Valid sequential inference on probability forecast performance

Johanna Ziegel

joint work with Alexander Henzi

University of Bern

INPUT research meeting

December 16, 2022

b
UNIVERSITÄT
BERN

Introduction

Comparing probability forecast performance

We want to know if r_t or q_t is a better forecast for a binary outcome $Y_{t+h} \in \{0, 1\}$.

Leading example

- Probability of rain in h days

Classical approach

Take a *proper scoring rule*, $S = S(\text{forecast}, \text{observation})$,

$$\mathbb{E}_\pi[S(\pi, Y)] \leq \mathbb{E}_\pi[S(q, Y)] \text{ for all } \pi, q \in [0, 1].$$

Here, $\mathbb{P}(Y = 1) = \pi$.

Examples

- Brier score $S(q, y) = (q - y)^2$
- Logarithmic score $S(q, y) = -y \log(q) - (1 - y) \log(1 - q)$

Interest could be in rejecting the null hypothesis

$$\mathcal{H}_0 : \quad \mathbb{E}_{\mathbb{P}}[S(r_t, Y_{t+h}) - S(q_t, Y_{t+h}) \mid \mathcal{F}_t] \leq 0 \text{ for all } t. \qquad (1)$$

In words:

> *"Given the information at the time of forecasting, $\mathcal{F}_t$, forecast r_t achieves a lower S-error than q_t, for all t."*

Mathematically:

- $(\mathcal{F}_t)_t$ is a filtration, for example, $\mathcal{F}_t = \sigma(r_s, q_s, Y_s; s \leq t)$.
- $(r_t, q_t, Y_t)_t$ is adapted.
- Null hypothesis $\mathcal{H}_0$ is set of all distributions $\mathbb{P}$ satisfying (1).

With data $(r_t, q_t, Y_{t+h})_{t=1,\dots,T}$, check if q_t attains a smaller error,

$$\frac{1}{T}\sum_{t=1}^{T} S(q_t, Y_{t+h}) \overset{?}{\leq} \frac{1}{T}\sum_{t=1}^{T} S(r_t, Y_{t+h}).$$

Is q_t *significantly better* than r_t? Compute p-value for $\mathcal{H}_0$.

$$d_T = \frac{1}{T}\sum_{t=1}^{T} \{S(r_t, Y_{t+h}) - S(q_t, Y_{t+h})\}, \ \frac{d_T}{\hat{\sigma}_T/\sqrt{T}} \sim_{\text{approx.}} \mathcal{N}(0,1)$$

(asymptotic) p-value:

$$p_T = 1 - \Phi\Big(\frac{d_T}{\hat{\sigma}_T/\sqrt{T}}\Big)$$

Diebold and Mariano (1995); Giacomini and White (2006)

Anytime valid inference

Why do we need another significance test for score differences?

Anytime valid inference

Why do we need another significance test for score differences?

Often, we have seen parts of the data before we test score differences.

- Ongoing comparisons (waiting for new data to come in)
- Data augmentation (years 2015, 2016, maybe later 2017, 2018)
- Monitoring of calibration

Data arrives *sequentially*, but the sample size T must be *fixed* in advance for (asymptotic) validity of p_T!

Is this problem really severe? Yes.

Simulation example: Let $\mu \in (0,1)$.

$$r_t, q_t \overset{\text{i.i.d.}}{\sim} \text{UNIF}(0,1), \quad \pi_t = \mu q_t + (1-\mu) r_t,$$
$$\mathbb{P}(Y_{t+1} = 1 \mid r_t, q_t) = \pi_t.$$

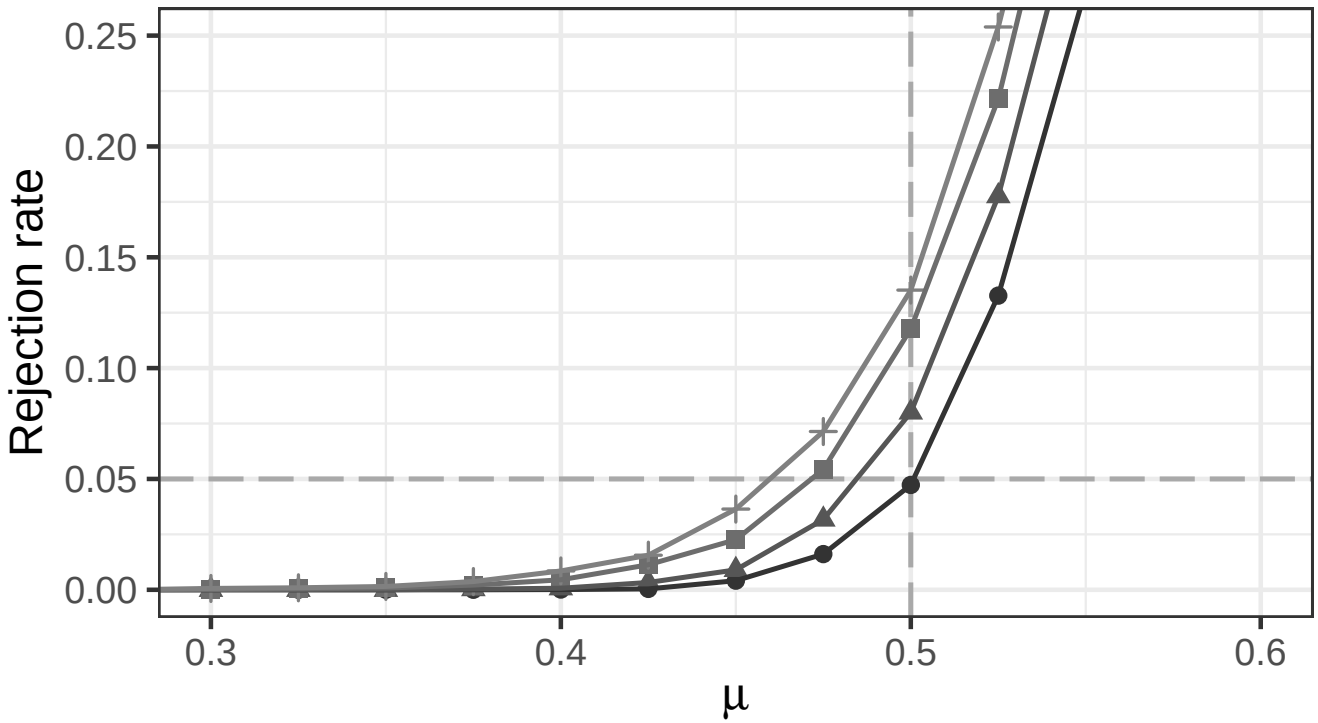
- Forecasters r_t, q_t only have partial information.
- Forecasters are not calibrated, that is

$$\mathbb{P}(Y_{t+1} = 1 \mid r_t) \neq r_t \quad \mathbb{P}(Y_{t+1} = 1 \mid q_t) \neq q_t.$$

- Use Brier score for comparison: r_t better than q_t if and only if

$$\pi_t \in \begin{cases} [0, (r_t + q_t)/2], & \text{if } r_t < q_t, \\ [(r_t + q_t)/2, 1], & \text{if } r_t > q_t, \end{cases} \quad \text{if and only if } \mu \leq 0.5.$$

Rejection rates

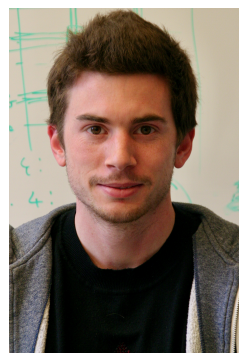


Rejection rates for Student's t-test for the hypothesis that r_t dominates q_t with respect to the Brier score with optional stopping after 1, 3, 5 equispaced time points (triangles, squares, crosses) and without stopping (dots). Sample size is $T = 600$.

Goal

For probability forecasts for binary events $Y_t \in \{0, 1\}$, one can construct forecast dominance tests which ...

- ... are valid in finite samples,
- ... do not require any assumptions on the data generating processes,
- ... anytime valid, that is, optional stopping is allowed.



Henzi and Ziegel (2022)

Anytime valid inference

Anytime validity

Let $\mathcal{H}_0$ be a null hypothesis concerning an adapted process $(X_t)_t$.

Example

$$\mathcal{H}_0 : \quad \mathbb{E}_{\mathbb{P}}(S(r_t, Y_{t+h}) - (S(q_t, Y_{t+h}) \mid \mathcal{F}_t) \leq 0 \text{ for all } t$$

Anytime validity

Let $\mathcal{H}_0$ be a null hypothesis concerning an adapted process $(X_t)_t$.

Example

$$\mathcal{H}_0 : \quad \mathbb{E}_{\mathbb{P}}(S(r_t, Y_{t+h}) - (S(q_t, Y_{t+h}) \mid \mathcal{F}_t) \leq 0 \text{ for all } t$$

An adapted sequence $(p_t)_t$ with values in $[0, 1]$ is

- a 'standard' sequence of p-values for $\mathcal{H}_0$ if for all $\mathbb{P} \in \mathcal{H}_0$,

$$\text{for all } t, \quad \mathbb{P}(p_t \leq \alpha) \leq \alpha;$$

Anytime validity

Let $\mathcal{H}_0$ be a null hypothesis concerning an adapted process $(X_t)_t$.

Example

$$\mathcal{H}_0: \quad \mathbb{E}_{\mathbb{P}}(S(r_t, Y_{t+h}) - (S(q_t, Y_{t+h}) \mid \mathcal{F}_t) \leq 0 \text{ for all } t$$

An adapted sequence $(p_t)_t$ with values in $[0, 1]$ is

- a 'standard' sequence of p-values for $\mathcal{H}_0$ if for all $\mathbb{P} \in \mathcal{H}_0$,

$$\text{for all } t, \quad \mathbb{P}(p_t \leq \alpha) \leq \alpha;$$

- an anytime valid sequence of p-values for $\mathcal{H}_0$ if

$$\mathbb{P}(\exists t : p_t \leq \alpha) \leq \alpha$$

$$\iff \quad \text{for any stopping time } \tau, \quad \mathbb{P}(p_\tau \leq \alpha) \leq \alpha.$$

History: Wald's sequential probability ratio test (SPRT)

$(X_t)_t$ iid

$$\mathcal{H}_0 : X_t \sim P_0 \text{ for all t} \qquad \mathcal{H}_1 : X_t \sim P_1 \text{ for all t}$$

Assume that P_0, P_1 have densities f_0, f_1.

History: Wald's sequential probability ratio test (SPRT)

$(X_t)_t$ iid

$$\mathcal{H}_0 : X_t \sim P_0 \text{ for all t} \qquad \mathcal{H}_1 : X_t \sim P_1 \text{ for all t}$$

Assume that P_0, P_1 have densities f_0, f_1.
Fix test level $\alpha \in (0,1)$ and desired power $1-\beta \in (0,1)$.
Define

$$M_T = \prod_{t=1}^{T} \frac{f_1(X_t)}{f_0(X_t)}.$$

For $T = 1, \ldots$

- If $M_T > (1-\beta)/\alpha$, accept $\mathcal{H}_1$;
- If $M_T < \beta/(1-\alpha)$, accept $\mathcal{H}_0$;
- Otherwise, continue sampling.

History: Wald's sequential probability ratio test (SPRT)

$(X_t)_t$ iid

$$\mathcal{H}_0 : X_t \sim P_0 \text{ for all t} \qquad \mathcal{H}_1 : X_t \sim P_1 \text{ for all t}$$

Assume that P_0, P_1 have densities f_0, f_1.
Fix test level $\alpha \in (0,1)$ and desired power $1-\beta \in (0,1)$.
Define

$$M_T = \prod_{t=1}^{T} \frac{f_1(X_t)}{f_0(X_t)}.$$

For $T = 1, \ldots$

- If $M_T > (1-\beta)/\alpha$, accept $\mathcal{H}_1$;
- If $M_T < \beta/(1-\alpha)$, accept $\mathcal{H}_0$;
- Otherwise, continue sampling.

Observation: $(M_t)_t$ is a non-negative martingale under $\mathcal{H}_0$.

Wald (1945)

Generalizations of Wald's SPRT

- Allow for composite nulls and alternatives
- Tools: Test supermartingales, e-values, e-processes

Recent surge of interest in statistics and machine learning

Carnegie Mellon University

CWI Centrum Wiskunde & Informatica

ims

Safe, Anytime-Valid Inference (SAVI) and Game-theoretic Statistics.
~~May 25-29, 2020 (Covid), Jun 28 to Jul 2, 2021 (Covid)~~ **May 30-Jun 3, 2022 in Eindhoven, Netherlands**

Official workshop page at EURANDOM

A large fraction of published research in top journals in applied sciences such as medicine and psychology has been claimed as irreproducable. In light of this 'replicability crisis', traditional methods for hypothesis testing, most notably those based on p-values, have come under intense scrutiny. One central problem is the following: if our test result is promising but nonconclusive (say, p = 0.07) we cannot simply decide to gather a few more data points. While this practice is ubiquitous in science, it invalidates p-values and error guarantees and makes the results of standard meta-analyses very hard to interpret. This issue is not unique for p-values: other approaches, such as replacing testing by estimation with confidence intervals, suffer from similar optional stopping/continuation problems. Over the last few years several distinct but closely related solutions have been

`https://www.stat.cmu.edu/~aramdas/SAVI/savi20.html`

Betting interpretation of testing

Null hypothesis $\mathcal{H}_0 = \{\mathbb{P}\}$ concerning an adapted process $(X_t)_{t\in\mathbb{N}}$

How can we find evidence against $\mathcal{H}_0$ (if we believe that $\mathbb{Q}$ is true)?

Betting interpretation of testing

Null hypothesis $\mathcal{H}_0 = \{\mathbb{P}\}$ concerning an adapted process $(X_t)_{t\in\mathbb{N}}$

How can we find evidence against $\mathcal{H}_0$ (if we believe that $\mathbb{Q}$ is true)?

Win money by betting against it!

- Start with capital $K_1 = 1$.

Betting interpretation of testing

Null hypothesis $\mathcal{H}_0 = \{\mathbb{P}\}$ concerning an adapted process $(X_t)_{t\in\mathbb{N}}$

How can we find evidence against $\mathcal{H}_0$ (if we believe that $\mathbb{Q}$ is true)?

Win money by betting against it!

- Start with capital $K_1 = 1$.
- At t, invest K_t and bet $E_t = E_t(y) \geq 0$ such that

$$\mathbb{E}_{\mathbb{P}}(E_t(X_{t+1}) \mid \mathcal{F}_t) \leq 1 \quad \text{BUT} \quad \mathbb{E}_{\mathbb{Q}}(E_t(X_{t+1}) \mid \mathcal{F}_t) \gg 1$$

- At $t+1$, receive $K_t E_t(X_{t+1})$ and reinvest for next bet.

Betting interpretation of testing

Null hypothesis $\mathcal{H}_0 = \{\mathbb{P}\}$ concerning an adapted process $(X_t)_{t\in\mathbb{N}}$

How can we find evidence against $\mathcal{H}_0$ (if we believe that $\mathbb{Q}$ is true)?

Win money by betting against it!

- Start with capital $K_1 = 1$.
- At t, invest K_t and bet $E_t = E_t(y) \geq 0$ such that

$$\mathbb{E}_{\mathbb{P}}(E_t(X_{t+1}) \mid \mathcal{F}_t) \leq 1 \quad \text{BUT} \quad \mathbb{E}_{\mathbb{Q}}(E_t(X_{t+1}) \mid \mathcal{F}_t) \gg 1$$

- At $t+1$, receive $K_t E_t(X_{t+1})$ and reinvest for next bet.

Resulting capital process

$$M_T = \prod_{t=1}^{T} E_t(X_{t+1})$$

is a test supermartingale for $\mathcal{H}_0$.

Shafer (2021, JRSSA)

Betting interpretation of testing

Null hypothesis $\mathcal{H}_0 = \{\mathbb{P}\}$ concerning an adapted process $(X_t)_{t\in\mathbb{N}}$

How can we find evidence against $\mathcal{H}_0$ (if we believe that $\mathbb{Q}$ is true)?

Win money by betting against it!

- Start with capital $K_1 = 1$.
- At t, invest K_t and bet e-value $E_t = E_t(y) \geq 0$ such that

$$\mathbb{E}_{\mathbb{P}}(E_t(X_{t+1}) \mid \mathcal{F}_t) \leq 1 \quad \text{BUT} \quad \mathbb{E}_{\mathbb{Q}}(E_t(X_{t+1}) \mid \mathcal{F}_t) \gg 1$$

- At $t+1$, receive $K_t E_t(X_{t+1})$ and reinvest for next bet.

Resulting capital process

$$M_T = \prod_{t=1}^{T} E_t(X_{t+1})$$

is a test supermartingale for $\mathcal{H}_0$.

Shafer (2021, JRSSA)

Test (super)martingales

Let $(M_t)_t$ be a non-negative adapted process with $M_0 = 1$. It is a test supermartingale for $\mathcal{H}_0$ if

$$\mathbb{E}_{\mathbb{P}}(M_{t+1} \mid \mathcal{F}_t) \leq M_t \quad \text{for all } t \text{ and any } \mathbb{P} \in \mathcal{H}_0.$$

Test supermartingales lead to anytime valid sequences of p-values:

Ville's inequality (Ville, 1939)

For any $\mathbb{P} \in \mathcal{H}_0$

$$\mathbb{P}\left(\exists t : \frac{1}{M_t} < \alpha\right) = \mathbb{P}\left(\exists t : M_t > \frac{1}{\alpha}\right) \leq \alpha$$

Example (Wald's SPRT)

Likelihood ratio process $M_T = \prod_{t=1}^{T} \frac{f_1(X_t)}{f_0(X_t)}$ is a test martingale.

Constructing test martingales

E-value

A random variable $E \geq 0$ is an e-value if it has expectation ≤ 1 under the null hypothesis.

- E is a "bet against the null hypothesis"
- Markov's inequality:
  $\mathbb{P}(1/E \leq \alpha) \leq \alpha$ for $\mathbb{P}$ in the null hypothesis.

An adapted sequence $(E_t)_t$ of conditional e-values,

$$\mathbb{E}_{\mathbb{P}}(E_t \mid \mathcal{F}_t) \leq 1 \quad \text{for all } \mathbb{P} \in \mathcal{H}_0,$$

leads to a test supermartingale

$$M_T = \prod_{t=1}^{T} E_t.$$

Power of sequential tests

Test supermartingales yield tests that are anytime valid.

How can we ensure that the test using $(M_t)_t$ also has power against alternatives?

Maximize the growth rate under the alternative

$$\text{maximize} \quad \mathbb{E}_{\mathbb{Q}}\Big[\log M_{t+1} \mid \mathcal{F}_t\Big] \quad \text{for } \mathbb{Q} \in \mathcal{H}_1.$$

Power of sequential tests

Test supermartingales yield tests that are anytime valid.

How can we ensure that the test using $(M_t)_t$ also has power against alternatives?

Maximize the growth rate under the alternative

$$\text{maximize} \quad \mathbb{E}_{\mathbb{Q}}\Big[\log M_{t+1} \mid \mathcal{F}_t\Big] \quad \text{for } \mathbb{Q} \in \mathcal{H}_1.$$

(One) motivation:
If $M_T = \prod_{t=1}^{T} E_t$ with $(E_t)_{t\in\mathbb{N}}$ iid and $\mathbb{E}_{\mathbb{Q}}(\log E_t) \geq L$, then, by the law of large numbers

$$M_T = \exp(TL + o(T)), \quad \text{almost surely, } T \to \infty.$$

Power of sequential tests

Test supermartingales yield tests that are anytime valid.

How can we ensure that the test using $(M_t)_t$ also has power against alternatives?

Maximize the growth rate under the alternative

$$\text{maximize} \quad \mathbb{E}_{\mathbb{Q}}\Big[\log M_{t+1} \mid \mathcal{F}_t\Big] \quad \text{for } \mathbb{Q} \in \mathcal{H}_1.$$

(One) motivation:
If $M_T = \prod_{t=1}^{T} E_t$ with $(E_t)_{t\in\mathbb{N}}$ iid and $\mathbb{E}_{\mathbb{Q}}(\log E_t) \geq L$, then, by the law of large numbers

$$M_T = \exp(TL + o(T)), \quad \text{almost surely, } T \to \infty.$$

- We do not (necessarily) maximize the power of the test.
- Maximize the capital we can make under the alternative.
- Derive empirical analogues of above criterion for optimal betting against the null hypothesis.

probability forecasts for binary events

We want to reject the null hypothesis

$$\mathbb{E}_{\mathbb{P}}[S(r_t, Y_{t+h}) - S(q_t, Y_{t+h}) \mid \mathcal{F}_t] \leq 0 \text{ for all } t.$$

In words:

> *"Given the information at the time of forecasting, $\mathcal{F}_t$, forecast r_t achieves a lower S-error than q_t, for all t."*

Possible extension

Test for forecast dominance under an adapted condition
$c_t \in \{0, 1\}$ (weather regimes, seasons, ...)

$$c_t \mathbb{E}_{\mathbb{P}}[S(r_t, Y_{t+h}) - S(q_t, Y_{t+h}) \mid \mathcal{F}_t] \leq 0 \text{ for all } t.$$

Construction, part I: One period setting

For fixed r, q, test null hypothesis H_S stating that

$$\mathbb{E}_\pi[S(r, Y) - S(q, Y)] \leq 0,$$

that is,

$$H_S = \{\pi \in [0, 1] \mid \mathbb{E}_\pi[S(r, Y) - S(q, Y)] \leq 0\}.$$

Direct computation shows that

$$H_S = \begin{cases} [0, \kappa_S([r, q))], & \text{if } r < q, \\ [\kappa_S([q, r)), 1], & \text{if } r > q \end{cases}$$

with $\kappa_S([a, b))]$ depending on S.

Examples

Brier score: $S(q, y) = (q - y)^2$

- $k_S([r, q)) = (r + q)/2$

Logarithmic score: $S(q, y) = -y \log(q) - (1 - y) \log(1 - q)$

- $k_S([r, q)) = \log\left(\frac{1-r}{1-q}\right) / \log\left(\frac{q(1-r)}{r(1-q)}\right)$

Theorem (Characterization of e-values)

Let S be a scoring function. Under some regularity conditions, $E = E(Y)$ is an e-value with null hypothesis H_S and alternative $[0,1] \backslash H_S$ if and only if for some $\lambda \in [0,1]$,

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

Theorem (Characterization of e-values)

Let S be a scoring function. Under some regularity conditions, $E = E(Y)$ is an e-value with null hypothesis H_S and alternative $[0,1] \backslash H_S$ if and only if for some $\lambda \in [0,1]$,

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

Theorem (Growth rate optimal (GRO) e-values)

For any $\pi \notin H_S$, $\mathbb{E}_\pi(\log(E_{r,q;\lambda}(Y)))$ can be maximized in λ and the resulting e-value is

$$E^\pi_{r,q}(y) = \frac{\pi^y (1-\pi)^{1-y}}{\kappa^y (1-\kappa)^{1-y}}$$

with $\kappa = \kappa_S([\min(r,q), \max(r,q)))$.

Construction, part II: From $t = 1$, to all t

Lag $h = 1$

If $(\lambda_t)_{t\in\mathbb{N}}$ is adapted, then

$$M_T = \prod_{t=1}^{T} E_{r_t,q_t;\lambda_t}(Y_{t+1})$$

is a test supermartingale for $\mathcal{H}_0$. Hence, for any stopping time τ:

$$\mathbb{E}_{\mathbb{P}} M_\tau \leq 1, \quad \mathbb{P} \in \mathcal{H}_0.$$

- No assumptions (like stationarity,. . .)

Lag $h > 1$

Similar combination formula:

$$M_T = \frac{1}{h} \sum_{k=1}^{h} \prod_{\ell \in I_k} E_{r_\ell, q_\ell; \lambda_\ell}(Y_{\ell+h})$$

with $I_k = \{k + hs : s = 0, \ldots, \lfloor (T-k)/h \rfloor - 1\}$ is an e-value for each T, and for any stopping time τ

$$\mathbb{E}_{\mathbb{P}} M_{\tau+h-1} \leq 1, \quad \mathbb{P} \in \mathcal{H}_0.$$

- No assumptions (like stationarity,. . .)

The e-values depend on tuning parameters λ_t.

- *Validity* is guaranteed for *all* choices of λ_t.
- *Power* requires a *good* choice of λ_t.

Inspiration from one-period setting

If $\mathbb{P}(Y = 1) = \pi$ and λ^* is the maximizer of $\mathbb{E}_\pi[\log(E_{r,q;\lambda}(Y))]$, then

$$E_{r,q;\lambda^*}(y) = E^\pi_{r,q} = \frac{\pi^y(1-\pi)^{1-y}}{\kappa^y(1-\kappa)^{1-y}}$$

for $\kappa \in (0,1)$ depending on S, r and q.

- If $S(q,y) = (q-y)^2$, then $\kappa = (r+q)/2$.
- Instead of λ, choose "alternative hypothesis" η as close as possible to π and use $E^\eta_{r,q}$.

Inspiration from one-period setting

If $\mathbb{P}(Y = 1) = \pi$ and λ^* is the maximizer of $\mathbb{E}_\pi[\log(E_{r,q;\lambda}(Y))]$, then

$$E_{r,q;\lambda^*}(y) = E^\pi_{r,q} = \frac{\pi^y(1-\pi)^{1-y}}{\kappa^y(1-\kappa)^{1-y}}$$

for $\kappa \in (0, 1)$ depending on S, r and q.

- If $S(q, y) = (q - y)^2$, then $\kappa = (r + q)/2$.
- Instead of λ, choose "alternative hypothesis" η as close as possible to π and use $E^\eta_{r,q}$.
  - Very confident that q is better than r: $\eta = q$
  - Less bold: $\eta = 0.25r + 0.75q$

Other betting strategies: Waudby-Smith and Ramdas (2023)

Optional stopping

If we want to assess predictive performance at a significance level $\alpha \in (0, 1)$, we gain power by optional stopping:

Lag $h = 1$

$$\tau_\alpha = inf\{t \geq 2 \mid M_t \geq 1/\alpha\}$$

Lag $h \geq 2$

Slightly more complicated ...

$$\tau_{\alpha,h} = \min\left\{T, \inf\left\{t \geq h+1 : M_t \geq \max_{j=t-h+1,\ldots,t-1} E_{r_j,q_j;\lambda_j}(\mathbb{1}\{r_j > q_j\})^{-1}/\alpha\right\}\right\}$$

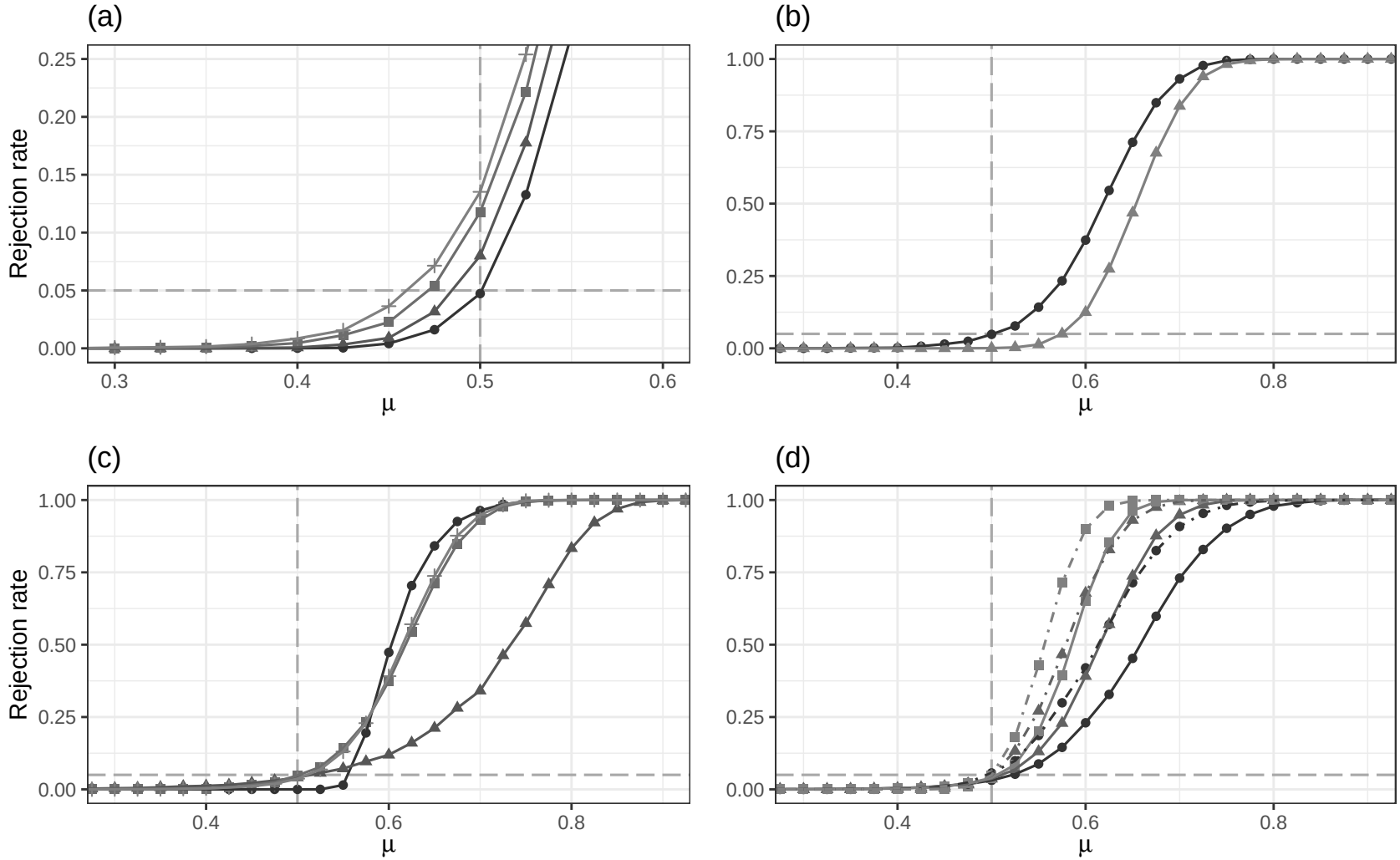
Simulation examples

Simulation example: Let $\mu \in (0, 1)$.

$$r_t, q_t \overset{\text{i.i.d.}}{\sim} \text{UNIF}(0, 1), \quad \pi_t = \mu q_t + (1 - \mu) r_t,$$
$$\mathbb{P}(Y_{t+1} = 1 \mid r_t, q_t) = \pi_t.$$

Betting strategies/alternative hypotheses:

- $\eta_t = q_t$
- $\eta_t = \pi_t$ (oracle necessary)
- $\eta_t = \xi(r_t + q_t)/2 + (1 - \xi)q_t$ for some $\xi \in (0, 1)$.
- Average over k choices of ξ



Rejection rates at 5% level. $T = 600$ for (a)-(c). (b) Stopped and unstopped e-value with $k = 1$. (c) E-values with q_t: triangles, π_t: dots, $k = 1$: crosses, $k = 5$: squares. (d) E-value ($k = 5$; normal lines) and t-test (without stopping; dot-dashed lines) for $T = 300, 600, 1200$.

Data application

Compare postprocessing methods for probability of precipitation forecasts.

- ECMWF ensemble forecasts, airport station observations

  (2007-2017, 50% as training data)
- Heteroscedastic censored logistic regression (HCLR: Messner et al., 2014)
- Isotonic distributional regression (IDR: Henzi et al., 2021)

Hypotheses:

- HCLR outperforms the more generic method IDR
  (Null: IDR achieves a lower Brier score than HCLR)
- IDR outperforms HCLR$_-$
  (Null: HCLR$_-$ achieves a lower Brier score than IDR)
- HCLR outperforms the simpler method HCLR$_-$
  (Null: HCLR$_-$ achieves a lower Brier score than HCLR)

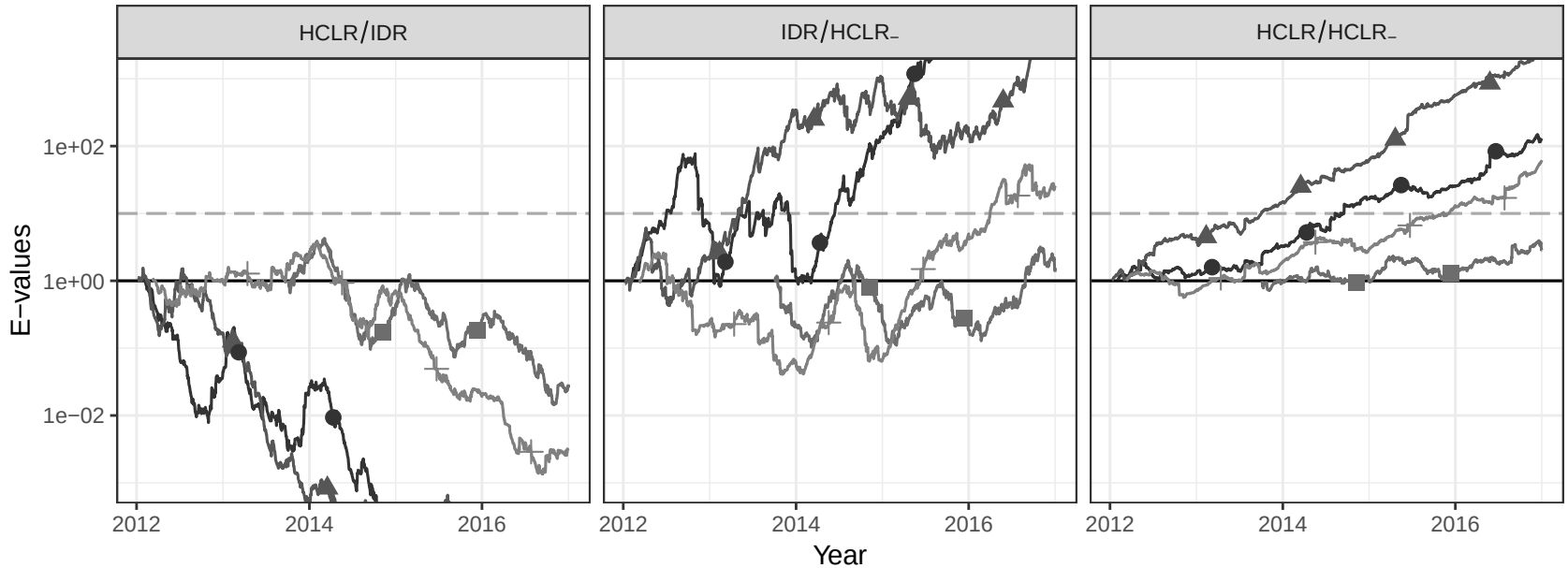
Numerical weather prediction models

- Physical model of the atmosphere is run with current (measured) inital conditions
- Initial conditions are measured with error: Several model runs with slightly perturbed inital conditions yields *ensemble of forecasts*
- Forecast ensembles are interpreted as random draws from the conditional distribution of the outcome
- Ensembles are usually biased and underdispersed: Statistical postprocessing

Bauer et al. (2015)

Table: E-values (alternative probability $\eta = 0.25q_{\mathrm{IDR}} + 0.75q_{\mathrm{HCLR}}, \ldots$) and P-values (one-sided Diebold-Mariano test)

		HCLR/IDR		IDR/HCLR_		HCLR/HCLR_	
	Lag	E	p	E	p	E	p
BRU	1	0	0.9998	> 100	$< 10^{-4}$	> 100	0.0702
	2	0.01	0.9471	> 100	0.0101	13.602	0.0294
	3	0.425	0.4405	> 100	0.1916	15.185	0.0019
	4	4.804	0.0138	1.943	0.9358	5.165	0.0074
	5	16.969	0.0002	0.415	0.9965	3.436	0.0003
FRA	1	0	0.7784	> 100	0.0213	> 100	$< 10^{-4}$
	2	0.054	0.9643	> 100	0.0002	> 100	0.0004
	3	0.078	0.9352	> 100	0.0001	26.569	$< 10^{-4}$
	4	2.291	0.0966	9.618	0.5245	5.54	0.0001
	5	1.526	0.0305	2.362	0.8871	3.227	0.0051



E-values for the hypotheses tests at lag 1 for Brussels (dots), Frankfurt (triangles), London (squares), and Zurich (crosses)

Extensions for probabilistic forecasts (work in progress)

Suppose that Y_{t+h} is a real-valued outcome, and F_t, G_t are predictive cdfs for Y_{t+h}.
Let S be a *proper scoring rule*.

Continuous ranked probability score (CRPS)

$$S(F, Y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}\{Y \leq x\})^2 \, \mathrm{d}x$$

Tests for the null hypothesis

$$\mathbb{E}_{\mathbb{P}}(S(F_t, Y_{t+h}) - S(G_t, Y_{t+h}) \mid \mathcal{F}_t) \leq 0 \text{ for all } t$$

can be constructed with the same method.

- Intuitive and successful betting strategy is not obvious.

Discussion and outlook

- Forecast evaluation is usually sequential: Inference methods should account for this.
- Extensions for comparison of probabilistic forecasts (for real-valued, vector-valued, function-valued, measure-valued outcomes) are possible, work in progress.
- Calibration can also be monitored sequentially (Arnold et al., 2022).
- More principled way for betting would be desirable. Room for improvement?
- Better combination formulae for higher lags?

References

S. Arnold, A. Henzi, and J. F. Ziegel. Sequentially valid tests for forecast calibration. *Annals of Applied Statistics*, 2022. To appear. Preprint available at arXiv:2109.11761.

P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 2015.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13: 253–263, 1995.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74:1545–1578, 2006.

A. Henzi and J. F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, 109:647–663, 2022.

A. Henzi, J. F. Ziegel, and T. Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B*, 85:963–993, 2021.

J. W. Messner, G. J. Mayr, D. S. Wilks, and A. Zeileis. Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, 142:3003–3014, 2014.

G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184:407–431, 2021.

J. Ville. Étude critique de la notion de collectif. *Thèses de l'entre-deux-guerres*, (218), 1939. URL http://www.numdam.org/item?id=THESE_1939__218__1_0.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 2023. To appear as Discussion Paper. Preprint available at arXiv:2010.09686.

Time series example

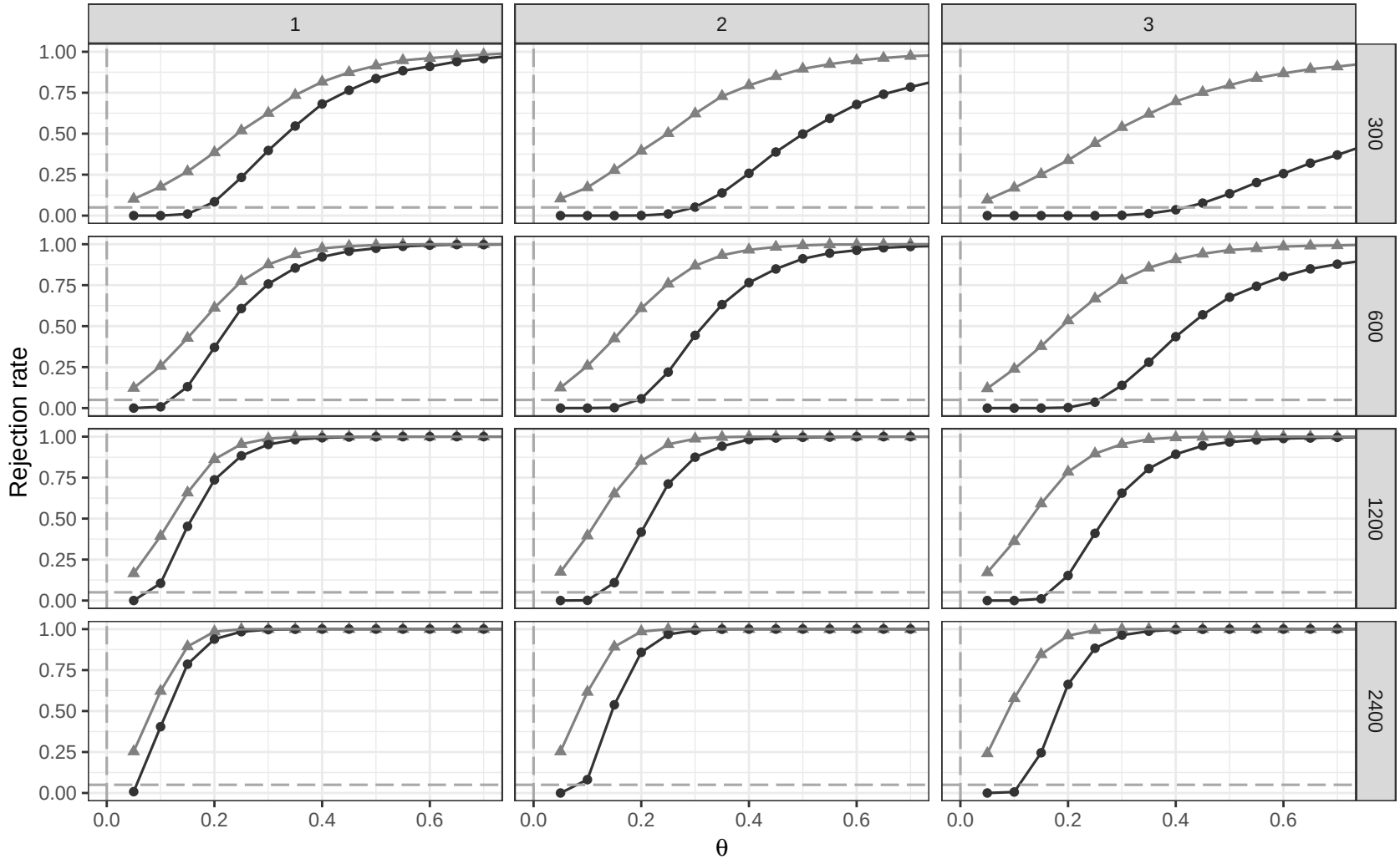
Let $h \in \{1, \ldots, 4\}$, $Z_t = \epsilon_t + \theta \sum_{j=1}^{4} \epsilon_{t-j}$, $\theta > 0$

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

Compare

$$q_t = \pi_{t;h} \quad \text{and} \quad r_t = \pi_{t;h+1}$$

- q_t outperforms r_t.
- For small θ, performance of q_t and r_t is similar.
- Best bet on alternative hypothesis: $\eta_t = q_t$
- Comparison to Diebold-Mariano test



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).

Anytime valid assessment of calibration

Anytime valid assessment of calibration

A probabilistic forecast F_t for $Y_{t+h} \in \mathbb{R}$ should be calibrated.
Most popular notion of calibration: *Flat PIT histogram*, that is,

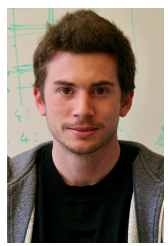
$$F_t(Y_{t+h}) \sim \mathrm{UNIF}(0,1).$$

- Suitable randomization if F_t is not continuous.
- Rank histograms for ensemble forecasts.

Null hypothesis of interest

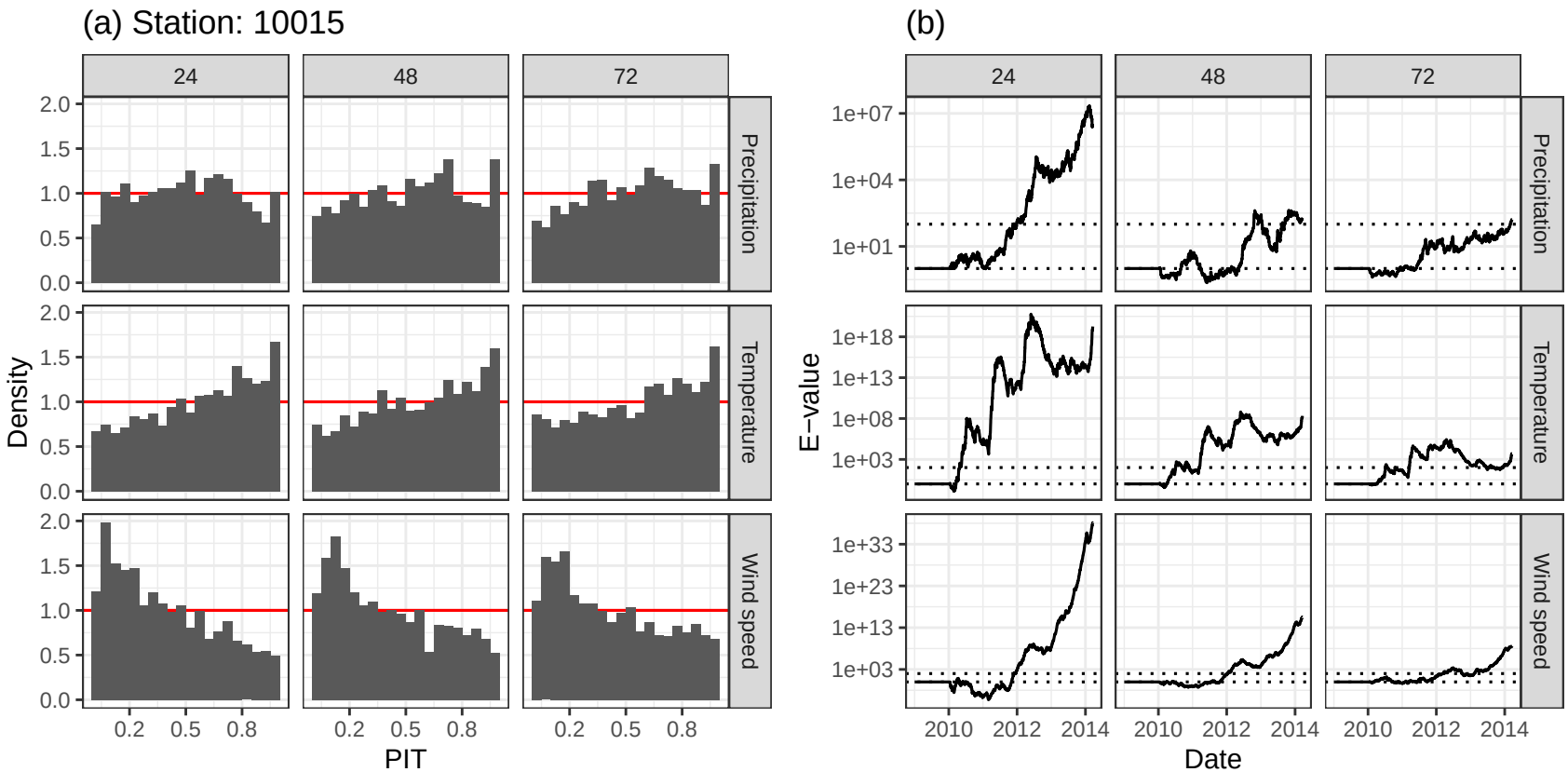
$$\mathcal{H}_0: \quad \mathcal{L}\big(F_t(Y_{t+h}) \mid F_j(Y_{j+h}), j \le t-h\big) = \mathrm{UNIF}(0,1) \quad \text{for all } t$$

- Sequentially valid test based on test martingales available.

Arnold et al. (2022)

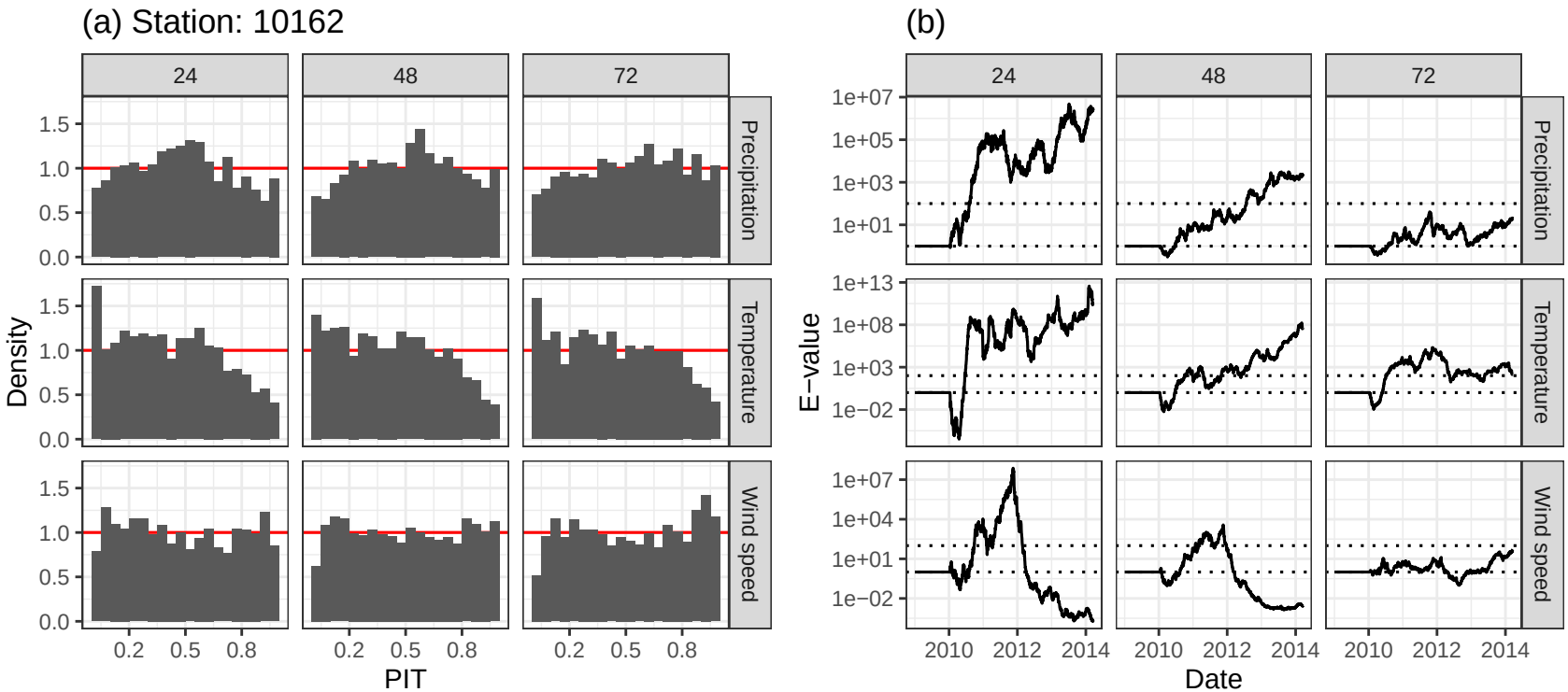
Helgoland



Focus: 24h temperature forecasts

- Not calibrated
- Seasonal pattern in miscalibration

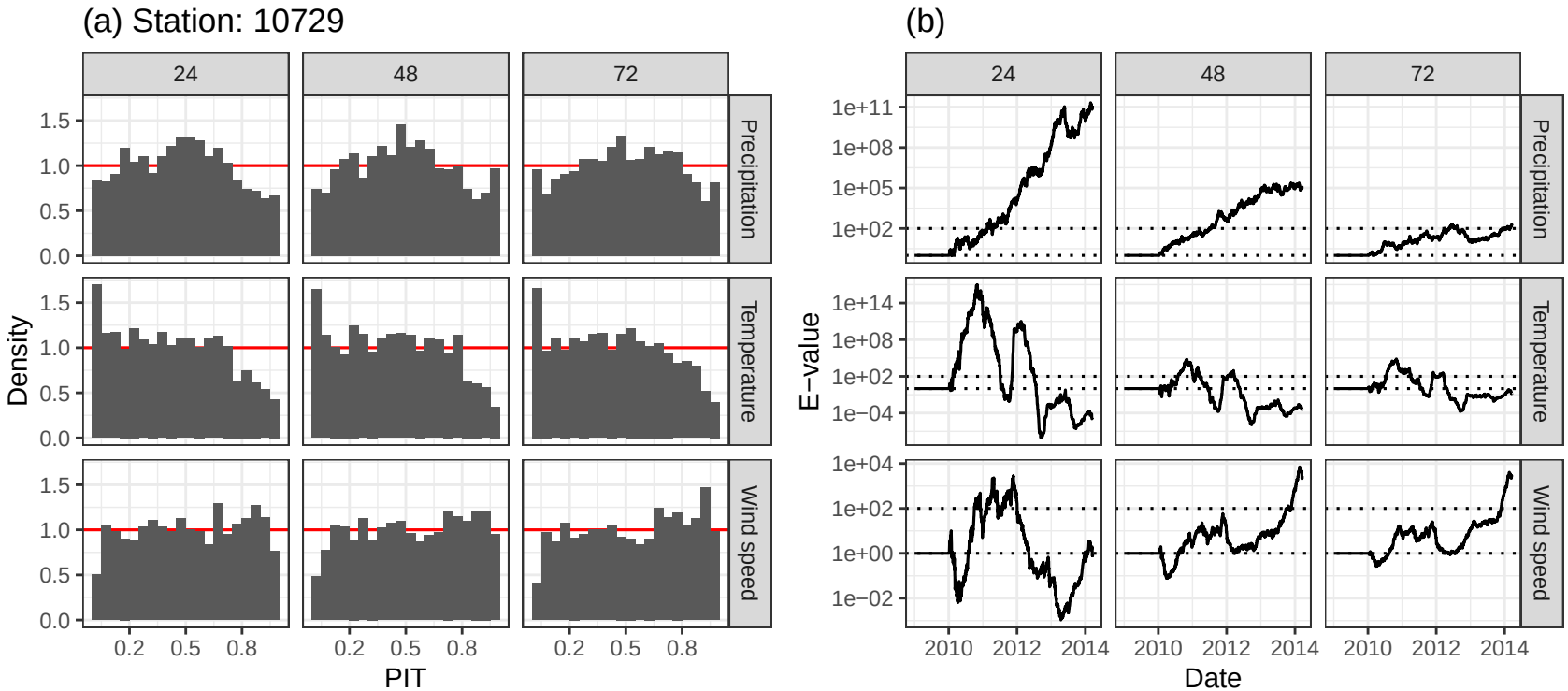
Schwerin



Focus: 24h wind speed forecasts

- PIT histogram relatively flat
- Strong evidence against calibration at the end of 2011
- Time-varying forecast bias

Mannheim



Focus: 48h precipitation forecasts

- Not calibrated
- Steadily growing evidence over time
- Forecasts have consistent dispersion error