ONLINE FORECAST AGGREGATION WITH QUANTILE-ANCHORED TAIL PROTECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A source forecast can improve expected continuous ranked probability score (CRPS) while substantially distorting the probability assigned to extreme events. We propose a quantile-anchored source transfer method (QAST), which lets source models reshape a baseline forecast only within cells defined by its tail quantiles, while preserving the baseline's exceedance probability and protected quantiles. Any convex combination of these forecasts keeps the combined-ratio supremum distance (CRSD) at a fixed threshold within a user-chosen radius of the baseline's CRSD, for every outcome sequence and every choice of weights. Within the protected class, we establish sublinear and constant regret for the aggregate forecast's finite-node cell-conditional square score against fixed transferred experts. An explicit iid construction isolates within-cell tail-shape misspecification and shows improvement in expected CRPS and limiting CRSD. On EUPP-Bench, QAST reduces annual proper-score losses relative to unprotected mixtures using the same weights under two baselines.

1 INTRODUCTION

Transferring a probabilistic forecast changes more than its average accuracy. A source model may improve predictions for common outcomes while assigning too much probability to extreme events. Example 2.1 shows that a forecast can have lower expected continuous ranked probability score (CRPS) while assigning twice the true exceedance probability. We study how to borrow information from source forecasts while preserving specified features of a baseline's tail distribution.

Our approach is to constrain the forecast distribution before learning how to combine the sources. We use the baseline's tail quantiles to partition the outcome space, retain the probability of each cell, and transfer the source distribution within each cell. We also retain a positive baseline component. This last step ensures that the transferred cumulative distribution function (CDF) crosses each protected probability level at the same point as the baseline, even when a source CDF has a flat region. We call this construction *quantile-anchored source transfer* (QAST). Figure 1 illustrates the construction: the baseline fixes cell probabilities, sources supply within-cell shapes, and online learning combines the protected CDFs. We measure tail changes using the combined-ratio supremum distance (CRSD), a finite-threshold diagnostic defined in Section 2.

Our main contribution is a baseline-anchored forecast class that is closed under convex aggregation and admits a weight-independent pathwise CRSD bound. Every aggregate forecast preserves the baseline exceedance probability and protected tail counts, while its CRSD remains within $(1-\kappa)h$ of the baseline's at every finite prefix, for every outcome path. Here κ controls baseline retention and h is the largest probability spacing between protected quantiles. The bound is sharp for uniform grids with $J \geq 2$.

We next show how to learn within the protected class. Since the candidates agree at the protected cuts, we use a finite-node cell-conditional Brier score to compare their conditional probabilities inside the cells. QAST-E schedules its learning rate by the number of observed exceedances using successive epochs containing $1, 2, 4, \ldots$ exceedances (the event-doubling schedule). Exponential weighting then gives a regret bound of order $\sqrt{N}$ after N exceedances. For the same square score, a fixed-rate rule without restarts gives a constant regret bound against each fixed transferred-expert label. Both results apply to the score of the aggregate forecast. The protection guarantee comes from the CDF construction and is common to the two learning rules.

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.

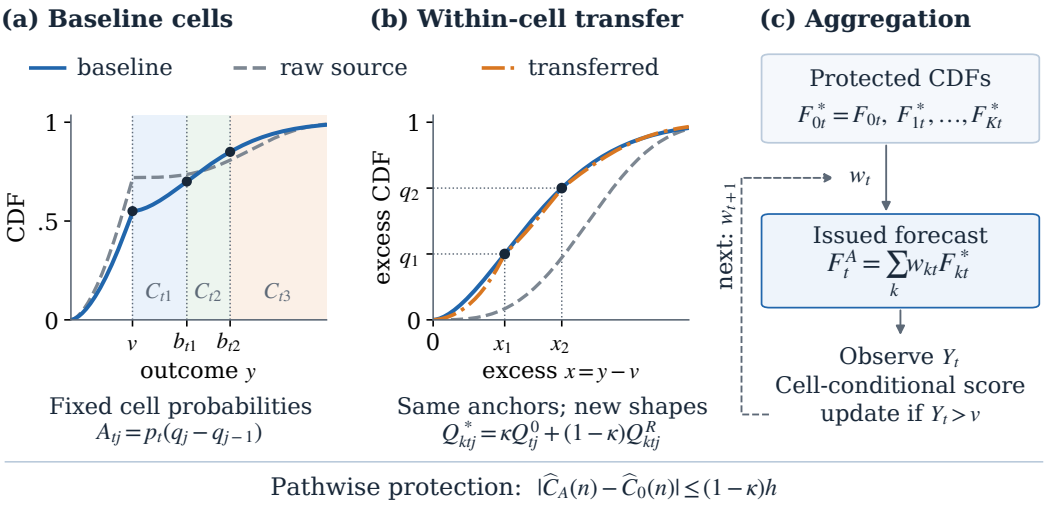

Figure 1: QAST with three tail cells and $\kappa = 0.5$ (illustrative CDFs). (a) Baseline cuts define cells C_{tj} with probabilities $A_{tj} = p_t(q_j - q_{j-1})$; the final cell extends to infinity. (b) Transfer changes within-cell shapes while preserving the anchors $(b_{tj} - v, q_j)$. (c) Weights chosen before the outcome combine the protected CDFs, and exceedance feedback updates the weights. Here $h = \max_j(q_j - q_{j-1})$. The conditional bodies coincide in this illustration; Section 3.1 gives the general construction.

Protection still leaves room for useful transfer. We construct an iid example in which the baseline and source have identical, correct bodies but imperfect tail shapes. QAST-E learns to use the transferred expert in a time-averaged sense and improves both expected CRPS and CRSD. A 2018 EUPPBench case study compares three tail partitions at the common protection radius $(1-\kappa)/J = 0.05$. Under both ensemble model output statistics (EMOS) and distributional regression network (DRN) baselines, QAST substantially lowers proper-score losses relative to unprotected mixtures with matched nominal baseline retention, and fixed protected mixtures already account for most of this reduction. The case study illustrates protection against negative transfer from regional source models (Section 5).

1.1 Related work

Proper scoring rules evaluate complete predictive distributions (Gneiting & Raftery, 2007); however, tail behavior needs separate attention. Brehmer & Strokorb (2019) show that near-optimal expected proper scores need not identify tail properties such as the extreme-value index. Threshold- and quantile-weighted scores emphasize regions of interest, including tails, while retaining propriety (Gneiting & Ranjan, 2011). Allen et al. (2025) develop tail calibration and diagnostics based on occurrence, severity, and combined ratios. We use their conditional excess distributions and excess probability integral transform (PIT) framework at a fixed finite threshold, asking how source transfer and convex aggregation preserve specified baseline probabilities and bound changes in CRSD pathwise. Wessel et al. (2026) incorporate tail-calibration quantities into training objectives. QAST addresses class-level preservation by constraining candidate forecasts before learning their weights.

Forecast combination involves choices about both distributional representation and weighting. Gneiting & Ranjan (2013) study the calibration and dispersion of combined forecasts, while Schulz et al. (2026) compare probability- and quantile-based aggregation for deep ensembles. Berrisch & Ziel (2023) learn combination weights that vary over time and quantile level, and Aastveit et al. (2026) use quantile-specific weights to exploit heterogeneous predictive performance across the distribution. QAST instead constructs a baseline-anchored forecast class before aggregation and then combines the transferred CDFs using weights common across CDF query positions.

For online probabilistic aggregation, V'yugin & Trunov (2019) derive horizon-independent CRPS regret for finitely many experts whose forecasts and outcomes lie in a fixed bounded interval. Their

confidence-based extension allows expert-specific confidence levels supplied before the outcome is observed (V'yugin & Trunov, 2022). QAST-E keeps all transferred experts in the pool at each date and updates their weights only after observed exceedances. Its regret bound concerns the finite-node cell-conditional Brier/square score in Section 4, whereas the baseline-relative protection follows from the forecast construction independently of the weighting rule.

The partition-reweighting step is closely related to Jeffrey's rule, which reassigns partition probabilities while retaining within-cell conditional laws (Diaconis & Zabell, 1982). QAST uses the same partition-based decomposition but fixes baseline-quantile cell probabilities and modifies conditional shapes subject to the constraints required for pathwise protection. Spatial postprocessing motivates borrowing source information across locations (Dabernig et al., 2017; Friedli et al., 2021).

Our weight updates use standard prediction-with-expert-advice tools (Freund & Schapire, 1997). Online conformal prediction (Bhatnagar et al., 2023; Angelopoulos et al., 2023; 2024a) and conformal risk control (Angelopoulos et al., 2024b) provide related sequential-coverage and expected-risk guarantees, respectively. These approaches adjust prediction sets or risk-control parameters. QAST operates on complete predictive CDFs and preserves selected baseline probabilities during aggregation.

2 Forecasts and CRSD

Forecasting protocol. Let $K \geq 1$ be the fixed number of sources. At each date t, the baseline CDF F_{0t}, source CDFs $P_{1t}, \ldots, P_{Kt}$, and aggregation weights are determined from past observations and currently available inputs, before $Y_t \geq 0$ is observed. For the online-learning results, Y_t is returned before date $t+1$ and is used to score every candidate issued at date t. Let $\mathscr{F}_0$ denote the information available before the run; Appendix A.1 gives the formal information and feedback protocol.

We fix a threshold $v \geq 0$ before evaluation (formally, v is $\mathcal{F}_0$-measurable).

Assumption 2.1 (Forecast regularity)**.** *At every issue time t, the baseline and raw-source forecasts are probability CDFs on $[0,\infty)$, continuous on (v,∞), and the baseline exceedance probability satisfies $p_t = 1 - F_{0t}(v) > 0$. Its excess CDF*

$$E_{0t}(x) = \frac{F_{0t}(v+x) - F_{0t}(v)}{p_t}, \qquad x \geq 0,$$

is continuous and strictly increasing on $[0, x_t)$ for some $x_t \in (0,\infty]$, with $E_{0t}(0) = 0$ and $E_{0t}(x) \to 1$ as $x \uparrow x_t$. If $x_t < \infty$, then $E_{0t}(x) = 1$ for all $x \geq x_t$.

To measure forecast behavior above v, let $M_t = \mathbf{1}\{Y_t > v\}$ and $N_n = \sum_{t \leq n} M_t$. On an exceedance, forecast i has excess PIT $Z_{it} = E_{it}(Y_t - v)$. For the aggregate forecast ($i = A$) and baseline ($i = 0$), define

$$A_i(q;n) = \sum_{t \leq n} M_t \mathbf{1}\{Z_{it} \leq q\}, \qquad D_i(n) = \sum_{t \leq n} [1 - F_{it}(v)],$$

$$H_i(q;n) = A_i(q;n) - qD_i(n), \qquad \widehat{C}_i(n) = \sup_{q \in [0,1]} \left| \frac{A_i(q;n)}{D_i(n)} - q \right|. \tag{1}$$

The combined ratio $A_i(q;n)/D_i(n)$ compares the observed number of exceedances with PIT at most q to the total forecast exceedance mass. When $N_n > 0$, it factors into the occurrence ratio $N_n/D_i(n)$ and the severity ratio $A_i(q;n)/N_n$. Following Allen et al. (2025, Section 4.2), we use $\widehat{C}_i(n)$ to summarize its distance from the line $q \mapsto q$; CRSD is evaluated at the fixed finite threshold. If no exceedances occur, the combined ratio is zero and this distance is one.

Example 2.1 (CRPS improvement with tail deterioration)*.* Let $U_{[a,b]}$ denote the CDF of the uniform distribution on $[a,b]$. Fix $0 < \varepsilon < 1/6$ and $v = 1$, and consider the target and raw-source CDFs

$$G_\varepsilon = (1-\varepsilon)U_{[0,1]} + \varepsilon U_{[1,2]}, \qquad T_\varepsilon = (1-2\varepsilon)U_{[0,1]} + 2\varepsilon U_{[1,2]}.$$

The source has the correct conditional distributions on the two intervals but doubles the exceedance probability. Let the baseline retain the target tail while misspecifying the body:

$$B_\varepsilon(y) = \begin{cases} G_\varepsilon(y) + 5\varepsilon y(1-y), & 0 \leq y \leq 1, \\ G_\varepsilon(y), & \text{otherwise.} \end{cases}$$

For $R(F) = \mathbb{E}_{G_\varepsilon}[\mathrm{CRPS}(F, Y)]$,

$$R(T_\varepsilon) - R(G_\varepsilon) = \frac{2}{3}\varepsilon^2 < \frac{5}{6}\varepsilon^2 = R(B_\varepsilon) - R(G_\varepsilon).$$

At $v = 1$, the baseline and source have population combined ratios q and $q/2$, and hence population CRSD values 0 and $1/2$, respectively. Thus expected CRPS favors the source despite its incorrect tail occurrence probability. Appendix C.1 gives the calculations and a protected transfer that improves the body while retaining the baseline tail.

3 Quantile-anchored transfer and pathwise protection

We construct protected forecasts in two steps: transfer each source within baseline quantile cells, then combine the transferred CDFs.

3.1 Transfer within quantile cells

The partition-reweighting step builds on Jeffrey's rule, which reassigns probabilities to a partition while retaining the conditional laws within its cells (Diaconis & Zabell, 1982). Here the baseline determines both the quantile partition and its cell probabilities, while each raw source contributes conditional shape within those cells. We revert to the baseline conditional law when a source assigns insufficient probability to a cell, and retain a positive baseline component to preserve strict crossing at the protected probability levels. Each source is transferred separately before the resulting CDFs are aggregated.

We fix $J \geq 1$, $\kappa \in (0, 1)$ and the source-cell mass threshold $\epsilon_{\mathrm{cell}} \in [0, 1)$. The protected probability grid is deterministic or $\mathscr{F}_0$-measurable and remains fixed during a run:

$$\mathcal{Q} = \{0 = q_0 < q_1 < \cdots < q_J = 1\}, \qquad h = \max_j (q_j - q_{j-1}).$$

Let $b_{t0} = v$, $b_{tJ} = \infty$, and let b_{tj} be the baseline quantile at level $F_{0t}(v) + p_t q_j$ for $1 \leq j < J$. These protected quantile cuts define the body cell $C_{t0} = (-\infty, v]$ and tail cells $C_{tj} = (b_{t,j-1}, b_{tj}]$. Their baseline probabilities are $A_{t0} = F_{0t}(v)$ and $A_{tj} = p_t(q_j - q_{j-1})$ for $j \geq 1$.

Write Q^0_{tj} for the baseline law conditional on a positive-mass cell. We use the raw-source conditional law Q^R_{ktj} when its cell probability exceeds ϵ_{cell}, and use Q^0_{tj} otherwise. Our transferred expert is

$$\mu^*_{kt} = \sum_{j: A_{tj} > 0} A_{tj}\{\kappa Q^0_{tj} + (1 - \kappa) Q^R_{ktj}\}, \qquad F^*_{kt}(y) = \mu^*_{kt}((-\infty, y]). \tag{2}$$

Thus the source determines a fraction $1 - \kappa$ of each conditional shape, while the baseline determines the cell probabilities and supplies the remaining fraction κ. Cells with zero baseline probability contribute no term. Setting $P_{0t} = F_{0t}$ includes the baseline itself, with $F^*_{0t} = F_{0t}$. Appendix A gives the conditional CDF formulas and an implementation using CDF evaluations.

Example 3.1 (Anchor values and PIT events). Take $v = 0$, baseline $F_0 = U_{[0,2]}$, and protected grid $\{0, 1/2, 1\}$, whose interior cut is $b_1 = 1$. The source $S = \frac{1}{2}U_{[0,1]} + \frac{1}{2}U_{[3/2,2]}$ already assigns baseline mass $1/2$ to each tail cell. It satisfies $S(1) = F_0(1) = 1/2$, but its support gap leaves S flat on $[1, 3/2]$. Since both forecasts have exceedance probability one, their CDF values are also their excess PITs. At $y = 5/4$, $S(y) = 1/2$ and $F_0(y) = 5/8$, giving

$$\mathbf{1}\{S(5/4) \leq 1/2\} = 1, \qquad \mathbf{1}\{F_0(5/4) \leq 1/2\} = 0.$$

With $\epsilon_{\mathrm{cell}} = 0$, transfer in (2) retains a positive baseline component:

$$F^*(y) = \kappa F_0(y) + (1 - \kappa) S(y), \qquad F^*(5/4) = \frac{1}{2} + \frac{\kappa}{8} > \frac{1}{2}.$$

The baseline supplies mass throughout the gap. More generally, $F^*(y) \leq 1/2$ for $0 < y \leq 1$ and $F^*(y) > 1/2$ for $y > 1$, so $\mathbf{1}\{F^*(y) \leq 1/2\} = \mathbf{1}\{y \leq 1\}$ for every $y > 0$. Positive retention therefore preserves both the protected CDF value and its crossing event, which is the property used in the pathwise bound.

We combine these experts using a simplex vector chosen before the outcome:

$$F_t^A(y) = \sum_{k=0}^{K} w_{kt} F_{kt}^*(y), \qquad w_{kt} \geq 0, \quad \sum_k w_{kt} = 1. \tag{3}$$

The same weights apply at every query position y. We transfer each source, including its fallback, before averaging. The aggregate forecast retains the common cell probabilities and positive baseline component by convexity (Lemma A.2).

3.2 A PATHWISE BOUND FOR THE AGGREGATE FORECAST

Our first result shows that this construction preserves the empirical counts at every protected cut and controls CRSD between cuts. The bound depends only on the baseline fraction and grid spacing.

Theorem 3.1 (Pathwise protection relative to the baseline)**.** *Under Assumption 2.1, construct the candidates as in* (2)*. Let F_t^A be any aggregate forecast defined by* (3)*. For every outcome sequence, every weight sequence, and every $n \geq 1$, we have*

$$D_A(n) = D_0(n) = \sum_{t \leq n} p_t, \quad A_A(q_j; n) = A_0(q_j; n),$$
$$H_A(q_j; n) = H_0(q_j; n), \quad |\widehat{C}_A(n) - \widehat{C}_0(n)| \leq (1-\kappa)h. \tag{4}$$

If $N_n = 0$, both CRSD values equal one.

Because the bound is uniform over the source forecasts and all pre-outcome weight sequences, any online rule that selects simplex weights before observing the outcome inherits the same guarantee.

Proof sketch. Convex closure preserves the baseline cell masses, while the positive baseline component enforces strict crossing at every protected cut; hence the protected excess-PIT indicators and count residuals agree with the baseline. Within each tail cell, the aggregate excess CDF differs from the baseline by at most $(1-\kappa)(q_j - q_{j-1}) \leq (1-\kappa)h$. Lemma A.3 converts this uniform PIT displacement, under the common denominator, into the CRSD bound. Appendix A.4 gives the full proof. □

3.3 PROTECTION–TRANSFER TRADEOFF

The radius $(1-\kappa)h$ decreases as the maximum probability spacing h decreases or the retained baseline fraction κ increases. For a uniform J-cell grid, $h = 1/J$, and the bound is sharp for $J \geq 2$ (Proposition A.2). Finer partitions preserve more baseline quantiles, while a larger κ restricts the admissible within-cell deformation. The baseline supplies the exceedance probability and the probabilities of all protected cells; source information adapts conditional shapes within those cells. In particular, a source sharing the baseline's within-cell conditional laws maps back to the baseline (Proposition C.4).

4 ONLINE LEARNING AND PREDICTIVE IMPROVEMENT

Within the protected class, we learn source weights using a common cell-conditional score. Since candidates agree at the protected cuts, we compare their predictions inside the cells. We first give regret guarantees for two exponential-weighting rules, then use an explicit iid construction to demonstrate predictive improvement under additional distributional assumptions.

4.1 CELL-CONDITIONAL SCORING

Before evaluation, fix $R \geq 1$, interior fractions $0 < u_1 < \cdots < u_R < 1$ and nonnegative node weights θ_ℓ summing to one. All experts use common cells, nodes and node weights. The fractions stay fixed, while the numerical nodes may change with the baseline. In tail cell j, $z_{tj\ell}$ is the baseline

quantile at level $F_{0t}(v) + p_t\{q_{j-1} + (q_j - q_{j-1})u_\ell\}$. If Q^*_{ktj} is candidate k's conditional CDF in this cell, its finite-node cell-conditional Brier score on an exceedance $Y_t \in C_{tj}$ is

$$c_{kt} = \sum_{\ell=1}^{R} \theta_\ell \big(Q^*_{ktj}(z_{tj\ell}) - \mathbf{1}\{Y_t \le z_{tj\ell}\}\big)^2, \qquad g_{kt} = M_t c_{kt}. \tag{5}$$

All costs are zero on a non-exceedance. This finite weighted sum of binary Brier scores follows the proper-scoring framework of Gneiting & Raftery (2007) and is proper for the selected conditional probabilities (Proposition A.1). The score targets the vector of within-cell conditional probabilities at the declared nodes, aligning the learning signal with the conditional shapes available for transfer. Section 4.2 bounds regret for this score; CRPS, threshold-weighted continuous ranked probability score (twCRPS), and CRSD evaluate aggregate forecasts.

Because the candidates share their cell probabilities, the aggregate conditional CDF is $Q^A_{tj} = \sum_k w_{kt} Q^*_{ktj}$. Expanding the square gives

$$g_{At} = \langle w_t, g_t \rangle - V_t, \qquad V_t \ge 0,$$

where V_t is the weighted dispersion of candidate probabilities at the score nodes (Lemma B.3). The aggregate forecast incurs no greater loss than the weighted average of the expert losses.

4.2 Exponential weighting and regret

We use standard exponential weighting from prediction with expert advice (Freund & Schapire, 1997; Cesa-Bianchi & Lugosi, 2006), applied here to the transferred experts and the cell-conditional score. We choose a prior $\pi_k > 0$ for $0 \le k \le K$ with $\sum_k \pi_k = 1$, using only initial information. Write $\pi_{\min} = \min_k \pi_k$ and $a = \log(1/\pi_{\min})$.

Event-doubling QAST-E. Initialize the event-doubling QAST-E weights at π. Within epoch e, we issue (3) and, after observing an exceedance, update the weights by

$$w_k^{\text{post}} = \frac{w_k e^{-\eta_e c_{kt}}}{\sum_r w_r e^{-\eta_e c_{rt}}}.$$

We start at $e = 0$, use event budgets $T_e = 2^e$ and rates $\eta_e = \sqrt{8a/T_e}$, and reset the weights to π when the epoch's T_e events have been observed. Counts for forecast evaluation continue across resets.

Write $G_A(n) = \sum_{t \le n} g_{At}$ for the cumulative aggregate score and $G_k(n) = \sum_{t \le n} g_{kt}$ for expert k.

Theorem 4.1 (Event-scale regret of the aggregate forecast)**.** *Under the full-feedback forecasting protocol in Section 2 and Assumption 2.1, use the common score in Section 4.1 and the event-doubling QAST-E rule with the positive prior in Section 4.2. For every realized history, every $n \ge 1$ and every label k, we have*

$$G_A(n) - G_k(n) \le \mathcal{R}(N_n) - \sum_{t \le n} V_t \le \mathcal{R}(N_n),$$

$$\mathcal{R}(N) = (1 + \sqrt{2})\sqrt{aN}, \qquad N \ge 0. \tag{6}$$

Here $V_t \ge 0$ is the conditional-probability diversity term in Lemma B.3. At $N_n = 0$, all losses and diversity terms vanish. At a fixed horizon, $\mathbb{E}[G_A - G_k \mid \mathscr{F}_0] \le (1 + \sqrt{2})\sqrt{a}\,\mathbb{E}[\sqrt{N_n} \mid \mathscr{F}_0]$.

The proof combines epochwise exponential weighting with the mixture-loss identity; Appendix B gives the details. For fixed a, the upper bound on regret per event tends to zero as $N_n \to \infty$. The comparator is a fixed transferred-expert label evaluated on the same realized history (Remark B.1).

Fixed-rate alternative. With a fixed learning rate $0 < \eta \le 1/2$ and no restarts, exp-concavity of the conditional square score gives the sharper aggregate bound

$$G_A(n) - G_k(n) \le \eta^{-1} \log(1/\pi_k)$$

for every fixed expert and finite path (Theorem B.1). Under the fixed two-expert iid conditions stated there, this rule also has almost-sure weight convergence. Under a cell-score gap, QAST-E's linear weighted-expert regret gives an explicit finite-sample bound on source use (Corollary B.1), which we use in Theorem 4.2. Figure 3 compares the twCRPS of QAST-E with that of Fixed QAST.

4.3 STRICT IMPROVEMENT IN AN IID CONSTRUCTION

An explicit iid construction isolates the predictive value of within-cell tail transfer. With the body and protected-cell probabilities correctly specified, both forecasts have misspecified tail shapes; under the construction below, QAST-E improves expected CRPS and limiting CRSD.

Fix $p \in (0,1)$, $r > 1$, $J \geq 1$ and $\kappa, \nu \in (0,1)$. The target G_p has mass $1-p$ uniformly on $[0,1]$ and mass p uniformly on $[1,2]$. Use $v = 1$ and uniform protected probability levels with $h = 1/J$. For $x \in [(j-1)h, jh]$, write $s = (x-(j-1)h)/h$ and define

$$E_0(x) = (j-1)h + hs^r, \qquad F_0(y) = (1-p)y \ (0 \leq y \leq 1), \qquad F_0(1+x) = 1-p+pE_0(x).$$

Both F_0 and G_p are zero below zero and one above two. The source is $P_1 = \nu F_0 + (1-\nu)G_p$. The baseline and source have the correct body and protected-cell probabilities, with misspecification confined to within-cell tail shape. Outcomes are iid from G_p, the initial information and forecast bank are fixed, and there are no informative current covariates. We use $\epsilon_{\text{cell}} = 0$ and the common interior score nodes of Section 4.1.

Write $\zeta = \kappa + (1-\kappa)\nu \in (0,1)$, so the transferred source is $F_1^* = \zeta F_0 + (1-\zeta)G_p$. Define the positive constants

$$\Delta_\nu = (1-\zeta^2)\sum_{\ell=1}^{R} \theta_\ell (u_\ell^{1/r} - u_\ell)^2,$$

$$d_{\text{tail}} = \frac{p^2(1-\zeta^2)}{2J^2}\left(\frac{1}{3} + \frac{1}{2r+1} - \frac{2}{r+2}\right), \qquad c_0 = \frac{(1-1/r)r^{-1/(r-1)}}{J}.$$

Here Δ_ν is the cell-score advantage of the transferred source, d_{tail} its scaled CRPS gain, and c_0 the baseline's population CRSD. Let $W_1^{\text{cal}}(n) = \sum_{t \leq n} w_{1t}$ and $\bar{w}_{1,n} = W_1^{\text{cal}}(n)/n$ denote the cumulative and time-averaged source weights.

Theorem 4.2 (Learned improvement from tail shape)**.** *Under the iid construction above, QAST-E with the event-doubling rule and positive prior of Section 4.2 satisfies, for $b(F,y) = \text{CRPS}(F,y)/2$ and every $n \geq 1$,*

$$\mathbb{E}\left[\frac{1}{n}\sum_{t \leq n}\{b(F_0, Y_t) - b(F_t^A, Y_t)\}\right] \geq d_{\text{tail}}\,\mathbb{E}[\bar{w}_{1,n}]$$

$$\geq d_{\text{tail}} \max\left\{0, 1 - \frac{(1+\sqrt{2})\sqrt{a}}{\Delta_\nu\sqrt{np}}\right\}. \tag{7}$$

As $n \to \infty$ with all construction, scoring, and prior parameters fixed,

$$\bar{w}_{1,n} \xrightarrow{p} 1, \qquad \widehat{C}_0(n) \xrightarrow{p} c_0, \qquad \widehat{C}_A(n) \xrightarrow{p} \zeta c_0 < c_0.$$

Appendix B.6.1 gives the proof.

From learned weights to predictive improvement. In this construction, the source weight directly controls the aggregate's CDF error:

$$F_t^A - G_p = \alpha_t(F_0 - G_p), \qquad \alpha_t = 1 - (1-\zeta)w_{1t}.$$

For an issued CDF F, write $R_p(F) = \mathbb{E}_{Y \sim G_p}[\text{CRPS}(F,Y)]$ for its risk under a fresh draw from the target distribution. The CRPS risk identity gives

$$R_p(F_t^A) - R_p(G_p) = \alpha_t^2\{R_p(F_0) - R_p(G_p)\}.$$

Since $0 < \zeta < 1$, increasing the source weight reduces the excess CRPS risk while leaving the body and protected-cell probabilities unchanged. The positive cell-score gap makes QAST-E's time-averaged source weight approach one, connecting its learned preference for within-cell shape to the predictive improvement in the theorem.

For fixed r, κ, ν and score nodes, Δ_ν is independent of p and J, whereas d_{tail} is proportional to p^2/J^2. Thus a positive conditional-score advantage can accompany a small global CRPS improvement when events are rare or the grid is fine. Proposition C.3 gives the corresponding integrated-Brier perturbation bound.

Table 1: Protected and unprotected aggregation at p95, $J = 5$, and $\kappa = 0.75$. All rows use the same 42,502 observed cases. Proper scores are in mm; CRSD is averaged equally across folds. The fixed-weight and same-weight unprotected controls retain QAST's nominal baseline coefficient; the latter uses QAST-E's weights. The independently learned raw mixture uses its own weights. Lower values are preferred. Scores are rounded to three significant figures; Table 5 reports small paired differences at higher precision.

	EMOS baseline			DRN baseline		
Forecast	CRPS	twCRPS	CRSD	CRPS	twCRPS	CRSD
Baseline	0.894	0.247	0.0912	0.897	0.246	0.0903
Unprotected mixture (fixed weights)	0.990	0.341	0.152	0.992	0.340	0.149
Unprotected mixture (same weights)	0.989	0.340	0.152	0.989	0.338	0.148
Raw mixture (independently learned)	0.998	0.353	0.172	0.995	0.351	0.168
Fixed QAST	0.894	0.247	0.0909	0.897	0.246	0.0901
QAST-E	0.894	0.247	0.0910	0.897	0.246	0.0901

5 A PRECIPITATION FORECASTING CASE STUDY

We use EUPPBench precipitation forecasts (Demaeyer et al., 2023) to study protection against negative transfer from regional source models.

5.1 SETUP

The 2018 evaluation covers 42,502 observed station-days at 117 stations. Five spatial folds hold out target stations from training (1997–2015) and validation (2016–2017), grouping stations at the same forecast grid point. Missing observations are excluded from scoring and learning. Global EMOS and a DRN are baselines. The common bank has forty regional models: eight each from EMOS, DRN, quantile regression forests with extended generalized Pareto distributions (QRF-EGP), Bayesian model averaging (BMA), and distributional forests.

We compare $(J, \kappa) = (1, 0.95), (3, 0.85), (5, 0.75)$, holding the protection radius fixed at $(1 - \kappa)/J = 0.05$. The threshold is a source-training precipitation quantile, including dry days; the main analysis uses p95 (1,485 observed exceedances), with p90 and p99 reported in Appendix D.1. QAST leaves the baseline CDF unchanged at and below the threshold. Fixed mixtures give half the weight to the baseline and divide the remainder equally among sources; QAST-E updates weights after observed exceedances. The fixed-weight and same-weight unprotected controls use the same source bank and either fixed weights or the same weights as QAST-E, with nominal baseline coefficient $\kappa + (1-\kappa)w_{0t}$. We also include a raw mixture that learns its own weights using the mean Brier score at 16 unconditional baseline quantiles chosen before the outcome. It updates after every released observed outcome and uses the same source bank without QAST's nominal baseline-retention constraint. Appendix D.1.2 gives the full definitions of these controls.

We report case-weighted CRPS and twCRPS in millimeters and the mean of the five fold-level CRSD values. Paired 95% intervals use 999 spatial cluster and seven-day block bootstrap replicates, conditional on the fitted models and aggregate forecasts. Appendix D.1.4 describes scoring and inference.

5.2 PROTECTION AGAINST NEGATIVE TRANSFER

QAST substantially reduces the proper-score losses from unprotected pooling (Table 1). At p95 with $J = 5$ and $\kappa = 0.75$, QAST-E reduces twCRPS relative to the unprotected mixture with the same weights by 27.5% under EMOS, from 0.340 to 0.247 mm, and by 27.2% under DRN, from 0.338 to 0.246 mm. The corresponding 95% CIs for the paired differences are $[-0.137, -0.054]$ mm and $[-0.133, -0.054]$ mm. CRPS decreases by 9.64% and 9.33%, respectively (Tables 5 and 8). Fixed protected mixtures already achieve most of these reductions.

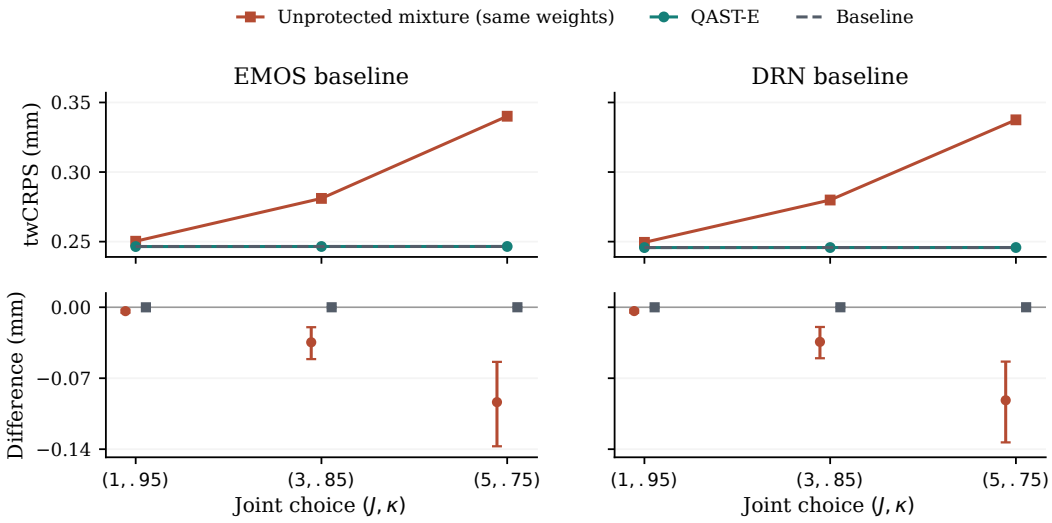

Figure 2: Aggregation at p95 with common protection radius $(1-\kappa)/J = 0.05$. Upper panels: twCRPS for QAST-E, its baseline, and the unprotected mixture (same weights). Lower panels: QAST-E minus the unprotected mixture (circles) and baseline (squares), with pointwise 95% paired intervals. Negative differences favor QAST-E; each position changes J and κ jointly.

QAST-E also improves CRPS and twCRPS over the independently learned raw mixture in this configuration, with pointwise 95% paired intervals below zero under both baselines (Table 5). This control uses its own learning objective and update schedule.

Figure 2 shows lower twCRPS for QAST-E than for its same-weight unprotected mixture across the three displayed p95 partitions under both baselines. QAST preserves baseline occurrence probabilities. At p95 with $J = 5$ and $\kappa = 0.75$, its proper-score and CRSD differences from the baseline have 95% intervals including zero (Table 5).

The proper-score advantage over unprotected pooling persists at p99, with 357 observed exceedances. At $(J, \kappa) = (5, 0.75)$, QAST-E reduces twCRPS relative to the unprotected mixture (same weights) by 53.5% (0.0936 mm) under EMOS and 54.0% (0.0951 mm) under DRN; both paired 95% intervals lie below zero. Changes relative to the respective baselines remain small, with intervals including zero (Table 6).

6 Discussion

QAST makes tail protection a property of the forecast class. Baseline quantile cells and positive baseline retention preserve selected probabilities and quantile-crossing events under convex aggregation. The resulting CRSD bound holds for any weights chosen before the outcome, including fixed and learned weights. Appendix C.4 extends the construction to approximate quantile cuts.

Within this class, exponential weighting gives sublinear or constant regret for the finite-node cell-conditional square score against fixed transferred experts. An explicit iid construction shows that exact protection can coexist with improved expected CRPS and limiting CRSD. On the 2018 EUPP-Bench sample, QAST reduces proper-score losses relative to unprotected mixtures using the same weights. Fixed protected mixtures deliver most of this benefit, with small additional adjustments from online learning (Figure 3).

QAST is suited to settings where selected baseline probabilities should remain fixed while source models refine within-cell shapes. Errors in those probabilities are inherited from the baseline. Future work could select the grid and baseline retention from data before evaluation, or relax probability preservation to allow joint correction of occurrence probabilities and conditional shapes.

AI USE STATEMENT

In this work, we independently developed the research idea and argument structure, selected the evaluation metrics, designed the experiments, and interpreted the results. ChatGPT and Claude assisted with literature searches, copy-editing, code execution and debugging, mathematical proof drafting and checking, and the supplementary Lean formalization. All AI-assisted mathematical and formal work was independently verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available archived forecasts and station observations (EUPPBench, WeatherBench 2) and involves no human subjects or personal data. Operational use would still require validating the baseline and the observation pipeline in the target setting.

REPRODUCIBILITY STATEMENT

Appendix A.1 records the information available before each outcome and the feedback schedule. Complete proofs of the theoretical results are given in Appendices A–C. The QAST-E procedure is summarized in Appendix A.3, the fixed-rate result is proved in Appendix B.4, and the approximate-cut extension is given in Appendix C.4. Appendix D.1 documents the EUPPBench sample, spatial folds, source-training thresholds, model selection, protected and unprotected mixtures, feedback timing, numerical scoring, and paired inference. The supplementary empirical package provides the implementation used for the reported EUPPBench analysis, the reported configurations and software requirements, and scripts for generating the paper's figures and tables. The supplement also provides a Lean 4 formalization of the main theoretical results and associated appendix results, together with a paper-to-Lean theorem map.

REFERENCES

Knut Are Aastveit, Saskia ter Ellen, and Giulia Mantoan. Quantile forecast combination: An application to forecasting US GDP growth. *International Journal of Forecasting*, 2026. doi: 10.1016/j.ijforecast.2026.06.003. In press.

Sam Allen, Jonathan Koh, Johan Segers, and Johanna Ziegel. Tail calibration of probabilistic forecasts. *Journal of the American Statistical Association*, 120(552):2796–2808, 2025. doi: 10.1080/01621459.2025.2506194.

Anastasios N. Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024a.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024b.

Jonathan Berrisch and Florian Ziel. CRPS learning. *Journal of Econometrics*, 237(2):105221, 2023. doi: 10.1016/j.jeconom.2021.11.008. Part C.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2337–2363, 2023.

Jonas R. Brehmer and Kirstin Strokorb. Why scoring functions cannot assess tail properties. *Electronic Journal of Statistics*, 13(2):4015–4034, 2019. doi: 10.1214/19-EJS1622.

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

Markus Dabernig, Georg J. Mayr, Jakob W. Messner, and Achim Zeileis. Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*, 143:909–916, 2017. doi: 10.1002/qj.2975.

J. Demaeyer, J. Bhend, S. Lerch, C. Primo, B. Van Schaeybroeck, A. Atencia, Z. Ben Bouallègue, J. Chen, M. Dabernig, G. Evans, J. Faganeli Pucer, B. Hooper, N. Horat, D. Jobst, J. Merše, P. Mlakar, A. Möller, O. Mestre, M. Taillardat, and S. Vannitsem. The EUPPBench postprocessing benchmark dataset v1.0. *Earth System Science Data*, 15(6):2635–2653, 2023. doi: 10.5194/essd-15-2635-2023. URL https://essd.copernicus.org/articles/15/2635/2023/.

Persi Diaconis and Sandy L. Zabell. Updating subjective probability. *Journal of the American Statistical Association*, 77(380):822–830, 1982. doi: 10.1080/01621459.1982.10477893.

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(1):119–139, 1997. doi: 10.1006/jcss.1997.1504.

Lea Friedli, David Ginsbourger, and Jonas Bhend. Area-covering postprocessing of ensemble precipitation forecasts using topographical and seasonal conditions. *Stochastic Environmental Research and Risk Assessment*, 35:215–230, 2021. doi: 10.1007/s00477-020-01928-4.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Tilmann Gneiting and Roopesh Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011. doi: 10.1198/jbes.2010.08110.

Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782, 2013. doi: 10.1214/13-EJS823.

Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20(83):1–28, 2019. URL https://jmlr.org/papers/v20/18-869.html.

Benedikt Schulz, Lutz Köhler, and Sebastian Lerch. Aggregating distribution forecasts from deep ensembles. *Machine Learning*, 115:120, 2026. doi: 10.1007/s10994-025-06946-3.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. doi: 10.1214/ECP.v16-1624.

Vladimir V'yugin and Vladimir Trunov. Online aggregation of probability forecasts with confidence. *Pattern Recognition*, 121:108193, 2022. doi: 10.1016/j.patcog.2021.108193.

Vladimir V. V'yugin and Vladimir G. Trunov. Online learning with continuous ranked probability score. In *Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 105 of *Proceedings of Machine Learning Research*, pp. 163–177. PMLR, 2019. URL https://proceedings.mlr.press/v105/v-yugin19a.html.

Jakob Benjamin Wessel, Maybritt Schillinger, Frank Kwasniok, and Sam Allen. Enforcing tail calibration when training probabilistic forecast models. *International Journal of Forecasting*, 42 (4):1336–1356, 2026. doi: 10.1016/j.ijforecast.2026.04.009.

APPENDIX

Appendix A states the forecasting protocol, defines the protected construction and score, and proves Theorem 3.1. Appendix B proves the event-clock and fixed-rate learning results and gives the proof of Theorem 4.2. Appendix C gives a score–tail counterexample, characterizes the protection–transfer tradeoff, and treats approximate quantile cuts. Appendix D begins with EUPPBench details (Section D.1) and then reports the supplementary WeatherBench 2 panel with meteorological aerodrome report (METAR) observations (Section D.2).

Symbol	Meaning
$F_{0t}, P_{kt}, F^*_{kt}, F^A_t$	Baseline, raw source, transferred source, aggregate forecast
q_j, b_{tj}, h	Protected probability levels, quantile cuts, maximum spacing
$M_t, N_n, D_i(n)$	Event indicator, event count, forecast exceedance mass
$\widehat{C}_i(n)$	Pathwise CRSD
$g_{kt}, g_{At}, \widetilde{G}(n), V_t$	Expert and mixture losses, weighted loss sum, diversity
$W^{\mathrm{cal}}_k(n)$	$\sum_{t\le n} w_{kt}$
T_e, L_{grid}	Epoch budget and concentration-grid size
$\overline{\mu}_{kt}, \overline{P}_{kt}$	Cell-reweighted measure and its CDF
r	Tail-shape exponent in the iid construction

A FORMAL SETUP AND PATHWISE PROTECTION

A.1 FORECASTS, INFORMATION, AND SCORING

This appendix formalizes the information available before each outcome (the pre-outcome information) and the feedback protocol used in the proofs.

Pre-outcome information and feedback. Work on $(\Omega, \mathscr{A}, \mathbb{P})$ with initial information $\mathscr{F}_0$, and assume that the conditional expectations and regular conditional laws used below exist. At deterministic issue times $t = 1, 2, \ldots$, let X_t collect the information released before issue t, and define

$$\mathscr{G}_t = \mathscr{F}_{t-1} \vee \sigma(X_t), \qquad \mathscr{P}_t = \mathscr{G}_t, \qquad \mathscr{F}_t = \mathscr{G}_t \vee \sigma(Y_t), \quad Y_t \ge 0.$$

At date t, the baseline and source CDFs, transferred candidates, quantile cuts, score nodes, and simplex weights are determined from $\mathscr{G}_t$; the weights are common to all CDF query positions. Under the main full-feedback protocol, Y_t is observed after issuance, is used to score every stored candidate, and is available before issue $t + 1$.

Parameters. Use the fixed threshold and finite source count of Section 2, the construction parameters of Section 3.1, the scoring parameters of Section 4.1, and the prior and learning rule of Section 4.2. Choices declared before evaluation use only $\mathscr{F}_0$ and remain fixed during a run, except for the learning rate specified by its update schedule. Forecasts and numerical quantiles may change with the permitted history. Put $P_{0t} = F_{0t}$.

Baseline regularity. Assumption 2.1 ensures positive baseline exceedance mass, unique tail cuts, and strict crossing at the protected levels. A zero-censored logistic CDF with finite pre-outcome location and positive scale is one admissible example. Raw discrete ensembles are not admissible baselines under the continuity condition, while zero source-cell mass is handled by baseline fallback below.

Definition A.1 (Baseline cells and probabilities)**.** *Let $b_{t0} = v$, $b_{tJ} = +\infty$, and for $1 \le j < J$ let*

$$b_{tj} = F^{-1}_{0t}\{F_{0t}(v) + p_t q_j\}.$$

Define $C_{t0} = (-\infty, v]$ and $C_{tj} = (b_{t,j-1}, b_{tj}]$, $1 \le j \le J$, interpreting the last cell as unbounded. If μ_{0t} is the baseline measure, its cell masses are

$$A_{t0} = F_{0t}(v), \qquad A_{tj} = p_t(q_j - q_{j-1}) > 0 \quad (j \ge 1).$$

For every positive-mass cell, write $Q^0_{tj} = \mu_{0t}(\cdot \mid C_{tj})$. *For raw source measure* μ^R_{kt}, *let* $B_{ktj} = \mu^R_{kt}(C_{tj})$ *and*

$$Q^R_{ktj} = \begin{cases} \mu^R_{kt}(\cdot \mid C_{tj}), & B_{ktj} > \epsilon_{\text{cell}}, \\ Q^0_{tj}, & B_{ktj} \le \epsilon_{\text{cell}}. \end{cases}$$

Zero baseline-mass cells are omitted. The transferred measure and its complete CDF are given by (2)*; for* $k = 0$, *this recovers* $F^*_{0t} = F_{0t}$.

What the transfer preserves. When fallback is inactive, cell reweighting is a Jeffrey partition update (Diaconis & Zabell, 1982, Eq. (1.1)). QAST uses the baseline tail-quantile partition, retains its cell masses, and combines each source conditional law with a positive baseline component. Baseline fallback supplies the baseline conditional law in cells with insufficient source mass. Thus the construction preserves the baseline exceedance and cell probabilities while allowing within-cell conditional shapes to change.

Computing the conditional distributions. For a cell $(l, u]$ and a CDF P with $B = P(u) - P(l) > 0$, its conditional CDF is

$$Q_P(y; l, u) = \begin{cases} 0, & y \le l, \\ [P(\min\{y, u\}) - P(l)]/B, & y > l. \end{cases}$$

We use $P(-\infty) = 0$ and $P(+\infty) = 1$ and apply the same formula to the baseline CDF when the fallback is active. The construction therefore uses only CDF evaluations at cell boundaries and score nodes. Once the cuts are available, evaluating all inner-node forecast probabilities requires on the order of $(K + 1)JR$ CDF evaluations per date.

Definition A.2 (Aggregate forecast)**.** *For a simplex vector* w_t *chosen before the outcome, transfer each source and execute its fallback before issuing the CDF in* (3). *Its probability measure is* $\mu^A_t = \sum_{k=0}^K w_{kt}\mu^*_{kt}$. *Write* $Q^*_{ktj} = \mu^*_{kt}(\cdot \mid C_{tj})$ *for every positive-mass cell, including the body. The aggregate conditional law is* $Q^A_{tj} = \sum_k w_{kt}Q^*_{ktj}$, *since all candidates share* A_{tj}. *The weights are constant across query positions* y *at a given date. The construction transfers each source before aggregation.*

Definition A.3 (Empirical combined ratio and supremum distance)**.** *For a complete forecast sequence* (F_{it}) *with* $1 - F_{it}(v) > 0$ *at every date, on each exceedance use its own excess probability integral transform (PIT)* $Z_{it} = [F_{it}(Y_t) - F_{it}(v)]/[1 - F_{it}(v)]$. *Put*

$$A_i(q) = \sum_{t \le n} M_t \mathbf{1}\{Z_{it} \le q\}, \quad D_i = \sum_{t \le n} [1 - F_{it}(v)], \quad H_i(q) = A_i(q) - qD_i.$$

If $D_i > 0$, *define*

$$\widehat{R}_i(q) = A_i(q)/D_i, \quad \widehat{C}_i = \sup_{q \in [0,1]} |\widehat{R}_i(q) - q|, \quad \widehat{C}^{\text{grid}}_i = \max_j |H_i(q_j)|/D_i.$$

Here $\widehat{R}_i$ *is the combined ratio,* $\widehat{C}_i$ *is CRSD, and* $\widehat{C}^{\text{grid}}_i$ *is its restriction to the protected grid. The baseline uses* F_{0t}. *The algorithm uses* $F_{At} = F^A_t$ *from* (3). *Counts on non-exceedances are defined as zero without evaluating* Z_{it}.

A.2 Cell-conditional scoring

Definition A.4 (Finite-node cell-conditional Brier score)**.** *Use the fixed fractions and normalized node weights from Section 4.1. For each tail cell, set*

$$z_{tj\ell} = F^{-1}_{0t}[F_{0t}(v) + p_t\{q_{j-1} + (q_j - q_{j-1})u_\ell\}].$$

These are strictly inside cell j, *even for the unbounded last cell. For* $z \in C_{tj}$, *let*

$$Q^*_{ktj}(z) = \frac{F^*_{kt}(z) - F^*_{kt}(b_{t,j-1})}{A_{tj}}.$$

Let $M_t = \mathbf{1}\{Y_t > v\}$ and $N_n = \sum_{t \le n} M_t$. If $M_t = 1$ and $Y_t \in C_{tj}$, define

$$c_{kt} = \sum_{\ell=1}^{R} \theta_\ell \left(Q^*_{ktj}(z_{tj\ell}) - \mathbf{1}\{Y_t \le z_{tj\ell}\} \right)^2, \qquad g_{kt} = M_t c_{kt}. \tag{8}$$

If $M_t = 0$, set $c_{kt} = g_{kt} = 0$ without evaluating an excess PIT. Evaluate the aggregate forecast by the same formula with Q^A_{tj} in place of Q^*_{ktj} to obtain c_{At} and $g_{At} = M_t c_{At}$. Define

$$G_A(n) = \sum_{t \le n} g_{At}, \qquad G_k(n) = \sum_{t \le n} g_{kt}, \qquad \widetilde{G}(n) = \sum_{t \le n} \langle w_t, g_t \rangle.$$

The score targets the declared vector of within-cell conditional probabilities at common nodes. Proposition A.1 proves properness for this finite vector; complete-distribution performance is evaluated separately by CRPS, twCRPS, and CRSD.

A simple choice for the analytic example is $R = 1$ with $u_1 = 1/2$. A practical alternative is to fix $R = 16$ equally weighted midpoint fractions before evaluation. Interior nodes $u_\ell \in (0, 1)$ matter: evaluating only at cell boundaries would assign identical scores to all transferred experts.

Proposition A.1 (Properness on the protected conditional class)**.** *Fix the history available before the outcome and one tail cell C_j of positive true conditional probability. Let its true conditional CDF be G_j. Then for every $z \in C_j$,*

$$\mathbb{E}[(Q(z) - \mathbf{1}\{Y \le z\})^2 \mid Y \in C_j, \mathscr{G}_t] = (Q(z) - G_j(z))^2 + G_j(z)(1 - G_j(z)).$$

Thus the excess expected loss in (8)*, relative to the true conditional probabilities at the common nodes, is a nonnegative weighted sum of squared conditional-CDF errors. Summing over true cell probabilities preserves this ordering. Cells with zero true probability contribute zero.*

Proof. We condition on the common history available before the outcome and the event that the outcome lies in cell C_j. Expanding the square and using $G_j(z)$ for the conditional indicator mean gives the displayed identity; its second term is the Bernoulli variance. We then multiply by each nonnegative θ_ℓ and sum over the nodes. Multiplying by $\mathbb{P}(Y \in C_j \mid \mathscr{G}_t)$ and summing over cells proves the result. The cells and nodes are shared by all experts and fixed before the outcome. □

This is a finite-node version of integrating binary proper scores against a node measure (Gneiting & Raftery, 2007, Section 6.4, Eq. (49)), written as a loss. Since the node weights are nonnegative and sum to one, the squared Bernoulli loss gives $0 \le g_{kt} \le M_t$, even with unbounded precipitation or misspecified experts.

A.3 The QAST-E algorithm

QAST-E combines quantile-anchored source transfer with event-doubling exponential weighting. Event counts, rather than calendar time, determine the learning-rate epochs.

Initialization. Choose v, $\mathcal{Q}$, κ, ϵ_{cell}, the score nodes and weights, and a positive prior π as in Sections 2–4.2. A concrete choice is $\pi_0 = 1/2$ and $\pi_k = 1/(2K)$ for $k \ge 1$. Set

$$a = \log(1/\pi_{\min}), \qquad e = 0, \qquad r_e = 0, \qquad T_e = 2^e, \qquad \eta_e = \sqrt{8a/T_e}, \qquad w = \pi.$$

At each issue date t:

1. From the pre-outcome information $\mathscr{G}_t$, generate $F_{0t}, P_{1t}, \ldots, P_{Kt}$ and compute the baseline cells, transferred CDFs in (2), and common score nodes.
2. Issue and store
$$F^A_t = \sum_{k=0}^{K} w_{kt} F^*_{kt}.$$

3. After observing Y_t, set $M_t = \mathbf{1}\{Y_t > v\}$. If $M_t = 0$, set all current losses to zero, set $w^{\text{post}} = w$, and leave r_e unchanged. If $M_t = 1$, identify the realized baseline cell, compute the costs in (8), and update
$$w_k^{\text{post}} = \frac{w_k e^{-\eta_e c_{kt}}}{\sum_{r=0}^{K} w_r e^{-\eta_e c_{rt}}}, \qquad r_e \leftarrow r_e + 1.$$
4. If $r_e < T_e$, use w^{post} at the next issue. If $r_e = T_e$, start epoch $e+1$ with
$$e \leftarrow e+1, \qquad T_e = 2^e, \qquad r_e = 0, \qquad \eta_e = \sqrt{8a/T_e}, \qquad w \leftarrow \pi.$$
Evaluation statistics are retained across resets.
5. Update the baseline and source procedures according to their declared rules using the returned information. These procedures continue to update regardless of their current aggregation weights.

Proposition C.5 treats approximate numerical quantile cuts.

A.4 Pathwise protection: proof of Theorem 3.1

We prove the safety bound in two steps. The cell construction preserves the protected levels and bounds the displacement of each excess PIT. We then show that the same displacement bounds the change in the combined-ratio supremum distance, even when its common denominator is small.

Lemma A.1 (Validity, mass preservation, and strict crossing)**.** *Under Assumption 2.1 and the construction in Section 3.1,* (2) *is a valid complete CDF with a continuous tail. For every t, k,*
$$F_{kt}^*(v) = F_{0t}(v), \qquad 1 - F_{kt}^*(v) = p_t > 0, \qquad F_{kt}^*(b_{tj}) = F_{0t}(v) + p_t q_j \quad (1 \le j < J).$$
For $y > v$ and $0 < j < J$,
$$\mathbf{1}\{E_{kt}^*(y - v) \le q_j\} = \mathbf{1}\{y \le b_{tj}\}. \tag{9}$$
At $q_0 = 0$ the indicator is zero, and at $q_J = 1$ it is one. With $E_{kt}^*(x) = [F_{kt}^*(v + x) - F_{kt}^*(v)]/p_t$,
$$\sup_{y > v} |E_{kt}^*(y - v) - E_{0t}(y - v)| \le (1 - \kappa)h. \tag{10}$$

Proof. We first verify that the construction defines a distribution with the required cell masses. The cells partition the outcome space, their nonnegative baseline masses sum to one, and each conditional measure is a probability measure supported in its cell. Their weighted sum is therefore a probability measure, with a nondecreasing, right-continuous CDF and the usual endpoint limits. Continuity of the baseline and raw-source tails excludes atoms at finite tail cuts, so the pieces join continuously above v. By construction, μ_{kt}^* assigns mass A_{tj} to each cell, including the body. Therefore
$$F_{kt}^*(v) = A_{t0} = F_{0t}(v), \; 1 - F_{kt}^*(v) = \sum_{r=1}^{J} A_{tr} = p_t,$$
and, for $1 \le j < J$,
$$F_{kt}^*(b_{tj}) = A_{t0} + \sum_{r=1}^{j} A_{tr} = F_{0t}(v) + p_t q_j = F_{0t}(b_{tj}).$$
For $k = 0$, both conditional components are Q_{tj}^0, and the construction recovers the baseline.

The positive baseline component gives strict crossing. If $y \le b_{tj}$, monotonicity yields $F_{kt}^*(y) \le F_{kt}^*(b_{tj})$. If $y > b_{tj}$, the baseline places positive mass in $(b_{tj}, y]$, including when y lies beyond a finite baseline endpoint. Because the transferred law contains $\kappa\mu_{0t}$, we have $F_{kt}^*(y) > F_{kt}^*(b_{tj})$. Dividing by $p_t > 0$ proves (9). Applying the same argument at v shows that the excess CDF is positive for every $y > v$; as a CDF, it is also at most one. These observations give the endpoint indicators.

It remains to bound displacement between the cuts. Define $\overline{\mu}_{kt} = \sum_{j: A_{tj} > 0} A_{tj} Q_{ktj}^R$ and let $\overline{P}_{kt}$ be its CDF. This measure has the same cell masses as the baseline, and we can write
$$F_{kt}^* = \kappa F_{0t} + (1 - \kappa)\overline{P}_{kt}.$$

Within tail cell j, the values $\overline{P}_{kt}(y)$ and $F_{0t}(y)$ lie between the same cumulative lower and upper cell masses. Their difference is thus at most $A_{tj} = p_t(q_j - q_{j-1})$, giving

$$|F^*_{kt}(y) - F_{0t}(y)| \le (1-\kappa)p_t(q_j - q_{j-1}).$$

Since the two CDFs agree at v, subtracting their common value and dividing by p_t proves (10).

The last cell extends to infinity even if the baseline endpoint $v+x_t$ is finite. To see the bound directly in this case, take $y \ge v + x_t$ and write $Q^R_{ktJ}(y)$ for the source conditional CDF in that cell. Then

$$1 - E^*_{kt}(y - v) = (1-\kappa)(1-q_{J-1})\{1 - Q^R_{ktJ}(y)\} \in [0, (1-\kappa)h].$$

The displacement bound therefore includes source mass beyond the baseline endpoint. Every internal cut lies below that endpoint, and the baseline increment beyond b_{tj} is $p_t(1-q_j) > 0$ when $y \ge v + x_t$, which also proves strict crossing there. At $q = 1$, both sublevel indicators are one because both PITs are at most one, whether or not both equal one. □

Lemma A.2 (Convex closure of the protected distribution class)**.** *Under Assumption 2.1 and the construction in Section 3.1, fix a date and its baseline partition. Let $\mathcal{S}_t$ consist of the laws*

$$\sum_{j: A_{tj} > 0} A_{tj}\{\kappa Q^0_{tj} + (1-\kappa)R_{tj}\},$$

where every R_{tj} is a probability measure supported in C_{tj} and has no atoms at finite points above v, including finite cell endpoints. This class is convex, contains all transferred candidates, and contains their aggregate (3)*. The aggregate satisfies every identity and bound in Lemma A.1.*

Proof. For simplex weights, let $\bar{R}_{tj} = \sum_k w_{kt} Q^R_{ktj}$, where $Q^R_{0tj} = Q^0_{tj}$. This is again a probability measure supported in the cell, with continuous tail pieces. Rearranging the finite sums gives

$$\mu^A_t = \sum_j A_{tj}\{\kappa Q^0_{tj} + (1-\kappa)\bar{R}_{tj}\}.$$

The same calculation holds for any finite collection of laws in $\mathcal{S}_t$, proving the convexity of the class. The aggregate has exactly the form used in Lemma A.1, so its mass identities, strict crossing, and displacement bound follow from that lemma, including beyond a finite baseline endpoint. □

Lemma A.3 (Stability of CRSD under bounded PIT perturbations)**.** *Let $M_t \in \{0,1\}$ be common event indicators. On events let two PIT arrays $z_t, z'_t \in [0,1]$ satisfy $|z_t - z'_t| \le r$ for some $r \in [0,1]$. Let both combined ratios divide their counts by exactly the same $D > 0$. Then their combined-ratio supremum distances satisfy*

$$|C - C'| \le r. \tag{11}$$

Proof. Write R, R' for the ratios and $N = \sum_t M_t$. We first bound $R(q) - q$ from above. When $q + r \le 1$, every event counted in $R(q)$ is also counted in $R'(q+r)$, hence

$$R(q) - q \le R'(q+r) - (q+r) + r \le C' + r.$$

When $q + r > 1$, the endpoint identity $R(q) \le N/D = R'(1)$ instead gives

$$R(q) - q \le R'(1) - 1 + 1 - q \le C' + r.$$

For the lower bound, take first $q \ge r$. Every event counted in $R'(q-r)$ is counted in $R(q)$, so

$$R(q) - q \ge R'(q-r) - (q-r) - r \ge -C' - r.$$

For $q < r$, nonnegativity gives $R(q) - q \ge -q \ge -r \ge -C' - r$. Together the four cases cover all q, including both endpoints, and show $C \le C' + r$. Interchanging the arrays proves the reverse bound. □

Proof of Theorem 3.1. By Lemma A.2, for every date t,

$$1 - F^A_t(v) = 1 - F_{0t}(v) = p_t.$$

Hence

$$D_A(n) = D_0(n) = \sum_{t \le n} p_t.$$

Moreover, for every protected level q_j, Lemmas A.1–A.2 give

$$M_t \mathbf{1}\{Z_{At} \le q_j\} = M_t \mathbf{1}\{Z_{0t} \le q_j\},$$

with both sides equal to $\mathbf{1}\{v < Y_t \le b_{tj}\}$ at interior levels. Summing over $t \le n$ yields

$$A_A(q_j; n) = A_0(q_j; n).$$

Therefore,

$$H_A(q_j; n) = A_A(q_j; n) - q_j D_A(n) = A_0(q_j; n) - q_j D_0(n) = H_0(q_j; n).$$

By (10), the aggregate PIT lies within $(1-\kappa)h$ of the baseline PIT on every event. Lemma A.3 now proves the supremum-distance bound in (4). The argument holds for every finite prefix of the same sequence, which establishes the simultaneous claim. Finally, if $N_n = 0$, both count numerators are zero and the supremum of $|-q|$ is one. □

Remark A.1 (The endpoint residual at $q_J = 1$). For complete CDFs, $A_A(1) = A_0(1) = N_n$ and $H_A(1) - H_0(1) = -(D_A - D_0)$; exact transfer makes both sides zero.

Sharpness of the protection radius.

Proposition A.2 (Sharpness of the protection radius)**.** *For uniform grids with $J \ge 2$, the constant $(1-\kappa)h$ in Theorem 3.1 cannot be uniformly reduced over its stated class.*

Proof. Fix an integer $m \ge 2$, put $p = 1/m$, $v = 1$, and $n = m$, and set $\epsilon_{\text{cell}} = 0$. The baseline has mass $1-p$ uniformly on $[0,1]$ and mass p uniformly on $[1,2]$. Its excess CDF is the identity on $[0,1]$, so it is continuous and strictly increasing, with uniform protected probability levels $q_j = j/J$ and $h = 1/J$.

For $0 < \varepsilon < h/2$, let the raw source have the same body as the baseline, place the first tail cell's mass ph uniformly on $[1, 1+\varepsilon/2]$, and retain the baseline conditional laws and masses in the other tail cells. The source CDF is continuous, all used cells have positive source mass, and cell reweighting leaves this law unchanged. Use the same fixed bank at every date and the predictable, admissible vertex weights $w_1 = 1$.

Take $m-1$ observations in $(0,1)$ and one observation $1+\varepsilon$. Then $N_n = 1$ and the common forecast exceedance mass is $np = 1$. The unique event has baseline PIT $z_0 = \varepsilon$ and transferred PIT $z_* = \kappa\varepsilon + (1-\kappa)h$. Both are below $1/2$, since $J \ge 2$, $0 < \kappa < 1$, and $\varepsilon < h/2$. For a single event with denominator one and PIT $z < 1/2$, the CRSD is $1 - z$. Consequently the two CRSD values differ by $(1-\kappa)(h-\varepsilon)$, tending to $(1-\kappa)h$ as $\varepsilon \downarrow 0$. This proves sharpness over the full protected aggregate class, which includes these vertex weights. □

B Online learning and usefulness proofs

The online learning analysis uses standard expert-advice and exp-concavity tools (Cesa-Bianchi & Lugosi, 2006, Sections 2.2–2.3 and 3.3). We verify their applicability within the protected conditional forecast class and derive the event-clock and predictive consequences below.

B.1 Event-clock regret of the aggregate forecast

We begin with the exponential-weights bound on the observed-event sequence. The square score then gives an exact relation between weighted expert loss and the loss of the aggregate forecast. Combining these two arguments proves Theorem 4.1 pathwise. We subsequently study what a conditional source advantage implies for the learned weights. We first recall Hoeffding's standard lemma in the form stated by Cesa-Bianchi & Lugosi (2006, Lemma 2.2, p. 16).

Lemma B.1 (Hoeffding's lemma (conditional form))**.** *If W is supported on an interval of width c, then for every $s \in \mathbb{R}$,*

$$\log \mathbb{E} e^{s(W - \mathbb{E}W)} \le s^2 c^2 / 8.$$

The same assertion holds for regular conditional laws.

The cited formulation uses endpoints a, b; here $c = b - a$, and we subtract $s\mathbb{E}W$ from its log-MGF bound. The conditional version follows by applying the same result to the regular conditional law.

Lemma B.2 (Exponential weights and event doubling)**.** *With the positive prior of Section 4.2, for any finite, possibly adaptive realized sequence $d_s \in [0,1]^{K+1}$, standard exponential weights initialized at π and updated with a fixed $\eta > 0$ satisfy, simultaneously for every k,*

$$\sum_{s=1}^{q} \langle w_s, d_s \rangle - \sum_{s=1}^{q} d_{ks} \le \frac{\log(1/\pi_k)}{\eta} + \frac{\eta q}{8}.$$

For the event-doubling recurrence in Section 4.2, with each cost vector processed before the next prediction on that learner's sequence, simultaneously for every realized cost history and every n, k,

$$\sum_{t \le n} \langle w_t, g_t \rangle - G_k(n) \le (1 + \sqrt{2})\sqrt{aN_n}. \tag{12}$$

When $N_n = 0$, both sides vanish.

Proof. Fixed-rate bound. We first derive the standard exponential-weights bound for completeness; see also Cesa-Bianchi & Lugosi (2006, Theorem 2.2 and Exercise 2.5). For a fixed learning rate $\eta > 0$, define the exponential-weights potential

$$Z_s = \sum_{k=0}^{K} \pi_k \exp\left(-\eta \sum_{r=1}^{s} d_{kr}\right), \qquad Z_0 = 1.$$

The weights used before observing d_s are therefore

$$w_{ks} = \frac{\pi_k \exp\left(-\eta \sum_{r=1}^{s-1} d_{kr}\right)}{Z_{s-1}}.$$

Hence

$$\frac{Z_s}{Z_{s-1}} = \sum_{k=0}^{K} w_{ks} e^{-\eta d_{ks}}.$$

Since $d_{ks} \in [0,1]$, Lemma B.1 applied to the random variable that takes value d_{ks} with probability w_{ks} gives

$$\log \frac{Z_s}{Z_{s-1}} = \log \sum_{k=0}^{K} w_{ks} e^{-\eta d_{ks}} \le -\eta \langle w_s, d_s \rangle + \frac{\eta^2}{8}.$$

Summing over $s = 1, \ldots, q$ yields

$$\log Z_q \le -\eta \sum_{s=1}^{q} \langle w_s, d_s \rangle + \frac{\eta^2 q}{8}.$$

On the other hand, for every fixed expert label k,

$$Z_q \ge \pi_k \exp\left(-\eta \sum_{s=1}^{q} d_{ks}\right).$$

Combining the preceding two displays and rearranging gives

$$\sum_{s=1}^{q} \langle w_s, d_s \rangle - \sum_{s=1}^{q} d_{ks} \le \frac{\log(1/\pi_k)}{\eta} + \frac{\eta q}{8}.$$

This argument is deterministic conditional on the realized loss sequence, so the bound also applies to adaptive loss sequences.

Event-clock adaptation. For QAST-E, $g_{kt} = 0$ for every expert on a non-exceedance date. Such a date contributes neither loss nor a weight update, so deleting all non-exceedance dates leaves exactly the exponential-weights recursion on the observed-event sequence.

Let q_e denote the number of observed events from epoch e that occur by horizon n. By construction,

$$0 \le q_e \le T_e, \qquad T_e = 2^e.$$

Each epoch is restarted from the same prior π and uses the fixed rate

$$\eta_e = \sqrt{\frac{8a}{T_e}}, \qquad a = \log(1/\pi_{\min}).$$

Applying the fixed-rate bound within epoch e, for every fixed expert label k,

$$R_e(k) := \sum_{s=1}^{q_e} \langle w_{es}, d_{es} \rangle - \sum_{s=1}^{q_e} d_{kes} \le \frac{\log(1/\pi_k)}{\eta_e} + \frac{\eta_e q_e}{8}.$$

Since $\log(1/\pi_k) \le a$ and $q_e \le T_e$,

$$R_e(k) \le \frac{a}{\eta_e} + \frac{\eta_e T_e}{8} = \sqrt{\frac{aT_e}{2}}.$$

Thus the same bound also covers a partially completed final epoch.

Suppose now that $N_n > 0$, and let E be the last epoch containing an observed event by time n. Epochs $0, \dots, E-1$ are complete and contain

$$\sum_{e=0}^{E-1} 2^e = 2^E - 1$$

events, while epoch E contains at least one. Therefore $N_n \ge 2^E$. Summing the epochwise bounds gives

$$\begin{aligned} \sum_{t \le n} \langle w_t, g_t \rangle - G_k(n) &\le \sum_{e=0}^{E} \sqrt{\frac{a2^e}{2}} \\ &\le \sqrt{\frac{a}{2}} \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{2^E} \\ &= (1+\sqrt{2})\sqrt{a2^E} \\ &\le (1+\sqrt{2})\sqrt{aN_n}. \end{aligned}$$

If $N_n = 0$, all expert losses are zero and both sides vanish. This proves (12). □

Lemma B.3 (Actual-mixture loss and diversity identity)**.** *For the construction and common score in Sections 3.1 and 4.1, on an exceedance in cell j, put $u_{kt\ell} = Q^*_{ktj}(z_{tj\ell})$ and $\bar{u}_{t\ell} = \sum_k w_{kt} u_{kt\ell}$. Define*

$$V_t = M_t \sum_\ell \theta_\ell \sum_k w_{kt}(u_{kt\ell} - \bar{u}_{t\ell})^2 \ge 0,$$

with $V_t = 0$ on non-exceedances without evaluating cell quantities. Then

$$g_{At} = \langle w_t, g_t \rangle - V_t, \qquad G_A(n) = \widetilde{G}(n) - \sum_{t \le n} V_t. \tag{13}$$

Proof. Because the experts have common cell masses, $Q^A_{tj} = \sum_k w_{kt} Q^*_{ktj}$. Let $B_\ell = \mathbf{1}\{Y_t \le z_{tj\ell}\}$. Expanding the squares at a common node gives

$$\sum_k w_{kt}(u_{kt\ell} - B_\ell)^2 - (\bar{u}_{t\ell} - B_\ell)^2 = \sum_k w_{kt}(u_{kt\ell} - \bar{u}_{t\ell})^2.$$

We multiply by θ_ℓ, sum over nodes, and multiply by M_t to obtain the one-step identity. When $M_t = 0$, every term is zero by definition. Summing over dates proves the cumulative result. □

Proof of Theorem 4.1. Combining Lemmas B.3 and B.2, we obtain, simultaneously for all n, k,

$$G_A(n) - G_k(n) = \widetilde{G}(n) - G_k(n) - \sum_{t \le n} V_t \le (1+\sqrt{2})\sqrt{aN_n} - \sum_{t \le n} V_t.$$

This inequality holds for every realized sequence. At any fixed horizon, the loss and diversity terms are bounded. Since $a < \infty$ is $\mathcal{F}_0$-measurable, we may localize to the events $\{a \le m\}$, take conditional expectations given $\mathcal{F}_0$, and let $m \to \infty$. This yields the stated conditional bound. If no event has occurred, the losses and diversity terms all vanish. □

For $N_n > 0$, dividing by N_n shows that the upper bound on regret per observed event tends to zero whenever $N_n \to \infty$.

B.2 Comparator scope

Remark B.1 (Transferred, raw-source, and policy comparators). Theorem 4.1 compares the aggregate forecast with a fixed transferred-expert label evaluated on the same realized history. If a raw-source score G_k^R is also defined on that history, then

$$G_A - G_k^R \le \mathcal{R}(N) + (G_k - G_k^R),$$

so the second term records the effect of transfer. A separately rerun source procedure coincides with the expert evaluated on the aggregate forecast's observed history only when its updates do not depend on past aggregation weights and forecasts. Otherwise, the theorem compares experts on that observed history, not on the history of a separately run procedure.

B.3 A source advantage in cell-conditional score

We now add a condition under which we can control the weight placed on experts other than the best transferred expert. Following the conditional expected loss gap of Mourtada & Gaïffas (2019, Corollary 6, Eq. (10)), we formulate the advantage for the QAST loss, which is zero on non-exceedances, conditional on the information available before the outcome. We retain the QAST-E restart schedule and derive the conclusion from its regret bound.

Assumption B.1 (Cell-conditional score advantage)**.** *Let $k^\star \in \{0, \dots, K\}$ and $0 < \Delta < \infty$ be $\mathscr{F}_0$-measurable and fixed throughout evaluation. Put $\lambda_t = \mathbb{P}(M_t = 1 \mid \mathscr{G}_t)$. Almost surely, for every t and every $k \ne k^\star$,*

$$\mathbb{E}[g_{kt} - g_{k^\star t} \mid \mathscr{G}_t] \ge \lambda_t \Delta. \tag{14}$$

The label $k^\star$ remains fixed while its transferred forecast may update along the realized history; $k^\star \ge 1$ represents an advantage for a transferred source. The assumption imposes a persistent conditional-score ordering beyond properness alone.

Corollary B.1 (Expected event-weighted nonoptimal mass)**.** *Under the full-feedback protocol in Section 2, Assumptions 2.1 and B.1, use the common score of Section 4.1 and QAST-E with the positive prior and event-doubling rule of Section 4.2. Define $W_{\ne}(n) = \sum_{t \le n} M_t(1 - w_{k^\star t})$. For every fixed finite n,*

$$\mathbb{E}[W_{\ne}(n) \mid \mathscr{F}_0] \le \frac{(1+\sqrt{2})\sqrt{a}}{\Delta} \mathbb{E}[\sqrt{N_n} \mid \mathscr{F}_0] \quad \text{a.s.} \tag{15}$$

Proof. The weights are $\mathscr{G}_t$-measurable. We can therefore apply Assumption B.1 inside the weighted conditional expectation to obtain

$$\begin{aligned}\mathbb{E}[\langle w_t, g_t \rangle - g_{k^\star t} \mid \mathscr{G}_t] &= \sum_{k \ne k^\star} w_{kt} \mathbb{E}[g_{kt} - g_{k^\star t} \mid \mathscr{G}_t] \\ &\ge \Delta \lambda_t (1 - w_{k^\star t}) = \Delta \mathbb{E}[M_t(1 - w_{k^\star t}) \mid \mathscr{G}_t].\end{aligned}$$

When the label is initially measurable, we first decompose over the finitely many events $\{k^\star = k\}$. Summing over dates and using the tower property then gives the lower bound $\Delta \mathbb{E}[W_{\ne} \mid \mathscr{F}_0]$ on $\mathbb{E}[\widetilde{G} - G_{k^\star} \mid \mathscr{F}_0]$. Lemma B.2 bounds the same weighted-loss difference pathwise by

$(1+\sqrt{2})\sqrt{aN_n}$. We take conditional expectations and divide by the positive, initially measurable Δ to obtain (15).

Lemma B.2 controls the weighted-expert difference $\widetilde{G} - G_{k^\star}$ used in this argument. □

The corollary controls event-weighted use of nonoptimal transferred experts. Theorem 4.2 adds the distributional assumptions that connect this conditional-score advantage to CRPS and CRSD.

B.4 Aggregation with a fixed learning rate

The conditional square score also admits a bound independent of the number of events when the learning rate is fixed and the weights are not restarted. We give this alternative separately because its weight convergence differs from that of QAST-E. Its regret bound applies standard exp-concave aggregation (Cesa-Bianchi & Lugosi, 2006, Section 3.3, Theorem 3.2 and Corollary 3.1) to the conditional square score.

Theorem B.1 (Fixed-rate aggregation for the conditional square score)**.** *Under the full-feedback protocol in Section 2 and Assumption 2.1, use the construction, common score and positive prior in Sections 3.1, 4.1 and 4.2. Issue* (3) *with weights*

$$w_{kt} = \frac{\pi_k e^{-\eta \sum_{s<t} g_{ks}}}{\sum_r \pi_r e^{-\eta \sum_{s<t} g_{rs}}}, \qquad 0 < \eta \le \tfrac{1}{2},$$

using one fixed learning rate and no resets. Then, on every admissible path, simultaneously for all n, k,

$$G_A(n) - G_k(n) \le \frac{\log(1/\pi_k)}{\eta}. \tag{16}$$

The protection radius is unchanged. For the additional weight-convergence conclusion only, assume a fixed iid two-expert bank with event probability $p > 0$, conditional score gap $\Delta > 0$ favoring expert 1, and no informative current covariates. Then $w_{1t} \to 1$ almost surely.

Proof. We first establish exponential concavity of the conditional score. Fix an event cell and its outcome indicators, and let $f(u) = \sum_\ell \theta_\ell (u_\ell - B_\ell)^2$ on $[0,1]^R$, where $\theta_\ell \ge 0$ and $\sum_\ell \theta_\ell = 1$, so $0 \le f \le 1$. For any direction d, weighted Cauchy–Schwarz gives

$$(Df(u)[d])^2 = 4\left(\sum_\ell \theta_\ell (u_\ell - B_\ell) d_\ell\right)^2 \le 4 f(u) \sum_\ell \theta_\ell d_\ell^2 \le 4 \sum_\ell \theta_\ell d_\ell^2.$$

Also, $D^2 f(u)[d,d] = 2\sum_\ell \theta_\ell d_\ell^2$. Hence the second directional derivative of $e^{-\eta f}$ is $e^{-\eta f}\{\eta^2 (Df[d])^2 - \eta D^2 f[d,d]\} \le 0$ when $\eta \le 1/2$. Restricting to each line segment proves concavity, even when some node weights are zero. On a non-event all losses are zero and the exponential equals one. Common cell masses give $Q^A_{tj} = \sum_k w_{kt} Q^*_{ktj}$, so the aggregate forecast uses the same convex combination of conditional probabilities. Consequently,

$$e^{-\eta g_{At}} \ge \sum_k w_{kt} e^{-\eta g_{kt}}.$$

This is precisely the one-step exp-concave aggregation inequality of Cesa-Bianchi & Lugosi (2006, Eq. (3.2)). Applying their standard log-sum argument (Cesa-Bianchi & Lugosi, 2006, Theorem 3.2 and Corollary 3.1) with prior π and no restarts gives

$$G_A(n) \le -\eta^{-1} \log \sum_k \pi_k e^{-\eta G_k(n)} \le G_k(n) + \eta^{-1} \log(1/\pi_k),$$

proving (16) for the aggregate forecast.

For the iid conclusion, the differences $g_{0s} - g_{1s}$ are bounded iid variables with mean $p\Delta > 0$. Their average converges almost surely to $p\Delta$ by the strong law. The weight odds satisfy

$$\log \frac{w_{0t}}{w_{1t}} = \log \frac{\pi_0}{\pi_1} - \eta \sum_{s<t} (g_{0s} - g_{1s}) \longrightarrow -\infty.$$

It follows that $w_{1t} \to 1$. For event doubling, each reset restores the prior, so the odds identity applies within an epoch rather than across all epochs. □

B.5 SHARED LEMMAS FOR THEOREM 4.2

The CRPS identity, martingale concentration bound, and inverse-CDF bridge support the proof of Theorem 4.2. The CRPS identity is also used in Proposition C.1. We recall the standard CRPS risk identity in the loss orientation; see Gneiting & Raftery (2007, Section 4.2, p. 367). Their divergence is $S(G, G) - S(F, G)$ for a reward S; our CRPS is its negative, so the excess risk below has the same positive sign.

Lemma B.4 (CRPS risk identity)**.** *If G is the law of Y and F and G have finite first moments, then*

$$\mathbb{E}_G\{\mathrm{CRPS}(F, Y)\} - \mathbb{E}_G\{\mathrm{CRPS}(G, Y)\} = \int_{\mathbb{R}} \{F(z) - G(z)\}^2 \, dz.$$

The next bound adapts the scalar Freedman inequality in Tropp (2011, Theorem 1.1, p. 263) to our interleaved information fields and conditioning on $\mathscr{F}_0$.

Lemma B.5 (A conditional martingale Bernstein bound)**.** *Let $\mathscr{F}_{t-1} \subseteq \mathscr{P}_t \subseteq \mathscr{F}_t$ be nested interleaved fields, with X_t measurable with respect to $\mathscr{F}_t$, $\mathbb{E}[X_t \mid \mathscr{P}_t] = 0$ and $|X_t| \le c$ a.s. Put $\sigma_t^2 = \mathbb{E}[X_t^2 \mid \mathscr{P}_t]$. For deterministic $v_* > 0, x > 0$,*

$$\mathbb{P}\left\{ \sum_{t \le n} X_t \ge b_c(v_*, x), \quad \sum_{t \le n} \sigma_t^2 \le v_* \,\middle|\, \mathscr{F}_0 \right\} \le e^{-x}, \qquad b_c(v_*, x) = \sqrt{2v_* x} + \frac{2c}{3} x. \tag{17}$$

The same inequality holds for $-\sum X_t$.

Proof. Set $\mathcal{H}_0 = \mathscr{F}_0$, $\mathcal{H}_{2t-1} = \mathscr{P}_t$ and $\mathcal{H}_{2t} = \mathscr{F}_t$, and give the partial-sum process increments $D_{2t-1} = 0$, $D_{2t} = X_t$. Indeed,

$$\mathbb{E}[D_{2t-1} \mid \mathcal{H}_{2t-2}] = 0, \qquad \mathbb{E}[D_{2t} \mid \mathcal{H}_{2t-1}] = \mathbb{E}[X_t \mid \mathscr{P}_t] = 0,$$

and

$$\sum_{s \le 2n} \mathbb{E}[D_s^2 \mid \mathcal{H}_{s-1}] = \sum_{t \le n} \sigma_t^2.$$

For any $A \in \mathscr{F}_0$ with $\mathbb{P}(A) > 0$, work under $\mathbb{P}_A = \mathbb{P}(\cdot \mid A)$. Since $A \in \mathscr{P}_t$, on A we still have $\mathbb{E}_A[X_t \mid \mathscr{P}_t] = 0$ and $\mathbb{E}_A[X_t^2 \mid \mathscr{P}_t] = \sigma_t^2$. Apply scalar Freedman under $\mathbb{P}_A$ and restrict its crossing event to the fixed step $2n$. For deterministic $v_* > 0$ and $u > 0$, this gives

$$\mathbb{P}_A\{\textstyle\sum_{t \le n} X_t \ge u, \ \sum_{t \le n} \sigma_t^2 \le v_*\} \le \exp\Big\{ -\frac{u^2}{2(v_* + cu/3)} \Big\}.$$

Because this holds for every such A, the same deterministic bound holds for the conditional probability given $\mathscr{F}_0$, almost surely. For $u = \sqrt{2v_* x} + 2cx/3$, direct expansion gives $u^2 \ge 2x(v_* + cu/3)$, yielding (17). Apply the same argument separately to $-X_t$, which has the same conditional variances and upper bound c, to obtain the other tail. □

Lemma B.6 (A direct inverse-CDF bridge for the analytic mixtures)**.** *Fix a threshold v and a continuous, strictly increasing excess CDF E_0 on $[0, 1]$, with endpoints zero and one. Let $E_1 = \alpha E_0 + (1 - \alpha)\mathrm{id}$ for fixed $\alpha \in (0, 1)$, and put $K_1 = E_1^{-1}$ and $c_0 = \|E_0 - \mathrm{id}\|_\infty$. The full baseline and source CDFs both assign tail probability $p > 0$. Assume, conditionally on each prediction field $\mathscr{P}_t$, that the event probability is p and the excess given an event is uniform on $[0, 1]$. At date t issue $E_t^A = w_{0t} E_0 + (1 - w_{0t}) E_1$, where $w_{0t} \in [0, 1]$ is $\mathscr{P}_t$-measurable. All combined ratios use the common denominator np. Put $W_0^{\mathrm{cal}}(n) = \sum_{t \le n} w_{0t}$. For every fixed $n \ge 1$, integer $L_{\mathrm{grid}} \ge 1$ and $x > 0$, with probability at least $1 - 2(L_{\mathrm{grid}} + 1)e^{-x}$,*

$$\|\widehat{R}_A - K_1\|_\infty \le c_0 \frac{W_0^{\mathrm{cal}}(n)}{n} + \frac{b_1(np, x)}{np} + \frac{1}{L_{\mathrm{grid}}}. \tag{18}$$

The same right side bounds $|\widehat{C}_A - \alpha c_0|$. For triangular runs with fixed E_0, α, $np_n \to \infty$ and $W_0^{\mathrm{cal}}(n)/n \to 0$ in probability imply $\|\widehat{R}_A - K_1\|_\infty \to 0$ and $\widehat{C}_A \to \alpha c_0$ in probability.

Proof. We first control how the inverse CDF changes with the baseline weight. For fixed $w \in [0,1]$, let $E_w = wE_0 + (1-w)E_1$ and $K_w = E_w^{-1}$; the inverse is continuous on the closed unit interval. For $y \geq x$, $E_1(y) - E_1(x) \geq (1-\alpha)(y-x)$, so K_1 is $1/(1-\alpha)$-Lipschitz. Now $E_w - E_1 = w(1-\alpha)(E_0 - \mathrm{id})$. Evaluating at $x = K_w(q)$ gives

$$|K_w(q) - K_1(q)| = |K_1(E_1(K_w(q))) - K_1(q)| \leq c_0 w \quad (q \in [0,1]).$$

The inequality also holds at the endpoints, where both inverses equal zero or one.

Next, define $Z_t^A = E_t^A(Y_t - v)$ on an event and assign it any value otherwise. Strict monotonicity and the conditional uniform excess law imply

$$\mathbb{E}[M_t \mathbf{1}\{Z_t^A \leq q\} \mid \mathscr{P}_t] = pK_{w_{0t}}(q).$$

The increments $X_t(q) = M_t \mathbf{1}\{Z_t^A \leq q\} - pK_{w_{0t}}(q)$ are conditionally centered, bounded in absolute value by one, and have conditional variance at most p. Event observations fall at neither endpoint almost surely. Apply both tails of Lemma B.5 on the deterministic grid $\xi_j = E_1(j/L_{\mathrm{grid}})$, $j = 0, \ldots, L_{\mathrm{grid}}$. A union bound gives

$$\max_{0 \leq j \leq L_{\mathrm{grid}}} \left| \widehat{R}_A(\xi_j) - \frac{1}{n} \sum_{t \leq n} K_{w_{0t}}(\xi_j) \right| \leq \frac{b_1(np, x)}{np}$$

outside an event of probability $2(L_{\mathrm{grid}} + 1)e^{-x}$. Our inverse-CDF bound gives $|n^{-1} \sum_t K_{w_{0t}}(q) - K_1(q)| \leq c_0 W_0^{\mathrm{cal}}/n$ for every q. Because $K_1(\xi_j) = j/L_{\mathrm{grid}}$, monotonicity of $\widehat{R}_A$ and K_1 extends the grid inequality to the full interval with error at most $1/L_{\mathrm{grid}}$. This proves (18). Finally, $\|K_1 - \mathrm{id}\|_\infty = \|E_1 - \mathrm{id}\|_\infty = \alpha c_0$, so the reverse triangle inequality yields the combined-distance bound.

For the triangular sequence, orthogonality of the centered martingale increments gives $\mathbb{E}[(\sum_{t \leq n} X_t(\xi_j)/(np_n))^2] \leq 1/(np_n)$ at each fixed grid point. Chebyshev's inequality and a finite union bound make the grid error vanish for every fixed L_{grid}. Together with $W_0^{\mathrm{cal}}/n \to 0$, this proves uniform convergence: choose $1/L_{\mathrm{grid}}$ smaller than a prescribed tolerance, let $n \to \infty$, and interpolate by monotonicity. The conditions $np_n \to \infty$ and $W_0^{\mathrm{cal}}(n)/n \to 0$ thus suffice for this limit. □

B.6 Analytic usefulness construction

B.6.1 Identical bodies and imperfect tails

Use the target, forecast bank, parameters, and score nodes defined in Section 4.3. The positive prior is $\pi = (\pi_0, \pi_1)$ and the source-cell mass threshold is zero. With fixed initial information and no informative new covariates, $\mathscr{P}_t = \sigma(Y_s : s < t)$ up to completion. Thus the current outcome has law G_p conditional on the predictable weights.

Recall that $W_1^{\mathrm{cal}}(n) = n\bar{w}_{1,n}$. Within the proof, we suppress the horizon argument n when unambiguous.

Proof of Theorem 4.2. We begin by identifying the transferred source. Both F_0 and G_p assign mass $1-p$ to the body and p/J to every tail cell. Their convex combination has the same masses, so cell rescaling leaves P_1 unchanged. Adding the positive baseline component gives $F_1^* = \kappa F_0 + (1-\kappa)P_1 = \zeta F_0 + (1-\zeta)G_p$, which agrees with G_p on the body.

For the CRPS gain, put

$$A_r = \int_0^1 (s^r - s)^2 \, ds = \frac{1}{3} + \frac{1}{2r+1} - \frac{2}{r+2}, \qquad H_{\mathrm{tail}} = \frac{p^2 A_r}{J^2}.$$

Apply Lemma B.4; all laws have bounded support $[0,2]$ and hence finite first moments. The integral over $[0,1]$ vanishes. In tail cell j, the baseline CDF error is $ph(s^r - s)$ and $dy = h\,ds$. Summing over the J cells gives $p^2 J h^3 A_r = H_{\mathrm{tail}}$. Transfer multiplies this CDF error by ζ, so its scaled CRPS gain is $d_{\mathrm{tail}} = H_{\mathrm{tail}}(1-\zeta^2)/2$.

We next relate the cell-conditional score gain to the learned weights. The node with baseline within-cell probability u_ℓ has physical coordinate $s_\ell = u_\ell^{1/r}$. Its true conditional event probability is s_ℓ,

while the forecast probabilities are u_ℓ and $\zeta u_\ell + (1-\zeta)s_\ell$. The Bernoulli square identity gives the same improvement Δ_ν in every tail cell. Hence the loss difference $Z_t = g_{0t} - g_{1t}$ satisfies $\mathbb{E}[Z_t \mid \mathscr{P}_t] = p\Delta_\nu$, and the weighted expert loss obeys

$$\widetilde{G} - G_1 = \sum_{t \le n} w_{0t} Z_t.$$

The pathwise weighted-expert regret bound (12) in Lemma B.2 and Jensen's inequality imply

$$p\Delta_\nu \mathbb{E}(n - W_1^{\mathrm{cal}}) \le (1+\sqrt{2})\sqrt{a}\,\mathbb{E}\sqrt{N_n} \le (1+\sqrt{2})\sqrt{anp}.$$

Together with $W_1^{\mathrm{cal}} \ge 0$, this proves the mean-weight lower bound, clipped at zero. Markov's inequality then gives $W_1^{\mathrm{cal}}/n = \bar{w}_{1,n} \to 1$ in probability. To evaluate the aggregate forecast, put

$$\alpha_t = \zeta + (1-\zeta)w_{0t} = 1 - (1-\zeta)w_{1t}.$$

Then

$$F_t^A = \alpha_t F_0 + (1-\alpha_t)G_p, \qquad F_t^A - G_p = \alpha_t(F_0 - G_p).$$

The CRPS risk identity, conditional on $\mathscr{P}_t$, gives

$$\mathbb{E}\{b(F_0, Y_t) - b(F_t^A, Y_t) \mid \mathscr{P}_t\} = \frac{1}{2} H_{\mathrm{tail}}(1 - \alpha_t^2).$$

Write $w = w_{1t}$ and expand:

$$1 - [1-(1-\zeta)w]^2 = (1-\zeta^2)w + (1-\zeta)^2 w(1-w) \ge (1-\zeta^2)w.$$

Taking expectations proves (7). The extra nonnegative term quantifies the benefit of averaging the CDFs. Here the explicit iid construction allows us to calculate ordinary-score improvement directly, in addition to the weighted-expert regret bound in Lemma B.2.

It remains to study the combined ratio. Since the target conditional excess is uniform, a continuous, strictly increasing excess CDF E has population combined-ratio supremum distance $\sup_q |E^{-1}(q) - q| = \sup_x |x - E(x)|$. The maximum of $s - s^r$ is $(1-1/r)r^{-1/(r-1)}$, giving distances c_0 and ζc_0 for the two forecasts. A law of large numbers on a finite grid, followed by monotone interpolation, gives uniform convergence of their empirical combined ratios. To pass to the learned mixture, put $W_0^{\mathrm{cal}} = n - W_1^{\mathrm{cal}}$. We have $W_0^{\mathrm{cal}}/n \to 0$ in probability, and Lemma B.6 applies with $E_1 = \zeta E_0 + (1-\zeta)x$. The lemma controls inverse-CDF displacement and the centered counts on a fixed grid. It gives uniform convergence for the aggregate forecast whenever $np \to \infty$ and $W_0^{\mathrm{cal}}/n \to 0$, as holds here for fixed $p > 0$. The finite-path protection radius follows from Theorem 3.1, since the bank satisfies its assumptions. □

All gains in this example come from the tail. Their scale depends on the score: d_{tail} is of order p^2/J^2, whereas the cell-conditional score gap Δ_ν is independent of p and J. The convergence result keeps p, J fixed. Along a rare-event triangular sequence, the conditional mixture-gain identity remains exact within each row, but the ordinary-tail-score gain can vanish. Occurrence probabilities and protected cuts remain those of the baseline throughout.

Source-cell mass threshold and the rare-event limit. The strict-gain construction above uses $\epsilon_{\mathrm{cell}} = 0$. Each raw-source tail cell has mass p/J. With a fixed positive mass threshold, if $p/J \le \epsilon_{\mathrm{cell}}$, every tail cell falls back to its baseline conditional law. Because the bodies already agree, the transferred CDFs then coincide and the gain is zero. This happens eventually when $p_n \to 0$ with fixed J and a fixed positive mass threshold. Thus the frequency of fallback is part of the interpretation of a positive-threshold implementation; the strict gain proved here uses the zero mass threshold.

C Structural limitations and selected extensions

C.1 Score improvement with tail deterioration

The example below shows that an unconstrained forecast can improve expected CRPS while worsening CRSD. For a continuous upper forecast tail, write $Z_{F,v} = [F(Y) - F(v)]/[1 - F(v)]$ on $Y > v$.

For a target law G, deterministic forecast CDF F, and threshold v with $1 - F(v) > 0$, define the population combined-ratio supremum distance at the finite threshold

$$C_G(F; v) = \sup_{q \in [0,1]} \left| \frac{\mathbb{P}_G\{Y > v,\ Z_{F,v} \le q\}}{1 - F(v)} - q \right|.$$

For a predeclared proper score, with smaller values preferred, we say that an unconstrained comparator exhibits *tail deterioration masked by score improvement* if it lowers the target expected score relative to the baseline while increasing $C_G(F; v)$. We use $R_{\mathrm{CRPS}}(F) = \mathbb{E}_G\{\mathrm{CRPS}(F, Y)\}$ for expected CRPS under the target law.

Proposition C.1 (A compact-support example with tail deterioration)**.** *For every $\varepsilon \in (0, 1/6)$, let the target G_ε mix uniforms on $[0, 1]$ and $[1, 2]$ with masses $1 - \varepsilon$ and ε. Let T_ε use the same components with masses $1 - 2\varepsilon$ and 2ε. Let the baseline B_ε retain the target tail component and use, conditional on the body, the CDF*

$$H_\varepsilon(y) = y + \frac{5\varepsilon}{1 - \varepsilon} y(1 - y), \qquad 0 \le y \le 1.$$

Then all three CDFs are continuous and

$$R_{\mathrm{CRPS}}(B_\varepsilon) - R_{\mathrm{CRPS}}(G_\varepsilon) = 5\varepsilon^2/6,$$
$$R_{\mathrm{CRPS}}(T_\varepsilon) - R_{\mathrm{CRPS}}(G_\varepsilon) = 2\varepsilon^2/3,$$

so expected CRPS strictly prefers the unconstrained comparator. At $v = 1$, however,

$$C_{G_\varepsilon}(B_\varepsilon; 1) = 0, \qquad C_{G_\varepsilon}(T_\varepsilon; 1) = 1/2.$$

Proof of Proposition C.1. We first check that the body distribution is well defined. Put $a = 5\varepsilon/(1 - \varepsilon)$. The assumption $0 < \varepsilon < 1/6$ gives $0 < a < 1$, and

$$H_\varepsilon(0) = 0, \qquad H_\varepsilon(1) = 1, \qquad H'_\varepsilon(y) = 1 + a(1 - 2y) \ge 1 - a > 0.$$

Thus the body law and B_ε are continuous CDFs. The target and comparator forecasts are continuous CDFs by their mixture construction.

For the unconstrained comparator, $T_\varepsilon - G_\varepsilon = -\varepsilon y$ on $[0, 1]$ and equals $-\varepsilon(2 - y)$ on $[1, 2]$. Applying Lemma B.4 gives

$$R_{\mathrm{CRPS}}(T_\varepsilon) - R_{\mathrm{CRPS}}(G_\varepsilon) = \varepsilon^2 \left\{ \int_0^1 y^2 \, dy + \int_1^2 (2 - y)^2 \, dy \right\} = \frac{2\varepsilon^2}{3}.$$

The baseline error is confined to $[0, 1]$, where its CDF difference from the target is $5\varepsilon y(1 - y)$. Its excess risk is therefore

$$R_{\mathrm{CRPS}}(B_\varepsilon) - R_{\mathrm{CRPS}}(G_\varepsilon) = 25\varepsilon^2 \int_0^1 y^2 (1 - y)^2 \, dy = \frac{5\varepsilon^2}{6}.$$

Comparing the two expressions shows that CRPS prefers the unconstrained comparator.

For the tail criterion, take $v = 1$. All conditional laws of Y above the threshold are uniform on $[1, 2]$, with excess law uniform on $[0, 1]$. Consequently, for every $q \in [0, 1]$, the target numerator is $\mathbb{P}_{G_\varepsilon}(Y > 1, Z \le q) = \varepsilon q$. The baseline and comparator forecasts assign threshold exceedance probabilities ε and 2ε. Their combined ratios are therefore q and $q/2$, yielding the stated distances. □

Protected transfer in this example. Use (2) with baseline B_ε, raw source T_ε, $v = 1$, and $\epsilon_{\mathrm{cell}} = 0$, allowing transfer in the body as well as the tail. On the body and on every baseline tail cell, T_ε has the same conditional shape as G_ε. The baseline cell masses are those of G_ε, so replacing the source masses by these masses gives G_ε. Hence $F_* = \kappa B_\varepsilon + (1 - \kappa) G_\varepsilon$, and Lemma B.4 yields

$$R_{\mathrm{CRPS}}(F_*) - R_{\mathrm{CRPS}}(G_\varepsilon) = \kappa^2 \int_{\mathbb{R}} (B_\varepsilon(z) - G_\varepsilon(z))^2 \, dz = \kappa^2 \frac{5\varepsilon^2}{6}, \qquad C_{G_\varepsilon}(F_*; 1) = 0.$$

The last equality follows because its full tail agrees with the target. This calculation evaluates the transferred source under the general body-transfer construction; the weather implementation freezes the body.

C.2 THE COST OF EXACT PROTECTION

Our protection constraints preserve a finite set of tail cuts while leaving the conditional shape between them free. The next result characterizes this freedom and shows what happens as more cuts are protected.

Proposition C.2 (Characterization of the exact protected class)**.** *Fix one baseline, threshold and grid. Among CDFs with positive tail mass and the level-crossing convention in Lemma A.1, equality of every grid residual for every possible single outcome is equivalent to equality of tail mass and of all excess-PIT sublevel sets at the grid. For CDFs with unique interior crossings, these conditions are equivalent to equality of the tail probabilities together with equality of the corresponding physical quantile cuts. Adding protected grid levels can only shrink this class. If two continuous strictly increasing excess CDFs have the same cuts at every level in a dense subset of* $[0, 1]$*, they coincide.*

Proof. We recover the constraints one at a time from the residuals. At $q = 1$, the residual is $\mathbf{1}\{y > v\} - p$, so equality for even one outcome requires equal p. With the same positive p, residual equality at each remaining q_j means that the sublevel-set indicators agree on $y > v$. Under unique crossings, those sets are $(v, b_j]$; their equality is exactly equality of the cuts. Reversing the argument proves sufficiency. Each additional protected level imposes another equality, so the feasible class can only shrink.

For the dense-grid claim, suppose E, F differ at x, with $E(x) < F(x)$. Density provides a protected level q strictly between the two values. Then x belongs to one sublevel set and not the other, contradicting equality of the cuts. Hence $E = F$. The argument leaves body laws below v unrestricted, but shows that increasingly dense exact protection removes the freedom to make a fixed nonzero tail-CDF correction. □

Resolution and score scale.

Proposition C.3 (A resolution bound for integrated excess-CDF scores)**.** *For an outcome* y *and a probability measure* ν *on the excess scale, put* $s(E, y) = \int (E(x) - \mathbf{1}\{y - v \le x\})^2 \nu(dx)$*. Then every transferred expert and every aggregate forecast satisfy, on an exceedance,*

$$|s(E^*_{kt}, y) - s(E_{0t}, y)| \le 2(1 - \kappa)h.$$

Consequently, at fixed κ*, the difference from the baseline under this integrated excess-CDF score vanishes with* h*. The cell-conditional score in* (8) *uses a different within-cell normalization.*

Proof. For $a, b, z \in [0, 1]$, factor the difference of squares to obtain $|(a-z)^2 - (b-z)^2| = |a-b||a+b-2z| \le 2|a-b|$. Integrating against the probability measure, then applying (10) and Lemma A.2, proves the bound. A cell-conditional CDF divides by its cell mass, so the corresponding rescaled difference need not be of order h. The two score normalizations therefore have different resolution limits. □

Retained occurrence probabilities.

Proposition C.4 (Occurrence-only source differences vanish under transfer)**.** *If the baseline has tail probability* $2p$ *and a raw source has tail probability* p *but the two have identical conditional distributions in every baseline cell, their transferred complete CDFs coincide. In this setting the aggregate forecast class therefore retains the baseline occurrence ratio.*

Proof. Conditioning removes the raw occurrence multiplier, so Q^R_{ktj} equals Q^0_{tj} on every used cell, including the tail cells. Transfer in (2) therefore gives $\sum_j A_{tj} Q^0_{tj} = \mu_{0t}$. More generally, (4) always preserves the occurrence ratio $N_n/D_A = N_n/D_0$. Let λ denote the true exceedance probability. The population occurrence ratio remains $\lambda/(2p)$. In particular, if the source occurrence probability is correct, so that $\lambda = p$, this ratio equals $1/2$. Assigning positive average weight to the source leaves that error unchanged. □

C.3 STRICT CROSSING

Verification of Example 3.1. The source places no mass between 1 and $3/2$, leaving its CDF at $1/2$. The baseline CDF is $y/2$, which gives the stated PIT values. After the baseline component is added, the second cell has positive conditional mass in $(1, 5/4]$, and the transferred CDF is strictly above $1/2$. The positive component therefore repairs the indicator equality by restoring strict crossing across the source's support gap. □

C.4 APPROXIMATE QUANTILE CUTS AND FORECAST MASS

Pathwise protection extends to numerically computed quantiles that do not attain the nominal probability levels exactly, provided that cell masses are defined using the baseline probabilities at the computed cuts. The resulting bound is governed by the largest actual spacing between these probability levels.

Proposition C.5 (Protection with approximate quantile cuts)**.** *Under Assumption 2.1, retain the parameters of Section 3.1. Let $\widehat{b}_{tj}$ be ordered finite interior cuts chosen before the outcome, with $\widehat{b}_{t0} = v$ and $\widehat{b}_{tJ} = +\infty$. Write $r_{tj} = E_{0t}(\widehat{b}_{tj} - v)$ for their actual baseline excess probabilities and set $r_{t0} = 0, r_{tJ} = 1$. Construct* (2) *using the actual cell masses $p_t(r_{tj} - r_{t,j-1})$, omitting cells of zero mass. Issue a convex combination of the constructed CDFs with common weights $w_{kt} \geq 0$ chosen before the outcome, $\sum_k w_{kt} = 1$, independent of the CDF query argument. If $|r_{tj} - q_j| \leq \epsilon_q$ at every interior cut and date, then*

$$D_A = D_0, \qquad |\widehat{C}_A - \widehat{C}_0| \leq (1 - \kappa) \min\{1, h + 2\epsilon_q\}.$$

For learning under the full-feedback protocol in Section 2, use the score fractions and weights of Section 4.1 and the positive prior and event-doubling rule of Section 4.2. Compute the score nodes on the actual positive probability intervals $[r_{t,j-1}, r_{tj}]$. When an exceedance falls in a tail cell with zero baseline mass, assign every expert the common cost $c_{kt} = 1$ and give the aggregate forecast the same cost. A conditional-CDF score is undefined on such a cell, but the common penalty keeps loss differences bounded. On positive-mass cells, the mixture's conditional CDF is the weighted average of the expert conditional CDFs. The square-loss identity then bounds the aggregate loss by the weighted expert loss, with equality for a common penalty, so the event-doubling regret result continues to hold. Protection applies at the actual cuts; exact residual equality in (4) *need not hold at the nominal q_j. The constructed CDFs also give the sharper pathwise protection radius $(1 - \kappa) \max_{t,j}(r_{tj} - r_{t,j-1})$.*

Proof. The finite-sum construction preserves every actual cell mass, including the body mass, and gives a probability measure. Write each transfer as $\kappa\mu_{0t} + (1 - \kappa)\overline{\mu}_{kt}$, where $\overline{\mu}_{kt}$ has the baseline cell masses. The aggregate forecast is $\kappa\mu_{0t} + (1 - \kappa) \sum_k w_{kt}\overline{\mu}_{kt}$, whose last component has those masses as well.

Within an actual tail cell, the baseline CDF and the mixture CDF $\sum_k w_{kt}\overline{P}_{kt}$ lie between the same cumulative masses. After adding the baseline component, the normalized PIT displacement of the aggregate forecast from the baseline is therefore at most $(1 - \kappa)(r_{tj} - r_{t,j-1})$. Subtracting the two quantile-probability error bounds shows that each spacing is at most $h + 2\epsilon_q$; it is also at most one. Lemma A.3 now gives the claimed bound. □

D EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

D.1 EUPPBENCH PROTOCOL AND SUPPLEMENTARY RESULTS

D.1.1 DATA, SPATIAL SEPARATION, AND MODEL FITTING

We form 24-hour precipitation totals by summing the four six-hour forecast increments at lead times 6, 12, 18, and 24 hours, converting meters to millimeters. Historical fitting uses the 11-member reforecasts; evaluation uses the 51-member operational ensemble. The training and validation years are 1997–2015 and 2016–2017, respectively. Both endpoints of a historical accumulation window must belong to the same declared year. Evaluation is indexed by the 2018 issue year, so the last

verification window ends on 1 January 2019. Feedback is assumed available at the end of the accumulation window.

The five folds start from 122 stations and 118 forecast-grid groups. Stations sharing a grid point remain together. A shuffled stratified group split uses a 2×2 latitude–longitude stratification; the folds need not be contiguous geographic regions. All target-fold labels are excluded from model fitting and selection. The calendar contains $122 \times 365 = 44,530$ issuances, of which 42,502 at 117 stations have observed outcomes. The remaining 2,028 outcomes are neither scored nor used for weight updates. The thresholds are the linearly interpolated p90, p95, and p99 quantiles of the uncapped source-training labels, including dry days (Table 2). Source-model fitting and selection use the same protocol across thresholds and protection settings.

Table 2: Observed cases, source-training thresholds in mm, and observed exceedances by spatial fold. Across folds, p90, p95, and p99 have 3,081, 1,485, and 357 exceedances, respectively.

Fold	Cases	v_{90}	Events	v_{95}	Events	v_{99}	Events
0	8,335	7.9	640	12.5	319	25.6	80
1	7,986	8.0	583	12.8	302	26.0	82
2	8,359	8.1	539	13.0	242	26.7	43
3	8,747	7.9	727	12.5	349	25.0	91
4	9,075	8.0	592	12.9	273	26.3	61

The predictors comprise six dynamic predictors and four static station attributes from the preprocessed panel. The six dynamic predictors consist of four ensemble summaries—the fractions of members above 10 mm and at or below zero, $\log(1 + \bar{x})$, and $\log(1 + s_x)$—and two seasonal terms, the sine and cosine of day of year with a period of 365.25 days. Here, $\bar{x}$ and s_x denote the ensemble mean and standard deviation, respectively. The fixed 10 mm predictor is distinct from the evaluation threshold. The four static attributes are station latitude and longitude (in degrees), $\log\{1 + \max(h_{\text{station}}, 0)\}$, and $(h_{\text{model}} - h_{\text{station}})/1000$, where h_{station} and h_{model} are the station and corresponding model-grid elevations in meters. The elevation difference uses the original, unclipped station elevation. Feature standardization is fitted on source-training data. Eight regional source groups are formed by K-means using standardized source-station attributes and historical precipitation summaries. Each group supplies the same five model families, giving forty regional CDFs.

DRN and quantile regression forest (QRF) specifications are selected on source validation data with a one-standard-error rule based on 14-day block CRPS in physical mm. We choose the least complex eligible model, with p95 twCRPS as a tie-breaker. All folds select a one-layer, 32-unit DRN and a 200-tree QRF. The DRNs use ReLU activations and Adam; early stopping uses analytic CRPS on the $\sqrt{\text{mm}}$ scale, whereas model selection uses physical-mm CRPS. Global and regional DRNs retain their source-training checkpoints selected by validation early stopping. The other families are finally fitted on source training and validation data. Unlike QRF, which uses an empirical response distribution, QRF-EGP fits an extended generalized Pareto law from forest-weighted probability-weighted moments.

All seeds, model-search grids, numerical settings, and fold-specific training and validation row counts are provided in the accompanying configuration, `configs/euppbench2018_manuscript_settings.json`.

D.1.2 Protected mixtures and unprotected controls

We compare three partitions on the contour $(1-\kappa)/J = 0.05$: $(1, 0.95)$, $(3, 0.85)$, and $(5, 0.75)$. For each choice, the conditional grid is $q_j = j/J$, and the source-cell mass threshold is $\epsilon_{\text{cell}} = 10^{-10}$. A source with insufficient probability in a cell uses the baseline's conditional distribution there. The CDF at and below the precipitation threshold is unchanged. The prior is $\pi_0 = 1/2$ and $\pi_k = 1/80$ for each regional source. Fixed QAST retains this prior throughout evaluation.

QAST-E uses the cell-conditional score in (5), with 16 midpoint nodes $u_\ell = (\ell - 1/2)/16$ and equal weights. The score is evaluated within the cell containing the observed exceedance. Event epochs have lengths $1, 2, 4, \ldots$, learning rate $\eta = \sqrt{8 \log(80)/T_{\text{epoch}}}$, and a reset to the prior at the end of

each epoch. Each station has a separate chronological state. Before issuing a forecast, the learner releases all previously issued outcomes available by that time, updates its weights, and then freezes the current weights. Only observed exceedances update QAST-E. The aggregate forecast is the convex CDF mixture, rather than a sampled expert. The fixed-rate alternative proved in Appendix B.4 is not used in these comparisons.

Let $P_{0t} = F_{0t}$ and let w_t be QAST-E's weights chosen before the outcome. The unprotected mixture (same weights) is

$$F_t^{\mathrm{raw,match}} = \{\kappa + (1-\kappa)w_{0t}\}F_{0t} + (1-\kappa)\sum_{k=1}^{40} w_{kt}P_{kt}. \tag{19}$$

The unprotected mixture (fixed weights) replaces w_t with π. These controls hold the source bank and nominal baseline coefficient constant while removing protected transfer. Cell fallback can increase QAST's effective baseline retention beyond this nominal coefficient, so the comparison measures the complete protected construction, including fallback.

Two further controls use the unshrunk raw pool $\sum_{k=0}^{40} \pi_k P_{kt}$ or learn their own weights in $\sum_{k=0}^{40} v_{kt}P_{kt}$. The latter uses the mean Brier score at 16 unconditional baseline quantiles chosen before the outcome and updates on every released observed outcome. It therefore provides an independently learned pooling comparison, whereas (19) isolates the effect of protection at QAST-E's weights.

D.1.3 A TAIL-CALIBRATION TRAINING COMPARISON

We adapt the tail miscalibration (TMCB) penalty of Wessel et al. (2026) to the precipitation DRN. For N observed exceedances, predicted event mass $D = \sum_t p_t$, occurrence ratio $O = N/D$, and sorted conditional excess PITs $Z_{(1)}, \ldots, Z_{(N)}$, the finite-sample penalty is

$$\mathrm{TMCB} = \frac{1}{N}\sum_{i=1}^{N}\left|OZ_{(i)} - \frac{i}{N}\right|.$$

Each candidate starts from the same selected DRN and minimizes mean physical-mm CRPS plus $\gamma\,\mathrm{TMCB}$, with $\gamma \in \{0, 1, 5, 20\}$. Fine-tuning uses full-data Adam updates with a globally evaluated penalty. Among candidates whose validation CRPS is within 1% of that of the $\gamma = 0$ candidate, we select the candidate with the lowest validation TMCB, breaking ties in favor of smaller γ. This is a precipitation adaptation using a censored-logistic DRN on the square-root scale.

At p90 and p95, all ten fold–threshold selections choose $\gamma = 0$ (Table 3). At these thresholds, the positive-penalty candidates satisfy the CRPS budget but have higher validation TMCB, so the selected and zero-penalty forecasts coincide. Their additional CRPS fine-tuning distinguishes them from the original DRN baseline. At p99, fold 3 selects $\gamma = 20$, while the other four folds select $\gamma = 0$.

D.1.4 SCORING AND PAIRED INFERENCE

For threshold v, we use

$$\mathrm{twCRPS}_v(F, y) = \int_v^\infty \big(F(z) - \mathbf{1}\{y \le z\}\big)^2\, dz.$$

CRPS and twCRPS include every observed case, with annual means weighted by the number of cases in each fold. Continuous forecasts use 128-node Gauss–Legendre integration per segment, splitting at the observation, threshold, and protected cuts; an infinite-tail rational transformation is applied on the square-root scale. A 256-node calculation supplies numerical refinement. QRF uses 399 midpoint quantiles and a 799-quantile refinement; the raw ensemble uses exact empirical-distribution scores. CDF inversion uses bisection on square-root brackets.

Tail diagnostics follow Allen et al. (2025). Alongside CRSD, we report the occurrence ratio $O = N/D$, whose target is one, and the severity distance $S = \sup_{z\in[0,1]} |\widehat{G}(z) - z|$, where $\widehat{G}$ is the conditional excess-PIT empirical CDF. For a continuous forecast, the excess PIT is $1 - \overline{F}(y)/\overline{F}(v)$,

Table 3: Validation CRPS (mm) / TMCB for every penalty candidate at p90, p95, and p99. Bold entries indicate the selected candidate: $\gamma = 20$ for fold 3 at p99 and $\gamma = 0$ in all other rows.

Fold	Threshold	$\gamma = 0$	$\gamma = 1$	$\gamma = 5$	$\gamma = 20$
0	p90	**1.10849 / 0.02732**	1.10988 / 0.03821	1.10974 / 0.03885	1.10972 / 0.03911
0	p95	**1.10849 / 0.01323**	1.10998 / 0.03013	1.10996 / 0.03068	1.10988 / 0.03100
0	p99	**1.10849 / 0.07271**	1.10969 / 0.10679	1.10937 / 0.10698	1.10928 / 0.10696
1	p90	**1.11399 / 0.02476**	1.11474 / 0.03042	1.11461 / 0.03001	1.11456 / 0.02989
1	p95	**1.11399 / 0.01105**	1.11435 / 0.02024	1.11418 / 0.02015	1.11413 / 0.02010
1	p99	**1.11399 / 0.05649**	1.11432 / 0.08716	1.11382 / 0.08723	1.11358 / 0.08721
2	p90	**1.14943 / 0.02611**	1.15035 / 0.03508	1.14989 / 0.03484	1.14969 / 0.03474
2	p95	**1.14943 / 0.00978**	1.15041 / 0.03094	1.15011 / 0.03117	1.15002 / 0.03122
2	p99	**1.14943 / 0.09267**	1.14896 / 0.10421	1.14854 / 0.10430	1.14840 / 0.10431
3	p90	**1.09064 / 0.03175**	1.09171 / 0.03882	1.09172 / 0.03886	1.09172 / 0.03893
3	p95	**1.09064 / 0.02800**	1.09238 / 0.04762	1.09243 / 0.04763	1.09244 / 0.04766
3	p99	1.09064 / 0.04817	1.09210 / 0.04337	1.09229 / 0.04306	**1.09234 / 0.04302**
4	p90	**1.13392 / 0.03449**	1.13565 / 0.04492	1.13581 / 0.04528	1.13580 / 0.04538
4	p95	**1.13392 / 0.02340**	1.13517 / 0.03862	1.13511 / 0.03876	1.13510 / 0.03880
4	p99	**1.13392 / 0.04668**	1.13426 / 0.09293	1.13395 / 0.09278	1.13386 / 0.09275

where $\overline{F}(y) = 1 - F(y)$ denotes the survival function. Discrete forecasts use randomized PITs with reproducible, case-specific uniforms shared across methods. CRSD, occurrence, severity, and TMCB are calculated separately in each fold and then averaged equally. An observed exceedance with zero forecast tail probability makes the corresponding conditional-PIT diagnostics undefined. Occurrence and proper scores remain reportable. QAST's event probabilities and occurrence scores coincide with those of its own baseline.

Paired confidence intervals use 999 bootstrap replicates. Within each spatial fold, we resample forecast-grid station clusters and nonoverlapping seven-day date blocks, stratifying dates by verification year and quarter. Temporal draws are shared across stations and methods. The cluster and temporal multiplicities jointly weight each case, and nonlinear calibration functionals are recomputed within each replicate. Reported 95% intervals are the percentile intervals of paired differences. They condition on fitted models, selected specifications, and the realized sequence of aggregate forecasts; there is no refitting or relearning. Intervals are pointwise, without adjustment for multiple comparisons or selection among configurations.

D.1.5 Results across thresholds and partitions

Table 4 reports QAST-E's twCRPS differences from the unprotected mixture with the same weights at p90, p95, and p99, for three configurations with radius 0.05 under both baselines. All 18 displayed differences and their pointwise 95% paired intervals lie below zero. As J increases across these configurations, κ decreases: the unprotected mixture receives more source weight while the protection radius remains fixed.

At p95 with $J = 5$ and $\kappa = 0.75$, QAST-E's proper-score differences from its baseline are 8.0×10^{-6} mm under EMOS and 1.6×10^{-5} mm under DRN. The 95% intervals for both proper-score and CRSD differences include zero (Table 5). At p99 with the same J and κ, the twCRPS differences are -6.6×10^{-6} mm under EMOS and -4.7×10^{-6} mm under DRN; the intervals for both proper-score differences also include zero (Table 6). In these displayed comparisons, the substantial gains concern protection against unprotected pooling, while the retained occurrence probabilities continue to reflect the chosen baseline.

At p95 with $J = 5$ and $\kappa = 0.75$, fixed protected mixtures already avoid most of the losses incurred by matched unprotected aggregation (Table 1). Figure 3 shows small negative online-minus-fixed twCRPS point estimates across the displayed configurations. At p99, the five-cell differences are -4.3×10^{-7} mm under EMOS and -2.9×10^{-7} mm under DRN, with intervals including zero (Table 6). This separation between the substantial gain from protected aggregation and the smaller

Table 4: Proper-score comparisons for QAST-E across thresholds and partitions, with $(1-\kappa)/J = 0.05$ throughout. Absolute scores and paired twCRPS differences are in mm. Δ denotes QAST-E minus the unprotected mixture (same weights), followed by its pointwise 95% paired interval. Results are rounded to three significant figures.

Threshold	J	κ	Baseline	CRPS	twCRPS	ΔtwCRPS [95% CI]
p90	1	0.95	EMOS	0.894	0.388	-0.00380 $[-0.00542, -0.00225]$
p90	1	0.95	DRN	0.897	0.387	-0.00371 $[-0.00536, -0.00216]$
p90	3	0.85	EMOS	0.894	0.388	-0.0344 $[-0.0503, -0.0197]$
p90	3	0.85	DRN	0.897	0.388	-0.0334 $[-0.0481, -0.0192]$
p90	5	0.75	EMOS	0.894	0.388	-0.0931 $[-0.135, -0.0537]$
p90	5	0.75	DRN	0.897	0.388	-0.0904 $[-0.131, -0.0528]$
p95	1	0.95	EMOS	0.894	0.246	-0.00382 $[-0.00550, -0.00224]$
p95	1	0.95	DRN	0.897	0.246	-0.00375 $[-0.00539, -0.00218]$
p95	3	0.85	EMOS	0.894	0.247	-0.0346 $[-0.0511, -0.0197]$
p95	3	0.85	DRN	0.897	0.246	-0.0341 $[-0.0503, -0.0195]$
p95	5	0.75	EMOS	0.894	0.247	-0.0936 $[-0.137, -0.0539]$
p95	5	0.75	DRN	0.897	0.246	-0.0917 $[-0.133, -0.0536]$
p99	1	0.95	EMOS	0.894	0.0814	-0.00371 $[-0.00537, -0.00216]$
p99	1	0.95	DRN	0.897	0.0810	-0.00372 $[-0.00538, -0.00216]$
p99	3	0.85	EMOS	0.894	0.0814	-0.0333 $[-0.0483, -0.0192]$
p99	3	0.85	DRN	0.897	0.0810	-0.0332 $[-0.0480, -0.0192]$
p99	5	0.75	EMOS	0.894	0.0814	-0.0936 $[-0.135, -0.0545]$
p99	5	0.75	DRN	0.897	0.0810	-0.0951 $[-0.137, -0.0547]$

incremental gain from learning is consistent with the short event histories available at individual stations.

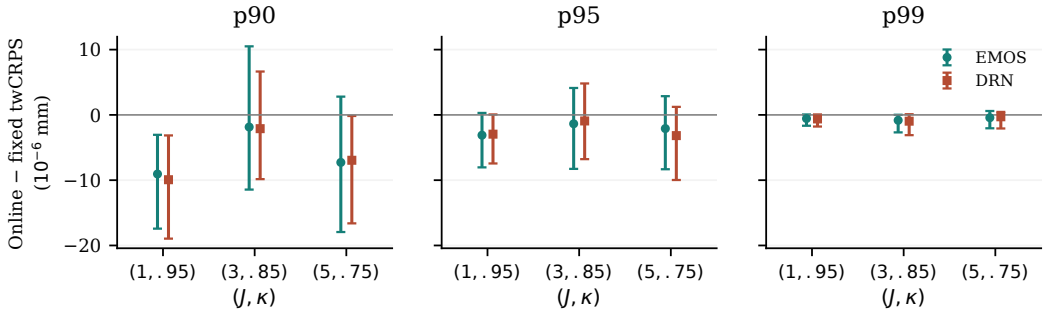

Figure 3: QAST-E minus Fixed QAST twCRPS at p90, p95, and p99, with pointwise 95% paired intervals. Negative values favor online weighting. Source pool, baseline, threshold, J, and κ are held fixed within each comparison; all displayed configurations have radius 0.05. The panels share the same vertical scale.

Tables 7–8 place these comparisons alongside the underlying forecast systems and the TMCB-adapted DRN. At p95, QRF and QRF-EGP attain lower twCRPS than the global EMOS and DRN baselines, with higher CRPS. The selected TMCB-adapted DRN at p95 has a lower CRSD point estimate than QAST-E, although the 95% intervals for their paired CRSD differences include zero (Table 5). The source protection comparison and the choice of a global forecasting model therefore address complementary decisions.

Table 5: Paired comparisons for QAST-E at p95, $J = 5$, and $\kappa = 0.75$. Differences and pointwise 95% intervals appear in separate columns. All entries are multiplied by 10^3 (proper-score units are 10^{-3} mm); negative values favor QAST-E.

QAST-E minus	Metric	Difference	95% interval
EMOS baseline			
Unprotected mixture (same weights)	CRPS	-95.348	$[-138.943, -55.098]$
	twCRPS	-93.590	$[-137.146, -53.881]$
	CRSD	-60.551	$[-69.971, 25.720]$
Baseline	CRPS	0.008	$[-0.079, 0.074]$
	twCRPS	0.008	$[-0.079, 0.074]$
	CRSD	-0.258	$[-1.748, 1.316]$
Fixed QAST	CRPS	-0.002	$[-0.008, 0.003]$
	twCRPS	-0.002	$[-0.008, 0.003]$
	CRSD	0.038	$[-0.137, 0.130]$
Raw mixture (independently learned)	CRPS	-104.195	$[-225.840, -33.867]$
	twCRPS	-106.929	$[-225.683, -41.693]$
	CRSD	-80.853	$[-95.765, 14.017]$
Selected TMCB DRN	CRPS	-2.517	$[-13.729, 6.522]$
	twCRPS	0.790	$[-3.216, 5.391]$
	CRSD	11.682	$[-31.729, 32.465]$
DRN baseline			
Unprotected mixture (same weights)	CRPS	-92.320	$[-133.949, -53.482]$
	twCRPS	-91.743	$[-133.304, -53.607]$
	CRSD	-58.392	$[-69.507, 27.281]$
Baseline	CRPS	0.016	$[-0.063, 0.081]$
	twCRPS	0.016	$[-0.063, 0.081]$
	CRSD	-0.278	$[-1.536, 1.593]$
Fixed QAST	CRPS	-0.003	$[-0.010, 0.001]$
	twCRPS	-0.003	$[-0.010, 0.001]$
	CRSD	0.005	$[-0.095, 0.125]$
Raw mixture (independently learned)	CRPS	-98.653	$[-218.778, -26.614]$
	twCRPS	-105.473	$[-224.804, -39.872]$
	CRSD	-78.359	$[-97.364, 21.102]$
Selected TMCB DRN	CRPS	0.172	$[-1.374, 1.477]$
	twCRPS	0.044	$[-0.782, 0.789]$
	CRSD	10.782	$[-17.961, 18.334]$

Table 6: Paired comparisons for QAST-E at p99, $J = 5$, and $\kappa = 0.75$. Each entry gives the difference followed by its pointwise 95% paired interval. All entries are multiplied by 10^3 (proper-score units are 10^{-3} mm) and rounded to three significant figures. Negative differences favor QAST-E. The unprotected mixture uses QAST-E's weights and nominal baseline coefficient.

QAST-E minus	ΔCRPS [95% CI]	ΔtwCRPS [95% CI]
EMOS baseline		
Unprotected mixture (same weights)	-96.0 $[-138, -57.1]$	-93.6 $[-135, -54.5]$
Baseline	-0.00656 $[-0.0291, 0.00437]$	-0.00656 $[-0.0291, 0.00437]$
Fixed QAST	-0.000430 $[-0.00205, 0.000598]$	-0.000430 $[-0.00205, 0.000598]$
DRN baseline		
Unprotected mixture (same weights)	-96.1 $[-138, -55.5]$	-95.1 $[-137, -54.7]$
Baseline	-0.00471 $[-0.0240, 0.00596]$	-0.00471 $[-0.0240, 0.00596]$
Fixed QAST	-0.000291 $[-0.00208, 0.000457]$	-0.000291 $[-0.00208, 0.000457]$

Table 7: Forecast systems and aggregation methods at p90, $J = 5$, and $\kappa = 0.75$. All methods use the same 42,502 observed cases. CRPS and twCRPS are in mm; CRSD, severity S, and TMCB have target zero, and occurrence O has target one. The selected TMCB DRN coincides with its zero-penalty candidate.

Forecast	CRPS	twCRPS	CRSD	O	S	TMCB
Raw ensemble	1.04463	0.41017	–	0.98179	–	–
Global EMOS	0.89410	0.38769	0.07489	0.98456	0.05132	0.04729
Global DRN	0.89678	0.38744	0.07616	0.98921	0.05257	0.04831
QRF	0.90761	0.38790	0.06485	1.00190	0.03976	0.03763
QRF-EGP	0.91177	0.38954	0.07651	0.99300	0.05507	0.04524
BMA	0.93284	0.40498	0.13632	1.07807	0.07502	0.04768
Distributional forest	0.93558	0.39797	0.16018	0.87205	0.09130	0.10191
TMCB DRN ($\gamma = 0$; selected)	0.89662	0.38742	0.07548	1.00899	0.04463	0.04353
EMOS: raw fixed	2.44211	1.92408	0.19040	0.80960	0.13634	0.15462
EMOS: unprotected mixture (fixed weights)	0.99138	0.48377	0.10097	0.93263	0.07961	0.06862
EMOS: unprotected mixture (same weights)	0.98848	0.48089	0.10094	0.93336	0.07898	0.06847
EMOS: Raw mixture (independently learned)	0.99971	0.49631	0.12285	0.92046	0.09447	0.08704
EMOS: Fixed QAST	0.89421	0.38780	0.07436	0.98456	0.05137	0.04730
EMOS: QAST-E	0.89420	0.38779	0.07434	0.98456	0.05137	0.04730
DRN: raw fixed	2.44055	1.92219	0.18828	0.81172	0.13584	0.15365
DRN: unprotected mixture (fixed weights)	0.99267	0.48300	0.09938	0.93671	0.07809	0.06539
DRN: unprotected mixture (same weights)	0.98758	0.47793	0.09892	0.93739	0.07767	0.06530
DRN: Raw mixture (independently learned)	0.99618	0.49292	0.11982	0.92357	0.09246	0.08478
DRN: Fixed QAST	0.89689	0.38756	0.07572	0.98921	0.05287	0.04830
DRN: QAST-E	0.89689	0.38755	0.07575	0.98921	0.05285	0.04830

Table 8: Forecast systems and aggregation methods at p95, $J = 5$, and $\kappa = 0.75$. All methods use the same 42,502 observed cases. CRPS and twCRPS are in mm; CRSD, severity S, and TMCB have target zero, and occurrence O has target one. The selected TMCB DRN coincides with its zero-penalty candidate.

Forecast	CRPS	twCRPS	CRSD	O	S	TMCB
Raw ensemble	1.04463	0.25829	–	1.19767	–	–
Global EMOS	0.89410	0.24651	0.09123	0.94717	0.06401	0.05086
Global DRN	0.89678	0.24576	0.09035	0.94764	0.06053	0.05524
QRF	0.90761	0.24435	0.09382	0.97732	0.06695	0.04949
QRF-EGP	0.91177	0.24492	0.09710	0.95176	0.05997	0.05531
BMA	0.93284	0.25313	0.13701	1.06345	0.08041	0.06009
Distributional forest	0.93558	0.25091	0.18572	0.85526	0.07884	0.09010
TMCB DRN ($\gamma = 0$; selected)	0.89662	0.24573	0.07929	0.97572	0.06154	0.05020
EMOS: raw fixed	2.42680	1.76295	0.33345	0.66655	0.12739	0.21595
EMOS: unprotected mixture (fixed weights)	0.99039	0.34105	0.15203	0.85266	0.06733	0.09567
EMOS: unprotected mixture (same weights)	0.98945	0.34011	0.15152	0.85298	0.06716	0.09556
EMOS: Raw mixture (independently learned)	0.99830	0.35345	0.17182	0.83283	0.07840	0.11434
EMOS: Fixed QAST	0.89411	0.24652	0.09093	0.94717	0.06387	0.05084
EMOS: QAST-E	0.89410	0.24652	0.09097	0.94717	0.06387	0.05084
DRN: raw fixed	2.42527	1.76144	0.33256	0.66744	0.13040	0.21688
DRN: unprotected mixture (fixed weights)	0.99168	0.34009	0.14862	0.85393	0.06727	0.09703
DRN: unprotected mixture (same weights)	0.98911	0.33752	0.14846	0.85413	0.06745	0.09698
DRN: Raw mixture (independently learned)	0.99545	0.35125	0.16843	0.83439	0.07940	0.11488
DRN: Fixed QAST	0.89680	0.24578	0.09006	0.94764	0.06040	0.05521
DRN: QAST-E	0.89679	0.24578	0.09007	0.94764	0.06041	0.05520

D.1.6 NUMERICAL RESOLUTION

Table 9 compares 128-node, 256-node, and adaptive integration for two selected station-day comparisons of QAST-E with the unprotected mixture using the same weights.

For each of the two inspected stations in Table 9, we select its largest saved daily twCRPS reduction from the unprotected mixture (same weights) to QAST-E. Adaptive integration on the square-root scale splits at source quantiles, logistic location-scale knots, the observation, threshold, and protected cuts. Table 9 retains a positive reduction in both cases. These targeted calculations use an independent integration routine with the same frozen CDF implementations. This check concerns the two displayed QAST-E versus unprotected-mixture comparisons.

Table 9: Daily twCRPS reduction (unprotected mixture (same weights) minus QAST-E, in mm) at two inspected stations. Dates index issuance. Both cases use the EMOS baseline, p95, $J = 5$, and $\kappa = 0.75$.

Fold	Station	Date	128 nodes	256 nodes	Adaptive
3	350	2018-10-30	0.145012	0.146840	0.146475
4	1443	2018-01-22	0.257396	0.241498	0.246021

D.2 A SUPPLEMENTARY WEATHERBENCH 2/METAR APPLICATION

This supplementary panel combines WeatherBench 2 forecasts with METAR-based observations to compare two source-model families across three protected partitions. The observations are obtained from NOAA's Meteorological Assimilation Data Ingest System (MADIS) via the WeatherBench 2 METAR archive. Their construction differs from the EUPPBench protocol.

D.2.1 DATA AND FORECASTS

The panel covers 100 stations over 2020–2022. Five spatial holdouts crossed with three future-year tests yield 15 folds and 109,600 station-day cases. Model fitting uses source-station data from the preceding three years, and the threshold is the source-training 95th percentile.

We use Global EMOS and Global DRN as alternative baselines, with eight regional EMOS or DRN source models. Small source groups are augmented with similar source stations. Three-fold spatial validation within the source sample selects the minimum group size from $\{1, 5, 10\}$ by mean source twCRPS, and both model families use the resulting memberships. QAST-E transfers source conditional laws within the baseline cells and issues a convex CDF average using weights chosen before the outcome and updated after exceedances.

For equal-probability tail cells, Theorem 3.1 gives $\delta = (1-\kappa)/J$. The three configurations $(J, \kappa) = (1, 0.90), (3, 0.90), (5, 0.75)$ have radii 0.100, $0.03\overline{3}$, and 0.050, respectively.

Table 10 evaluates the aggregate forecast and keeps all three metrics together for each source–baseline pair. CRPS and twCRPS are case-weighted averages in mm, computed by quantile approximation; CRSD is the equally weighted mean of the 15 fold-wise CRSD values.

D.2.2 PREDICTIVE-SCORE STABILITY AND CRSD IMPROVEMENTS

Across the twelve baseline–source–partition comparisons in Table 10, the reported proper-score point estimates remain close to those of the corresponding baselines: the absolute relative changes are below 0.032% for CRPS and 0.088% for twCRPS. CRSD point estimates are lower in nine of the twelve comparisons. These results describe numerical changes within this panel.

The largest gain is obtained with EMOS sources and the Global EMOS baseline at $(J, \kappa) = (1, 0.90)$, where CRSD decreases from 0.126261 to 0.124644, a 1.28% reduction, together with a 0.0318% reduction in CRPS. With the same sources and the Global DRN baseline, CRSD decreases by 0.75% and CRPS by 0.0228%.

The CRSD improvements are especially consistent for Global EMOS: both source families reduce CRSD at all three partitions. For Global DRN, both source families reduce CRSD at $J = 3$, and

Table 10: Performance of QAST-E on the WeatherBench 2/METAR panel. Columns group forecasts by their protected baseline, and source rows identify the regional model family. CRPS and twCRPS are in mm; CRSD is dimensionless. Boldface marks CRSD point estimates below the corresponding baseline value.

	EMOS baseline			DRN baseline		
Source family	CRPS	twCRPS	CRSD	CRPS	twCRPS	CRSD
Baseline only	1.621961	0.591182	0.126261	1.618339	0.588014	0.149917
$J = 1, \quad \kappa = 0.90, \quad \delta = 0.100$						
EMOS	1.621445	0.590666	**0.124644**	1.617970	0.587644	**0.148786**
DRN	1.621909	0.591130	**0.125968**	1.618247	0.587921	0.149923
$J = 3, \quad \kappa = 0.90, \quad \delta \simeq 0.0333$						
EMOS	1.621906	0.591127	**0.125963**	1.618280	0.587955	**0.149810**
DRN	1.621957	0.591179	**0.126241**	1.618310	0.587984	**0.149908**
$J = 5, \quad \kappa = 0.75, \quad \delta = 0.050$						
EMOS	1.621920	0.591141	**0.126128**	1.618291	0.587965	0.150026
DRN	1.621962	0.591183	**0.126185**	1.618308	0.587982	0.150062

EMOS sources also reduce it at $J = 1$; the remaining three configurations produce only modest increases. In this panel, CRSD point estimates are lower than their corresponding baselines in nine of twelve configurations, while proper-score point estimates remain close to baseline values.