# FORECASTING THE FORECASTERS: TRAINING-FREE TWO-PHASE AGGREGATION OF FROZEN TIME-SERIES FOUNDATION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Frozen time-series foundation models differ in accuracy across datasets and over time, making it difficult to decide which forecasts to trust. Combination weights fixed from historical backtests cannot adapt to new outcomes, while deployment-only aggregation has little feedback when few forecasts are scored. Learned routers require additional training. To address these limitations, we introduce TIMEHEDGE, a training-free aggregation procedure that combines unscored historical backtests, called *practice*, with deployment feedback. Practice sets initial weights from pooled and series-specific errors; deployment feedback updates them. The weights combine forecasts at each quantile level, followed by sorting. We extend classical expert-path analysis to this two-phase protocol and ordered quantile levels. An exact identity quantifies how practice changes a deployment regret bound, where regret measures cumulative excess loss relative to a sequence of model choices. Convexity and rearrangement bound forecast pinball loss by weighted member losses. On GIFT-Eval, TIME and fev-bench, TIMEHEDGE outperformed all 22 tested same-pool baselines: eight individual models and 14 selection or aggregation methods. These gains held for both mean absolute scaled error and continuous ranked probability score on all three benchmarks. Ablation studies indicate that practice drives most of the gains: pooled and series-specific evidence each help, and the same protocol benefits other online learners.

## 1 INTRODUCTION

Time-series foundation models (TSFMs) provide forecasts without fitting a new model to each dataset. Their relative performance, however, varies across domains, sampling frequencies and horizons (Aksu et al., 2024). A practitioner with several frozen models must therefore decide which forecasts to trust. Learned routers such as TimeRouter (Ning et al., 2026) address this decision by fitting a routing head. We study how to combine the same forecasts using observed forecasting losses, without training a separate predictor.

Sequential forecasting supplies two sources of feedback. First, we can test the models within the history already available. For example, before forecasting the next day, we can predict earlier days using only the observations preceding each. We call these backtests *practice*: their errors inform the combiner but do not enter the benchmark score. Second, during *deployment*, each new observed outcome supplies feedback for the next forecast. These forecasts do count toward the score. Some benchmarks allow only a few such updates; others allow many.

TIMEHEDGE uses both sources in one deterministic procedure (Figure 1). It first pools practice errors across series sharing a dataset, frequency and horizon, which we call a *configuration*. These errors define initial weights, or a *prior*, over the frozen models. It then adapts the weights using the target series' own practice and updates them as deployment outcomes arrive. The median and the tails can favor different models, so each quantile level has its own weights. Sorting the combined quantiles puts them in the required order. The aggregation settings are fixed before scored evaluation.

The theory asks how practice changes a guarantee on later forecasts. An *expert path* is a sequence of member choices: for example, using model A for two rounds, then model B. *Regret* is the combiner's cumulative loss minus the path's loss; positive values mean worse performance. We extend



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the classical path-based analysis to the two-phase protocol and the quantile axis. Practice changes the deployment bound by an exact amount: evidence in favor of a path improves its bound; misleading evidence can worsen it. This result allows deployment losses to change arbitrarily over time (§5). The bound uses normalized, capped member losses; we audit their connection to the reported quantile scores. Forecast weights also give member contributions and regret bounds (§6.4).

Our experiments separate the value of practice information from the choice of weight-update rule. TIMEHEDGE leads every member and every tested same-pool rule on GIFT-Eval, TIME and fev-bench (Table 1); all rule comparisons are significant across the 293 configurations combined. A two-by-two ablation shows that pooled and target-series practice each recover most of the gain over using neither, and using both improves further (Table 3). Changing the online learner while keeping this practice protocol has a smaller effect (Table 9). The protocol improves all five tested learners in both metrics; four gains remain significant after merging shared sources and correcting for multiple comparisons (Appendix F.3). TIMEHEDGE also leads this pooled comparison.

- We propose TIMEHEDGE, a two-phase method combining practice and deployment feedback without training its members or a router. Expert paths let the favored model change; separate quantile weights let the center and tails favor different models (§3, §4).
- We extend expert-path analysis to the two-phase protocol and quantile axis. We derive the exact practice correction and audit the link between member and forecast losses (§5, §6.4).
- We show that TIMEHEDGE leads all tested same-pool rules across three benchmarks. Controlled ablations separate the two practice sources and show that their benefit transfers to other online learners (§6).

## 2 Related work

**Forecast combination.** Combining forecasts is classical (Bates & Granger, 1969; Timmermann, 2006). Simple averages can outperform estimated weights, motivating work on shrinkage (Liu et al., 2024). Quantile-space pooling, also called Vincentization (Vincent, 1912; Lichtendahl Jr. et al., 2013), and batch Wasserstein-barycentric aggregation (Androulakis et al., 2025) are established approaches. We study frozen foundation-model forecasts and separate the value of practice information from the choice of aggregation rule.

**Online learning and expert tracking.** Online aggregation updates forecast weights after each observed outcome. Exponential weights (Cesa-Bianchi & Lugosi, 2006), Fixed-Share tracking (Herbster & Warmuth, 1998) and switch-rate mixtures (Koolen & de Rooij, 2013) provide the components of our deployment learner. Adaptive rates and Follow-the-Leader (FTL) (de Rooij et al., 2014; Koolen & van Erven, 2015) and smoothed-adversary regret analyses (Haghtalab et al., 2020) address related questions. We extend the expert-path analysis in two directions: practice before scoring begins, whose effect on the deployment bound is an exact identity (§5), and the ordered quantile levels, across which a path can also be defined (Proposition 5). Practice can be viewed as potentially unreliable advice (Purohit et al., 2018); a PAC-Bayes risk interpretation (Thiemann et al., 2017) requires the additional assumptions in Appendix A.

**Online aggregation across quantiles.** CRPS-Learning (Berrisch & Ziel, 2023; 2024) applies Bernstein online aggregation at each quantile level with smoothing across levels. TIMEHEDGE instead processes pooled and target-series practice before deployment. We measure deployment regret against member sequences and test both the complete protocols and learner substitutions under the same practice information (§6).

**Foundation models and routers.** Our members are public frozen checkpoints, listed with their sources and inference settings in Appendix C. TimeRouter (Ning et al., 2026) fits a routing head over such models, while large language model (LLM) agents can select or post-process forecasts. TIMEHEDGE obtains weights from observed member losses without fitting a router. We compare these approaches on the same member forecasts and practice labels.

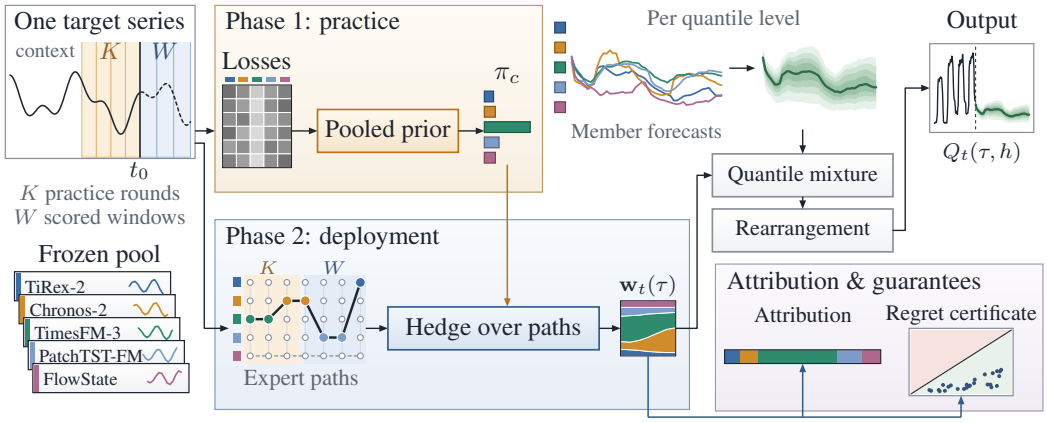


Figure 1: TIMEHEDGE combines frozen forecasts using past errors. Phase 1 pools practice errors across series in a configuration to set a prior over members. Phase 2 processes the target series' own $K$ practice rounds, then updates the weights as outcomes from the $W$ scored windows arrive. Each quantile level has its own weights; sorting the combined quantiles restores their order. The weights also describe member contributions and support regret bounds (§5). No model or router is trained, and no forecast uses an unrevealed scored target.

## 3 THE TWO-PHASE FORECASTING PROTOCOL

**Forecasts and scores.** A configuration contains series $s = 1, \ldots, S$ from one dataset, with sampling frequency $f$ and forecast horizon $H$. Each series has $W \geq 1$ scored forecast windows. At each forecast time, or *origin*, a frozen member $m$ maps the observed history $x_{1:t}$ to quantiles $Q_m(\tau, h) \in \mathbb{R}$, for horizon steps $h \leq H$ and nine levels $\tau \in \mathcal{T}$. For example, $\tau = 0.5$ predicts the median and $\tau = 0.9$ the 90th percentile. We call this collection a *quantile field*. The combiner returns the same format. MASE (mean absolute scaled error) evaluates the median; the benchmark's quantile score evaluates the full grid. Weight updates use the pinball loss, which penalizes underprediction and overprediction according to the quantile level. For prediction $q$ and outcome $y$, $\text{pinball}_\tau(q, y) = (\tau - \mathbf{1}\{y < q\})(y - q)$.

**Two sources of feedback.** Practice and deployment both update the weights, but only deployment forecasts enter the score. *Practice rounds* (L1) are $K$ backtests entirely inside the observed history. For a first scored origin $t_0$, they are $(x_{1:t_0-kH} \to x_{t_0-kH+1:t_0-(k-1)H})$, $k = K, \ldots, 1$. Every member predicts each held-out window using only the history before it, so all practice losses are available before scoring begins. *Deployment rounds* (L2) are the $W$ scored windows. Each forecast is issued before its target is revealed; after the reveal, its error is available for the next update. Origins advance by $H$ observations, and the outcome distribution may change over time.

This protocol covers both practice-dominated and update-rich settings. GIFT-Eval uses $K = 6$ and offers at most three scored windows for half its configurations ($W = 1$ for 17 of 97). A rolling benchmark can instead offer many deployment updates ($W \sim 20$), even with little or no practice. At every decision, only the configuration's practice and outcomes already revealed are available.

**Two uses of practice.** Pooling errors across a configuration supplies the prior; processing the target series' own errors adapts it to that series. The target's practice rounds enter both operations. The deployment analysis conditions on the resulting prior. Interpreting practice as a statistical estimate of future risk requires additional independence assumptions (Appendix A).

**Why retain several members?** The practice winner can change after deployment. On GIFT-Eval, the practice leader is the best member of the scored window for only 0.18 of series (uniform-choice reference 0.125 for eight members). This motivates keeping alternatives to the practice leader; §6 compares combining and selecting members.

# 4 TIMEHEDGE: TRAINING-FREE TWO-PHASE AGGREGATION

TIMEHEDGE must choose among frozen forecasts from limited feedback, while allowing preferences to vary across series, quantiles and time. The four steps below address these requirements (Figure 1; pseudocode in Appendix D). We include the members' equal-weight average as an additional expert, the *closure member*, so ordinary averaging remains available. Here and in §5, $M$ counts active experts including it.

**(1) Set initial weights from pooled practice.** A few backtests may not distinguish members reliably on one series. Pooling practice across a configuration supplies shared evidence before scoring. To make errors comparable across series, we divide each (series, round) loss vector $r$, averaged over quantile levels, by its median across members. From the resulting vectors $\tilde{r}$, we form $\pi_c(m) \propto \exp(-\eta_0 \sum_{\text{rows}} \tilde{r}(m))$ over valid rows, with $\eta_0$ given by Algorithm 3. This prior favors members with smaller relative practice errors and uses no scored outcomes.

**(2) Adapt to the target series before scoring.** A pooled ranking can hide a member's strengths on one series or quantile. We therefore process the target's own $K$ practice losses in time order at each level. The learner updates probabilities over expert paths, giving the *practice posterior*. A member's first scored weight sums the probabilities of paths using it on that round. Theorem 1 accounts for this adaptation in the deployment guarantee.

**(3) Update weights as outcomes arrive.** Practice rankings can persist or change during deployment. We therefore keep updating from newly revealed outcomes and retain both static and switching member sequences. A path $\sigma : \{1, \dots, T\} \to \{1, \dots, M\}$ selects one member $\sigma_t$ per round, with $T = K + W$. Its prior $\pi(\sigma)$ mixes two transition rules, or *kernels*: static paths retain one member; Herbster–Warmuth switching paths (Herbster & Warmuth, 1998) redistribute a fraction $1/T$ of member weight uniformly between rounds. To keep alternatives represented when pooled practice strongly favors a few members, each path type starts from either $\pi_c$ or uniform weights, giving four equally weighted components.

The update favors sequences that have accumulated smaller errors: a path with cumulative observed loss $L$ receives weight proportional to $\pi(\sigma)e^{-\eta L}$. A forward recursion computes next-round member weights without enumerating paths (Algorithm 2). The same recursion updates the component weights (Koolen & de Rooij, 2013); Corollary 1 bounds the cost of the mixture. This update is used throughout practice and deployment.

To compare relative accuracy across rounds and limit the influence of extreme errors, we subtract the minimum member loss, divide by the median absolute member loss with a numerical floor, and cap at $b = 4$. We fix $\eta = \sqrt{8 \ln M/T}$ before scoring and use it throughout. The bound uses these losses; Appendix A.9 audits their link to forecast scores.

**(4) Combine quantiles and restore their order.** Separate weights $w_m(\tau)$ let different members predict the center and tails, but can place a predicted 90th percentile below the median. To preserve this flexibility while producing ordered quantiles, we first form $Q(\tau, h) = \sum_m w_m(\tau) Q_m(\tau, h)$, then sort across levels at each horizon step $h$. This *monotone rearrangement* cannot increase the total grid pinball loss (Proposition 2). With weights shared across levels, quantile averaging gives a 2-Wasserstein ($W_2$) barycenter (a distributional average); separate weights require the ordering step.

**Inference settings and cost.** Members' multivariate feeding and input resampling modes are chosen by configuration-level practice loss (Appendix C). Mixing quantiles costs $O(MWH|\mathcal{T}|)$ per series. Replaying the weight recursion at each origin costs $O(n_\kappa W(K+W)M|\mathcal{T}|)$, with $n_\kappa = 4$ components. Both are small relative to member inference. In the measured setting, one member forecast takes tens to hundreds of milliseconds (Appendix C). The returned weights also describe each member's contribution (Appendix B).

# 5 THEORY

The analysis answers two questions: how practice changes the deployment regret bound, and how that bound relates to the combined quantile forecasts. At one quantile level, let $\ell_t(m) \in [0, b]$ be member $m$'s capped, normalized pinball loss and $w_t$ the weights used to forecast ($b = 4$). The member set is fixed and $\eta$ uses the planned $T = K + W$ rounds. Practice and scored sums below use local round indices. Proofs are in Appendix A.

## 5.1 HOW PRACTICE CHANGES THE DEPLOYMENT BOUND

Regret compares the combiner with any chosen member sequence, called the *comparator path*. It may be chosen *in hindsight*, after observing all outcomes. The combiner's member-weighted loss is $\langle w_t, \ell_t \rangle = \sum_m w_t(m)\ell_t(m)$. A path with prior probability $\pi(\sigma)$ has *code length* $\ln(1/\pi(\sigma))$: the less initial weight a path receives, the larger its contribution to the bound. We first state the standard bound without practice, then show exactly what practice changes.

**Proposition 1** (Deployment code-length regret; standard)**.** *Exponential weights over expert paths with the kernel-mixture prior $\pi$ over $n_\kappa$ components ($n_\kappa = 4$: static and switching kernels, each started from $\pi_c$ or from the uniform distribution) guarantees, for every path $\sigma$,*

$$\sum_{t \le W} \langle w_t, \ell_t \rangle - \sum_{t \le W} \ell_t(\sigma_t) \;\le\; \frac{\ln(1/\pi(\sigma))}{\eta} + \frac{\eta W b^2}{8}.$$

*The bound is worst-case. At the registered rate $\eta = \sqrt{8 \ln M/T}$, $T = K+W$, the second term is at most $b^2\sqrt{T \ln M/8}$ ($W \le T$); the first term is the comparator's path-dependent code length.*

Proposition 1 is the classical mixture-prior bound (Vovk, 1990; Herbster & Warmuth, 1998; Koolen & de Rooij, 2013). Our extension adds the practice phase, followed by the quantile axis in Proposition 2. The next theorem compares a path's practice loss $L_K(\sigma)$ with the prior's exponential-average practice score $M_K$, called the *mix loss*.

**Theorem 1** (Practice as prior)**.** *Run the path learner with prior $\pi$ through the $K$ unscored practice rounds and then the $W$ scored rounds. Write $L_K(\sigma) = \sum_{k \le K} \ell_k(\sigma_k)$ for a path's practice loss and $M_K = -\eta^{-1} \ln \sum_{\sigma'} \pi(\sigma') e^{-\eta L_K(\sigma')}$ for the prior's mix loss on the practice rounds. Then for every path $\sigma$, the regret on the scored rounds satisfies*

$$\textstyle\sum_{t \le W} \langle w_t, \ell_t \rangle - \sum_{t \le W} \ell_t(\sigma_t) \;\le\; \underbrace{\frac{\ln(1/\pi(\sigma))}{\eta}}_{\text{initial path cost}} + \underbrace{\big(L_K(\sigma) - M_K\big)}_{\text{practice correction}} + \underbrace{\frac{\eta W b^2}{8}}_{\text{deployment term}}.$$

*The correction is the comparator's practice loss minus the prior mix loss. It shortens the code length when the path outperformed the mixture on the unscored practice rounds. Its upper bound is the path's excess practice loss over the per-round best member, $\sum_{k \le K} \big(\ell_k(\sigma_k) - \min_m \ell_k(m)\big) \le Kb$.*

The middle term makes the effect of practice explicit. A path that did well in practice receives a smaller deployment bound; a path that did poorly receives a larger one, relative to the reference $M_K$. With $K = 0$ the correction vanishes. The change is exact: the practice posterior is $\rho_K(\sigma) \propto \pi(\sigma)e^{-\eta L_K(\sigma)}$, so $-\ln \rho_K(\sigma) = -\ln \pi(\sigma) + \eta(L_K(\sigma) - M_K)$. Summing over possible practice paths gives a *regret certificate*, a bound computed from observed losses and weights (Appendix A.9). Proposition 4 shows when the bound is approached. Switching allows a path to leave the practice leader at a cost included in the bound.

The corollary explains why we keep both pooled and uniform starts. Pooled starts favor members supported by practice; uniform starts give every member an initial route into the comparison.

**Corollary 1** (Consistency and robustness)**.** *Theorem 1 can be read through any component of the mixture, the correction $L_K(\sigma) - M_K$ (against the whole mixture) being the same. Through a uniform-start component every static comparator $m$ has scored-round regret at most $(\ln M + \ln n_\kappa)/\eta + \big(L_K(m) - M_K\big) + \eta W b^2/8$, the correction at most $m$'s excess practice loss over the per-round best member. The static-path code-length term is independent of $\pi_c$ and the practice correction is retained (robustness). Through the static $\pi_c$-start component, a favored member $m$ keeps its shorter code length $\ln(1/\pi_c(m))$ (consistency). Carrying both costs $\ln n_\kappa/\eta$, with the uniform-start alternatives accounting for $\ln 2/\eta$.*

### 5.2 FROM MEMBER LOSSES TO QUANTILE FORECASTS

The learner averages member losses, whereas the benchmark scores one combined forecast. Convexity connects the two; sorting cannot increase the combined forecast's total quantile loss.

**Proposition 2** (Quantile aggregation and rearrangement)**.** *(i)* Common weights. *In one dimension, for weights $w$ shared across levels, $\sum_m w_m Q_m$ is the Fréchet mean of the member distributions under the 2-Wasserstein metric, whereas mixing cumulative distribution functions (CDFs) adds between-member variance; with level-dependent weights $w_m(\tau)$ the combination is a convex combination of quantile functions at each level, a valid quantile function after (iii).*

*(ii)* Forecast loss. *For a forecast CDF $F$ with quantile function $Q$, $\mathrm{CRPS}(F, y) = 2\int_0^1 \mathrm{pinball}_\tau(Q(\tau), y)\, d\tau$; on the benchmark's equally weighted grid, convexity bounds the combined pinball loss by the member-weighted loss and sorting cannot increase the sum across levels (Appendix A.9). Clipping and scale corrections relate the capped normalized regret to the raw grid score; Appendix A.9 reports these measured corrections.*

*(iii)* Ordering. *The monotone rearrangement also weakly reduces the $L_p$ distance to every monotone target, $p \geq 1$ (Chernozhukov et al., 2010), and preserves the set of level values.*

Paths can also follow members across increasing quantile levels (Proposition 5). The deployment bound holds for any sequence of normalized, capped losses; the statistical interpretation of practice additionally requires independence (Appendix A).

## 6 EXPERIMENTS

### 6.1 SETUP

**Benchmarks and metrics.** **GIFT-Eval** (Aksu et al., 2024) has 97 configurations (dataset $\times$ frequency $\times$ horizon). **TIME** (Qiao et al., 2026) has 98 configurations with one to 186 scored windows per series (twelve at the median configuration). **fev-bench** (Shchur et al., 2025) has 100 tasks across seven domains; our members ignored its covariates. We used the same pool and pipeline throughout. Normalized MASE measures median-forecast error; scaled CRPS (continuous ranked probability score) evaluates the quantile forecasts. Lower is better for both. We aggregated configuration scores by their geometric mean. For GIFT-Eval we also obtained scores from the official `gluonts` harness and compared them with our engine (Appendix C).

**Pools.** Each pool rule was declared before its evaluation; the inclusive rule became primary after both were scored (Appendix E). The zero-shot pool (8 members) spanned five open model lineages: TiRex, TimesFM, PatchTST-FM, Chronos-2 and FlowState. It admitted every public checkpoint of those lineages whose own published entry is typed zero-shot (Table 5). The extended pool (TIMEHEDGE-P, 14 member slots) added the pretrained-category families Chronos-2, Moirai-2, Toto-2.0 and Timer-S1, and a second TimesFM-2.5 configuration with a shorter context. Member inference settings mirrored each checkpoint's published configuration (parity audit in Appendix C), and no zero-shot member was trained on a GIFT-Eval train split.

**Available information.** We preregistered every scored run (Appendix E). Weights, inference modes and member choices used only outcomes already revealed. Routers used practice labels for training (Table 1).

### 6.2 MAIN COMPARISON: ONE POOL, EVERY METHOD FAMILY

Table 1 compares complete aggregation protocols over the same frozen pool, with each protocol's use of practice stated in the caption. TIMEHEDGE had the lowest aggregate MASE and CRPS on all three benchmarks, ahead of every member and every rule from five families, including TimeRouter's own pipeline. Across the 293 configurations combined, every rule was behind it ($p<0.001$). These comparisons survived Holm and Bonferroni corrections with configurations clustered by dataset or source (Appendix F). Direct runs of the authors' BOA/BEWA package, with practice and online selection over 18 settings, also favored TIMEHEDGE on all three benchmarks (Table 11).

Table 1: Same-pool comparison on GIFT-Eval (97 configurations), TIME (96 comparable settings), and fev-bench (100). Online baselines started uniformly and updated on scored reveals; static rules and routers used practice as stated. Geometric-mean normalized MASE and scaled CRPS; among displayed rows, best bold and second best underlined. Shared-prior learner substitutions are ablations (Table 9). BOA + fixed smoothing is our restricted implementation (Appendix F.4). †TimeRouter used its own pipeline trained on the same practice labels.

| | GIFT-Eval | | TIME | | fev-bench | |
|---|---|---|---|---|---|---|
| Method | MASE↓ | CRPS↓ | MASE↓ | CRPS↓ | MASE↓ | CRPS↓ |
| *Pool members, used alone* | | | | | | |
| TimesFM-3 (Google Research, 2026) | 0.6689 | 0.7553 | 0.6710 | 0.6831 | 0.6688 | 0.8399 |
| PatchTST-FM-r2 (IBM Research, 2026) | 0.6849 | 0.7721 | 0.6829 | 0.6960 | 0.6956 | 0.8707 |
| TiRex-2 (NX-AI, 2026) | 0.6976 | 0.7886 | 0.7009 | 0.7158 | 0.6658 | 0.8357 |
| FlowState-r1.1 (Graf et al., 2025) | 0.7019 | 0.8052 | 0.7229 | 0.7476 | 0.7083 | 0.9016 |
| TimesFM-2.5 (Das et al., 2024) | 0.7056 | 0.8016 | 0.7055 | 0.7248 | 0.6571 | 0.8326 |
| PatchTST-FM (IBM Research, 2026) | 0.7090 | 0.7996 | 0.7082 | 0.7236 | 0.6873 | 0.8664 |
| TiRex-1.1 (Auer et al., 2025) | 0.7097 | 0.8048 | 0.7213 | 0.7367 | 0.6778 | 0.8492 |
| Chronos-2-synth (Ansari et al., 2025) | 0.7243 | 0.8250 | 0.7240 | 0.7432 | 0.7211 | 0.9054 |
| *Simple pooling* | | | | | | |
| Uniform (Clemen, 1989) | 0.6770 | 0.7632 | 0.6745 | 0.6873 | 0.6561 | 0.8215 |
| Median (Stock & Watson, 2004) | 0.6764 | 0.7617 | 0.6725 | 0.6847 | 0.6539 | 0.8186 |
| Best member overall (practice) | 0.7097 | 0.8048 | 0.6710 | 0.6831 | 0.6778 | 0.8492 |
| Best member per config (practice) | <u>0.6543</u> | <u>0.7404</u> | 0.6731 | 0.6860 | 0.6259 | 0.7857 |
| *Static combination weights (fit on practice rounds)* | | | | | | |
| BMA (Hoeting et al., 1999) | 0.6634 | 0.7481 | 0.6700 | 0.6824 | 0.6325 | 0.7917 |
| Stacking (ridge) (Wolpert, 1992) | 0.6559 | 0.7405 | 0.6684 | 0.6804 | 0.6230 | 0.7807 |
| Best-3 mixture (Caruana et al., 2004) | 0.6589 | 0.7430 | <u>0.6675</u> | <u>0.6795</u> | 0.6292 | 0.7882 |
| *Online aggregation (uniform start; scored reveals)* | | | | | | |
| Hedge (Freund & Schapire, 1997) | 0.6719 | 0.7577 | 0.6718 | 0.6842 | 0.6367 | 0.7984 |
| AdaHedge (de Rooij et al., 2014) | 0.6623 | 0.7468 | 0.6701 | 0.6827 | <u>0.6188</u> | <u>0.7753</u> |
| BOA (Wintenberger, 2017) | 0.6826 | 0.7707 | 0.6814 | 0.6964 | 0.6665 | 0.8379 |
| FTL (Cesa-Bianchi & Lugosi, 2006) | 0.6640 | 0.7508 | 0.6761 | 0.6905 | 0.6240 | 0.7840 |
| BOA + fixed smoothing (Berrisch & Ziel, 2023) | 0.6818 | 0.7708 | 0.6811 | 0.6955 | 0.6667 | 0.8374 |
| *Learned routing (fit on practice rounds)* | | | | | | |
| TimeRouter† (Ning et al., 2026) | 0.6640 | 0.7500 | 0.6815 | 0.6951 | <u>0.6188</u> | 0.7790 |
| GBM router (Montero-Manso et al., 2020) | 0.6661 | 0.7548 | 0.6826 | 0.6976 | 0.6297 | 0.7937 |
| **TIMEHEDGE (ours)** | **0.6506** | **0.7329** | **0.6653** | **0.6770** | **0.6092** | **0.7643** |

**Practice information.** On GIFT-Eval, practice-informed static rules scored 0.654–0.663 MASE, compared with 0.662–0.683 for scored-only online learners and 0.676–0.677 for plain pooling. Learned routers trained on the same practice labels scored 0.664–0.666. Selecting one member globally from practice chose TiRex-1.1 (0.710), whereas the best member in hindsight was TimesFM-3 (0.669), which was behind TIMEHEDGE in the pooled comparison ($p<0.001$).

**Closest alternatives.** The strongest alternatives were close on individual benchmarks. Their gaps varied: stacking was behind on GIFT-Eval ($p=0.02$), Best-3 on fev-bench ($p=0.006$), and AdaHedge on GIFT-Eval and TIME ($p \leq 0.02$). Large per-series losses relative to Uniform (CRPS ratio above 1.25) occurred on 0.38%, 0.02% and 0.72% of the series, against 3.64%, 2.75% and 4.81% under the practice pick (Appendix F.9).

**Transfer of the practice protocol.** Giving the five reported online learners the same practice improved their pooled MASE and CRPS relative to their scored-only forms (Table 9), and the improvements survived Holm correction under benchmark–dataset clustering and after merging shared sources for four of the five learners (Appendix F.3). Gains over Uniform and the practice pick survived all five lineage removals on both metrics in every benchmark (Appendix F.7).

**Official-harness result.** Under the official harness, TIMEHEDGE scored 0.6502 and TIMEHEDGE-P 0.6540. Published individual-model scores were 0.6668 (zero-shot, TimesFM-3), 0.6660 (pretrained, Falcon-2.0) and 0.6582 (fine-tuned, LoRA-adapted Chronos-2); TimeRouter scored 0.6746. TIMEHEDGE-P used six additional members (Table 5); it required more inference and scored 0.004 worse than the zero-shot pool.

Table 2: Selected ablations on the 8-member pool, one change per row (metrics: Table 1). Among shown rows, best and second distinct displayed values are bold and underlined; ties share marks. Full form highlighted; constants: Table 16.

| | GIFT-Eval | | TIME | | fev-bench | |
|---|---|---|---|---|---|---|
| Configuration | MASE↓ | CRPS↓ | MASE↓ | CRPS↓ | MASE↓ | CRPS↓ |
| *Pool and prior* | | | | | | |
| w/o pooled practice prior (flat prior) | 0.6554 | 0.7377 | 0.6677 | 0.6795 | <u>0.6106</u> | <u>0.7660</u> |
| one checkpoint per family (5 members) | 0.6560 | 0.7391 | 0.6658 | <u>0.6774</u> | 0.6337 | 0.7924 |
| *Practice rounds* | | | | | | |
| practice selection only, no scored updates | 0.6523 | 0.7385 | 0.6708 | 0.6849 | 0.6136 | 0.7714 |
| w/o practice rounds (scored rounds only) | 0.6551 | 0.7389 | <u>0.6657</u> | 0.6776 | 0.6131 | 0.7686 |
| *Deployment rounds* | | | | | | |
| practice posterior frozen (no scored updates) | <u>0.6514</u> | <u>0.7342</u> | 0.6665 | 0.6784 | 0.6131 | 0.7696 |
| TIMEHEDGE (full, 8 members) | **0.6506** | **0.7329** | **0.6653** | **0.6770** | **0.6092** | **0.7643** |

Table 3: Separating practice information with deployment updates and rates fixed. Engine MASE; best in bold, second best underlined. Full TIMEHEDGE uses both sources (shaded). The permutation control averaged three seeds. Definitions and learner substitutions: Appendix F.5 and Table 9.

| Pooled prior | Own practice | GIFT-Eval ↓ | TIME ↓ | fev-bench ↓ |
|---|---|---|---|---|
| no | no | 0.6697 | 0.6707 | 0.6231 |
| yes | no | <u>0.6551</u> | <u>0.6657</u> | 0.6131 |
| no | yes | 0.6554 | 0.6677 | <u>0.6106</u> |
| yes | permuted labels | 0.6748 | 0.6743 | 0.6326 |
| yes | yes (full) | **0.6506** | **0.6653** | **0.6092** |

### 6.3 Ablations: practice and deployment updates

Table 2 reports changes to one component at a time. Removing the pooled practice prior had the largest single effect on GIFT-Eval and TIME. Keeping a weighted mixture after practice improved on selecting one member on all three benchmarks. Deployment updates helped most on fev-bench.

Table 3 separates pooled from target-series practice. Each source recovered most of the gain over using neither; using both improved further. Randomly permuting the member labels on the target's practice losses was worse than omitting practice altogether (GIFT-Eval $0.6748$ vs. $0.6697$). The information about which member made each error mattered, not just the number of updates.

We also asked when deployment updates helped. Across $155$ configurations, their gain over freezing the practice weights rose with practice–deployment agreement: the fraction of series whose practice leader remains best on the first scored window (Spearman $\rho{=}0.33$, $p{<}10^{-4}$; $+0.07\%$ MASE in the lowest agreement third, $+0.39\%$ in the highest). The gain from mixing rather than selecting after practice fell with agreement ($\rho{=}-0.31$). Within TIME, the number of scored windows was more informative (Figure 3). These were observed associations; Proposition 6 gives a separate sufficient condition for updates to improve on frozen weights in normalized, capped loss.

### 6.4 Reading the weights: attribution with certificates

The weights reveal which members contribute to a forecast (Figure 2; Appendix B). Within a static component, their ratios encode cumulative loss differences exactly (T1). We audited certificate T5 for normalized, capped losses on series with a fixed member set and complete loss history ($69\%$, $99\%$ and $61\%$ of GIFT-Eval, TIME and fev-bench; Appendix F.1). On GIFT-Eval, TimesFM-3 carried $23\%$ of the weight and led on a third of the windows; the equal-weight closure carried $15\%$ (Table 4, Figure 4).

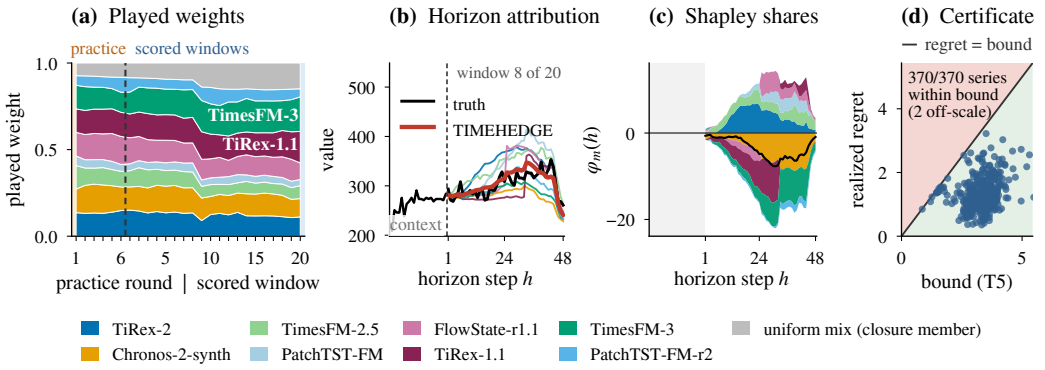


Figure 2: Reading TIMEHEDGE's weights (GIFT-Eval, electricity at 15 minutes). (a) One series' forecast weights: practice concentrated them, and scored outcomes updated them; no member exceeded a fifth. (b) Window 8: the combined median was closest to the truth (MASE 0.40 vs. 0.47 for the best member). (c) Member contributions $\varphi_m(h)$ to the combined median relative to uniform averaging (Shapley values; black: sum). (d) On all 370 series, regret against the hindsight-best member was below its computed bound (certificate T5, normalized capped losses).

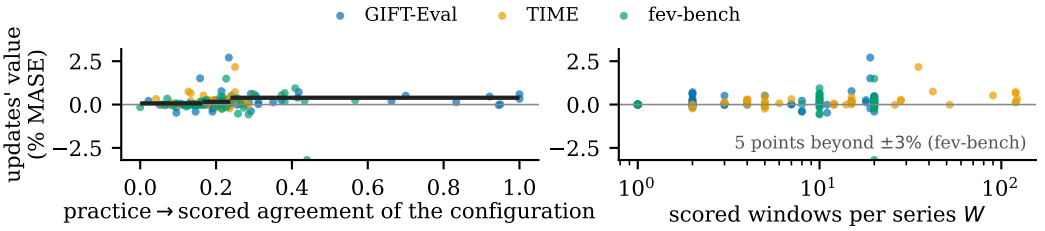


Figure 3: When deployment updates help (155 configurations). MASE improvement over frozen practice weights (positive: updating helped), plotted against agreement with practice (left; ticks: tercile means) and scored windows per series (right).

## 7 Discussion and Conclusion

**What the result says.** Observed forecasting errors can guide a frozen model pool without training a router. TIMEHEDGE led the same-pool comparisons across three benchmarks (Table 1). Both sources of practice recovered most of the gain, and their benefit transferred to other online learners (Tables 3 and 10). The theory accounts exactly for how practice changes the deployment bound, while the forecast weights and computed certificates make the combination inspectable (§5, §6.4).

**Limitations.** Treating practice as an estimate of future performance requires the assumptions in Appendix A; the practice leader remained best in 0.18 of GIFT-Eval series. We inherit the members' licenses and failure modes and use $P(K + W)$ forecasts per series for $P$ checkpoints. Individual member calls took tens to hundreds of milliseconds in the measured setting (Appendix C).

### Reproducibility Statement

All pool members are publicly released checkpoints used frozen; the aggregation is training-free and deterministic given the member forecasts. Appendix C lists every checkpoint, inference setting, and the member-level parity audit against the published benchmark rows; Appendix E describes run registrations and subsequent design choices. Supplementary material contains the anonymized aggregation library, a synthetic example, and tests of the prior, predictive path recursion, quantile ordering and loss certificates.

### AI USE STATEMENT

We used large language model assistants to propose and refine hypotheses, develop theoretical arguments and assist with proof writing, design experiments and synthetic diagnostics, implement methods and data-processing and evaluation code, and interpret results. They also assisted literature searches, translation, manuscript drafting and editing, figure and table preparation, and decisions about which exploratory analyses to include. Their role extended beyond language editing. TIME-HEDGE's forecasting procedure does not call a language model or train its member models.

Reported benchmark metrics were computed from saved forecasts and benchmark targets; synthetic diagnostics are identified separately. Verification included library–engine equivalence tests, saved-weight replays and predictive-loss audits (Appendices C and F). Related-work claims were checked against the cited sources. The authors are responsible for the correctness and originality of the final text, code, results and proofs, including all AI-assisted content.

### ETHICS STATEMENT

We evaluated publicly released benchmarks using frozen public checkpoints, without recruiting participants or collecting new participant data. Combining forecasts does not remove the member models' biases or failure modes; their behavior must be assessed for the intended application.

## REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Emmanouil Androulakis, Georgios I. Papayiannis, and Athanasios N. Yannacopoulos. Barycentric model aggregation in the Wasserstein space of distributions and a variational approach to consistency. *arXiv preprint arXiv:2507.11719*, 2025.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *arXiv preprint arXiv:2505.23719*, 2025. NeurIPS 2025.

John M. Bates and Clive W. J. Granger. The combination of forecasts. *Operational Research Quarterly*, 20(4):451–468, 1969.

Jonathan Berrisch and Florian Ziel. CRPS learning. *Journal of Econometrics*, 237(2):105221, 2023.

Jonathan Berrisch and Florian Ziel. Multivariate probabilistic CRPS learning with an application to day-ahead electricity prices. *International Journal of Forecasting*, 2024.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, 2004.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

Robert T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583, 1989.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

Google Research. TimesFM-3 model checkpoint. Hugging Face model card, `https://huggingface.co/google/timesfm-3.0-pytorch`, 2026.

Lars Graf, Thomas Ortner, Stanisław Woźniak, and Angeliki Pantazi. FlowState: Sampling-rate-equivariant time-series forecasting. *arXiv preprint arXiv:2508.05287*, 2025. ICML 2026.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. In *Advances in Neural Information Processing Systems*, 2020.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401, 1999.

IBM Research. PatchTST-FM: a patch-based time-series foundation model (release r2). Hugging Face model card, `https://huggingface.co/ibm-granite/granite-timeseries-patchtst-fm-r2`, 2026.

Wouter M. Koolen and Steven de Rooij. Universal codes from switching strategies. *IEEE Transactions on Information Theory*, 59(11):7168–7185, 2013.

Wouter M. Koolen and Tim van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, pp. 1155–1175, 2015.

Kenneth C. Lichtendahl Jr., Yael Grushka-Cockayne, and Robert L. Winkler. Is it better to average probabilities or quantiles? *Management Science*, 59(7):1594–1611, 2013.

Li Liu, Xianfeng Hao, and Yudong Wang. Solving the forecast combination puzzle using double shrinkages. *Oxford Bulletin of Economics and Statistics*, 86(3):714–741, 2024.

Pablo Montero-Manso, George Athanasopoulos, Rob J. Hyndman, and Thiyanga S. Talagala. FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, 36 (1):86–92, 2020.

Kanghui Ning, Yushan Jiang, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, and Dongjin Song. TimeRouter: Efficient and adaptive routing of time-series foundation models. *arXiv preprint arXiv:2606.11625*, 2026.

NX-AI. TiRex-2 model checkpoint. Hugging Face model card, `https://huggingface.co/NX-AI/TiRex-2-gifteval-zs`, 2026.

Marco Obermeier, Marco Pruckner, Florian Haselbeck, and Andreas Zeiselmair. FETS benchmark: Foundation models enable scalable and generalizable energy time series forecasting. *Energy and AI*, 26:100867, 2026. doi: 10.1016/j.egyai.2026.100867.

Manish Purohit, Zoya Svitkina, and Ravi Kumar. Improving online algorithms via ML predictions. In *Advances in Neural Information Processing Systems*, 2018.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's TIME: Towards the next generation of time series forecasting benchmarks. *arXiv preprint arXiv:2602.12147*, 2026. ICML 2026.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

James H. Stock and Mark W. Watson. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430, 2004.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *International Conference on Algorithmic Learning Theory*, pp. 466–492, 2017.

Allan Timmermann. Forecast combinations. In *Handbook of Economic Forecasting*, volume 1, pp. 135–196. Elsevier, 2006.

Stella Burnham Vincent. The function of the vibrissae in the behavior of the white rat. *Behavior Monographs*, 1(5), 1912.

Volodimir G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT)*, pp. 371–383, 1990.

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106 (1):119–141, 2017.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

# A Proofs

## A.1 Proposition 3

**Assumption 1** (Practice validity)**.** *The member skill process is approximately stationary across the $K \cdot H$ most recent steps of the context. This is the only place a statistical assumption on the series enters; the deployment guarantee is worst-case.*

The Gibbs posterior $\rho_\lambda(m) \propto \pi(m)\exp(-\lambda K \hat{L}_m/b)$ on the $K$ practice losses, with $L(\rho)$ the expected loss under the skill process in units of $b$, satisfies, for every temperature $\lambda \in (0, 2)$ in a fixed grid $\Lambda$ and with probability $1-\delta$ over $K$ i.i.d. practice loss vectors,

$$L(\rho_\lambda) \;\le\; \underbrace{\frac{\mathbb{E}_{\rho_\lambda}[\hat{L}]/b}{1-\lambda/2}}_{\text{empirical term}} \;+\; \underbrace{\frac{\mathrm{KL}(\rho_\lambda\|\pi) + \ln(2\sqrt{K}/\delta) + \ln|\Lambda|}{K\lambda(1-\lambda/2)}}_{\text{complexity term}}, \tag{1}$$

**Proposition 3** (Practice phase is a risk problem)**.** *Assume the $K$ practice loss vectors of a series are i.i.d. bounded draws from the recent skill process and that the prior $\pi$ is chosen without using that series' practice rounds. Then every Gibbs posterior $\rho_\lambda$ of equation 1 satisfies the PAC-Bayes-$\lambda$ risk bound at its temperature. If the practice leader $m^\star$ is unique with gap $\Delta = \min_{m\neq m^\star}(\hat{L}_m - \hat{L}_{m^\star}) > 0$ and $\pi(m^\star) > 0$, then $1-\rho_\lambda(m^\star) \le \frac{1-\pi(m^\star)}{\pi(m^\star)}\exp(-\lambda K\Delta/b)$: the posterior concentrates on the leader once the practice evidence $\lambda K\Delta/b$ exceeds the prior log-odds, and in the uninformative case ($\hat{L}$ constant across members) the bound is minimized at the prior.*

The PAC-Bayes-$\lambda$ bound of Thiemann et al. (2017) holds for i.i.d. bounded losses and for every posterior, in particular for every Gibbs posterior of the family at a fixed temperature; under the independence assumption the $K$ practice losses of each member are i.i.d. draws from the recent skill process, and when $\lambda$ is chosen from the data the stated union price over the $\lambda$ grid applies. For the limits, write the bound at posterior $\rho$ as $B(\rho,\lambda) = \frac{\mathbb{E}_\rho[\hat{L}]/b}{1-\lambda/2} + \frac{\mathrm{KL}(\rho\|\pi)+c_K}{K\lambda(1-\lambda/2)}$. For fixed $\lambda$ the optimal $\rho$ is the Gibbs posterior. For the concentration inequality, $\rho_\lambda(m) \propto \pi(m)e^{-\lambda K\hat{L}_m/b}$ and the normalizer is at least $\pi(m^\star)e^{-\lambda K\hat{L}_{m^\star}/b}$, so $1-\rho_\lambda(m^\star) = \sum_{m\neq m^\star}\rho_\lambda(m) \le \sum_{m\neq m^\star}\frac{\pi(m)}{\pi(m^\star)}e^{-\lambda K(\hat{L}_m-\hat{L}_{m^\star})/b} \le \frac{1-\pi(m^\star)}{\pi(m^\star)}e^{-\lambda K\Delta/b}$. Conversely if $\hat{L}$ is constant across members the first term is $\rho$-independent and $B$ is minimized at $\mathrm{KL}=0$, $\rho=\pi$. □

*Remark.* Assumption 1 (approximate stationarity) does not by itself deliver the independence used above, and the pooled prior $\pi_c$ of Algorithm 3 includes the target series' own practice rows, so the second assumption holds only for a leave-one-series-out prior. The proposition is therefore the idealized reading of the practice phase as a risk problem; the deployment guarantee (Proposition 1, Theorem 1) does not depend on it.

### A.2 PROPOSITION 1

The arguments in this subsection and the next are for the path-predictive marginals $v_t$, which the registered form plays ($w_t = v_t$, so the correction $\mathcal{E}_W$ of Appendix A.9 is zero). For an implementation that plays the filtered state, Appendix A.9 retains the signed correction.

Exponential weights over the finite comparator class of paths, with prior $\pi(\sigma) = \frac{1}{n_\kappa}\sum_\kappa \pi_\kappa(\sigma)$, has regret against any $\sigma$ at most $\eta^{-1}\ln\frac{1}{\pi(\sigma)} + \eta W b^2/8$ by the standard mix-loss argument with Hoeffding's lemma (Cesa-Bianchi & Lugosi, 2006); since $\pi(\sigma) \geq \pi_\kappa(\sigma)/n_\kappa$ for each component (kernel and start distribution), the code-length form follows. Corollary 1 substitutes this inequality into Theorem 1, whose correction $L_K(\sigma) - M_K$ is defined with the mix loss $M_K$ of the whole mixture prior and is therefore the same whichever component the code length is read through; for a static path through a uniform-start component $\ln(1/\pi_\kappa(\sigma)) = \ln M$. The correction cannot be replaced by the uniform-start component's own practice mix loss $M_K^u$: writing $Z_K = \frac{1}{n_\kappa}\sum_\kappa Z_K^\kappa$ gives $\min_\kappa M_K^\kappa \leq M_K \leq M_K^u + \ln n_\kappa/\eta$, so relative to running the uniform-start component alone the mixture pays up to $M_K^u - \min_\kappa M_K^\kappa$ more, the practice advantage of the best component over the uniform start, which is the posterior mass the uniform-start component lost during practice. The $\pi_c$-free bound of the corollary follows from $M_K \geq \sum_{k\leq K}\min_m \ell_k(m)$ as in the theorem. The path sum is computed exactly by the forward recursion for Markov kernels (static, Herbster–Warmuth, seasonal-phase are all Markov in the round index). □

### A.3 THEOREM 1

Exponential weights over paths is Bayesian: after the $K$ practice rounds its state is the posterior $\rho_K(\sigma) = \pi(\sigma)e^{-\eta L_K(\sigma)}/Z_K$ with $Z_K = \sum_{\sigma'}\pi(\sigma')e^{-\eta L_K(\sigma')} = e^{-\eta M_K}$, and the forward recursion computes exactly the marginals of this posterior (the proof of Proposition 1, applied round by round). The run over the scored rounds is therefore exponential weights over paths with prior $\rho_K$, and the mix-loss argument of Proposition 1 gives, for every $\sigma$, a scored-round regret of at most $\eta^{-1}\ln(1/\rho_K(\sigma)) + \eta W b^2/8$, where the component-selection cost is already represented inside $\pi$. Substituting $\ln(1/\rho_K(\sigma)) = \ln(1/\pi(\sigma)) + \eta L_K(\sigma) + \ln Z_K = \ln(1/\pi(\sigma)) + \eta\big(L_K(\sigma) - M_K\big)$ gives the bound. For the last claim, $L_K(\sigma') \geq \sum_{k\leq K}\min_m \ell_k(m)$ for every path $\sigma'$, so $Z_K \leq e^{-\eta\sum_k \min_m \ell_k(m)}$ and $M_K \geq \sum_{k\leq K}\min_m \ell_k(m)$; hence $L_K(\sigma) - M_K \leq \sum_{k\leq K}(\ell_k(\sigma_k) - \min_m \ell_k(m)) \leq Kb$. When the path outperformed the prior mixture on the practice rounds, $L_K(\sigma) < M_K$ and the correction is negative. □

*Remark (nonnegative code length).* Because $\rho_K$ is a probability distribution, $\ln(1/\rho_K(\sigma)) = \ln(1/\pi(\sigma)) + \eta(L_K(\sigma) - M_K) \geq 0$, so the practice correction is bounded below by $-\ln(1/\pi(\sigma))/\eta$: the unscored practice rounds can at most cancel the comparator's code length, never the $\eta W b^2/8$ term in this upper bound; the realized-gap certificate of Appendix A.9 replaces that term by the measured gaps.

**Proposition 4** (The practice correction is exact; the future-path code length is attained)**.** *Fix the practice losses and let $\rho_K$ be the posterior over complete paths after the $K$ practice rounds, $\ell_t^{\mathrm{mix}} = -\eta^{-1}\ln\sum_\sigma w_t(\sigma)e^{-\eta\ell_t(\sigma_t)}$ the mix loss of round $t$, and $L_W(\sigma) = \sum_{t\leq W}\ell_t(\sigma_t)$ a path's scored loss. For a member $m$ write $\rho_K^W(m) = \rho_K\{\sigma : \sigma_t = m \text{ for every scored } t\}$ for the posterior probability of the event that the scored part of the path is $m$ throughout; $\rho_K^W(m) \geq \rho_K(\sigma)$ for every complete path $\sigma$ with that scored part, with equality when every component is static. (i)* Identity. *For every scored loss sequence, $\sum_{t\leq W}\ell_t^{\mathrm{mix}} = -\eta^{-1}\ln\sum_\sigma \rho_K(\sigma)e^{-\eta L_W(\sigma)}$: on the scored rounds the learner is exponential weights over paths started from $\rho_K$, and the correction of Theorem 1 is the exact change of the comparator's complete-path code length, $L_K(\sigma) - M_K = \eta^{-1}\big[\ln(1/\rho_K(\sigma)) - \ln(1/\pi(\sigma))\big]$, which no scored round can alter. (ii)* Attainment. *On the scored loss sequence on which member $m$ loses $0$ on every scored round and every other member loses $b$, the scored-round mix-loss regret of the learner against the static comparator $m$ lies between*

$$\frac{\ln(1/\rho_K^W(m))}{\eta} - \frac{1}{\eta}\ln\Big(1 + \big(\tfrac{1}{\rho_K^W(m)} - 1\big)e^{-\eta b}\Big) \quad \textit{and} \quad \frac{\ln(1/\rho_K^W(m))}{\eta},$$

*and the played-loss regret is at least the lower value. The future-path code length $\ln(1/\rho_K^W(m))/\eta$, the quantity the certificate of Appendix B evaluates, is therefore attained up to a residual at most*

$(1/\rho_K^W(m) - 1)e^{-\eta b}/\eta$; *the complete-path code length of Theorem 1 bounds it from above and coincides with it when every component is static.*

*Proof.* (i) is the mix-loss telescoping of Proposition 1: over all $T = K+W$ rounds the mix loss sums to $-\eta^{-1} \ln Z_T$ with $Z_T = \sum_\sigma \pi(\sigma) e^{-\eta L_T(\sigma)}$, over the practice rounds to $-\eta^{-1} \ln Z_K = M_K$, and $Z_T/Z_K = \sum_\sigma \rho_K(\sigma) e^{-\eta L_W(\sigma)}$ because $L_T = L_K + L_W$ and $\rho_K(\sigma) = \pi(\sigma) e^{-\eta L_K(\sigma)}/Z_K$; the expression for the correction is $\ln \rho_K(\sigma)$ written out. (ii) On the stated sequence $L_W(\sigma') = b\, d(\sigma')$, where $d(\sigma')$ counts the scored rounds on which $\sigma'$ is not on $m$: $d = 0$ exactly on the event of probability $\rho_K^W(m)$ and $d \geq 1$ elsewhere, so $\rho_K^W(m) \leq \sum_{\sigma'} \rho_K(\sigma') e^{-\eta L_W(\sigma')} \leq \rho_K^W(m) + (1 - \rho_K^W(m)) e^{-\eta b}$. By (i) the scored mix loss lies between $-\eta^{-1} \ln \left[\rho_K^W(m)\big(1 + (1/\rho_K^W(m) - 1) e^{-\eta b}\big)\right]$ and $\eta^{-1} \ln(1/\rho_K^W(m))$, and the comparator's scored loss is 0. The played loss $\langle w_t, \ell_t \rangle$ is at least the mix loss by Jensen's inequality, so the lower value bounds the played-loss regret from below as well. The inequality $\rho_K^W(m) \geq \rho_K(\sigma)$ is the inclusion of the complete path in the event; when every component is static, complete paths are constant and the event is the single path. □

*Remark (the residual).* The residual is not zero at a finite cap: with $\eta = 1$, $b = 4$, $W = 1$ and $\rho_K^W(m) = \frac{1}{2}$ the attained regret is $\ln 2 - \ln(1 + e^{-4}) = 0.675$ against a code length of $0.693$. What is exact is the accounting: the scored rounds start from $\rho_K$, the correction of Theorem 1 is the change of the complete-path code length, and the future-path code length, which is what the audited certificate compares the realized regret with, is met on some scored sequence within that residual. The $\ln(1/\pi(\sigma))/\eta$ term of Proposition 1 is the standard price of a prior and of the minimax order for $M$ experts, $\sqrt{W \ln M}$ (Cesa-Bianchi & Lugosi, 2006, Thm. 3.7).

### A.4 Proposition 2(i)

For distributions on $\mathbb{R}$, $W_2^2(F, G) = \int_0^1 (Q_F - Q_G)^2 d\tau$, so the Fréchet functional $\sum_m w_m W_2^2(G, F_m)$ is minimized pointwise in $\tau$ by $Q_G = \sum_m w_m Q_{F_m}$ when $w$ does not depend on $\tau$. With level-dependent weights the pointwise minimizer at each $\tau$ is still $\sum_m w_m(\tau) Q_{F_m}(\tau)$, but the functional is no longer a Fréchet functional of the member distributions, so the barycenter reading holds for shared weights only. Overdispersion of the CDF mixture is the law of total variance: $\mathrm{Var}_{\text{mix}} = \sum_m w_m \mathrm{Var}_m + \sum_m w_m (\mu_m - \bar{\mu})^2$. □

### A.5 Proposition 2(ii)

The continuous CRPS identity is standard. The finite-grid inequalities follow from Jensen and the exchange argument in Appendix A.9. The clipping and scale identities there explain why a capped-regret guarantee is not automatically a raw-CRPS regret guarantee.

### A.6 Proposition 2(iii)

Immediate from Chernozhukov et al. (2010, Prop. 4): monotone rearrangement weakly reduces $\int |Q - Q^*|^p d\tau$ for every monotone $Q^*$ and every $p \geq 1$. This is an improvement property, not a projection identity: the rearranged vector is in general not the $L_2$ projection onto the monotone cone (isotonic regression; $(2, 0)$ rearranges to $(0, 2)$ while the projection is $(1, 1)$). The order-statistic corollary in §6 follows since sorting the per-level selections makes the reported $\tau$=0.5 value the median of the selected members' values at the central levels. □

### A.7 Proposition 5

**Proposition 5** (Level-path prior)**.** *The ordered level axis admits the same code-length construction as time: comparators are member fields $m(\tau)$ with switches priced by a Herbster–Warmuth kernel, and the forward–backward posterior is a hierarchical estimator sharing practice evidence across neighbouring levels with an $O(\ln M)$ budget per switch. Jointly, the comparator is a field $m(\tau, t)$ with code-length priors on both ordered axes.*

The level axis is a finite ordered index; a Herbster–Warmuth kernel over it defines a Markov prior on member fields $m(\tau)$, and the same mix-loss argument as Proposition 1 gives code-length regret

against any field. Because all levels' practice losses are available simultaneously, the forward–backward (rather than forward-only) posterior is legitimate at decision time; it is the exact posterior over fields, hence a hierarchical estimator with $O(\ln M)$ cost per switch. Combining the two kernels (time, level) as a product prior yields the field comparator $m(\tau, t)$. □

### A.8 THE VALUE OF THE DEPLOYMENT RECURSION

**Proposition 6** (Value of the scored updates)**.** *Let $\rho_K$ be the posterior over paths after the practice rounds, $\bar{\rho}_K$ the member marginal the frozen learner plays on every scored round (the mixture of the four components' marginals after the practice rounds) and $w_t$ the member marginal the full recursion plays. For every path $\sigma$, writing $\langle \rho_K, \ell_t \rangle$ for $\langle \bar{\rho}_K, \ell_t \rangle$,*

$$\sum_{t \le W} \langle \rho_K, \ell_t \rangle - \sum_{t \le W} \langle w_t, \ell_t \rangle \;\ge\; \Big[ \sum_{t \le W} \langle \rho_K, \ell_t \rangle - \sum_{t \le W} \ell_t(\sigma_t) \Big] - \frac{\ln(1/\rho_K(\sigma))}{\eta} - \sum_{t \le W} \delta_t,$$

*where $\delta_t \le \eta b^2/8$ are the mixability gaps of Appendix A.9. The updates therefore gain at least the frozen marginal's regret against $\sigma$ minus $\sigma$'s code length under $\rho_K$ and minus the accumulated mixability gaps.*

*Proof.* The scored rounds are exponential weights over paths with prior $\rho_K$, so $\sum_t \langle w_t, \ell_t \rangle \le \sum_t \ell_t(\sigma_t) + \ln(1/\rho_K(\sigma))/\eta + \sum_t \delta_t$ (the mix-loss argument of Proposition 1 with $\rho_K$ in place of $\pi$); subtract from the frozen loss. □

The bound certifies a positive update gain when some comparator path's advantage over the frozen marginal $\bar{\rho}_K$ exceeds its code length under $\rho_K$ plus the accumulated mixability gaps; a non-positive lower bound is inconclusive, since the realized gain can still be positive. A static path is cheap under $\rho_K$ when practice already favours its member, and may be ahead of $\bar{\rho}_K$ when that member stays the best across the scored windows (it need not be, if the posterior is already concentrated on it), which is a series on which practice and the scored future agree; Figure 3 tests empirically, and only empirically, how the realized update gains relate to the practice→scored agreement.

### A.9 FROM THE PATH CERTIFICATE TO THE REPORTED PREDICTIVE LOSS

All identities below are deterministic. They require a fixed member support, fixed rate and transition kernels, and the same loss history at each update. The audit reports coverage and excludes trajectories violating these conditions from the path certificate; it does not silently drop their missing rounds. No forecaster or aggregation weight is changed by this audit.

**Predictive versus filtered weights.** Let $c_t(m) \in [0, b]$ denote the shifted, median-normalized and capped emission loss ($\ell_t$ in Section 5). Write $v_t$ for the path model's predictive marginal, after applying its transition to the previous filtered state, and $w_t$ for the weights actually played by Algorithm 2. The registered implementation plays $w_t = v_t$ (the transition is applied before the round, and with an empty history the mixture of the start distributions is played), so the correction below is zero by construction; an earlier form of ours played the filtered state, and the identities are kept for implementations of that kind. Define

$$m_t(a) = -\eta^{-1} \log \sum_m a_m e^{-\eta c_t(m)}, \quad \delta_t = \langle w_t, c_t \rangle - m_t(w_t), \quad \epsilon_t = m_t(w_t) - m_t(v_t).$$

Then $0 \le \delta_t \le \eta \,\mathrm{range}(c_t)^2/8$ by Hoeffding's lemma. The signed correction $\mathcal{E}_W = \sum_{t \in \mathrm{scored}} \epsilon_t$ is zero for the registered implementation and must be retained for one that plays the filtered state; it then has a deterministic upper bound. After a nonempty history, with $\alpha = 1/T$, the mixture satisfies $v_t = (1-\alpha) w_t + \alpha \nu_t$ for a probability vector $\nu_t$. Since $e^{-\eta b} \le e^{-\eta c_t(m)} \le 1$,

$$\epsilon_t \le \eta^{-1} \log\big(1 + \alpha(e^{\eta b} - 1)\big).$$

For the empty-history return ($K = 0$, first scored round), use $\epsilon_t \le b$ instead. Thus retaining the measured correction does not require an unverified assumption about the transition; a worst-case alternative is available as well.

**Proposition 7** (Distributional practice certificate; standard variational form)**.** *Let $\rho_K$ be the distribution on future expert paths conditional on the practice observations under the fixed path model. For any distribution $q$ over those paths with finite $\mathrm{KL}(q\|\rho_K)$,*

$$\sum_{t\in\text{scored}} \langle w_t, c_t\rangle - \mathbb{E}_q L_{\text{scored}} \le \frac{\mathrm{KL}(q\|\rho_K)}{\eta} + \sum_{t\in\text{scored}} \delta_t + \mathcal{E}_W.$$

*For a comparison distribution on complete paths, the corresponding complete-path posterior satisfies*

$$\mathrm{KL}(q\|\rho_K) = \mathrm{KL}(q\|\pi) + \eta\big(\mathbb{E}_q L_K - M_K\big).$$

*The latter identity uses complete paths on both sides; marginalizing to future paths yields the first bound without requiring a choice of past path.*

*Proof.* The path evidence telescopes to $\sum_t m_t(v_t) = -\eta^{-1}\log \mathbb{E}_{\rho_K} e^{-\eta L_{\text{scored}}}$. The variational inequality bounds this by $\mathbb{E}_q L_{\text{scored}} + \mathrm{KL}(q\|\rho_K)/\eta$. Adding $\delta_t + \epsilon_t$ gives the actual played loss. Expanding the Gibbs posterior gives the second identity. This is the usual mix-loss analysis (de Rooij et al., 2014), with the implementation's marginal convention made explicit. □

The audit used the hindsight-best constant-member future path at each level, and also the uniform distribution on the two best such paths. Its probability is computed from the practice-conditioned component states, including all future self-transition probabilities; the first scored marginal alone is not a probability for an entire constant path. With this same comparator and signed correction, replacing $\sum_t \eta\,\mathrm{range}(c_t)^2/8$ by $\sum_t \delta_t$ cannot loosen the bound.

**Exact clipping and scale bridge.** Let $r_t(m)$ be the raw horizon-mean pinball loss, $s_t > 0$ the seasonal scale, $a_t = \min_m r_t(m)/s_t$, and $d_t = \max\{\mathrm{median}_m\, |r_t(m)/s_t|, 10^{-9}\}$. Set $x_t(m) = (r_t(m)/s_t - a_t)/d_t$, $c_t(m) = \min\{x_t(m), b\}$, and $e_t(m) = (x_t(m) - b)_+$. For any comparator path $\sigma$,

$$R_x(\sigma) = R_c(\sigma) + \sum_t \big[\langle w_t, e_t\rangle - e_t(\sigma_t)\big],$$
$$R_r(\sigma) = \sum_t s_t d_t \big[\langle w_t, c_t\rangle - c_t(\sigma_t) + \langle w_t, e_t\rangle - e_t(\sigma_t)\big].$$

Both are identities by substitution; the row shift cancels. A small clipping frequency alone does not control the residual magnitude. The time-varying factors $s_t d_t$ preclude identifying the unweighted capped-regret bound with a bound on raw CRPS. The bridge is reported as measured quantities (Appendix F.1).

**Convex combination and sorting in the scored loss.** For ordered levels $\tau_1 < \cdots < \tau_L$ with equal score weights, at every horizon step and outcome $y$, let $z_i = \sum_m w_{im} Q_m(\tau_i)$ and let $z^{\uparrow}$ be its sorted vector. Then

$$\sum_i \mathrm{pinball}_{\tau_i}(z_i^{\uparrow}, y) \le \sum_i \mathrm{pinball}_{\tau_i}(z_i, y) \le \sum_{i,m} w_{im}\mathrm{pinball}_{\tau_i}(Q_m(\tau_i), y).$$

The second inequality is Jensen. For the first, write pinball as $(z-y)_+ + \tau(y-z)$: only $-\sum_i \tau_i z_i$ depends on the permutation. Sorting maximizes $\sum_i \tau_i z_i$ by the rearrangement inequality. Equal spacing is unnecessary; equal score weights suffice. The improvement holds for the sum across levels, not necessarily for each individual level or for median MASE. Multiplying by $2/L$ gives the grid-CRPS relation. Positive outcome-only benchmark normalizations preserve this relation when applied consistently to all three stages.

**Why practice alone cannot guarantee the first scored decision.** Two environments can have identical practice observations and next-round expert losses $(0, b)$ and $(b, 0)$. If weight $p$ is placed on the first expert, excess loss over the best expert is respectively $b(1-p)$ and $bp$; their maximum is at least $b/2$. The same statement holds in expectation for randomized weights. Without a relationship between practice and future losses, no amount of practice removes this first-round ambiguity. This observation motivates the deployment guarantee.

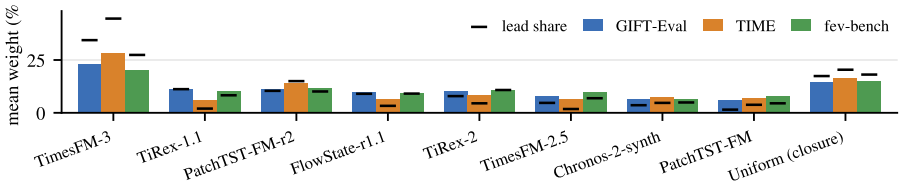


Figure 4: Mean played weight of each member at the median level (bars; averaged over series and scored windows, then configurations) and the share of scored windows on which it carried the largest weight (dashes), per benchmark; numbers in Table 4. The strongest member carried 20–28% of the weight; the margin over it came from the other two thirds.

## B ATTRIBUTION WITH CERTIFICATES

This appendix derives member contributions directly from the learner's weights. The identities describe aggregation before sorting; Appendix A.9 analyzes the sorting step.

**T1: weight–loss identity (exact).** Within a single static exponential-weights component at rate $\eta$ with prior $\pi$, for every pair $(m, m^\star)$,

$$L_m - L_{m^\star} = \tfrac{1}{\eta}\big[\ln(\pi_m/\pi_{m^\star}) - \ln(w_m/w_{m^\star})\big].$$

The played weights are a calibrated readout of realized cumulative loss differences; measured violation on the static kernel was $\leq 1.4 \times 10^{-14}$ (floating point).

**T2: path posteriors (exact).** Under the code-length prior over expert paths, the forward message is the filtered posterior after the observed rounds, which the implementation carries to the next forecast. The path-predictive marginal additionally applies the next transition; Appendix A.9 retains their distinction. The forward–backward marginal is the retrospective attribution. The two-slice posterior yields $P(\text{switch at } t \mid \text{all rounds})$ under the switching kernel: regime attribution. Series (and variate, where present) are explicit axes.

**T3: Shapley uniqueness.** The pooled quantile $q(\tau, h) = \sum_m w_m q_m(\tau, h)$ is an additive game in the members. By efficiency, linearity, null player, and symmetry, the Shapley value of member $m$ is exactly $w_m\,(q_m(\tau, h) - q_{\text{ref}}(\tau, h))$ for any reference $q_{\text{ref}}$, here the uniform average.

**T4: diversity gain (Jensen).** $\sum_m w_m L(q_m) - L(\sum_m w_m q_m) \geq 0$ for pinball loss $L$. This Jensen gap measures the loss reduction from combination relative to weighted member losses.

**T5: audited mix-loss certificate.** Proposition 7 bounds scored regret by the practice-conditioned future-path code length and the realized mixability gaps (the filtered-to-predictive correction of Appendix A.9 is zero for the registered implementation). Figure 2 uses this bound at the median level against the best constant future path. Replay, complete-history and fixed-support checks are required before a series is certified. Appendix A.9 gives the exact clipping, rescaling and quantile-pooling bridges; Appendix F.1 reports their measured magnitudes and coverage. A code length based only on the first scored member marginal is not substituted for a full path probability.

**Demonstration.** Figure 2 shows the three readings on one GIFT-Eval series: the played weights over the practice and scored rounds (T2, predictive marginal), the Shapley decomposition of the last window's median forecast by horizon step (T3), and the realized-regret certificate across all series of the configuration (T5). Figure 5 shows an air-quality series on which the practice rounds concentrated weight on TiRex-1.1; the scored updates retained most of this mass. The Shapley panel shows its contribution relative to the uniform forecast. Appendix G illustrates these views for representative configurations.

**Hardest windows.** Figure 7 shows the scored windows on which the strongest single member of the pool, TimesFM-3, had its largest MASE, one per dataset, with TIMEHEDGE drawn on the same window. The ranking excluded degenerate windows (near-constant context or a mostly-zero target) and used TimesFM-3's error independently of the combination's error. The panels show the member and the combination on the same level shifts and spikes.

Table 4: Mean played weight of each pool member (median level, averaged over series and scored windows, then over configurations; %) and the share of scored windows on which the member carried the largest weight (*lead*, %). The uniform mixture is the closure member of the class (§4).

| | GIFT-Eval | | TIME | | fev-bench | |
|---|---|---|---|---|---|---|
| Member | weight | lead | weight | lead | weight | lead |
| TimesFM-3 | 22.9 | 34.4 | 28.3 | 44.6 | 20.0 | 27.4 |
| TiRex-1.1 | 11.2 | 11.2 | 6.1 | 2.0 | 10.4 | 8.3 |
| PatchTST-FM-r2 | 10.9 | 10.4 | 14.1 | 15.0 | 11.4 | 10.1 |
| FlowState-r1.1 | 9.9 | 9.0 | 6.4 | 3.3 | 9.0 | 9.1 |
| TiRex-2 | 10.2 | 7.9 | 8.4 | 4.5 | 10.5 | 10.8 |
| TimesFM-2.5 | 7.9 | 4.7 | 6.3 | 1.8 | 9.9 | 6.9 |
| Chronos-2-synth | 6.4 | 3.6 | 7.2 | 4.7 | 6.4 | 4.9 |
| PatchTST-FM | 5.8 | 1.5 | 7.0 | 3.8 | 7.6 | 4.5 |
| Uniform mixture (closure member) | 14.6 | 17.4 | 16.2 | 20.4 | 14.8 | 18.1 |

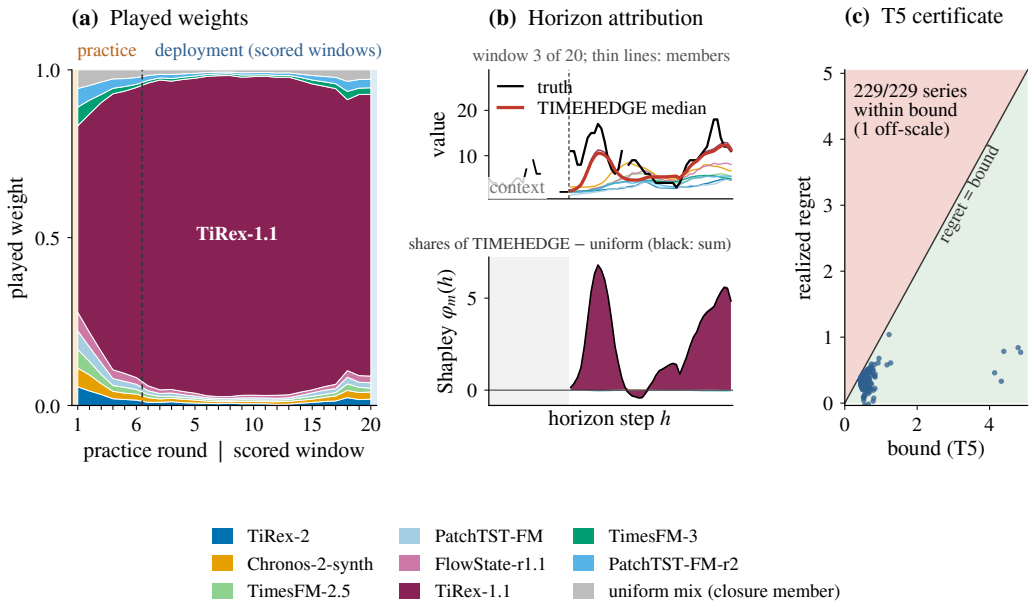


Figure 5: Attribution on GIFT-Eval, KDD-Cup-2018 hourly, one series, with the same panels as Figure 2. Practice concentrated weight on TiRex-1.1; the realized regret was below the certificate on all 229 eligible series of the configuration.

**Success cases.** Figures 8 and 9 are the complementary selections, again one window per dataset and with the same degenerate-window filter, restricted to windows where the combination was accurate and the series had visible structure (TIMEHEDGE MASE at most 0.6; the target's standard deviation over the horizon at least the seasonal scale). Figure 8 ranks such windows, with horizon at least 12 steps, by how much TIMEHEDGE improved on TimesFM-3 (its MASE minus TIMEHEDGE's, at least 0.3): the strongest member missed and the combination recovered, because other members carried the window and the played weights had already moved toward them. Figure 9 ranks windows with horizon at least 24 steps and a structured context (standard deviation over the last $4H$ steps at least 1.5 seasonal scales) by the relative margin by which TIMEHEDGE outperformed the *best* member of the window (every member with a forecast at the origin; margin at least 10%): here no single checkpoint would have done as well, and the gain came from pooling itself (Appendix B, T4). These selections condition on accurate TIMEHEDGE forecasts; Figure 7 conditions on large TimesFM-3 errors.

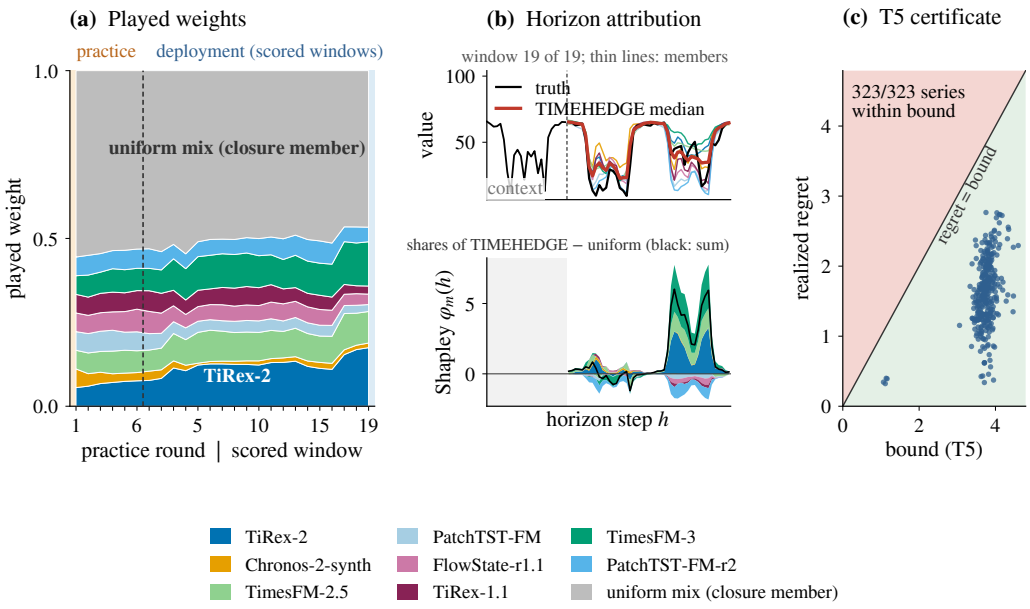


Figure 6: Attribution when the members disagree (GIFT-Eval, Loop Seattle hourly, one series), same panels as Figure 2. The pool was undecided on this series, so the closure member carried most of the weight; on the last scored window the members spread widely in the second dip and the played combination stayed between them, close to the truth. The realized regret was below the certificate on all 323 series of the configuration.

# C Experimental details

## C.1 Test panel

Series counts, horizons, and window counts of the 97 GIFT-Eval test configurations were read from the materialized test panel and agreed with the benchmark's declared metadata. TIMEHEDGE used $K=6$ unscored practice rounds before each series' first scored origin.

## C.2 Pools, checkpoints, and member parity

Table 5 lists every member, its public checkpoint, its own leaderboard lane, the inference-time operating modes TIMEHEDGE may select among, and the parity of our member reproduction against the published leaderboard row. Parity was audited by re-running each member under its official published settings (the replication notebook or model-card configuration) on pilot configurations and comparing per-configuration MASE to the public CSV; univariate-source configurations reproduced essentially exactly (differences at CSV-rounding or bf16 cross-hardware level). Members whose official rows use joint multivariate feeding differed on multivariate-source configurations when fed univariately; where the official mode was reproducible we exposed it as an operating mode and let practice-phase selection choose (below).

## C.3 Operating-mode selection is practice-phase selection

Members with multiple inference modes (univariate vs. multivariate feeding; joint batching; input resampling; context handling) were routed per configuration by the same pooled pre-origin practice comparator used everywhere else: for each configuration, each mode's pooled practice MASE was computed from rounds strictly before the first scored origin, and the better mode was used for all of that configuration's scored forecasts. No scored window contributed to any selection. The available modes are listed in Table 5.

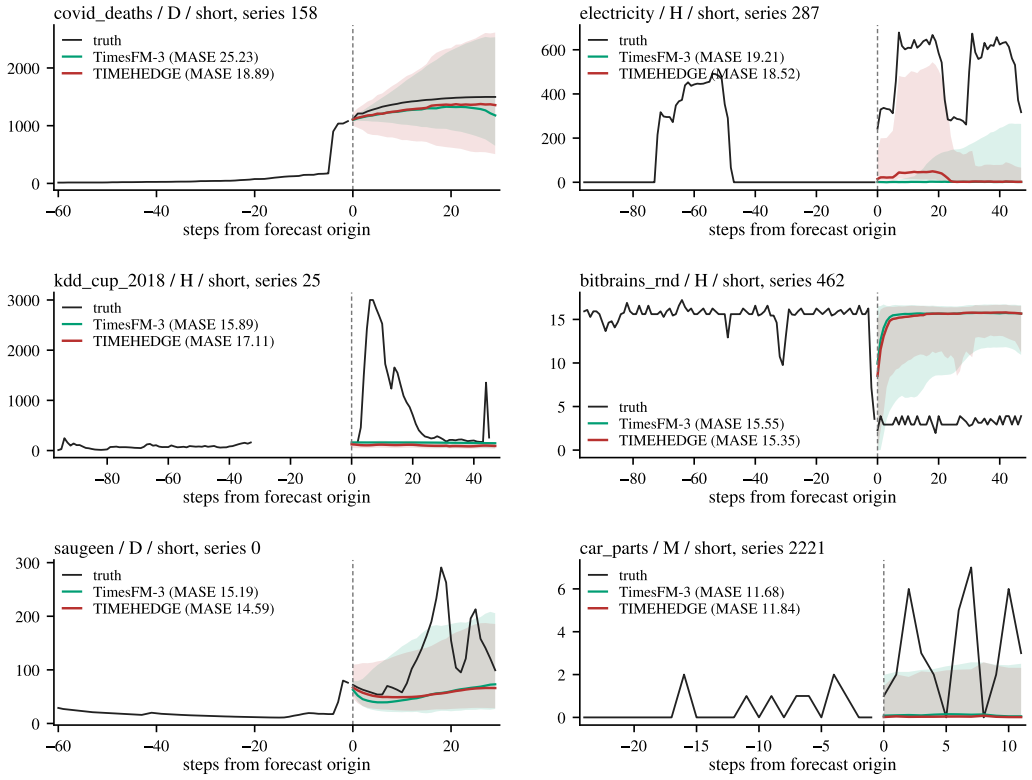


Figure 7: The hardest scored windows for the strongest single member (TimesFM-3), one per dataset, ranked by its own MASE with degenerate windows excluded. Black: context and truth; green: TimesFM-3 median with its 10–90% band; red: TIMEHEDGE median with its 10–90% band. MASE values in the legends are for the window shown.

### C.4 Evaluation harness and engine verification

The official-harness numbers of §6.2 (0.6502, 0.6540 and the one-checkpoint-per-family contrasts) were produced by the official `gluonts evaluate_model` harness over all 97 configurations (`mask_invalid_label=True`, `allow_nan_forecast=False`). Our remaining benchmark scores were produced by our engine, which was verified to reproduce the harness on member forecasts (agreement to $10^{-8}$ MASE on audit configurations). On the final TIMEHEDGE forecasts the two scores differed on 21 of the 97 configurations (engine geometric mean $0.6506$ against $0.6502$ on the harness; the largest differences on m4_yearly and bitbrains), where the harness's fallback and window handling applied to windows without a complete pool; both aggregates are reported and neither is presented as the other. Normalized scores divide each configuration's MASE by the official seasonal-naive row before geometric aggregation, matching the leaderboard's construction.

### C.5 Coverage rule

Member forecast dumps that had not completed by a preregistered deadline left that member simply absent from the affected configuration's pool (the combiner renormalized); the rule is deterministic (size-ordered processing) and disclosed per shot in Appendix E. Final entries had full 97/97 coverage for all members except where noted.

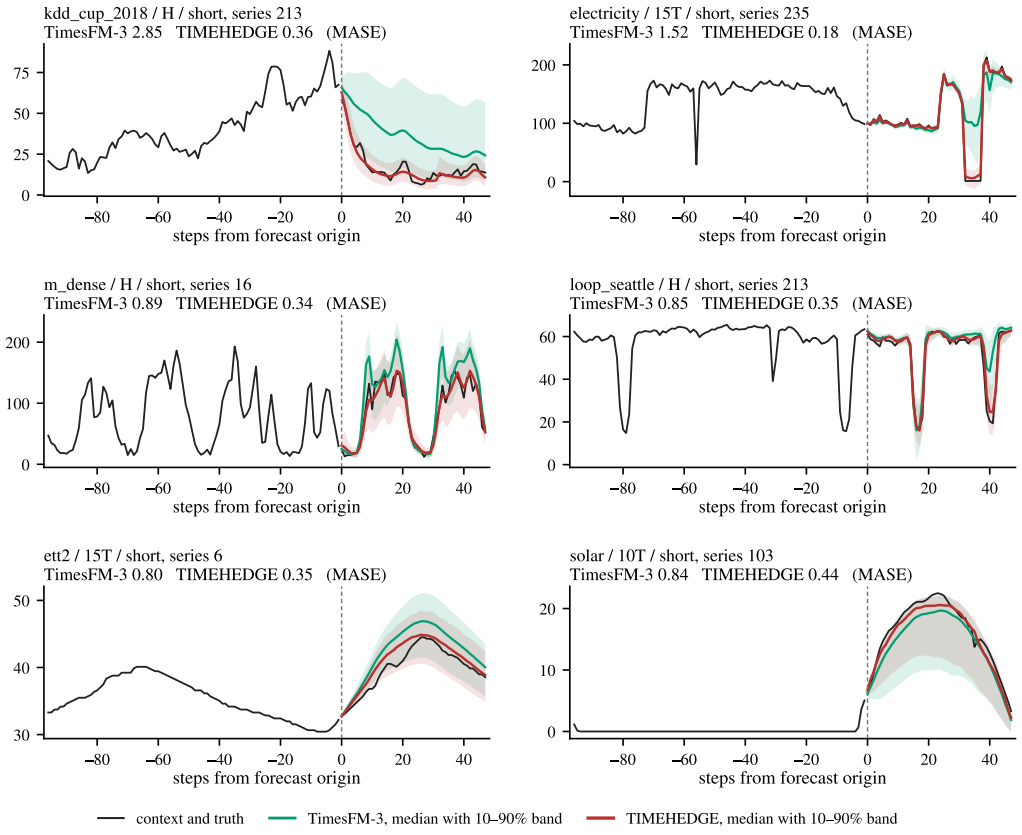


Figure 8: Windows on which the strongest single member (TimesFM-3) missed and TIMEHEDGE recovered: one per dataset, ranked by TimesFM-3's MASE minus TIMEHEDGE's among accurate, structured windows (selection rule in the text). Black: context and truth; green: TimesFM-3 median with its 10–90% band; red: TIMEHEDGE median with its 10–90% band. MASE values in the panel titles are for the window shown.

### C.6 Compute

Member inference dominates. A single member issues $W$ forecasts per series and the uniform mixture $MW$. TIMEHEDGE and practice-informed rules that mix all members require $M(K+W)$, including the $MK$ in-context backtests at earlier origins. After selecting one physical member, a practice pick requires only $MK + W$ forecasts. Selecting three physical members requires $MK + 3W$; selecting the closure still requires all $M$ members. The shared evaluation cache contains all members' forecasts, but this does not make their deployment costs equal. On GIFT-Eval ($K = 6$, $W$ mostly 1–2), TIMEHEDGE used 3.5–7 times as many forecasts as the uniform mixture, $8(K+W)$ per series for the eight-member pool (56 at $W=1$); on TIME ($W \sim 20$) the factor was about 1.3. The practice forecasts were computed once, cached, and shared by all practice-reading methods; in aggregate they amounted to GPU-days per pool across the benchmarks, dominated by long-context high-frequency configurations (the fev-bench dump took between twenty minutes and fourteen hours per member on a shared GPU). Aggregation is CPU arithmetic: per origin and level, $n_\kappa$ forward recursions of length $T \le K + W$ over $M+1$ members, $O(n_\kappa TM)$ multiply-adds, with $n_\kappa(M+1)$ floats of state per level; the recursion is replayed from the prior at every origin, so a series costs $O(n_\kappa W(K+W)M|\mathcal{T}|)$ multiply-adds in all; the full TIMEHEDGE pass over the 235,039 scored origins of fev-bench (nine levels each) took three minutes on twelve CPU cores, about 9 ms per origin per core.

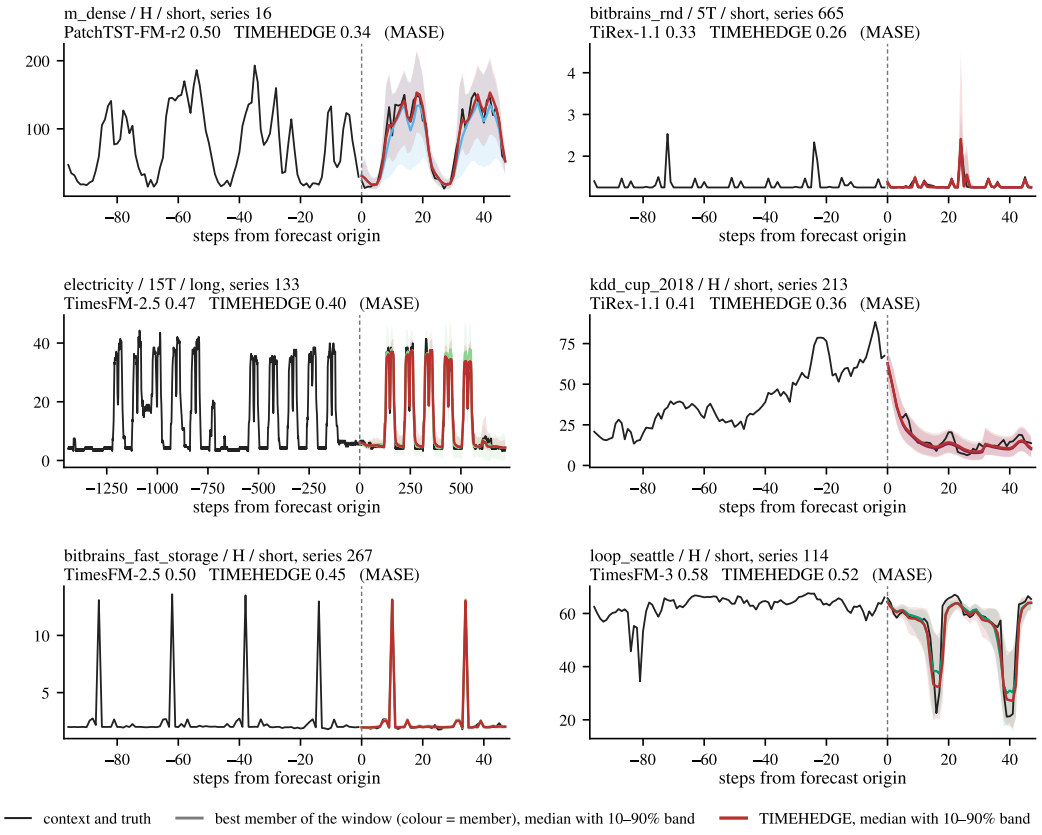


Figure 9: Windows on which TIMEHEDGE outperformed every member of the pool: one per dataset, ranked by the relative margin over the best member of the window among accurate, structured windows (selection rule in the text). The coloured line is the best member for that window, named in the panel title (colours as in Figure 2); red: TIMEHEDGE median with its 10–90% band.

**Actual serving cost.** We timed the original eight adapters and the public replay implementation on three single-series FETS panels (Obermeier et al., 2026): Mobilithek (public EV charging), Industrial VEA and PV Hong Kong (15-minute cadence, 24-hour horizon). The panels were fixed by metadata hashes before accuracy inspection. Table 6 measures serial execution on an otherwise idle RTX PRO 6000 Blackwell Server Edition GPU and an Intel Xeon Platinum 8559C host. Every timed run performed new member forecasts, including practice where required, and included loss/prior computation, causal updates, replay, mixing, sorting and CPU transfers. Pool startup from cached checkpoints took 28.3 s; explicit warmup and file I/O are also excluded from the table. TIMEHEDGE took 7.09, 2.20 and 1.30 times Uniform's wall time for $W = 1, 5, 20$, respectively. The practice pick queried only its selected member after practice. TIMEHEDGE incurred no parameter-training cost. On this GPU, the standalone TSFM forecasts took 0.064 s (TiRex-2) and 0.431 s (TimesFM-3). Including all eight members and $K = 6$ practice rounds, TIMEHEDGE issued the first forecast in 13.3 s and 20 scored forecasts in 49.4 s (Table 6). It combined these outputs directly, without an additional language-model inference stage.

Table 5: Pool composition. "Lane" is the member's own classification on the public leaderboard; the zero-shot pool spans five open lineages and admits every public zero-shot-typed checkpoint of those lineages (generations and revisions are separate members). "Parity" is our reproduction of the member's official leaderboard row under its published inference settings (relative MASE difference on pilot configurations; univariate-source configurations).

| Member | Checkpoint | Lane | Modes used | Parity |
|---|---|---|---|---|
| *Zero-shot pool (TIMEHEDGE, 8 members: every public zero-shot checkpoint of five lineages)* | | | | |
| TiRex-2 | `NX-AI/TiRex-2-gifteval-zs` | zero-shot | – | $\leq 0.1\%$ |
| TiRex-1.1 | `NX-AI/TiRex-1.1-gifteval` | zero-shot | resampling | $\leq 0.5\%$ |
| Chronos-2-synth | `autogluon/chronos-2-synth` | zero-shot | uni/MV/joint | $\leq 0.6\%$ |
| TimesFM-3 | `google/timesfm-3.0-pytorch` | zero-shot | MV attention | $\leq 0.3\%$ |
| TimesFM-2.5 | `google/timesfm-2.5-200m-pytorch` | zero-shot | – | exact |
| PatchTST-FM-r2 | `ibm-granite/granite-timeseries-patchtst-fm-r2` | zero-shot | – | $\leq 0.1\%$ |
| PatchTST-FM-r1 | `ibm-research/patchtst-fm-r1` | zero-shot | – | $\leq 0.1\%$ |
| FlowState-r1.1 | `ibm-granite/granite-timeseries-flowstate-r1` (r1.1) | zero-shot | – | $\leq 0.5\%$ |
| *Additional members in TIMEHEDGE-P (14 member slots)* | | | | |
| Chronos-2 | `amazon/chronos-2` | pretrained | uni/MV; joint | $\leq 0.1\%$ |
| Moirai-2 | `Salesforce/moirai-2.0-R-small` | pretrained | ctx mode | $\leq 1\%$ |
| TimesFM-2.5 (short context) | `google/timesfm-2.5-200m-pytorch` | zero-shot | context $\leq 4{,}096$ | – |
| Toto-2.0 | `Datadog/Toto-2.0-1B` | pretrained | MV, imputation | $\leq 0.7\%$ |
| Timer-S1 | `thuml/Timer-S1` | pretrained | – | $\leq 0.5\%$ |

Table 6: Actual wall time (seconds) to issue $W$ forecasts, charging the $K = 6$ practice forecasts where used. Each cell is the geometric mean over three fixed single-series panels of the median of three rotated repeats per panel (189 runs; $H = 96$). Models and input panels were resident; startup, explicit warmup and retrospective scoring were excluded.

| Method | $W = 1$ | $W = 5$ | $W = 20$ |
|---|---|---|---|
| TIMEHEDGE | 13.312 | 21.007 | 49.393 |
| Uniform | 1.878 | 9.549 | 37.865 |
| Practice pick | 11.520 | 12.174 | 13.942 |
| Hedge + own practice | 13.248 | 21.048 | 49.244 |
| AdaHedge + own practice | 13.332 | 20.986 | 49.015 |
| TiRex-2 | 0.064 | 0.314 | 1.252 |
| TimesFM-3 | 0.431 | 2.137 | 8.682 |

**Equivalent incremental implementation.** With a fixed prior, learning and switching rates, loss cap, and unchanged cached losses, the filtered messages and log normalizers can be retained between origins: the next emission updates the same forward state that replay would reconstruct. Thus the weight recursion can also be implemented in $O(n_\kappa(K{+}W)M|\mathcal{T}|)$ total work per series, without changing the predictive mixture; quantile mixing has its original cost. We audited this optional implementation on 18 cached rolling-origin panels ($K{=}6$, $W \in \{8, 32\}$; 746 series, 19,800 issued forecasts; eight frozen members plus closure, nine quantile levels). Every forecast agreed with the saved reference within $\mathrm{rtol} = 2 \times 10^{-5}$, $\mathrm{atol} = 2 \times 10^{-5}$; the largest $|\Delta q|/\max(1, |q|)$ was $5.96 \times 10^{-8}$. On one Intel Xeon Platinum 8559C CPU process with numerical-library threads set to one, three alternating timed runs after warmup gave a 23.0-fold geometric mean of panel-wise median CPU speedups (10.2–33.2-fold; 18 of 18 panels faster). Both paths rebuilt the pooled prior and loss caches and included loss/scale calculation, closure, weighting, mixing and sorting; model inference, disk I/O, metric evaluation and logging were excluded. The registered results above used replay.

**Algorithm 1** TIMEHEDGE for one series (registered form `TIMEHEDGE-TW-online`)

**Input:** target series $y$; first scored origin $t_0$; horizon $H$; number of scored windows $W$; seasonality $P$; member fields $Q_m^{(o)}$ at every origin used below; configuration prior $\pi_c$ over the $M{+}1$ pool members (Algorithm 3); practice budget $B = 6$

**Output:** combined fields $Q^{(o)}$ at the $W$ scored origins (an origin with no present member or no finite target is skipped); played weights $w_t(\tau)$

$K \leftarrow \min(B, \lfloor t_0/H \rfloor)$; origins $o_t \leftarrow t_0 - (K - t + 1)H$ for $t \le K$ (practice) and $o_{K+w} \leftarrow t_0 + (w-1)H$ for $w = 1, \dots, W$ (scored)

closure member at every origin: $Q^{(o)}_{M+1}(\tau, h) \leftarrow$ mean of $Q^{(o)}_m(\tau, h)$ over the members $m \le M$ with a finite forecast at $o$

**for** $t = 1, \dots, K$ **do**

  compute $\ell_{o_t}(\tau, m)$ for all $\tau \in \mathcal{T}$, $m \le M{+}1$ *practice rounds: observable, unscored*

**end for**

**for** $w = 1, \dots, W$ **do**

  $t \leftarrow K + w$; $A \leftarrow \{m \le M : Q_m^{(o_t)} \text{ finite}\} \cup \{M{+}1\}$; $\pi \leftarrow \pi_c$ restricted to $A$ and renormalized

  **for** $\tau \in \mathcal{T}$ **do**

    $w_t(\tau) \leftarrow \textsc{PathHedge}\big(\pi;\ (\ell_{o_u}(\tau, m))_{u<t,\, m\in A}\big)$ *Algorithm 2*

    $Q^{(o_t)}(\tau, h) \leftarrow \sum_{m\in A} w_t(\tau, m)\, Q_m^{(o_t)}(\tau, h)$ for $h \le H$ *per-level convex combination*

  **end for**

  for each $h$, sort $\big(Q^{(o_t)}(\tau, h)\big)_{\tau\in\mathcal{T}}$ increasingly *rearrangement*

  emit $Q^{(o_t)}$; observe $y_{o_t+1}, \dots, y_{o_t+H}$; compute $\ell_{o_t}(\tau, m)$ for all $\tau$, $m \le M{+}1$

**end for**

**return** $Q^{(o_{K+1})}, \dots, Q^{(o_{K+W})}$ and the played weights $w_{K+1}(\tau), \dots, w_{K+W}(\tau)$

## D PSEUDOCODE

Algorithms 1–3 list the procedure as executed by the registered strategy (`TIMEHEDGE-TW-online`) on the GIFT-Eval test split. Notation follows Sections 3 and 4: a configuration has series $s = 1, \dots, S$, horizon $H$, $W$ scored origins per series, and the seasonality $P$ declared for its frequency; the pool has $M$ members, indexed $m \le M$, plus the closure member $M{+}1$; the level grid is $\mathcal{T} = \{0.1, \dots, 0.9\}$. A round is one forecast origin $o$, and $Q_m^{(o)}(\tau, h)$ is member $m$'s quantile field issued at $o$. The loss of pool member $m$ at level $\tau$ on round $o$ is the mean pinball loss over the horizon steps with a finite target, divided by the seasonal-naive scale $s$ of the series' context (the MASE denominator),

$$\ell_o(\tau, m) = \frac{1}{s\,|\mathcal{H}_o|} \sum_{h\in\mathcal{H}_o} \text{pinball}_\tau\big(Q_m^{(o)}(\tau, h),\, y_{o+h}\big), \qquad \mathcal{H}_o = \{h \le H :\ y_{o+h} \text{ finite}\},$$

and is undefined when $Q_m^{(o)}$ is not finite in every entry. Every step below normalizes a round's loss vector by its own median across members, so a positive per-round rescaling of the losses (in particular the choice of $s$, which the code takes from the context before $t_0$ for the practice rounds and from the context before the following origin for the scored rounds) leaves all weights unchanged away from numerical floors. Numerical floors on probabilities, scales and the standard deviation in Algorithm 3 are omitted.

**Attribution outputs.** The weights $w_t(\tau)$ returned by Algorithm 1 are the predictive marginals actually played (T2 in Appendix B); for the practice-phase panel of Figure 2 the same recursion is evaluated on the first $k = 0, \dots, K-1$ practice rounds, giving the posterior entering practice round $k{+}1$ ($k = 0$ plays $\frac{1}{2}(\pi + u)$, where $\pi$ is $\pi_c$ restricted to the members present at $o_{K+1}$ and $u$ is uniform on that support); the panel shows $\tau = 0.5$. The Shapley decomposition of a scored window's median forecast is $\varphi_m(h) = w_t(0.5, m)\big(Q_m^{(o_t)}(0.5, h) - Q_{M+1}^{(o_t)}(0.5, h)\big)$ for $m \le M$, with the closure member as reference (its own share is zero), so that $\sum_m \varphi_m(h)$ is the combined median before rearrangement minus the uniform mixture. The regret certificate is evaluated at $\tau = 0.5$ over the scored rounds on which every pool member has a finite

**Algorithm 2** PATHHEDGE: exponential weights over expert paths by the forward recursion

**Input:** prior $\pi$ on $M_A$ members; loss vectors $\ell_1, \dots, \ell_n$ of the past rounds in time order; the series' planned number of rounds $T = K + W$

**Output:** weights $w$ to play on the next round

keep the rounds whose loss vector is finite in every entry; let $n' \le T$ be their number and re-index them $u = 1, \dots, n'$; **if** $n' = 0$ **return** $\frac{1}{2}(\pi + u)$ *no history: the mixture of the start distributions*

$\eta \leftarrow \sqrt{8 \ln M_A / T}$ *Registered rate on the planned $T = K + W$ rounds, fixed before the first scored round; not on $n'$*

$\mathcal{K} \leftarrow \{\text{STATIC } (\alpha = 0),\ \text{HW } (\alpha = 1/T)\} \times \{\pi,\ u\}$ *kernel × start distribution ($u$ uniform on $M_A$); $n_\kappa = 4$*

$a_\kappa \leftarrow$ the start distribution of $\kappa$ and $Z_\kappa \leftarrow 0$ for every $\kappa \in \mathcal{K}$

**for** $u = 1, \dots, n'$ **do**

  $e_u(m) \leftarrow \exp\big(-\eta \min\{(\ell_u(m) - \min_{m'} \ell_u(m')) \,/\, \text{median}_{m'} |\ell_u(m')|,\ b\}\big)$, $b$=4 *self-normalized, capped emission*

  **for** $\kappa \in \mathcal{K}$ **do**

    **if** $u > 1$: $a_\kappa \leftarrow (1 - \alpha_\kappa)\, a_\kappa + \alpha_\kappa \big(\sum_m a_\kappa(m)\big)/M_A$ *switch step between consecutive rounds; identity for* STATIC*; none before the first round*

    $a_\kappa \leftarrow a_\kappa \odot e_u$; $\quad Z_\kappa \leftarrow Z_\kappa + \ln \sum_m a_\kappa(m)$; $\quad a_\kappa \leftarrow a_\kappa / \sum_m a_\kappa(m)$

  **end for**

**end for**

$q_\kappa \leftarrow e^{Z_\kappa} / \sum_{\kappa' \in \mathcal{K}} e^{Z_{\kappa'}}$ *kernel mixture by marginal likelihood, uniform kernel prior*

$a_\kappa \leftarrow (1 - \alpha_\kappa)\, a_\kappa + \alpha_\kappa \big(\sum_m a_\kappa(m)\big)/M_A$ for every $\kappa$ *predictive marginal: the transition applied once more, towards the round to be played*

**return** $w \leftarrow \sum_\kappa q_\kappa a_\kappa$, normalized to sum to one

**Algorithm 3** Configuration prior $\pi_c$ from pooled practice rounds

**Input:** for every series $s$ of the configuration whose context before $t_0(s)$ has a finite positive scale: its $K_s$ practice-round loss tables $\ell_o(\tau, m)$ (lines 1 and 4–6 of Algorithm 1); the set $\mathcal{M}_c$ of pool members with a forecast dump for the configuration, including $M{+}1$ (all $M{+}1$ on the registered run), $M_c = |\mathcal{M}_c|$

**Output:** $\pi_c$ on $\mathcal{M}_c$ (zero mass elsewhere), used for every level $\tau$

one row per (series, practice round): $r_{s,k}(m) \leftarrow \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \ell_{o_{s,k}}(\tau, m)$ for $m \in \mathcal{M}_c$ *mean over levels*

drop rows with a non-finite entry; $N \leftarrow$ number of rows kept; **if** $N = 0$ **return** the uniform distribution on $\mathcal{M}_c$

$\tilde{r}(m) \leftarrow r(m)/\text{median}_{m'} r(m')$ for every kept row *per-round normalization*

$\eta_0 \leftarrow \sqrt{8 \ln M_c / N} \,/\, \text{std}(\tilde{r})$ *std over all entries of the normalized rows*

**return** $\pi_c(m) \propto \exp\big(-\eta_0 \sum_{\text{rows}} \tilde{r}(m)\big)$

loss (a series with fewer than two such rounds, or with a skipped scored origin, is omitted from the panel): with $\tilde{\ell}_t(m) = \min\{(\ell_t(m) - \min_{m'} \ell_t(m'))/\text{median}_{m'} |\ell_t(m')|,\ b\}$ the emission-normalized capped losses the learner plays ($b$=4), $m^\star$ the member with the smallest $\sum_t \tilde{\ell}_t(m)$, $\eta = \sqrt{8 \ln(M{+}1)/(K + W)}$ the rate the weights were played with, and $\rho_K^W(m^\star)$ the probability, under the practice-conditioned state of the four components (future self-transitions included), that the scored-round path stays on $m^\star$ throughout (Proposition 7, Appendix A.9), the realized regret $\sum_t \langle w_t(0.5), \tilde{\ell}_t \rangle - \sum_t \tilde{\ell}_t(m^\star)$ is compared with $\ln(1/\rho_K^W(m^\star))/\eta + \sum_t \delta_t$ (the gap certificate) and with $\ln(1/\rho_K^W(m^\star))/\eta + \frac{\eta}{8} \sum_t \max_m \tilde{\ell}_t(m)^2$ (the Hoeffding certificate); the first scored marginal alone is not a path probability and is not used.

**Correspondence to Section 4.** The listings are the registered form described in Section 4. Two details are implementation choices not spelled out there. (i) The learning rate $\eta = \sqrt{8 \ln M_A / T}$ and the switch rate $1/T$ use the series' planned number of rounds $T = K + W$, fixed before the first scored round; the recursion is re-run from the prior at each origin over the finite-loss rounds

seen so far (at most $T$), so Proposition 1 and Theorem 1 apply to the played weights as stated. (ii) On all three benchmarks of the main tables the configuration prior starts from a uniform base (Algorithm 3); TIME is scored through the same engine as GIFT-Eval and fev-bench, with each (series, variate) pair a panel series and the official rolling origins recovered from the benchmark's release. Appendix F additionally reports TIME scored by the benchmark's own evaluation code, where the rows of Algorithm 3 are formed per level from at most six rows per configuration, one per practice-round index, each the mean over the configuration's series of that round's loss vector at that level, and the exponent is chained from a pool-level leave-one-configuration-out prior of the same form entering as a log-prior base. The level-path variant (QP) was a development-panel variant. Members without a forecast at an origin receive weight zero and the prior is renormalized on the present members (coverage rule, Appendix C).

# E PREREGISTRATION AND FULL DISCLOSURE

Every scored evaluation in this project was *registered before execution*: an amendment freezing the exact configuration (pool, learner, modes, coverage rule, primary and secondary outcomes) was written to the project log before the run. We summarize the registration trajectory below.

**Learner shots (frozen 7–8-member pool).** Four registered shots varied the practice-phase learner (selection form; two-phase composition; empirical-Bayes shrunk selection; config-level practice-CV over a learner menu). Their observed scored-test results were close (0.6769–0.6837 under the development pool and its normalization, not comparable to Table 1). The registered validation-side learner ablation is Table 2.

**Member-quality shots.** Subsequent registered shots reproduced each member's official inference settings (context caps, multivariate feeding, joint batching, input resampling), with per-member parity audits (Table 5).

**Mode-routing decomposition.** Practice-CV routing of member operating modes was evaluated member-by-member in registered shots; routing was adopted where the practice comparator improved the (pre-registered) outcome and rejected otherwise. Both outcomes were logged.

**Pool rule.** The five lineages (TiRex, TimesFM, PatchTST-FM, Chronos-2, FlowState) were fixed before the campaign as the open lineages with public checkpoints and published benchmark rows we could reproduce. Two pool rules over these lineages were run, both declared before their evaluations: first the one-strongest-checkpoint-per-lineage rule (five zero-shot members, eight members for the pretrained pool), then the inclusive rule that admits every public checkpoint of the lineages typed for the lane (eight and fourteen). The inclusive rule was adopted as the primary after both had been scored; the lineage-unique pools are reported as contrasts in Table 2 and Appendix F, and the ordering of the entries was the same under either rule. Membership within each declared rule was fixed for its evaluation; selection between rules occurred afterward. No public zero-shot checkpoint outside the five lineages was evaluated in this campaign.

**Test-set hygiene.** At prediction time, every combination weight, mode choice, and member-mode route of TIMEHEDGE and TIMEHEDGE-P used configuration-level practice data and previously revealed outcomes, never an unrevealed scored target; both declared pools are reported. Category classification inherits members' training corpora by construction of the pools.

# F ADDITIONAL RESULTS

## F.1 AUDIT OF THE CERTIFICATE AND PREDICTIVE-LOSS BRIDGE

We replayed the saved weights on all three cached benchmark panels without new forecaster inference. The fixed-support audit required every planned practice and scored loss to be finite and the replay to agree within $10^{-5}$; exclusions are reported rather than treated as successful certificates. Six electricity trajectories failed this origin-scale replay check: the $10^{-9}$ normalization floor makes the engine's reveal-time cache scale relevant. Replaying with that scale resolved all six (maximum

Table 7: Cached-data audit. Certificate eligibility requires complete fixed support and verified replay. Code, gap and correction are mean scored-phase terms over eligible series and levels. Reduction is the mean of $1 - B_{\mathrm{gap}}/B_{\mathrm{H}}$ with identical comparator and correction.

| Panel | Configs | Eligible / total | Mean $W$ | Code | $\sum \delta_t$ | $\mathcal{E}_W$ | Reduction |
|---|---|---|---|---|---|---|---|
| GIFT-Eval | 97 | 109,448 / 159,376 | 2.86 | 1.636 | 0.064 | +0.0000 | 10.8% |
| TIME | 98 | 8,918 / 8,973 | 12.10 | 2.181 | 0.102 | +0.0000 | 14.7% |
| fev-bench | 100 | 76,405 / 125,344 | 2.11 | 1.612 | 0.025 | -0.0000 | 3.6% |

Table 8: Measured predictive-loss bridge, including the scored-only pass. Tail is the mean played normalized clipping tail per scored origin and level. Convex/member and sorted/convex are mean ratios of seasonal-scaled grid scores over configurations where the denominator is positive.

| Panel | Series with scores | Scored windows | Mean tail | Max tail | Convex/member | Sorted/convex |
|---|---|---|---|---|---|---|
| GIFT-Eval | 155,581 | 363,989 | 0.01282 | 1361 | 0.974124 | 0.999957 |
| TIME | 8,973 | 111,071 | 0.0009136 | 16.55 | 0.976702 | 0.999976 |
| fev-bench | 121,701 | 230,077 | 0.007046 | 265.3 | 0.966473 | 0.999818 |

weight error $2.8 \times 10^{-8}$); they remain excluded from the fixed-reference audit counts reported here. A separate scored-only pass measured the loss bridge on available scored origins of excluded series, without assigning them a path guarantee. The comparison was at each of nine levels, included the closure member, and used the best constant future member in capped loss, with a uniform-on-best-two distribution as an additional comparator. Clipping-tail columns combine eligible and scored-only groups by their numbers of scored origins and levels. All eligible single-member and best-two distributional certificates satisfy their bounds within $10^{-7}$. The nine-level score bridge checks Jensen and the summed sorting inequality in floating point. The code-length, gap and signed transition-correction terms were averaged over eligible series and levels. The gap-based and Hoeffding certificates used the same future-path probability and correction. Scoring bridges recorded raw pinball loss, the seasonal-scaled equal-grid score used in the main comparison, and the absolute-target-scaled grid score separately. Scores were recomputed after convex combination and after sorting.

### F.2 Exchanging the learner inside TIMEHEDGE

TIMEHEDGE is the recipe of Algorithm 1: practice rounds read as in-context backtests, the pooled practice prior, the closure member, per-level weights with rearrangement, and a fixed-rate path hedge over the scored reveals as the learner. Table 9 reports exchanges of only the last ingredient: the registered learner was replaced by each online learner of Table 1, run with the same prepended practice rounds and the same supplied prior $\pi_c$ (FTL used its own leader rule), under its own loss convention and, for Hedge, the horizon rate $\sqrt{8\ln(M{+}1)/(K{+}W)}$. Every exchange in Table 9 improved in the pooled comparison with the same learner's scored-only form (Hedge gained $3.3\%$ on GIFT-Eval, BOA and its smoothed variant $4$–$7\%$ on GIFT-Eval and fev-bench): the recipe carried the margin. The effect of choosing a learner within it was small and benchmark-dependent: on GIFT-Eval the differences were under $0.7\%$ (Hedge, AdaHedge and FTL showed no significant difference, $p \geq 0.3$; BOA and its smoothed variant behind, $p \leq 0.041$); on TIME every reported exchange was behind by $0.3$–$1.1\%$ ($p \leq 0.04$); on fev-bench these alternatives were behind by $0.8$–$1.5\%$ ($p \leq 0.009$). Pooled over the 293 configurations with dataset clusters, Hedge, AdaHedge, FTL, BOA and its smoothed variant were behind the registered learner ($+0.35$, $+0.39$, $+0.74$, $+1.10$ and $+1.04\%$; unadjusted $p \leq 0.015$). These comparisons place the main observed benefit in practice information, rather than learner choice. For the registered learner, Appendices B and G provide the evaluated certificate and attribution identities.

### F.3 Does practice transfer across update rules?

We compared each learner with itself, with and without the practice protocol, on the same 293 configurations. This comparison kept the update rule fixed and changed the available initialization and loss history. Hedge's horizon-based rate included practice when present; this is a comparison of complete protocols, distinct from the rates-fixed ablation in Table 12. All five reported

Table 9: Learner substitutions inside TIMEHEDGE. All rows used its pooled practice prior, own-series practice, closure member and rearrangement; each update rule kept its loss convention. These are ablations of the combination procedure, separate from the uniform-start online baselines in Table 1. Geometric-mean normalized MASE and scaled CRPS; lower is better. AdaHedge used the corrected log-domain implementation.

| | GIFT-Eval | | TIME | | fev-bench | |
|---|---|---|---|---|---|---|
| Learner | MASE | CRPS | MASE | CRPS | MASE | CRPS |
| Hedge | 0.6497 | 0.7348 | 0.6672 | 0.6797 | 0.6144 | 0.7722 |
| AdaHedge | 0.6494 | 0.7350 | 0.6688 | 0.6815 | 0.6142 | 0.7706 |
| BOA | 0.6551 | 0.7415 | 0.6725 | 0.6857 | 0.6184 | 0.7775 |
| FTL | 0.6522 | 0.7393 | 0.6708 | 0.6857 | 0.6161 | 0.7774 |
| BOA + fixed smoothing | 0.6544 | 0.7412 | 0.6717 | 0.6849 | 0.6185 | 0.7775 |
| TIMEHEDGE | 0.6506 | 0.7329 | 0.6653 | 0.6770 | 0.6092 | 0.7643 |

Table 10: Practice transferred across online learners. Pooled relative score reductions from giving each learner the pooled prior and own practice, versus its scored-only form (positive is better). Tests merged shared data sources across benchmarks and used paired sign flips on configuration log ratios. Five learners are reported; Holm correction retained the original twelve-comparison family.

| Learner | MASE reduction (%) | CRPS reduction (%) | $p_{\mathrm{Holm}}$ MASE | $p_{\mathrm{Holm}}$ CRPS |
|---|---|---|---|---|
| Hedge | 2.52 | 2.35 | 0.006 | 0.009 |
| AdaHedge | 0.96 | 0.78 | 0.176 | 0.176 |
| BOA | 4.26 | 4.24 | $< 0.001$ | $< 0.001$ |
| FTL | 1.28 | 1.03 | 0.004 | 0.022 |
| BOA + fixed smoothing | 4.28 | 4.25 | $< 0.001$ | $< 0.001$ |

learners improved in pooled MASE and CRPS. Two-sided sign-flip tests on configuration log ratios retained both improvements for four of the five learners after Holm correction over the original twelve-comparison family, under benchmark–dataset clusters and after merging shared source datasets across benchmarks (Table 10). We used 100,000 sign draws, a plus-one Monte Carlo correction, and the same source map as the main cluster sensitivity analysis. For the separate family comparing TIMEHEDGE with the practice-informed alternatives, Holm correction likewise retained the original twelve comparisons across both metrics.

### F.4 CRPS-Learning comparisons

The row called "BOA + fixed smoothing" applies our BOA implementation to member losses and smooths adjacent level weights with $[0.25, 0.5, 0.25]$. It is a restricted local baseline inspired by CRPS-Learning (Berrisch & Ziel, 2023), not a reproduction of the full author package or its smoothing and tuning options.

**Direct author-package comparison.** We also executed the pinned R package itself on all 293 main-panel configurations, calling `online` with BOA or BEWA and all other settings at their defaults. This used uniform initialization over the eight physical members, gradient regret, horizon- and quantile-specific weights, and the default rearrangement. Inputs were divided by the pre-origin seasonal scale. Scored-only and six-round own-practice histories were separate conditions; neither received TIMEHEDGE's pooled prior. Before scoring, a metadata hash selected up to 32 candidate series per configuration, from which we took the first 16 with complete forecasts, scales and target vectors and at least 64 context observations before practice. All rules, including TIMEHEDGE, used exactly those series and all their scored origins. This yielded 2,368 series in 79, 76 and 60 configurations, respectively; the other 18, 20 and 40 configurations had no complete eligible candidate and were not replaced. The four fixed author-package variants had larger aggregate errors on each benchmark (Table 11). Perturbing current and future outcomes left the current predictions unchanged in a separate causality check.

Table 11: Direct execution of the authors' `profoc` R package (BOA or BEWA). Relative reduction of TIMEHEDGE in configuration-geometric mean error (%, positive: TIMEHEDGE is lower). The four default variants and two uniform-start grid variants used identical complete-target subsets: 79 GIFT-Eval, 76 TIME and 60 fev-bench configurations. Own practice supplied six unscored rounds.

| | GIFT-Eval | | TIME | | fev-bench | |
|---|---|---|---|---|---|---|
| Author implementation | MASE | CRPS | MASE | CRPS | MASE | CRPS |
| BOA, scored only | 3.55 | 3.71 | 1.46 | 1.47 | 6.03 | 6.35 |
| BOA, own practice + scored | 3.33 | 3.33 | 0.83 | 0.93 | 3.55 | 3.86 |
| BEWA, scored only | 3.37 | 3.46 | 1.42 | 1.39 | 6.38 | 6.54 |
| BEWA, own practice + scored | 2.85 | 2.92 | 0.99 | 1.06 | 3.88 | 4.30 |
| BOA, own practice + grid | 2.55 | 2.66 | 0.53 | 0.64 | 1.73 | 2.15 |
| BEWA, own practice + grid | 2.16 | 2.29 | 0.78 | 0.84 | 2.01 | 2.44 |

Table 12: Practice information, separated. Configuration prior (uniform or pooled practice) crossed with whether the target series' own practice rounds enter the recursion; deployment updates, learning rate and switch rate identical in all cells (rates fixed from $T = K+W$). Bottom: the registered form with the member labels of the practice rounds permuted at random per series (practice information destroyed, its amount kept; mean over three seeds). Engine MASE, convention of Table 1.

| Configuration prior | Series' practice rounds | GIFT-Eval | TIME | fev-bench |
|---|---|---|---|---|
| uniform | no | 0.6697 | 0.6707 | 0.6231 |
| pooled practice | no | 0.6551 | 0.6657 | 0.6131 |
| uniform | yes | 0.6554 | 0.6677 | 0.6106 |
| **pooled practice** | **yes (registered)** | **0.6506** | **0.6653** | **0.6092** |
| pooled practice, labels permuted | yes | 0.6748 | 0.6743 | 0.6326 |

We next fixed an 18-cell grid before its evaluation: `gamma`$\in \{0.5, 1, 2\}$, `forget_regret`$\in \{0, 0.05\}$ and quantile-smoothing $\lambda \in \{-\infty, 0, 2\}$. The package selected its next cell from past combination losses; we do not report the best cell on scored targets. Both uniform-start algorithms received six own-practice rounds, and TIMEHEDGE retained lower aggregate MASE and CRPS on all three benchmarks; with source-cluster sign-flip tests and Holm correction over three variants and two metrics, all four uniform-start comparisons passed on GIFT-Eval, and on TIME both CRPS comparisons and BEWA's MASE comparison passed. For both uniform-start grid variants, every leave-one-source-out recomputation retained lower MASE and CRPS on each benchmark; the smallest remaining reductions were $0.34\%$ and $0.51\%$, respectively. Giving each source equal weight also preserved these signs.

### F.5 Practice information, separated

Table 12 separates the two ways TIMEHEDGE reads the practice rounds, holding the deployment updates and the constants fixed: the configuration prior (uniform or pooled practice) crossed with whether the target series' own practice rounds enter the recursion. Each source alone recovered most of the gap from no practice information: on GIFT-Eval the pooled prior brought $0.6697$ to $0.6551$ and the series' own rounds to $0.6554$, and the two together to $0.6506$ (the sources overlapped, the joint gain being smaller than the sum); on TIME the pooled prior carried most of it ($0.6707$ to $0.6657$, the own rounds to $0.6677$, both $0.6653$) and on fev-bench the own rounds ($0.6231$ to $0.6106$, the prior to $0.6131$, both $0.6092$). The last row is the control: the registered form run with the member labels of every practice round permuted at random within the series (the deployment forecasts and their labels untouched, three seeds, spread at most $0.004$). It scored $0.6748$, $0.6743$ and $0.6326$, worse than no practice information at all on every benchmark: the gain of the practice phase came from the identity of the members that were right on the unscored practice rounds, not from the number of rounds the recursion was run through, and practice that pointed to the wrong members cost more than practice that was absent.

Table 13: Pool dependence: engine MASE of TIMEHEDGE and of the main simple rules when the strongest member, or its whole family, is removed from the pool (same member forecasts, every rule refit on the reduced pool); last column: the best member of that pool in hindsight.

| Benchmark | Pool | TIMEHEDGE | per-config pick | stacking | Best-3 | AdaHedge | uniform | best member |
|---|---|---|---|---|---|---|---|---|
| GIFT-Eval | 8 members (registered) | **0.6506** | 0.6543 | 0.6559 | 0.6589 | 0.6623 | 0.6770 | 0.6689 (TFM-3) |
| | w/o TimesFM-3 (7) | **0.6578** | 0.6594 | 0.6620 | 0.6649 | 0.6680 | 0.6811 | 0.6849 (PTST-FM-r2) |
| | w/o TimesFM family (6) | **0.6583** | 0.6595 | 0.6619 | 0.6650 | 0.6680 | 0.6812 | 0.6849 (PTST-FM-r2) |
| TIME | 8 members (registered) | **0.6653** | 0.6731 | 0.6684 | 0.6675 | 0.6701 | 0.6745 | 0.6710 (TFM-3) |
| | w/o TimesFM-3 (7) | **0.6724** | 0.6849 | 0.6740 | 0.6747 | 0.6759 | 0.6788 | 0.6829 (PTST-FM-r2) |
| | w/o TimesFM family (6) | **0.6731** | 0.6840 | 0.6739 | 0.6749 | 0.6764 | 0.6792 | 0.6829 (PTST-FM-r2) |
| fev-bench | 8 members (registered) | **0.6092** | 0.6259 | 0.6230 | 0.6292 | 0.6188 | 0.6561 | 0.6571 (TFM-2.5) |
| | w/o TimesFM-3 (7) | **0.6141** | 0.6344 | 0.6272 | 0.6326 | 0.6224 | 0.6579 | 0.6571 (TFM-2.5) |
| | w/o TimesFM family (6) | **0.6297** | 0.6447 | 0.6384 | 0.6404 | 0.6362 | 0.6619 | 0.6658 (TiRex-2) |

### F.6 POOL-RULE CONTRASTS

The inclusive rule (every public checkpoint of the five lineages typed for the lane) is the primary. The one-checkpoint-per-family contrast (TiRex-2, Chronos-2-synth, TimesFM-3, PatchTST-FM-r2, FlowState-r1.1) scored 0.6559 on the official harness, against 0.6502 for the inclusive pool: older generations and revisions of a family were individually weaker members, yet the combination extracted a further $0.7$–$0.9\%$ from them, because the practice rounds identified the configurations on which a superseded checkpoint was still the better forecaster. The pretrained contrast also favored the inclusive pool (0.6584 vs. 0.6540).

### F.7 POOL DEPENDENCE

Table 13 reports removal of the strongest member (TimesFM-3, the best member in hindsight on GIFT-Eval and TIME) and then its whole family (TimesFM-3 and TimesFM-2.5) from the pool, with every rule refitted on the reduced pool from the same member forecasts. TIMEHEDGE remained first among these rows on every benchmark in both reduced pools, ahead of the best remaining member by $4.1\%$ on GIFT-Eval ($0.6578$ vs. PatchTST-FM-r2 $0.6849$), $1.5$–$1.6\%$ on TIME ($0.6724$ vs. $0.6829$) and $5.7$–$7.0\%$ on fev-bench ($0.6141$ vs. TimesFM-2.5 $0.6571$, and $0.6297$ vs. TiRex-2 $0.6658$ without the family), margins of the same order as in the registered pool ($2.8\%$, $0.9\%$ and $7.9\%$). The removal reduced the pool's headroom: TIMEHEDGE lost $0.007$–$0.008$ MASE on GIFT-Eval and TIME and $0.005$ on fev-bench without TimesFM-3, and $0.021$ on fev-bench without the family, in step with the simple rules. The margins over the best simple rule were $0.0008$–$0.0016$ on GIFT-Eval and TIME in the reduced pools (per-configuration pick and stacking) and $0.0083$ and $0.0065$ on fev-bench (AdaHedge), so the statement that survived the pool change was the one made in §6.2: several practice-informed rules were close, and TIMEHEDGE retained the lowest aggregate score on every benchmark in both reduced pools.

**Removing every lineage in turn.** We extended the check to all five lineages, removing each in turn and recomputing the closure, pooled prior, practice pick and physical Best-3. The fixed comparison included the original Hedge update warmed by the target's own practice, with shared physical-member weights and with per-level weights including the closure. No removed lineage was selected as the preferred pool. All pools and rules were scored on the same $701{,}867$ origins in $292$ configurations ($284{,}193$ series–configuration pairs), using the origin-scale replay and common-support criterion of Appendix F.9. TIMEHEDGE had lower aggregate MASE and CRPS than Uniform, the practice pick, Best-3 and own-practice Hedge in every removal–benchmark combination, with a single $0.02\%$ exception for Best-3 on GIFT-Eval MASE (Table 14). Pooled across benchmarks with shared sources merged, all $20$ contrasts (five removals, two comparators, two metrics) survived a joint Holm correction of source-cluster sign-flip tests ($p_{\mathrm{Holm}} < 0.001$). The signs also held when restricted to the $283{,}328$ series–configuration pairs with every scored origin present.

### F.8 ATTRIBUTION AND PRACTICE DIAGNOSTICS

Figures 2 and 10 show member weights and contributions for a representative series; Appendix B defines the corresponding quantities.

Table 14: All five lineage removals, on identical scored support across pools and rules. Entries are the minimum and maximum relative error reduction of TIMEHEDGE (%) across removing TiRex, Chronos, TimesFM, PatchTST or FlowState. Positive means lower TIMEHEDGE error. These are ranges across pools, not confidence intervals. Own-practice Hedge used either one shared weight vector over the remaining physical members or separate weights at each quantile level with a recomputed closure.

| Comparator | Metric | GIFT-Eval | TIME | fev-bench |
|---|---|---|---|---|
| Uniform | MASE | $[1.25, 3.93]$ | $[0.90, 1.60]$ | $[4.87, 7.63]$ |
| | CRPS | $[1.37, 3.93]$ | $[0.97, 1.75]$ | $[4.93, 7.37]$ |
| Pooled practice pick | MASE | $[0.24, 0.72]$ | $[1.04, 1.60]$ | $[1.95, 2.55]$ |
| | CRPS | $[0.47, 0.99]$ | $[1.15, 1.90]$ | $[2.02, 2.63]$ |
| Physical Best-3 | MASE | $[-0.02, 1.96]$ | $[0.39, 0.78]$ | $[2.02, 5.08]$ |
| | CRPS | $[0.30, 2.03]$ | $[0.44, 0.88]$ | $[1.93, 4.77]$ |
| Hedge, own (per-level) | MASE | $[0.70, 2.07]$ | $[0.21, 0.74]$ | $[1.71, 2.89]$ |
| | CRPS | $[0.76, 2.18]$ | $[0.22, 0.89]$ | $[1.85, 2.97]$ |
| Hedge, own (shared) | MASE | $[0.70, 1.27]$ | $[0.16, 0.48]$ | $[1.18, 1.87]$ |
| | CRPS | $[0.78, 1.27]$ | $[0.17, 0.57]$ | $[1.14, 1.62]$ |

**Misleading practice (stress test).** To see what each rule did when the practice rounds pointed to the wrong member, we relabelled the members on the scored origins of TIME by a cyclic shift ($m \to m{+}1$) and left the practice rounds untouched: the alignment of practice-fitted rankings with member identities was disrupted at deployment, while the set of forecasts at each origin, hence the uniform mixture, the median and the hindsight oracles, was unchanged. Relabelling from the first scored window degraded the practice-fitted static rules by 0.9–3.8% MASE (BMA $\times 1.009$, stacking $\times 1.014$, Best-3 $\times 1.019$, per-configuration practice pick $\times 1.038$), left the scored-only online learners unchanged ($\times 1.000$), and degraded TIMEHEDGE by $\times 1.016$ (without the uniform-start components: $\times 1.022$); relabelling from the tenth scored window left every rule within $\times 1.006$ of its unperturbed score (TIMEHEDGE $\times 1.003$, stacking $\times 1.003$, Best-3 $\times 1.001$). The uniform-start components of Corollary 1 reduced the degradation in the first intervention.

**The practice leader in numbers.** On the GIFT-Eval test panel (127,516 series, first scored window, level-mean pinball loss over the seasonal scale) the member with the lowest practice loss was the best member of the scored window for $0.176$ of the series (chance $0.125$). When it was, it outperformed the uniform mixture by $0.102$; when it was not, it lost $0.042$, so per series the practice leader was $0.017$ worse than the uniform mixture ($0.587$ against $0.570$; the hindsight best member scored $0.477$). Pooled over the series of a configuration and aggregated as the benchmark does (geometric mean over configurations), the same leader was $2.6\%$ better than the uniform mixture and won on $47\%$ of the configurations. The sign of the practice signal depended on the aggregation, which is why TIMEHEDGE hedges over it instead of following it, and why its prior is pooled. The hindsight second-best member of each series scored $0.517$, below the uniform mixture, but this is not a deployable selection rule. Holding the practice leader's conditional excess loss on wrong picks fixed ($+0.133$ over the best member), the diagnostic break-even agreement was $0.30$, against the $0.18$ observed.

**The code length in numbers.** The pooled prior $\pi_c$ was sharp: in $52$ of the $97$ configurations its largest weight exceeded $0.9$, and on average it left four of the nine members below $10^{-6}$, so the hindsight-best member of the first scored window sat at the numerical floor of $\pi_c$ on $29\%$ of series. This is what the uniform-start components of Corollary 1 are for: with them every member's code length entering the scored phase is finite, and after the series' own practice rounds the hindsight-best member's code length (median level) was $2.39$ nats (natural-log units) on average, median $2.29$, against $\ln 9 = 2.20$ under the uniform prior: $1.64$ nats when the practice leader was the eventual best member, $2.56$ when it was not, the signed correction of Theorem 1 in numbers.

**Played loss and raw loss.** The learner plays, at each level, the pinball loss of the round shifted by the round's best member, divided by the median absolute loss across members and capped at $b{=}4$ (Algorithm 2). Proposition 1, Theorem 1, Corollary 1 and the certificate T5 are statements about this

Table 15: Downside of per-series CRPS relative to Uniform on common scored origins. Rate: source-balanced percentage of series–configuration pairs with loss ratio above 1.25; tail: mean ratio in the worst 10% of that weighted distribution. Lower is better. All eleven fixed rules are shown.

| | GIFT-Eval | | TIME | | fev-bench | |
|---|---|---|---|---|---|---|
| Rule | Rate (%) | Tail | Rate (%) | Tail | Rate (%) | Tail |
| Uniform | 0.00 | 1.000 | 0.00 | 1.000 | 0.00 | 1.000 |
| Pooled practice pick | 3.64 | 1.352 | 2.75 | 1.201 | 4.81 | 1.310 |
| Physical Best-3 | 1.47 | 1.149 | 0.23 | 1.099 | 0.54 | 1.099 |
| Own-practice pick | 3.81 | 1.353 | 3.00 | 1.250 | 4.90 | 1.314 |
| Hedge, own practice (shared-8) | 1.44 | 1.134 | 0.31 | 1.059 | 1.08 | 1.140 |
| Hedge, own practice (level-9) | 0.86 | 1.078 | 0.22 | 1.031 | 0.56 | 1.085 |
| AdaHedge, own practice (shared-8) | 1.21 | 1.127 | 0.96 | 1.096 | 1.08 | 1.138 |
| AdaHedge, own practice (level-9) | 0.97 | 1.107 | 0.82 | 1.085 | 0.79 | 1.115 |
| AdaHedge, pooled-prior ablation | 1.98 | 1.200 | 1.34 | 1.130 | 1.68 | 1.177 |
| Frozen TIMEHEDGE | 0.39 | 1.062 | 0.16 | 1.062 | 0.85 | 1.099 |
| TIMEHEDGE | 0.38 | 1.059 | 0.02 | 1.056 | 0.72 | 1.083 |

capped, median-normalized surrogate. Proposition 2(ii), Jensen's inequality and the sorting step, is a statement about the pinball losses themselves at each level and outcome, not about the surrogate; the exact bridges from the surrogate regret to the raw and seasonal-scaled scores carry the round-wise factors $s_t d_t$ (seasonal scale times median normalizer) and the clipping residual, and are reported as measured quantities (Appendix A.9), not as a further guarantee. The shift by the round's minimum is common to all members and cancels in every regret. On the registered runs, counting only the most recent available loss round in each prediction-time recursion call, the cap was active on $0.79\%$ of the (round, member, level) emissions on GIFT-Eval, $0.27\%$ on fev-bench and $0.23\%$ on TIME.

**TIME in the benchmark's own normalization.** The TIME columns of Tables 1 and 2 score the benchmark through the same engine as GIFT-Eval and fev-bench. For reference, the TIME evaluation code itself (MASE with the benchmark's NaN-compacted seasonal scale, CRPS normalized by the window's mean absolute target, geometric means over all 98 configurations, and the configuration prior chained from a pool-level leave-one-configuration-out prior, Appendix D) yielded: TIMEHEDGE $0.8387$ MASE / $0.1877$ CRPS; the best member, TimesFM-3, $0.8459$ / $0.1881$; the uniform mixture $0.8506$ / $0.1918$; its own implementations of BMA $0.8442$, BOA $0.8474$ and follow-the-leader $0.8500$; the hindsight oracles $0.7122$ (per window) and $0.8314$ (per configuration). Every member had forecasts on every configuration; the engine admits a configuration when at least six members are present.

### F.9 Large per-series losses

**Main-panel diagnostic.** Small differences in aggregate error can conceal different distributions of per-series losses. We replayed all 293 materialized configurations on common scored origins: every physical member must have finite forecasts, the seasonal scale must be positive and finite, and at least one target step must be observed. This retained $284{,}193$ series–configuration pairs and $701{,}867$ scored origins in 292 configurations; weekly Saugeen had no common scored support. Each series was averaged over its retained origins. We then formed its CRPS ratio to Uniform and gave equal weight to sources, configurations within a source, and series within a configuration. Zero Uniform losses were excluded from ratios for every rule.

TIMEHEDGE exceeded Uniform's per-series CRPS by more than $25\%$ on $0.38\%$, $0.02\%$ and $0.72\%$ of this weighted support, versus $3.64\%$, $2.75\%$ and $4.81\%$ for the pooled practice pick. The reductions against the pick survived source-level sign-flip tests with Holm correction over the 24 prespecified contrasts per benchmark (four primary controls, two metrics, three downside summaries; $p_{\mathrm{Holm}} \leq 0.035$).

**How many practice rounds, and how sensitive are the constants.** Table 16 reports variations in the number of practice rounds and the three constants of the registered form, everything else fixed. With $k$ the number of most recent practice rounds that enter both the pooled prior and the series'

Table 16: Practice-round count and constant sensitivity on the final 8-member pool (engine MASE and scaled CRPS, convention of Table 1). Top: only the $k$ most recent practice rounds enter the pooled prior and the series' recursion ($k$=6 is the registered form; $k$=0 leaves a uniform prior and the scored rounds). Bottom: each row changes one constant and fixes the others.

| | GIFT-Eval | | TIME | | fev-bench | |
|---|---|---|---|---|---|---|
| Setting | MASE↓ | CRPS↓ | MASE↓ | CRPS↓ | MASE↓ | CRPS↓ |
| *Practice rounds used* | | | | | | |
| $k$=0 (uniform prior, scored rounds only) | 0.6684 | 0.7528 | 0.6705 | 0.6825 | 0.6216 | 0.7789 |
| $k$=1 | 0.6558 | 0.7399 | 0.6680 | 0.6800 | 0.6103 | 0.7660 |
| $k$=2 | 0.6538 | 0.7371 | 0.6670 | 0.6788 | 0.6081 | 0.7638 |
| $k$=3 | 0.6522 | 0.7349 | 0.6671 | 0.6787 | 0.6077 | 0.7630 |
| $k$=4 | 0.6522 | 0.7345 | 0.6665 | 0.6780 | 0.6091 | 0.7642 |
| $k$=5 | 0.6510 | 0.7334 | 0.6656 | 0.6770 | 0.6092 | 0.7640 |
| $k$=6 (registered form) | **0.6506** | **0.7329** | **0.6653** | **0.6770** | **0.6092** | **0.7643** |
| *Constants (registered: $b$=4, $\eta = \sqrt{8 \ln M/T}$, $\alpha = 1/T$, $T = K+W$)* | | | | | | |
| loss cap $b$=2 | 0.6506 | 0.7329 | 0.6653 | 0.6770 | 0.6088 | 0.7638 |
| loss cap $b$=8 | 0.6506 | 0.7329 | 0.6653 | 0.6770 | 0.6096 | 0.7649 |
| no loss cap | 0.6506 | 0.7329 | 0.6653 | 0.6770 | 0.6096 | 0.7649 |
| learning rate $\eta \times 0.5$ | 0.6527 | 0.7356 | 0.6653 | 0.6771 | 0.6113 | 0.7668 |
| learning rate $\eta \times 2$ | 0.6499 | 0.7319 | 0.6659 | 0.6776 | 0.6085 | 0.7637 |
| switch rate $\alpha \times 0.5$ | 0.6498 | 0.7322 | 0.6653 | 0.6770 | 0.6093 | 0.7645 |
| switch rate $\alpha \times 2$ | 0.6516 | 0.7340 | 0.6655 | 0.6772 | 0.6093 | 0.7645 |

recursion, the first round bought most of the value: on GIFT-Eval $k$=1 recovered 0.013 of the 0.018 MASE that separates no practice ($k$=0, a uniform prior updated on the scored rounds only) from the registered $k$=6, and the remaining rounds added 0.005 with diminishing returns; on TIME the curve was shallow (0.6705 → 0.6653, with one 0.0001 reversal at $k$=3); on fev-bench it saturated at $k$=2–3 and the oldest rounds cost 0.0014, the practice→scored agreement decaying with the distance of a round from the scored origin (§3). The registered $k$=6 was the practice budget of every panel, fixed before any test-set evaluation, not a tuned value; series with shorter histories used fewer rounds (such series occurred in 10, 5 and 30 configurations of GIFT-Eval, TIME and fev-bench). The loss cap was immaterial: $b \in \{2, 4, 8, \infty\}$ moved no cell by more than 0.0006, consistent with the cap binding on about 1% of the emissions or fewer (above). The learning rate was the only constant with a visible effect, and only when halved: $\eta \times 0.5$ cost 0.0022 on GIFT-Eval and 0.0022 on fev-bench and nothing on TIME (differences from unrounded values), while $\eta \times 2$ and the switch rate at $\alpha \times 0.5$ or $\alpha \times 2$ stayed within 0.0011 of the registered form in either direction. These constants were fixed in advance ($\eta$ and $\alpha$ from the planned $T$, $b$ as the cap of the bounded-loss analysis) and the table evaluates their sensitivity.

**What predicts the value of the scored updates.** For every configuration with at least 20 series that have complete practice and scored losses (155 of the 293: 65 GIFT-Eval, 41 TIME, 49 fev-bench) we computed three log-ratios of engine MASE from Table 2's rows: the value of the scored updates (frozen practice posterior over full), the value of hedging over the practice posterior rather than selecting from it (practice selection over frozen) and the value of the pooled prior (flat prior over full), and related them to the configuration's practice→scored agreement (the share of its series whose practice leader is the best member of the first scored window), the practice leader's excess loss, the sharpness of $\pi_c$ and the number of scored windows $W$ (Figure 3). Pooled over the three benchmarks the value of the updates rose with agreement (Spearman $\rho$=0.33, $p<10^{-4}$; GIFT-Eval 0.47, fev-bench 0.35, TIME 0.01) and with $W$ ($\rho$=0.28; TIME 0.47), and was unrelated to the sharpness of the prior ($\rho$=−0.03): in the lowest agreement third the updates were worth +0.07% MASE and helped on half of the configurations, in the highest +0.39% and on 79%. The value of hedging over the posterior moved the other way ($\rho$=−0.31 with agreement, $p<10^{-4}$), and so did the value of the prior in reverse ($\rho$=+0.27). Where the practice signal was right, prior and deployment rounds compounded it; where it was wrong, hedging rather than selecting was what protected the forecast. The number of windows mattered where the signal was reliable (TIME) and not where it was not (fev-bench, $\rho$=−0.04).

Table 17: Effect sizes behind the protocols in Table 1: relative MASE gap of each same-pool rule to TIMEHEDGE (geometric mean of the per-configuration ratio minus one, in percent; positive favors TIMEHEDGE) with 95% dataset-cluster bootstrap intervals, for each benchmark and all 293 configurations pooled.

| Rule | GIFT-Eval | TIME | fev-bench | Pooled |
|---|---|---|---|---|
| Best member per config (practice) | +0.6 [+0.0, +1.2] | +1.2 [+0.3, +2.1] | +2.8 [+1.6, +4.1] | +1.5 [+0.9, +2.2] |
| Best member overall (practice) | +9.1 [+5.6, +13.1] | +0.9 [+0.3, +1.6] | +11.3 [+6.4, +16.8] | +7.0 [+4.8, +9.8] |
| Uniform | +4.1 [+1.5, +7.4] | +1.4 [+0.5, +2.0] | +7.7 [+2.8, +13.7] | +4.4 [+2.5, +6.7] |
| Median | +4.0 [+1.0, +7.8] | +1.1 [+0.4, +1.7] | +7.4 [+2.3, +13.6] | +4.1 [+2.2, +6.8] |
| Inverse-MSE weights | +2.7 [+1.0, +5.0] | +1.1 [+0.3, +1.6] | +3.3 [+1.0, +6.5] | +2.4 [+1.3, +3.7] |
| BMA | +2.0 [+0.8, +3.4] | +0.7 [+0.2, +1.1] | +3.8 [+1.2, +7.3] | +2.2 [+1.1, +3.4] |
| Stacking (ridge) | +0.8 [+0.2, +1.5] | +0.5 [+0.2, +0.8] | +2.3 [+0.7, +4.5] | +1.2 [+0.6, +2.0] |
| Best-3 mixture | +1.3 [-0.0, +2.9] | +0.3 [-0.0, +0.6] | +3.3 [+0.9, +6.6] | +1.6 [+0.7, +2.9] |
| Hedge | +3.3 [+1.1, +6.0] | +1.0 [+0.3, +1.5] | +4.5 [+1.5, +8.7] | +2.9 [+1.6, +4.7] |
| AdaHedge | +1.8 [+0.5, +3.3] | +0.7 [+0.2, +1.1] | +1.6 [+0.0, +3.4] | +1.4 [+0.6, +2.2] |
| BOA | +4.9 [+2.1, +8.5] | +2.4 [+1.4, +3.3] | +9.4 [+4.2, +15.9] | +5.6 [+3.5, +8.1] |
| FTL | +2.1 [+1.1, +3.2] | +1.6 [+1.0, +2.1] | +2.4 [+1.2, +3.9] | +2.0 [+1.5, +2.8] |
| BOA + fixed smoothing | +4.8 [+2.0, +8.5] | +2.4 [+1.4, +3.2] | +9.4 [+4.1, +16.0] | +5.6 [+3.4, +8.1] |
| TimeRouter† | +2.1 [+0.6, +3.7] | +2.4 [+1.4, +3.7] | +1.6 [+0.7, +2.5] | +2.0 [+1.3, +2.8] |
| GBM router | +2.4 [+1.2, +3.9] | +2.6 [+1.8, +3.5] | +3.4 [+1.8, +5.5] | +2.8 [+2.0, +3.7] |

The same pattern held within configurations, series by series. Splitting the series with at least two scored windows by whether their practice leader is the best member of the first scored window, and measuring the value of the updates on the later windows (where frozen and full first differ), the updates were worth 0.0023 MASE on the series whose practice leader was right and 0.0011 on the others on GIFT-Eval (52,709 series, difference 0.0012 with configuration-cluster bootstrap interval $[0.0005, 0.0019]$; the updates helped 68% against 59% of the series); on TIME the difference had the same sign (0.0025 against 0.0015, interval $[-0.0010, 0.0023]$) and on fev-bench, with 7,955 eligible series and heavy tails, it was not resolved. Finally, the reliability of the practice signal could be read from the practice rounds themselves: the share of a configuration's series whose leader on practice rounds $1, \ldots, K-1$ is the best member on practice round $K$, a quantity that uses no scored data, correlated with the realized practice→scored agreement at Spearman $\rho$=0.46 (GIFT-Eval), 0.62 (TIME) and 0.54 (fev-bench), and with the realized value of the pooled prior at $\rho$=0.34 pooled (0.35, 0.35 and 0.30 per benchmark) and of the updates at $\rho$=0.20 ($p$=0.01).

**Effect sizes with uncertainty.** Table 17 reports relative MASE differences for the combination rules with a 95% dataset-cluster bootstrap interval. The clusters of the permutation tests used (benchmark, dataset name) pairs, 167 of them; the same source series appear at several sampling frequencies (fev-bench's ETT at four) and some sources in two benchmarks. Merging clusters by source across frequencies and benchmarks (117 clusters) left the reported scored-only protocol comparisons at $p \leq 0.002$, all surviving Holm and Bonferroni corrections (adjusted $p \leq 0.044$, retaining the original 28-comparison family). Pooled over the three benchmarks every interval excluded zero, the closest rules being stacking (+1.2%, $[0.6, 2.0]$), AdaHedge (+1.4%), the per-configuration practice pick (+1.5%) and Best-3 (+1.6%), and the median configuration gained 0.1–1.2% against each rule.

**Calibration and sharpness of the reported quantiles.** Table 18 evaluates the quantiles on the common observation set, the target points at which TIMEHEDGE and all eight members had finite forecasts (39.8, 7.7 and 8.9 million points on GIFT-Eval, TIME and fev-bench), with configurations weighted equally as the benchmarks aggregate. Without any calibration step, every level TIMEHEDGE reported was within 0.015 of its nominal value on all three benchmarks (0.014, 0.005, 0.015); the uniform mixture was similarly calibrated (0.018, 0.006, 0.011), and the strongest member under-covered the upper levels on GIFT-Eval and fev-bench (0.023, 0.033). Per configuration, the mean absolute coverage error over the nine levels was 0.053, 0.029 and 0.054 for TIMEHEDGE against 0.058, 0.031 and 0.055 for the uniform mixture and 0.044, 0.030 and 0.056 for the best member. Sharpness was where the combination differed: TIMEHEDGE's 80% interval was the narrowest on every benchmark, 7%, 4% and 6% narrower than the uniform mixture's and 10%, 3% and 11% narrower than the best member's, and its interval score was the best on TIME and fev-bench and within 0.1% of the best member's on GIFT-Eval. Hedging over the members therefore sharpened the distribution without paying in calibration.

Table 18: Calibration and sharpness on the common observation set (target points at which TIMEHEDGE and all eight members had finite forecasts; configurations weighted equally as the benchmarks aggregate). Coverage: share of observed points below the reported quantile at each level. $|\Delta|$: mean over configurations of the mean absolute coverage error over the nine levels. Width and score: the $80\%$ interval (levels $0.1$ and $0.9$), width and interval score (Winkler, $\alpha = 0.2$) divided by the series' seasonal scale, mean over windows then over configurations; lower is better. Best member: TimesFM-3 (GIFT-Eval, TIME), TimesFM-2.5 (fev-bench).

| Benchmark | Forecast | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | $\lvert\Delta\rvert$ | width | score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GIFT-Eval | TIMEHEDGE | 0.094 | 0.200 | 0.309 | 0.413 | 0.514 | 0.611 | 0.705 | 0.801 | 0.894 | 0.053 | **3.020** | **7.792** |
| | uniform mixture | 0.091 | 0.195 | 0.307 | 0.415 | 0.518 | 0.615 | 0.709 | 0.805 | 0.900 | 0.058 | 3.248 | 7.857 |
| | best member (TimesFM-3) | 0.088 | 0.179 | 0.277 | 0.378 | 0.479 | 0.581 | 0.685 | 0.791 | 0.892 | **0.044** | 3.358 | 7.797 |
| TIME | TIMEHEDGE | 0.098 | 0.200 | 0.303 | 0.402 | 0.500 | 0.598 | 0.696 | 0.795 | 0.895 | **0.029** | **3.063** | **4.764** |
| | uniform mixture | 0.094 | 0.197 | 0.300 | 0.401 | 0.500 | 0.598 | 0.698 | 0.798 | 0.899 | 0.031 | 3.182 | 4.823 |
| | best member (TimesFM-3) | 0.096 | 0.197 | 0.299 | 0.398 | 0.495 | 0.595 | 0.695 | 0.796 | 0.898 | 0.030 | 3.161 | 4.810 |
| fev-bench | TIMEHEDGE | 0.100 | 0.196 | 0.296 | 0.397 | 0.498 | 0.597 | 0.692 | 0.787 | 0.885 | **0.054** | **7.556** | **15.905** |
| | uniform mixture | 0.094 | 0.193 | 0.296 | 0.395 | 0.496 | 0.596 | 0.692 | 0.789 | 0.890 | 0.055 | 8.076 | 16.424 |
| | best member (TimesFM-2.5) | 0.089 | 0.179 | 0.268 | 0.367 | 0.476 | 0.580 | 0.678 | 0.774 | 0.880 | 0.056 | 8.476 | 17.452 |

# G POST-HOC ATTRIBUTION WITHOUT OPENING THE MEMBERS

TIMEHEDGE plays predictive marginals of an exponential-weights posterior over expert paths, so its explanations are identities of the learner rather than approximations of the members, and they are available for every quantile level and horizon step. Three views are exact and can be computed from the forecasts and saved weights (Figure 10); a fourth uses the practice phase as a sandbox for counterfactuals that never touch a scored round.

**Which member, at which level and step.** One learner runs per level, so the played weights $w_m(\tau)$ exist per level: for a prediction interval $[Q(0.1), Q(0.9)]$ the two learners say which members support its lower and its upper end (Figure 10(a)). Per horizon step the combination is linear before the rearrangement, so the shares $\varphi_m(\tau, h) = w_m(\tau)\big(Q_m(\tau, h) - Q_{M+1}(\tau, h)\big)$ (T3, Appendix B) decompose the reported quantile at every $(\tau, h)$ exactly; where the rearrangement permutes two levels the shares are reported for the levels before sorting together with the permutation.

**Which rounds moved the trust.** Within a static component $\ln w_m = \ln \pi_c(m) - \eta \sum_t \tilde{\ell}_t(m) + \mathrm{const}$ (T1), so the contribution of every practice or scored round to every member's log-weight is the number $-\eta\tilde{\ell}_t(m)$ (Figure 10(b)); for the switching components the forward messages give the same per-round evidence. This answers "why is this member trusted now" by pointing at the rounds on which it earned or lost that trust, without any access to the member.

**Which part of the input matters: the practice phase as a sandbox.** What the members do with their input is not visible to the learner, but the practice rounds are unscored and leakage-free, so any counterfactual on the input (a shortened context, an occluded segment, a removed covariate) can be priced on them: rerun the members on the modified practice contexts and compare the practice losses, member by member and for the combination played with its own weights. No scored round is used, so the procedure is admissible at deployment. Appendix G.1 reports one such experiment.

## G.1 A PRACTICE-PHASE SANDBOX: PRICING THE CONTEXT LENGTH

Figure 11 varies context length on one configuration, retaining the pooled prior fitted on full contexts. Panel (a) shows all member and combination losses at the five settings, whose member rankings changed with context length. TIMEHEDGE's practice loss ordered these settings in the same way as its scored MASE ($1.53$, $1.29$, $0.94$, $0.91$, $0.90$). The comparison used practice losses to assess the input intervention before any scored outcome was revealed. The same measurement procedure can be applied to an occluded window, a removed covariate or a rescaled history.

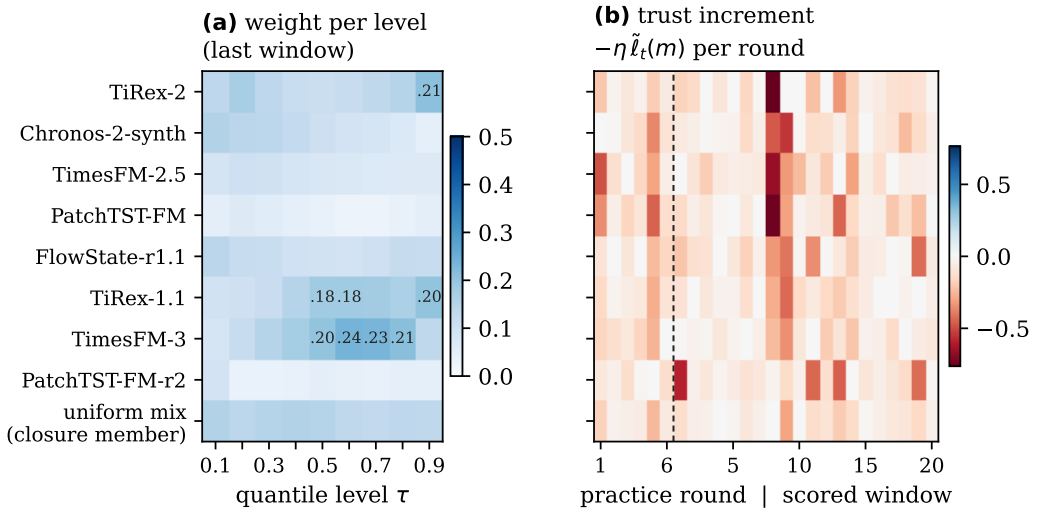


Figure 10: Exact attributions for one series (GIFT-Eval, electricity at 15-minute resolution, the series of Figure 2). (a) Played weight per quantile level on the last scored window: TimesFM-3 carried the centre of the distribution, TiRex-2 the upper tail, TiRex-1.1 the upper half. (b) Trust increment $-\eta\tilde{\ell}_t(m)$ of every round (practice rounds, then scored windows) for every member: the exact per-round contribution to the log-weights (T1); the deep-red column is the window on which most members missed a level shift and the closure member gained.

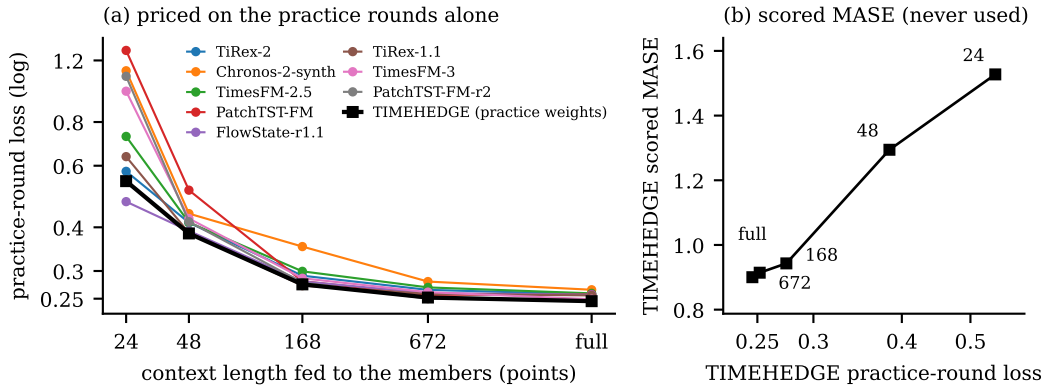


Figure 11: The practice phase as a sandbox (GIFT-Eval, electricity at hourly resolution, short horizon: 370 series, 6 practice rounds each, 2,190 rounds present at every setting). The eight members were rerun with their context truncated to the last $L$ points and the engine was rerun on each truncated pool. (a) Mean practice-round loss (level-mean pinball loss over the seasonal scale) of every member and of TIMEHEDGE played with its own practice weights, computed on the unscored practice rounds only. (b) TIMEHEDGE's scored MASE at the same settings, which the sandbox never saw: the practice-round ordering of the settings matched the scored ordering.