForecasters

Why Compare Forecasters?





Which one is better?

Forecast Comparison (or Evaluation) Has a Long History

including a long line of work from our very own department:

The Statistician 32 (1983)

The Comparison and Evaluation of Forecasters†

MORRIS H. DeGROOT and STEPHEN E. FIENBERG

Department of Statistics, Carnegie–Mellon University, Pittsburgh, PA 15213, USA

Abstract: In this paper we present methods for comparing and evaluating forecasters whose predictions are presented as their subjective probability distributions of various random variables that will be observed in the future, e.g. weather forecasters who each day must specify their own probabilities that it will rain in a particular location. We begin by reviewing the concepts of calibration and refinement, and describing the relationship between this notion of refinement and the notion of sufficiency in the comparison of statistical experiments. We also consider the question of interrelationships among forecasters and discuss methods by which an observer should combine the predictions from two or more different forecasters. Then we turn our attention to the concept of a proper scoring rule for evaluating forecasters, relating it to the concepts of calibration and refinement. Finally, we discuss conditions under which one forecaster can exploit the predictions of another forecaster to obtain a better score.

The Annals of Statistics
1989, Vol. 17, No. 4, 1856–1879

A GENERAL METHOD FOR COMPARING PROBABILITY ASSESSORS

BY MARK J. SCHERVISH

Carnegie Mellon University

A probability assessor or forecaster is a person who assigns subjective probabilities to events which will eventually occur or not occur. There are two purposes for which one might wish to compare two forecasters. The first is to see who has given better forecasts in the *past*. The second is to decide who will give better forecasts in the *future*. A method of comparison suitable for the first purpose may not be suitable for the second and vice versa. A criterion called calibration has been suggested for comparing the forecasts of different forecasters. Calibration, in a frequency sense, is a function of long

CALIBRATION, COHERENCE, AND SCORING RULES*

TEDDY SEIDENFELD†

Department of Philosophy
Washington University in St. Louis

Can there be good reasons for judging one set of probabilistic assertions more *reliable* than a second? There are many candidates for measuring "goodness" of probabilistic forecasts. Here, I focus on one such aspirant: calibration. Calibration requires an alignment of announced probabilities and observed relative frequency, e.g., 50 percent of forecasts made with the announced probability of .5 occur, 70 percent of forecasts made with probability .7 occur, etc.

*Received December 1983; revised June 1984.
†I thank Jay Kadane and Mark Schervish for helpful discussions about their important work on calibration, and Isaac Levi for his constructive criticism of this and earlier drafts. Also, I have benefited from conversations with M. De Groot and J. K. Ghosh.

But Is It Still a Relevant Problem?

If anything, it matters even more in modern ML.

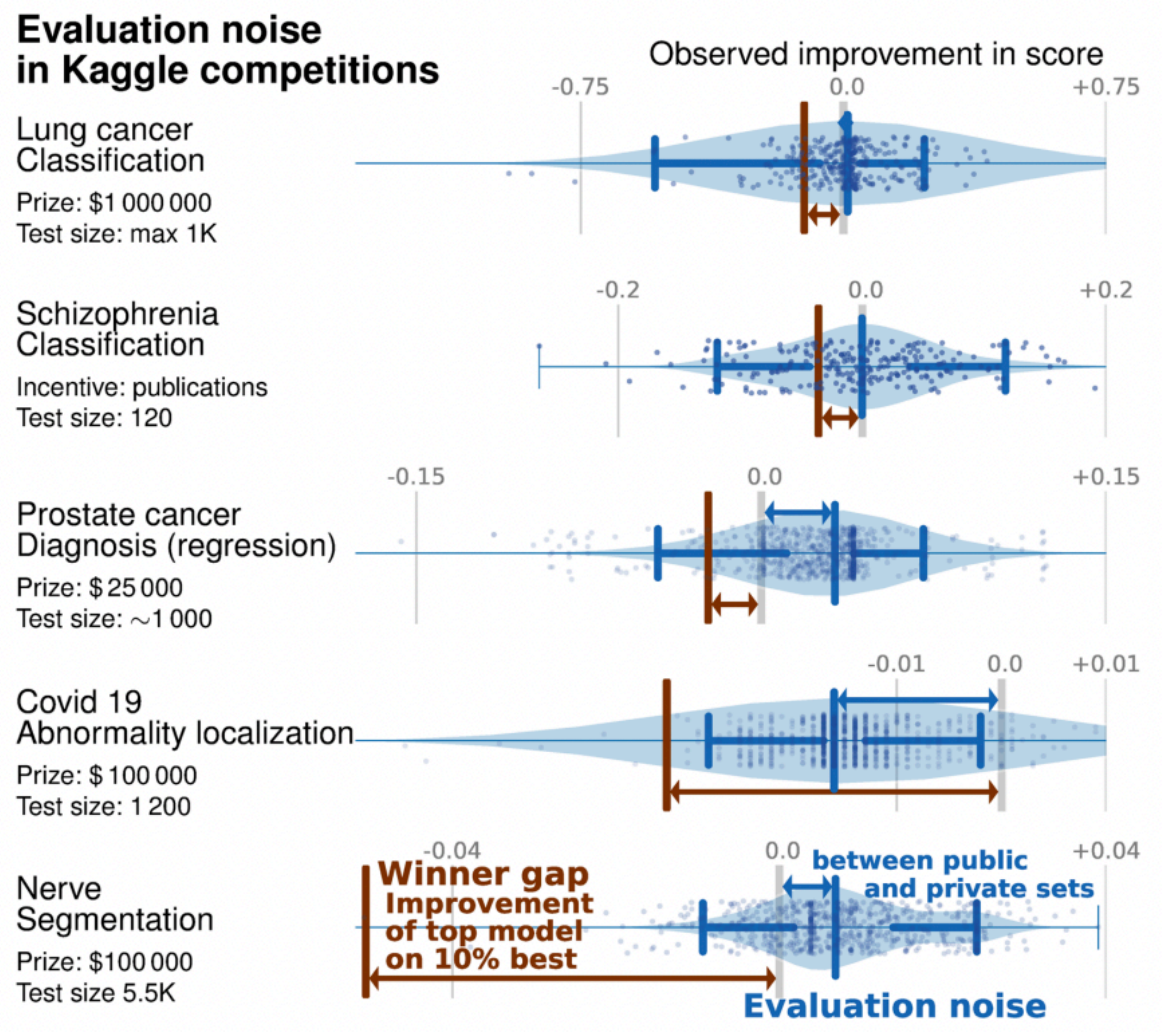


Figure from Varoquaux and Cheplygina (2022).

*"[…] overall medical imaging research seldom analyzes how likely empirical results are to be due to chance: **only 6%** of segmentation challenges surveyed[1], and **15%** out of 410 popular computer science papers published by ACM[2] use a statistical test."*

[1]Maier-Hein et al. (2018)
[2]Cockburn et al. (2020)

Forecast Comparison Meets Anytime-Valid Sequential Inference

Comparing Sequential Forecasters

2019 World Series (WSN vs. HOU)	Game 1	2	3	4	5	6	7
FiveThirtyEight	38%	41%	53%	59%	37%	41%	48%
Vegas Betting Odds	35%	38%	41%	51%	34%	37%	43%
Difference	*3%*	*3%*	***12%***	***8%***	*3%*	*4%*	*5%*
WSN Result	Win	Win	Loss	Loss	Loss	Win	Win

Probability forecasts, differences, and outcomes for the 2019 World Series.
Forecasts are provided as win percentages (%) for WSN.

Is one of the forecasters actually better than the other?

Can we answer this question repeatedly over time, and without making assumptions on the outcomes/forecasters?

Forecast Comparison as an Inference Problem

Especially Popular in Meteorology, Economics and Finance

- **Diebold and Mariano (1995)**
  - Asymptotic & exact finite-sample tests of equal forecast performance, assuming stationarity.
- **Giacomini and White (2006)**
  - Asymptotic tests of equal *conditional* forecast performance; allows non-stationarity (requires mixing).
- **Lai et al. (2011)**
  - Asymptotic tests of average scores & score differentials that have *linear equivalents*.
- **Henzi and Ziegel (2021)**
  - Valid sequential inference of conditional forecast *dominance* via e-processes.

A Game-Theoretic Setup

Let $\mathscr{P}$ (e.g., $[0, 1]$) denote the space of probability distributions on an outcome space $\mathscr{Y}$ (e.g., $\{0, 1\}$).

Consider the following protocol involving two forecasters:

Game (Comparing Sequential Forecasters). For rounds $t = 1, 2, \ldots$:

1. Forecaster 1 makes their probability forecast, $p_t \in \mathscr{P}$.
2. Forecaster 2 makes their probability forecast, $q_t \in \mathscr{P}$.
   (Steps 1 and 2 are in an arbitrary order.)
3. Reality chooses a probability $r_t \in \Delta(\mathscr{Y})$.
   (Note that r_t is not revealed to the forecasters.)

Game Filtration $\mathscr{G}_{t-1}$

4. $y_t \sim r_t$ is sampled and revealed.

p_t, q_t, r_t are predictable w.r.t. $\mathscr{G}_t$.
(i.e., $y_t \sim r_t$ is the only source of randomness)

How do we derive a valid sequential inference approach for comparing these two forecasters?

cf. Shafer & Vovk (2005; 2019); Lai et al. (2011); Henzi & Ziegel (2021)

Desiderata

<u>Game</u> **(Comparing Sequential Forecasters).**

For rounds $t = 1, 2, \ldots$:

1. Forecaster 1 makes their probability forecast, $p_t \in \mathscr{P}$.
2. Forecaster 2 makes their probability forecast, $q_t \in \mathscr{P}$.
   (Steps 1 and 2 are in an arbitrary order.)
3. Reality chooses a probability $r_t \in \Delta(\mathscr{Y})$.
   (Note that r_t is not revealed to the forecasters.)
4. $y_t \sim r_t$ is sampled and revealed.

1. **Time-Uniform & Anytime-Valid:** validity under continuous monitoring and at all (data-dependent) stopping times.
   - *Can we update our conclusions* ***as-we-go****, without sacrificing validity?*
2. **"Distribution-Free":** no assumptions on (the dynamics of) $(r_t)_{t \geq 1}$.
   - *The dynamics of real-world outcomes are probably not stationary or Markovian.*
3. **Model-Free:** No assumptions on the forecasts $(p_t)_{t \geq 1}$ and $(q_t)_{t \geq 1}$.
   - *We do not know the forecasting models of AccuWeather or Vegas betting odds.*
4. **Estimation/Test of the Average Conditional Predictive Ability:**
   - *Which forecaster has* ***usually*** *outperformed the other* ***so far****?*

SAVI Against Statistical Malpractice

- Classical inference methods (e.g., p-values & confidence intervals) do **NOT** guarantee validity under continuous monitoring and at data-dependent stopping times. (Susceptible to "p-hacking.")
- Safe, Anytime-Valid Inference (SAVI) approaches (Ramdas et al., 2022) have **statistical guarantees at arbitrary stopping times**, including data-dependent sample sizes. Confidence sequences (CS), in particular, can also be monitored continuously.
    - These methods are particularly suitable for *sequential* settings, *composite nulls*, and settings under *weaker assumptions*.
    - **Examples:** sequential tests, e-processes, p-processes, and confidence sequences.

	Classical	SAVI
Inference at Any Stopping Time ("peeking")	Invalid (requires correction)	**Valid**
Imprecise Probabilities (e.g., composite nulls)	No / Tricky	**Yes**
Game-Theoretic Interpretation	No	**Yes**

cf. Ville (1939); Wald (1945); Darling and Robbins (1967); Lai (1976); …
Ramdas, Grünwald, Vovk, and Shafer (2022): recent survey w/ many more references.

Evaluating Forecasters via Scoring Rules

- Given a probabilistic forecast $p \in \mathscr{P}$ for an outcome $y \in \mathscr{Y}$,
  a **scoring rule** $S : \mathscr{P} \times \mathscr{Y} \to \mathbb{R} \cup \{-\infty\}$ is any (quasi-integrable) function that assesses forecast quality.
  - Throughout this talk, **higher scores imply better forecasts**.
- A **proper** scoring rule elicits honest forecasts, and it measures both the calibration and sharpness of the forecaster.
  - Formally, S is proper if $\mathbb{E}_{y\sim q}[S(q, y)] \geq \mathbb{E}_{y\sim q}[S(p, y)], \forall p, q \in \mathscr{P}$. It is strictly proper when equality $\Leftrightarrow p = q$.

Brier	$S(p, y) = 1 - (p - y)^2$
Accuracy / Zero-One (proper but not strictly proper)	$S(p, y) = \mathbf{1}(p \geq 0.5)y + \mathbf{1}(p < 0.5)(1 - y)$
Winkler Score (relative to forecaster *q*)	$W(p, y; q) = \dfrac{S(p, y) - S(q, y)}{S(p, \mathbf{1}(p \geq q)) - S(q, \mathbf{1}(p < q))}$

Examples of proper scoring rules for binary forecasts.

cf. Gneiting and Katzfuss (2014); Dawid & Musio (2014); Gneiting and Raftery (2007); Winkler et al. (1996); Schervish (1989); Dawid (1986); Savage (1971); Brier (1950); …

Comparing Forecasts via Average Score Differentials

Let S be a scoring rule. The **average score differential** Δ_t between forecasts $(p_i)_{i \le t}$ and $(q_i)_{i \le t}$ is a time-varying parameter quantifying the expected difference in forecast quality up to time t:

$$\Delta_t = \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}_{i-1}[S(p_i, y_i) - S(q_i, y_i)],$$

where $\mathbb{E}_{i-1}[\,\cdot\,] = \mathbb{E}[\,\cdot \mid \mathscr{G}_{i-1}]$ is the conditional expectation w.r.t. $y_i \sim r_i$.

<u>**Goal:**</u> ***Estimate*** Δ_t ***at any time*** t***. (alternatively, test if*** $H_0 : \Delta_t \underset{(\ge)}{\le} 0$ ***for all times*** t***).***

Confidence Sequences for Estimating Time-Varying Parameters

Let $(\theta_t)_{t \geq 1}$ be a sequence of parameters indexed by time.

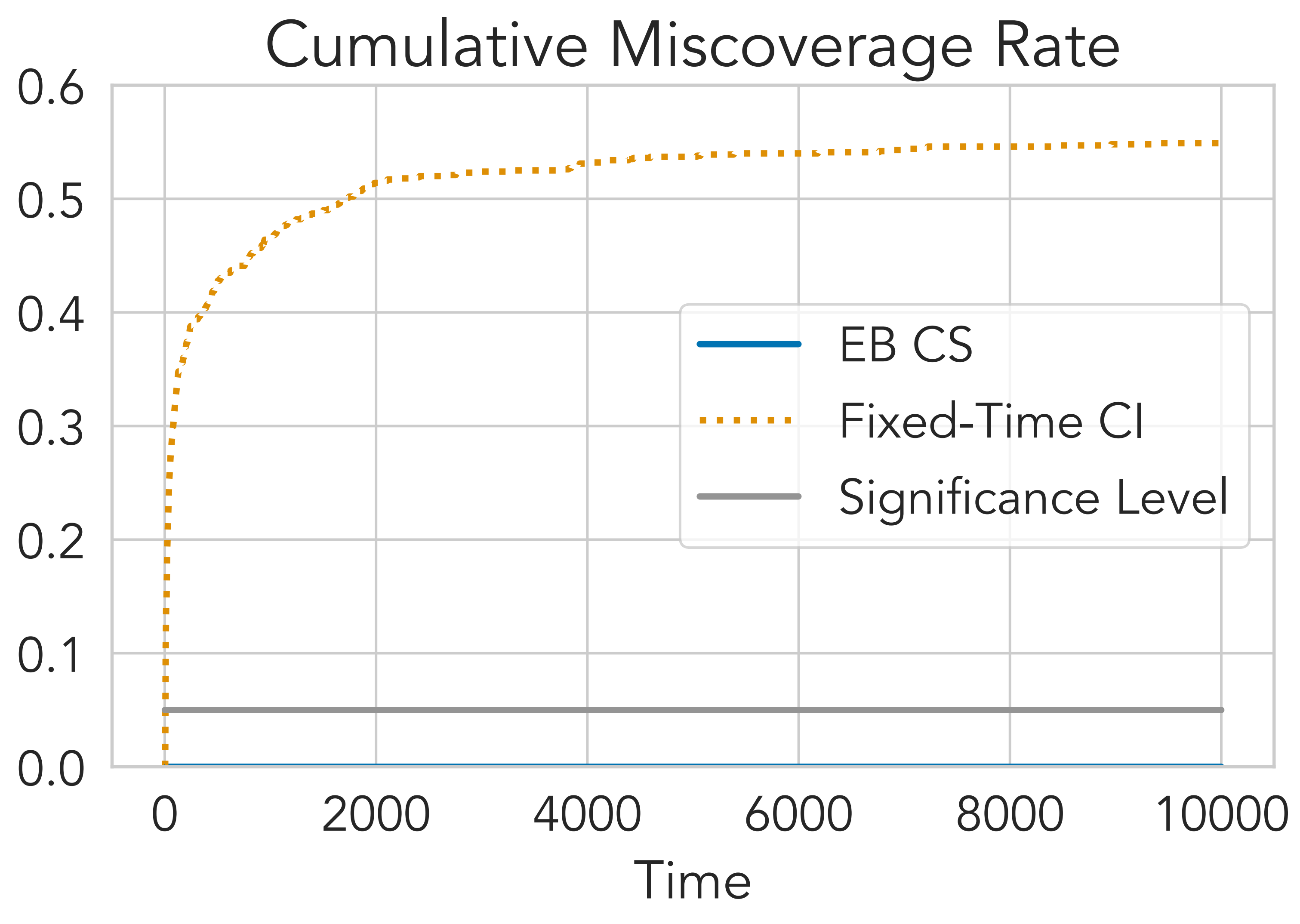
A **$(1-\alpha)$-level confidence sequence (CS)** $(C_t)_{t \geq 1}$ for $(\theta_t)_{t \geq 1}$ is a sequence of confidence intervals (CI) that has a **uniform coverage guarantee over time ("time-uniform")**:

$$\mathbb{P}(\forall t \geq 1 : \theta_t \in C_t) \geq 1 - \alpha.$$

The coverage guarantee is also valid at arbitrary & data-dependent stopping times ("anytime-valid"). *E.g., collecting additional data after estimation does not invalidate the guarantee.*

This is **not** true for the usual, fixed-time CI C_n, which only has coverage guarantees at a fixed sample size n:

$$\forall n \geq 1, \mathbb{P}(\theta_n \in C_n) \geq 1 - \alpha.$$



The fixed-time CI does not have a time-uniform coverage guarantee. A 95% CS has a cumulative miscoverage rate of ≤ 0.05 (zero in the above).

*Cumulative Miscoverage Rate: $\mathbb{P}(\exists i \leq t : \theta_i \notin C_i)$ (averaged over repeated simulations)

cf. Darling and Robbins (1967); Howard et al. (2021)

Main Result 1: CSs for Sequential Forecast Comparison

Theorem (Empirical Bernstein CS). Let $\hat{\delta}_i = S(p_i, y_i) - S(q_i, y_i)$ and $\hat{\Delta}_t = \frac{1}{t}\sum_{i=1}^{t} \hat{\delta}_i$. Suppose that $|\hat{\delta}_i|$ are bounded a.s. for each $i \geq 1$. Then, for each $\alpha \in (0, 1)$,

$$C_t := \left(\hat{\Delta}_t \pm c_\alpha \cdot \frac{\sqrt{\hat{V}_t \log\log \hat{V}_t}}{t} \right) \text{ forms a } (1-\alpha)\text{-level CS for } \Delta_t,$$

where $\hat{V}_t = \sum_{i=1}^{t} (\hat{\delta}_i - \hat{\Delta}_{i-1})^2$ denotes an empirical variance term and $c_\alpha \asymp \sqrt{\log(1/\alpha)}$ is a constant.

- **Asymptotically Zero Width.** The width of the CS shrinks to zero, at a $O(\sqrt{t^{-1}\log\log t})$ rate, achieving the same rate as a fixed-time CI up to logarithmic factors.

- **Variance-Adaptivity.** The width of this CS shrinks quickly as the variance stabilizes.

cf. Howard et al. (2020; 2021)

Main Result 2: E-Processes for Testing $H_0 : \Delta_t \leq 0$

Game-Theoretic Interpretation:

"The wealth of a gambler that bets against H_0."

(You're expected to lose money if H_0 is true & win money otherwise.)

Now consider the following (composite) null hypothesis:

$$H_0 : \Delta_t \leq 0, \quad \forall t \geq 1.$$

An **e-process** $(E_t)_{t \geq 0}$ for H_0 is a sequence of nonnegative random variables such that:

$$\text{for any stopping time } \tau, \qquad \mathbb{E}_{H_0}[E_\tau] \leq 1.$$

An e-process measures **the amount of accumulated evidence against the null hypothesis**.
If I observe an e-value (a realization at some stopping time τ) of ***100****, I would know that, if H_0 were true, the chance of it happening is* ***at most 1%*** *by Markov's inequality (or, in the sequential case, by Ville's inequality).*

Theorem (E-Process; Informal). Assume the same conditions as the previous Theorem.
Given the null $H_0 : \Delta_t \leq 0, \forall t$, there exists an **e-process** that corresponds to the (UCB of) the EB CS.

cf. Shafer (2011); Grünwald et al. (2019); Vovk and Wang (2021); Ramdas et al. (2022)

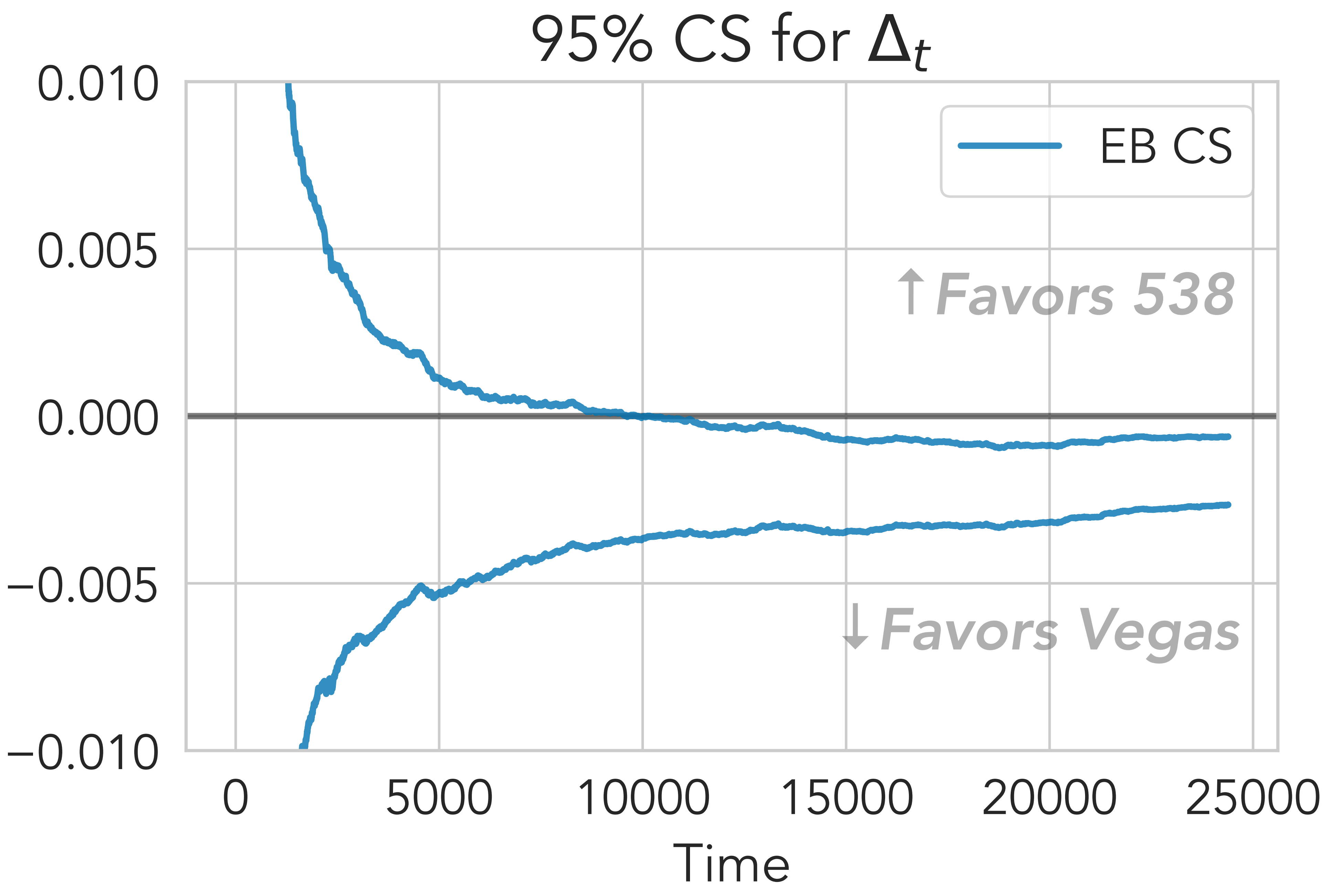
What's "Game-Theoretic" About It?

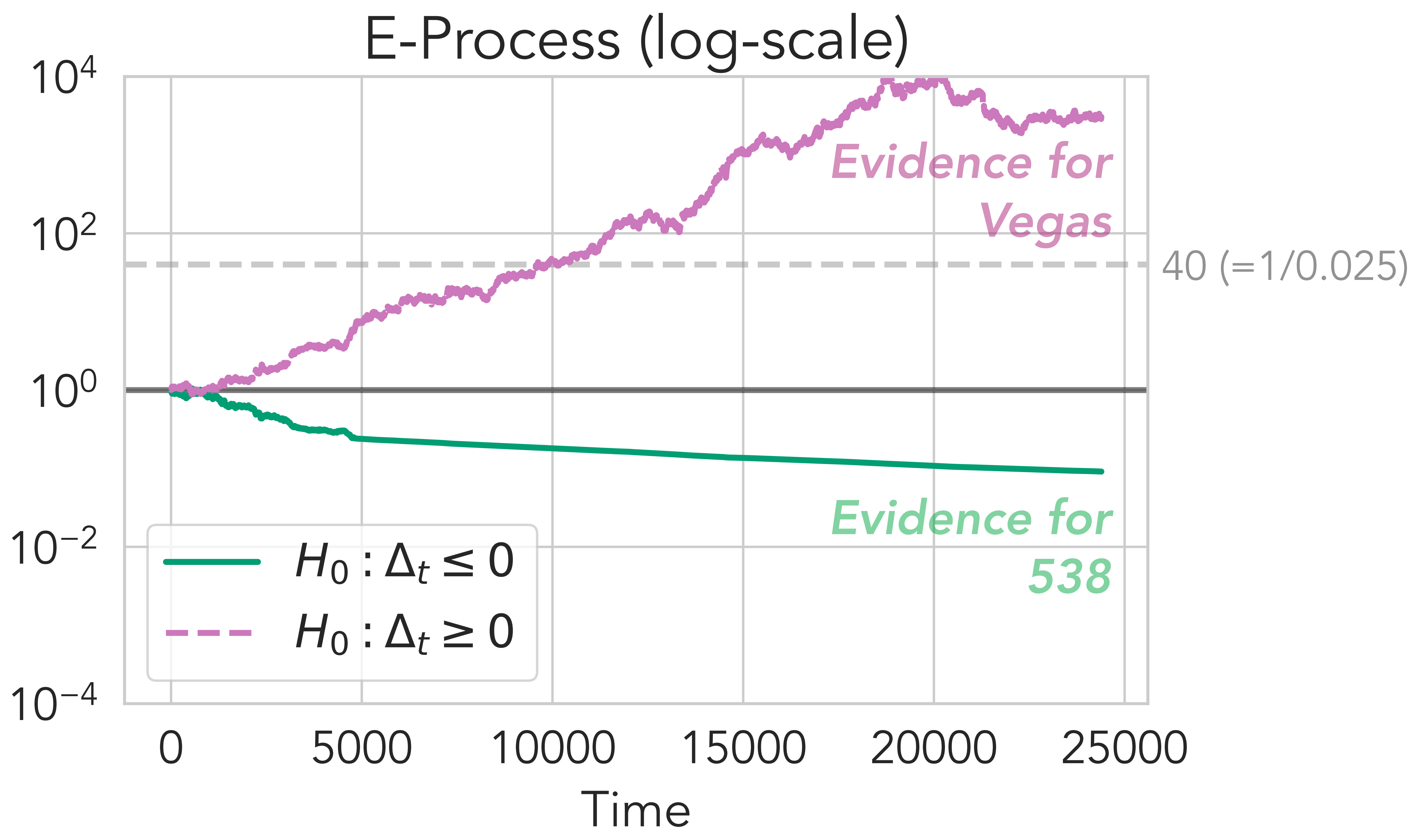
- Recall that a **supermartingale** $(L_t)_{t \geq 0}$ w.r.t. a distribution P (think: a point null) satisfies $\mathbb{E}_P[L_t \mid \mathcal{G}_{t-1}] = L_{t-1}\ \forall t \geq 1$.
    - *A nonnegative supermartingale (NSM) for* P *is the wealth of a gambler who bets on a game with odds determined (possibly unfairly) w.r.t.* P*.*
- An **e-process** for a set of distributions $\mathscr{P}$ (think: composite null) is any nonnegative process that is *upper-bounded* by a NSM for every $P \in \mathscr{P}$.
    - *An e-process for* $\mathscr{P}$ *is the minimum wealth of a gambler who places bets on all games determined by* $P \in \mathscr{P}$*.*
- **Game-theoretic statistics** sits in between game-theoretic probability and online learning, with a focus on **valid inference under weaker assumptions**.
    - *Key references include* Shafer; Grünwald; Ramdas et al.;
      Earlier references include Wald, Robbins, Darling, Siegmund, and Lai.

Experiments

Comparing Major League Baseball Forecasters

FiveThirtyEight vs. Vegas betting odds, using the Brier score

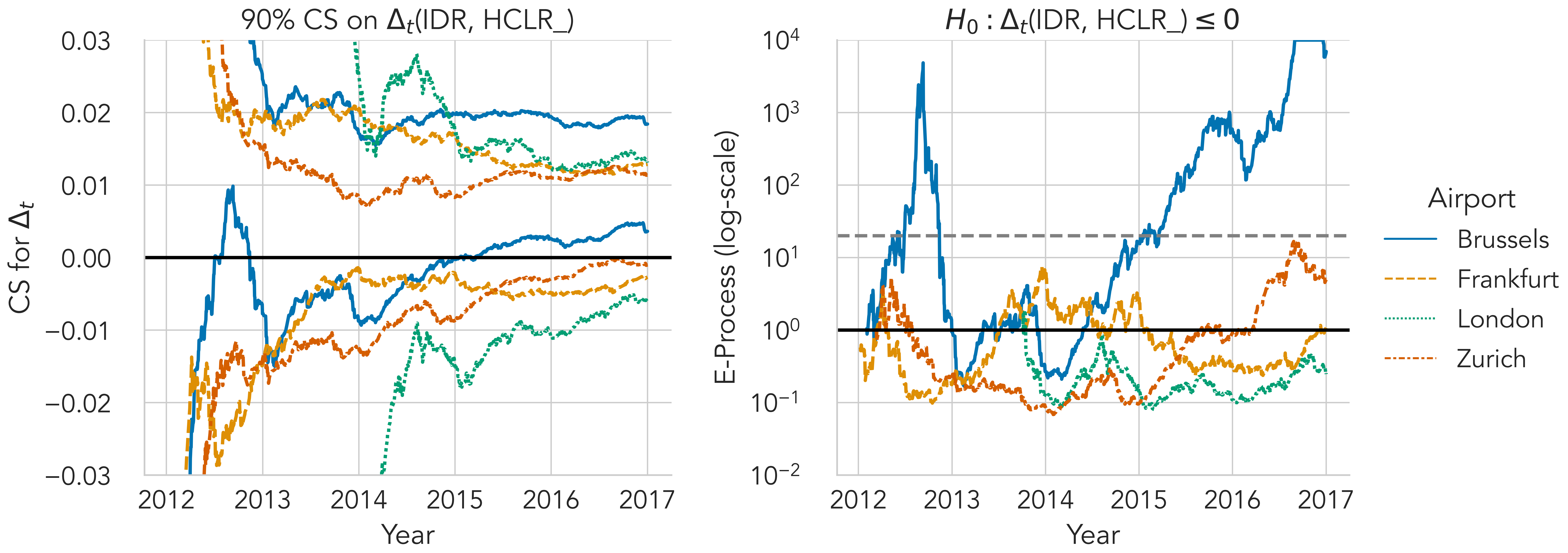




Data: Every MLB game's win/loss outcomes from 2010 to 2019.
See paper for further experiment details.

Comparing Ensemble Weather Forecasts

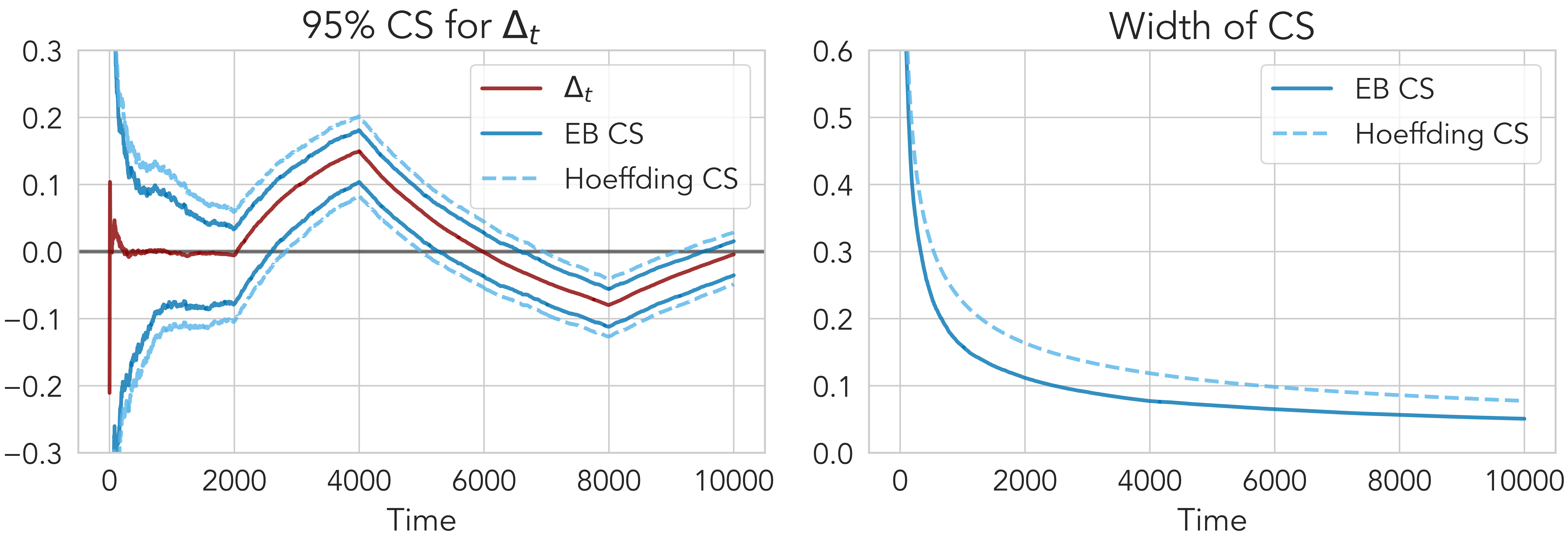
Experiment Adapted from Henzi & Ziegel (2022).



Data: ECMWF; Henzi et al. (2021)

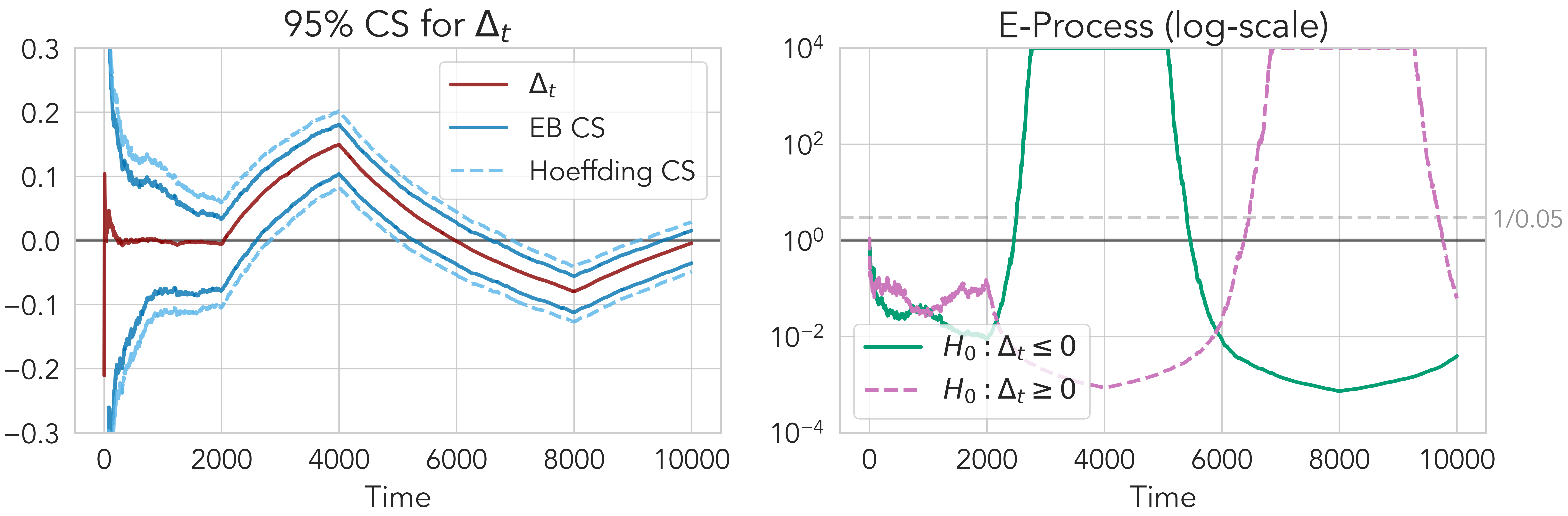
Simulated Experiments

CSs Uniformly Cover Time-Varying Means; EB CS (Variance-Adaptive) Is Tighter.



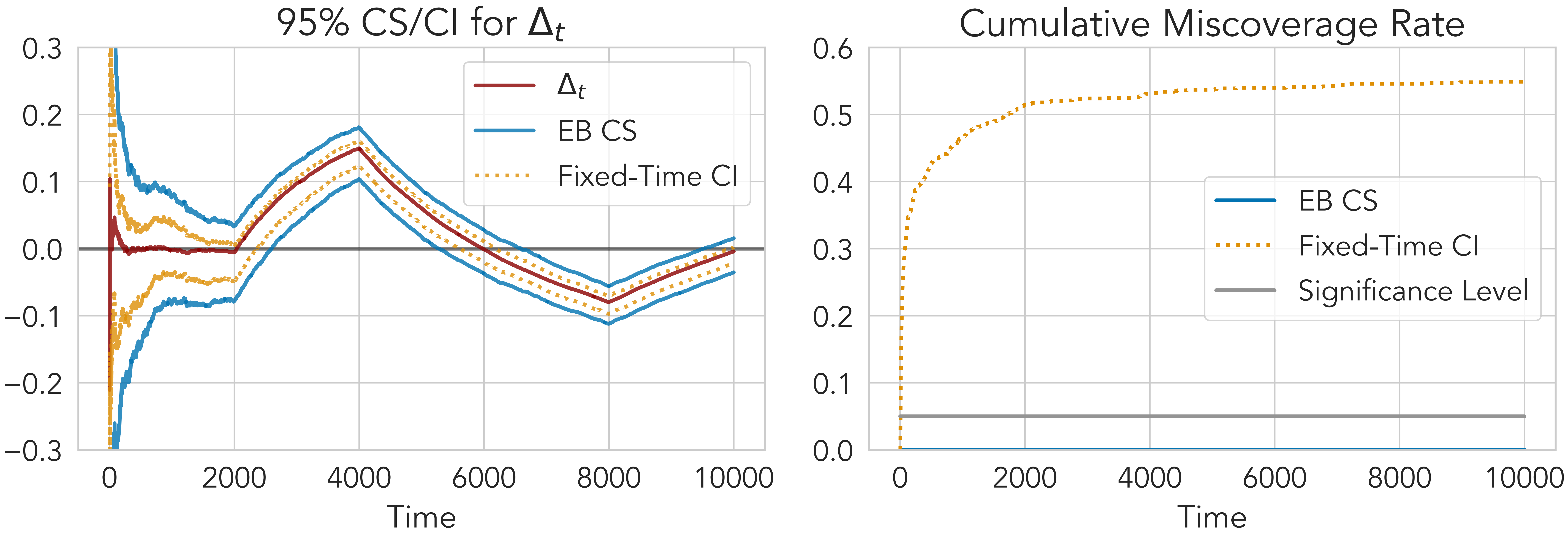
Simulated Experiments

E-Processes for p and q Match the LCB and UCB of EB CS Crossing the Zero Line.



Simulated Experiments

Fixed-Time CI Does Not Have the Time-Uniform Coverage Guarantee of CS.



*Cumulative Miscoverage Rate: $\mathbb{P}(\exists i \le t : \Delta_i \notin C_i)$
(averaged over repeated simulations)

Fixed-time CI is based on the martingale CLT (Lai et al., 2011).

Some Theory

Main Result 1: CSs for Sequential Forecast Comparison

<u>**Theorem (Empirical Bernstein CS).**</u> Let $\hat{\delta}_i = S(p_i, y_i) - S(q_i, y_i)$ and $\hat{\Delta}_t = \frac{1}{t}\sum_{i=1}^{t} \hat{\delta}_i$. Suppose that $|\hat{\delta}_i|$ are bounded a.s. for each $i \geq 1$. Then, for each $\alpha \in (0, 1)$,

$$C_t := \left(\hat{\Delta}_t \pm c_\alpha \cdot \frac{\sqrt{\hat{V}_t \log\log \hat{V}_t}}{t} \right) \text{ forms a } (1-\alpha)\text{-level CS for } \Delta_t,$$

where $\hat{V}_t = \sum_{i=1}^{t} (\hat{\delta}_i - \hat{\Delta}_{i-1})^2$ denotes an empirical variance term and $c_\alpha \asymp \sqrt{\log(1/\alpha)}$ is a constant.

- **Asymptotically Zero Width.** The width of the CS shrinks to zero, at a $O(\sqrt{t^{-1}\log\log t})$ rate, achieving the same rate as a fixed-time CI up to logarithmic factors.
- **Variance-Adaptivity.** The width of this CS shrinks quickly as the variance stabilizes.

cf. Howard et al. (2020; 2021)

Main Result 2 (More Formally): E-Processes for Testing $H_0 : \Delta_t \leq 0$

Theorem (E-Process). Assume the same conditions* as the previous Thm. Then, for each $\lambda \in [0, \lambda_{max})$,

$$E_t(\lambda) := \exp\left\{\lambda t \hat{\Delta}_t - \psi_E(\lambda)\hat{V}_t\right\} \text{ is an e-process for } H_0 : \Delta_t \leq 0, \forall t,$$

where $\psi_E(\lambda) = -\log(1-\lambda) - \lambda$ ("the sub-exponential CGF").

- **Method of Mixtures for E-Processes (& CSs).** For any distribution F on $[0, \lambda_{max})$, the mixture $E_t^{mix}(F) := \int E_t(\lambda) dF(\lambda)$ is also an e-process. (F can be chosen to be a "conjugate" distribution such that $E_t^{mix}(F)$ has a closed form.)
- **P-Process (Anytime-valid p-value).** If you'd prefer getting a p-value, then the e-process can be converted into a p-process via $p_t = E_t^{-1}$ or $p_t = (\max_{i \leq t} E_i)^{-1}$.

*In the case of e-processes, these conditions can further be weakened to pointwise score differentials being bounded-from-above only.

Underlying Theory:
Exponential Time-Uniform Boundaries for Sub-ψ Processes

One key underlying technique for constructing CSs is to derive a **nonnegative supermartingale (NSM)** that uniformly bounds the deviations of the sum.*

Define, for each $t \geq 1$:

- $S_t = \sum_{i=1}^{t} (\hat{\delta}_i - \delta_i)$, the (cumulative) "sum process" of deviations from conditional means, and
- $\hat{V}_t = \sum_{i=1}^{t} (\hat{\delta}_i - \gamma_i)^2$, its "variance process" (also called the "intrinsic time").

Then, we say that $(S_t)_{t\geq 1}$ is **sub-ψ_E ("sub-exponential") with variance process** $(\hat{V}_t)_{t\geq 1}$ if

$$L_t(\lambda) = \exp\left\{\lambda S_t - \psi_E(\lambda)\hat{V}_t\right\}$$

is bounded by a *supermartingale*. Here, $\psi_E(\lambda) = -\log(1-\lambda) - \lambda$ is the "CGF-like" function of an exponential r.v.

*More generally, all CSs are constructed (explicitly or implicitly) using e-processes, which strictly generalize NSMs. In our case, the above form of NSM suffices.

cf. Howard et al. (2020; 2021)

Underlying Theory:
Exponential Time-Uniform Boundaries for Sub-ψ Processes

If $(S_t)_{t\geq 1}$ is sub-ψ with variance process $(\hat{V}_t)_{t\geq 1}$ (i.e., $\mathbb{E}\left[\exp\left\{\lambda S_t - \psi(\lambda)\hat{V}_t\right\} \mid \mathcal{F}_{t-1}\right] \leq 1\ \forall t$), then for any $\alpha \in (0,1)$, we denote any boundary function $u_{\alpha/2}$ that satisfies the property

$$\mathbb{P}\left(\forall t \geq 1 : S_t \leq u_{\alpha/2}(\hat{V}_t)\right) \geq 1 - \alpha$$

as a **sub-ψ uniform boundary**. There are different options for forming tight uniform boundaries $u_{\alpha/2}$. Dividing the sum by t gives a CS for the time-varying average (e.g., of score differentials).

Furthermore, if $S_t = \sum_{i=1}^{t}(X_i - \mu_i)$ for an adapted sequence $(X_i)_{i\geq 0}$ with conditional means $\mu_i = \mathbb{E}_{i-1}[X_i]$, then we immediately obtain an **e-process for** $H_0 : \bar{\mu}_t := \frac{1}{t}\sum_{i=1}^{t}\mu_i \leq 0$:

$$E_t(\lambda) = \exp\left\{\lambda \sum_{i=1}^{t} X_i - \psi_E(\lambda)\hat{V}_t\right\}.$$

cf. Howard et al. (2020; 2021)

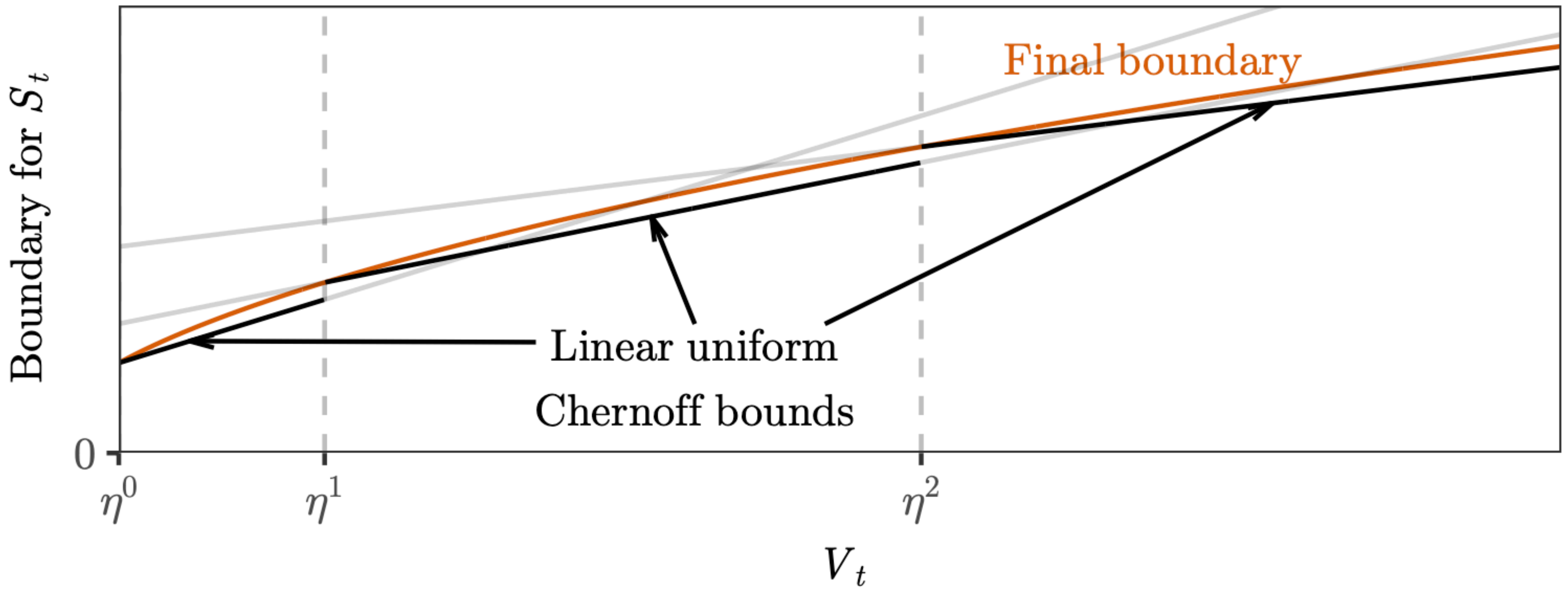
Uniform Boundary Option #1: Conjugate Mixture (CM), a.k.a. Method of Mixtures

- In our context, choose $F(\lambda)$ to be a suitable conjugate distribution for $(S_t)_{t\geq 0}$.
  - **Normal Mixture:** If $(S_t)_{t\geq 0}$ is sub-Gaussian, then choose F to be Gaussian.
  - **Gamma-Exponential Mixture:** If $(S_t)_{t\geq 0}$ is sub-exponential, then choose F to be Gamma.
  - *Betting interpretation:* mix bets over all λ-e-processes (and make it tractable).

- The CM boundary leads to a CS of width $O(\sqrt{t^{-1}\log t})$ (assuming $\hat{V}_t = O(t)$) and is usually tight in practice.
- Empirically, the mixture e-processes can be computed in closed-form; the corresponding uniform boundaries can be computed numerically or analytically depending on the mixture.

cf. Robbins and Siegmund (1970); Lai (1976); …; Howard et al. (2021); Kaufmann & Koolen (2021)

Uniform Boundary Option #2: Polynomial Stitching



$$\hat{\Delta}_t \pm 2 \cdot \frac{1.7\sqrt{\left(\hat{V}_t \vee 1\right)\left(\log\log\left(2\left(\hat{V}_t \vee 1\right)\right) + 3.8\right)} + 3.4 \log\log\left(2\left(\hat{V}_t \vee 1\right)\right) + 13}{t}$$

$\longleftarrow O(\sqrt{t^{-1} \text{ loglog } t})$

(assuming $\hat{V}_t = O(t)$)

cf. Howard et al. (2021)

Illustration: A Hoeffding-Style E-Process

Let $\hat{\delta}_i = S(p_i, y_i) - S(q_i, y_i)$ and $\delta_i = \mathbb{E}_{i-1}[\hat{\delta}_i] = S(p_i; r_i) - S(q_i; r_i)$.

Suppose that, for $i \geq 1$, $\hat{\delta}_i$ is sub-Gaussian (e.g., bounded scores) conditional on $\mathcal{G}_{i-1}$:

$$\mathbb{E}_{i-1}\left[\exp\{\lambda(\hat{\delta}_i - \delta_i) - \psi_N(\lambda)\}\right] \leq 1 \quad \forall \lambda \in \mathbb{R},$$

where $\psi_N(\lambda) = \lambda^2/2$ is the Gaussian cumulant generating function (CGF).

It then follows immediately that, for each $\lambda \in [0,\infty)$, the process $(L_t^H(\lambda))_{t\geq 0}$ defined by

$$L_t^H(\lambda) = \prod_{i=1}^{t} \exp\left\{\lambda(\hat{\delta}_i - \delta_i) - \lambda^2/2\right\} = \exp\left\{\lambda \sum_{i=1}^{t} (\hat{\delta}_i - \delta_i) - \psi_N(\lambda)t\right\}$$

is a NSM.

We also say that the cumulative sums $S_t = \sum_{i=1}^{t} (\hat{\delta}_i - \delta_i)$ are **sub-ψ_N ("sub-Gaussian")** with variance process $V_t = t$.

Illustration: A Hoeffding-Style E-Process

Now suppose that the weak null holds, i.e., $H_0^w : \Delta_t = \frac{1}{t}\sum_{i=1}^{t} \delta_i \leq 0$.

Under H_0^w, for any $\lambda \in [0,\infty)$ we have that $\exp\left\{-\lambda \sum_{i=1}^{t} \delta_i\right\} \geq 1$, so

$$L_t^H(\lambda) = \exp\left\{\lambda \sum_{i=1}^{t} (\hat{\delta}_i - \delta_i) - \psi_N(\lambda)t\right\} \geq \exp\left\{\lambda \sum_{i=1}^{t} \hat{\delta}_i - \psi_N(\lambda)t\right\} =: E_t^H(\lambda).$$

Since $(L_t^H(\lambda))_{t\geq 0}$ is a supermartingale, it follows from the supermartingale optional stopping theorem that, for any stopping time $\tau \leq \infty$,

$$\mathbb{E}_{H_0^w}[E_\tau^H(\lambda)] \leq \mathbb{E}_{H_0^w}[L_\tau^H(\lambda)] \leq \mathbb{E}_{H_0^w}[L_0^H(\lambda)] = 1.$$

In other words, $(E_t^H(\lambda))_{t\geq 0}$ **is an e-process for** H_0^w. The mixture over λ is also an e-process for H_0^w.

Additional Results in the Paper

- **An Asymptotic CS (Waudby-Smith et al., 2021) that requires only $(2 + \delta)$ bounded moments.**
  - Useful for estimating differences in unbounded scores.
- **A one-sided CS and e-process for Winkler's normalized score.**
  - Applicable to any proper scores for binary forecasts, such as the logarithmic score.
- **An approach for comparing lagged forecasts.**
  - More powerful tests or CSs remain an open problem.
- **Detailed comparisons with existing forecast comparison methods.**
  - Comparable power with fixed-time tests (DM'95, GW'06) in simulated examples.

Thank You

Preprint: https://arxiv.org/abs/2110.00115
Python Package (`comparecast`): https://github.com/yjchoe/ComparingForecasters
YJ's Webpage: https://yjchoe.github.io/

Questions?

Appendix

What is a "good" forecast?

Allan H. Murphy, in his 1993 essay, suggested three types of "goodness" in the context of weather forecasting. In his view, good forecasters achieve high levels of:

1. **Consistency:** correspondence between their forecasts and judgments;
    - Proper scoring rules encourage forecasters to achieve this consistency.
2. **Quality:** correspondence between their forecasts and the actual observations;
    - *Multifaceted*: not just accuracy or skill, but also reliability, resolution, and sharpness.
3. **Value:** incremental benefits of their forecasts to decision makers who use them.

The Testing-by-Betting Analogy

- I propose to you a game, which costs \$0.5 to enter. I'll pay you:
  - **\$1** if the roulette ball lands on a red slot ($\mathrm{P}(\text{red}) = 0.46$), and
  - **\$0** if it does not.
- This is an "unfair" game where I'm expected to earn \$0.04 for every round played. ($\mathbb{E}[\text{profit}] = 0.46 \cdot (-0.5) + 0.54 \cdot (+0.5) = +0.04$)
- Suppose you start with some budget and keep playing this game according to some rule. Then, your wealth at the end of each round forms a **nonnegative supermartingale (NSM)** w.r.t. $\mathrm{P} = 0.46$, as you're not expected to increase your wealth by playing this game.
- Yet, if the roulette is "hacked" in your favor and the actual probability is higher than $\mathrm{P} = 0.46$, then over time you'll make more money!
- Finally, replace P with the null hypothesis (possibly composite) and your wealth in the game quantifies the evidence the null.

At each round, a roulette ball lands on a red (or black) slot with probability ~46%.

cf. Shafer (2021); Ramdas et al. (2022); *inter alia.*

From Measure-Theoretic Probability To Game-Theoretic Probability

Events of small probability = Events for which the skeptic's capital grows large

Ville's Theorem (1939)

- An event A (a set of many sequences) has probability $P(A) = 0$ **if and only if** there exists a nonnegative supermartingale (NSM) $(L_t)_{t \geq 0}$ w.r.t. P such that $L_0 = 1$ and $\lim_{t \to \infty} L_t = \infty$ on A.

Ville's Inequality (1939)

- For any value $\alpha \in (0, 1)$, an event A has probability at most α, i.e., $P(A) \leq \alpha$, **if and only if** there exists a NSM $(L_t)_{t \geq 0}$ w.r.t. P such that

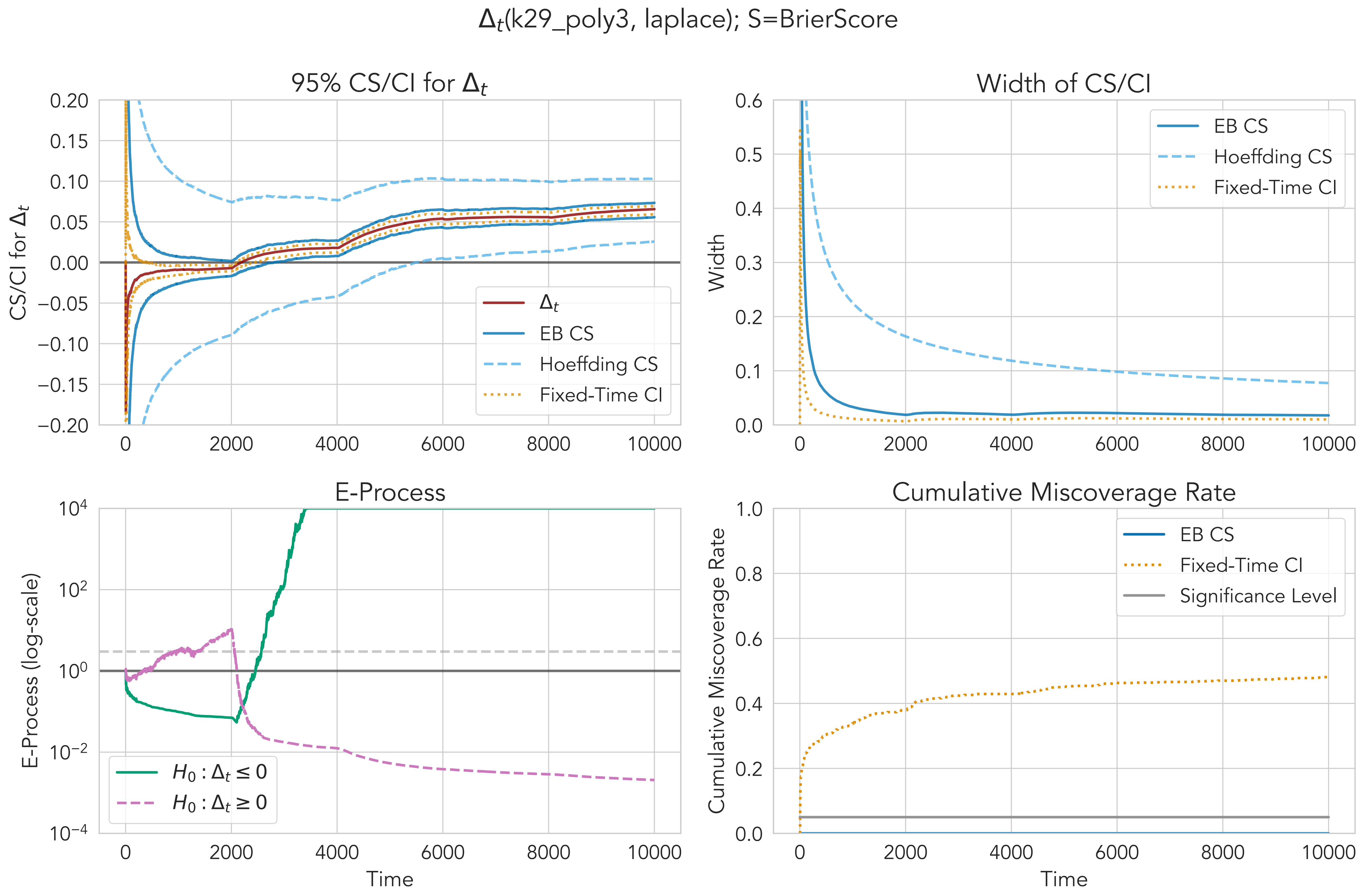
$$P(\exists t \geq 1 : L_t \geq 1/\alpha) \leq \alpha\,.$$

A Composite Generalization (Ruf et al., 2022)

- For composite sets of probabilities, the generalization corresponding to Ville's NSM is an **e-process** (after defining a proper outer measure).

More Simulated Experiments

Case: p eventually dominates q



Takeaway Message:

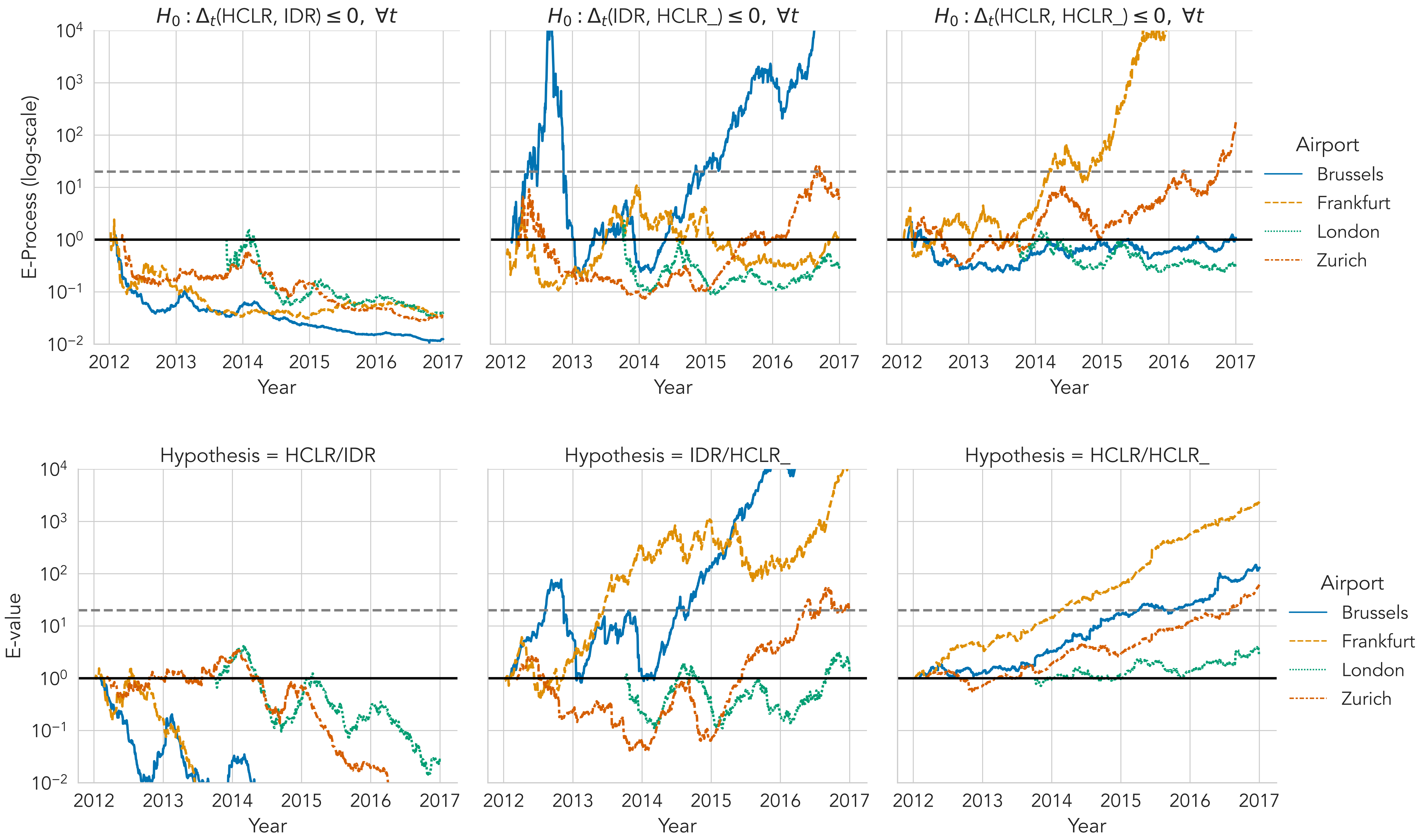
The fixed-time CI does NOT have a time-uniform guarantee.

E-Process Comparison with Henzi & Ziegel (2022)

Comparing Postprocessing Methods for Ensemble Weather Forecasts

Ours
(Weak null)

HZ'22
(Strong null)



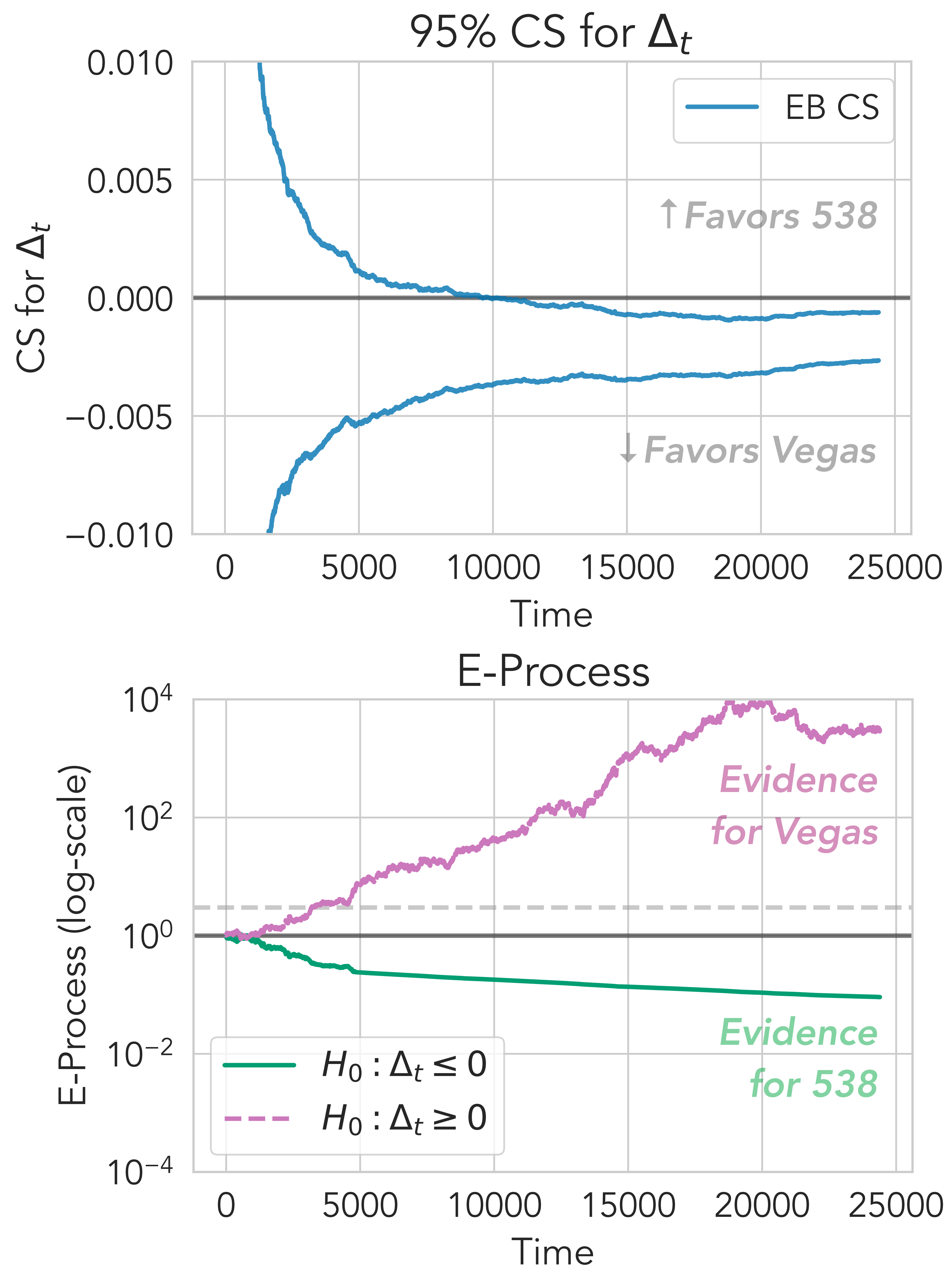
Methodology Comparison with HZ’22

	Ours	HZ’22
Anytime-Valid	**Yes**	**Yes**
Distribution-/Model-Free	**Yes**	**Yes**
Null Hypothesis	**Weak**	Strong
Estimation (Confidence Sequences)	**Yes**	No (not obvious)
E-Process Form	Exponential; variance-adaptive (Betting: mixture)	Product (Betting: GROW in the batch sense)
Outcome Type	**General**	Binary
Scoring Rule Type	Bounded or sub-Gaussian	**Any consistent scoring function (induces proper scoring rule)**
k-Step Forecasts	**Yes** (less power)	**Yes**

Why Use CSs & E-Processes in Practice?

An Easy-To-Use & Worry-Free Comparison Framework

- Especially in a sequential setting (think: A/B testing), the **graphical expressions** of CSs and e-processes provide a lot more information than CIs and p-values.
- Visualizations of e-processes also help **alleviate dichotomous thinking**, which is a contributing factor to the "replication crisis" in science (Helske et al., 2021).
- The anytime-validity of these methods ensure that the methods can be used **"worry-free"** and are less prone to misinterpretation.



End of Slides