A Quantitative Framework for Macroeconomic Forecasting and Uncertainty Measurement

Supervisor

Nisarg Kamdar

Portfolio Manager

Prepared by

Nicholas Wong

Graduate Student, MIT

Summer 2025

Contents

1 Introduction

We present a comprehensive, generalizable pipeline that delivers real-time point and distributional forecasts for key U.S. macroeconomic releases. Our goal is to create a streamlined framework that (i) ingests historical macroeconomic data, (ii) produces point/directional predictions, and (iii) defines calibrated prediction bands to quantify uncertainty to assist with portfolio construction, position sizing, and risk management decisions ahead of notable macroeconomic releases.

1.1 Scope

This study focuses on the **Change in Nonfarm Payrolls (NFP TCH)**, released monthly on the first Friday at 8:30 a.m. ET. NFP measures the net change in U.S. non-farm employment and is generally regarded as the most influential labor-market indicator for financial markets. It is a primary driver of interest rate expectations and FX volatility, particularly in the immediate aftermath of its release where large surprises have historically produced substantial knee-jerk market reactions.

By concentrating on a single, high-impact indicator, we conduct a targeted evaluation of the proposed forecasting framework. While the empirical results presented are specific to NFP, the analytical methodology is generalizable to other macroeconomic releases with similar characteristics and sufficient forecast coverage.

2 Analytical Framework

We break down our unified analytical pipeline into five stages.

- **Data:** Each workbook is reshaped into a long panel and written to *parquet*. Schema harmonization happens here, so all downstream code is indicator-agnostic and lends to our framework's generalizability. We restrict the analysis window to begin in June 2003 (2003–06), coinciding with the introduction of the CES birth–death adjustment; pre-2003 vintages are excluded to avoid a methodology break.
- **Exploration:** Diagnostics such as rolling error plots, distribution tests (Ljung-Box, Kolmogorov-Smirnov), and analytical regressions are run to detect regime shifts and

inform subsequent modeling decisions.

- **Point and Directional Forecast Ensembles:** We deploy five forecast engines: (i) static inverse-error weighting, (ii) exponentially weighted moving averages, (iii) soft Bayesian model averaging with Student-t likelihoods, (iv) multiplicative weights update and (v) a *robust majority-vote ensemble*. For each method, hyperparameter grids are traversed in walk-forward loops that yield out-of-sample smart predictions and directional calls relative to the consensus median.
- **Distributional Engines:** Four distributional forecasting methods are deployed to quantify uncertainty: Student-t, t-GARCH, Gaussian mixture models (GMM), and Bayesian model averaging (BMA); an optional spread-elastic crisis multiplier dynamically adjusts to time-varying volatility. For each distributional engine, the result is a full predictive density and calibrated prediction intervals for every release.
- **Evaluation:** A common rubric scores point, directional, and interval performance (RMSE, DM; hit rate, binomial, PT; coverage targets and AC). Formal definitions appear in §A.1.

3 Exploratory Analysis

3.1 Median-Error Distribution and the COVID Shock

We begin by inspecting the stability of consensus accuracy through time. Figure 1 plots the 6-month rolling RMSE of the crowd median. The pandemic period features a vertical spike that dwarfs the surrounding history, indicating that a handful of months dominate squared-error risk. This motivates maintaining *two* evaluation panels throughout the paper: a **full** panel (complete history, including the pandemic shock) and a **COVID-filtered** panel that excludes the 2020–2022 extremes when we want to study typical regimes. These findings also suggest that models have to be *regime-aware*.

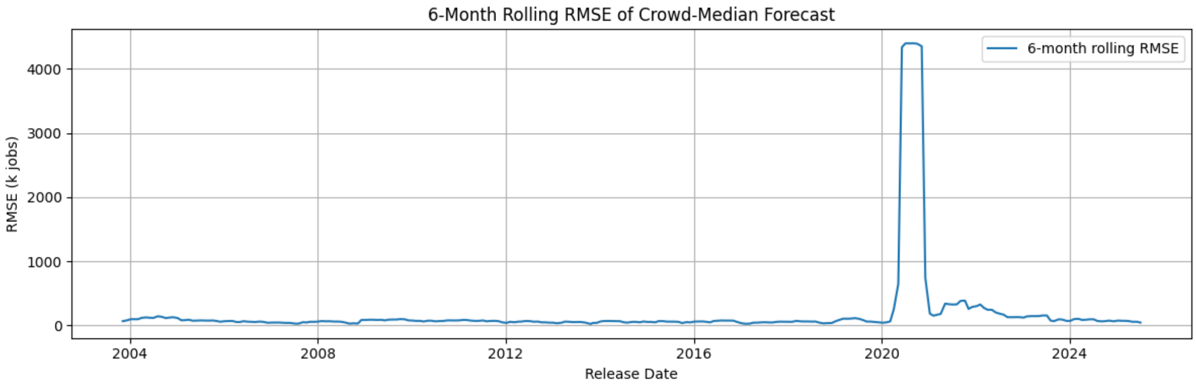

Figure 1: Six-month rolling RMSE of the crowd-median NFP forecast. The pandemic shock generates a discontinuous jump in error magnitude, motivating the use of both full and COVID-filtered panels.

3.2 Distributional form of median errors.

On the COVID-filtered panel, the distribution of median-forecast errors is well captured by a symmetric, heavy-tailed Student-t. Figure 5 overlays the fitted $\mathrm{t}_\nu(\mu, \sigma)$ density on the histogram of median errors; Figure 3 shows a QQ-plot against the Normal, highlighting tail deviations consistent with excess kurtosis. A formal goodness-of-fit check against the *fitted* Student-t does not reject at conventional levels:

$$\text{K–S vs. fitted } t^{1}\text{: } D = 0.036,\ p = 0.9143 \qquad \text{and} \qquad \text{CvM}^{2}\text{: } W^2 = 0.022,\ p = 0.9947.$$

These diagnostics support using a Student-t baseline for error modeling.

[2] One-sample Kolmogorov–Smirnov goodness-of-fit test comparing the empirical CDF of the median errors to the fully specified Student-t CDF fitted by MLE; $D = \sup_x |F_n(x) - F(x)|$. The null is that the data are i.i.d. draws from that distribution.

[2] Cramér–von Mises goodness-of-fit test using the integrated squared difference $\int (F_n(x) - F(x))^2 \, dF(x)$, emphasizing overall shape rather than the single worst deviation. Same null as above.

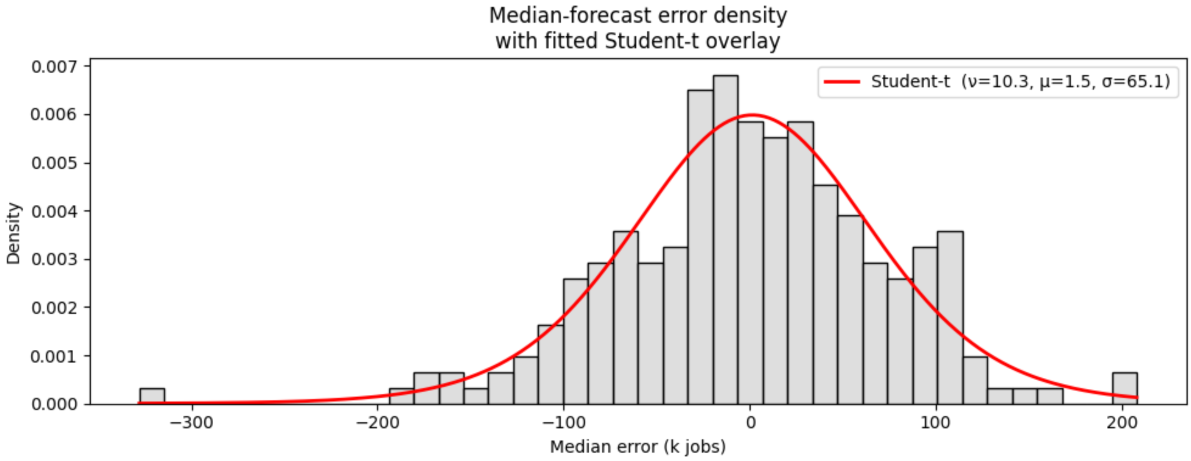

Figure 2: Histogram of median-forecast errors with fitted Student-t overlay (COVID-filtered panel). The fit captures central mass and tails.

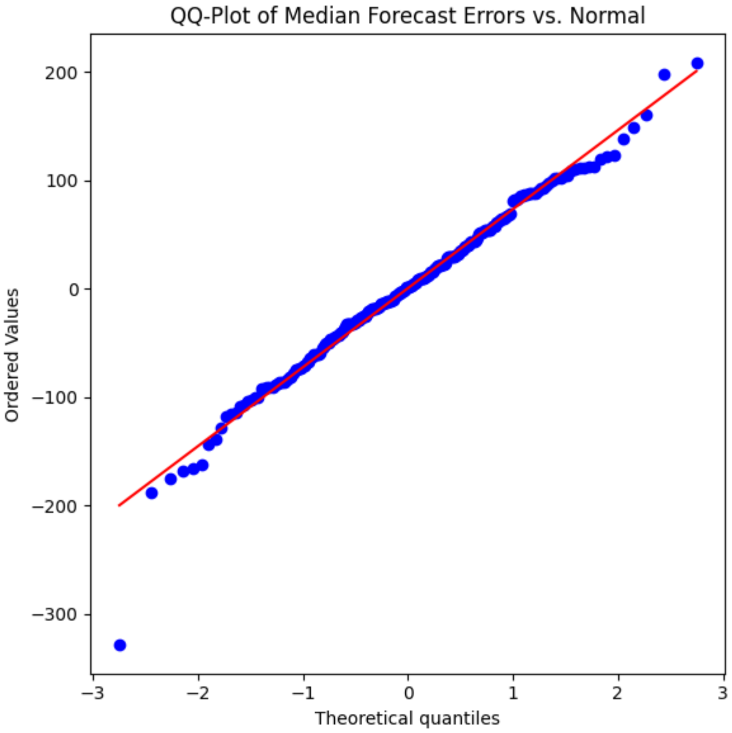

Figure 3: QQ-plot of median-forecast errors vs. Normal (COVID-filtered panel). Systematic tail departures motivate heavy-tailed modeling.

Two consequences flow from this exploration:

1. *Panel design.* Because the pandemic months dominate RMSE, we report results on

both the full and COVID-filtered panels to separate typical calibration from crisis behavior.

2. *Distributional choice.* The adequacy of the Student-t fit informs our distributional engines (§4.2.1, §4.2.2, §4.2.3, §4.2.4) and our *soft-BMA* weighting scheme (§4.1.3), which explicitly leverages Student-t likelihoods to remain robust to fat tails.

3.3 Cross-Sectional Spread as an Ex-Ante Proxy of Forecast Risk

Before the print we observe the *cross-sectional spread* of submitted forecasts, $s_t = \text{stdev}\{f_{i,t}\}$. After the print we observe the *realized miss* of the crowd median, $|e_t| = |f_t^{\text{med}} - y_t|$. If disagreement contains information about event risk, s_t should co-move with $|e_t|$. This gives us an ex-ante knob to widen (or not) our prediction bands when quantifying uncertainty.

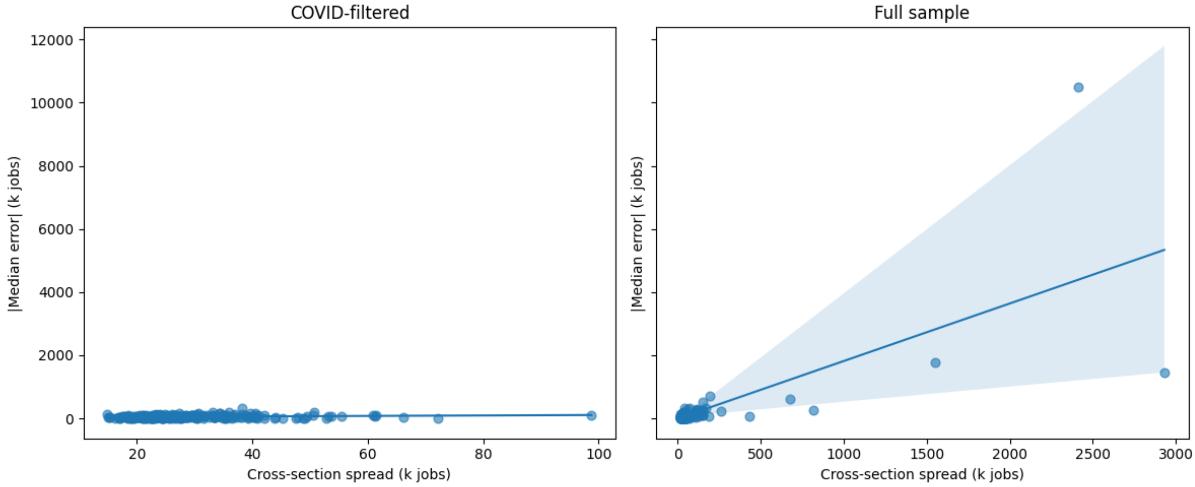

Figure 4: Cross-section spread vs. absolute median error (COVID-filtered vs. full sample).

Table 1 reports linear (Pearson) and rank (Spearman) correlations between s_t and $|e_t|$. The COVID-filtered panel shows small but statistically significant associations, while the full sample shows a very strong Pearson correlation and a moderate Spearman correlation. The gap between the Pearson and Spearman statistics suggests that a handful of extreme months bend the relationship, which is precisely the regime where interval width matters most.

Table 1: Correlation between cross-sectional spread s_t and absolute median error $|e_t|$.

Panel	**Pearson r (p-val)**	**Spearman ρ (p-val)**
COVID-filtered	0.168 (0.0105)	0.182 (0.0057)
Full sample	0.710 ($< 10^{-4}$)	0.347 ($< 10^{-4}$)

We formalize the slope via a Newey–West OLS on logs (leveraging properties of elasticity in a log-log regression),

$$\ln |e_t| \;=\; \beta_0 + \beta_1 \ln s_t + \varepsilon_t,$$

so β_1 is the *elasticity* of the miss with respect to spread. Results:

Table 2: Log–log regression $\ln |e_t| = \beta_0 + \beta_1 \ln s_t$ (HAC).

Panel	**N**	$\hat{\beta}_0$	$\hat{\beta}_1$	**SE**($\hat{\beta}_1$)	**p-val**	R^2
COVID-filtered	228	2.596	0.312	0.224	0.164	0.009
Full sample	264	1.019	**0.792**	0.074	$< 10^{-22}$	0.236

The COVID-filtered elasticity is small and statistically indistinguishable from zero; the full-sample elasticity is ≈ 0.79, highly significant. Interpretation: outside crises, disagreement adds little incremental information; in crisis months, disagreement scales errors almost proportionally on a log scale. The discrepancy in statistical significance of the coefficients implies a dominance by a few enormous months and informs our choice of a gated, crisis-dependent adjustment.

The elasticity $\hat{\beta}_1 \approx 0.79$ suggests a transparent rescaling of any baseline half-width $h_{L,t}$:

$$m_t \;=\; \left(\frac{s_t}{\text{median}\{s_{t-k} : 1 \le k \le 24\}}\right)^{\beta_t}, \qquad h^{\text{adj}}_{L,t} \;=\; m_t \cdot h_{L,t}.$$

To avoid unnecessary widening in ordinary months (where the COVID-filtered slope is weak), we *gate* the exponent:

$$\beta_t \;=\; \begin{cases} 0, & s_t \le \text{Pct}_{95}\{s_{t-k} : 1 \le k \le 24\}, \\ 0.80, & s_t > \text{Pct}_{95}\{s_{t-k} : 1 \le k \le 24\}. \end{cases}$$

In doing so, we are able to (i) use the crowd's contemporaneous disagreement s_t as an ex-ante stress proxy; (ii) anchor the curvature of the adjustment to the estimated elasticity (≈ 0.8); (iii) activate only in the right tail of disagreement where both correlation and regression signal are strongest; (iv) select the gating percentile via walk-forward coverage validation (we use the 95th percentile heuristic; this can be scaled to account for systematic under- or overcoverage). The result is a crisis-elastic interval: tight in tranquil regimes, wider exactly when disagreement telegraphs tail risk.

3.4 Cross-Section vs. Rolling Time-Series Student-t Coverage

Given the correlation between disagreement in the cross-section and median errors, we might ask: can we calibrate reliable prediction intervals straight from the *cross-section* of economists' forecasts at each release (XS-T), or do we need a *rolling time-series* fit to past median-forecast errors (TS-T)? In simpler terms: does the *cross-section* contain enough information to properly quantify uncertainty, or do we necessarily need information from errors in the *time-series*?

For each release t and nominal level $L \in \{50, 60, 70, 80, 90, 95\}\%$, we test whether the realized print falls inside a two-sided Student-t band:

- XS-T: fit $\mathrm{t}_\nu(\mu, \sigma)$ to the contemporaneous cross-section of forecasts; use $\hat{\mu}$ as the center and $\hat{\sigma}, \hat{\nu}$ for the half-width.
- TS-T$_W$: fit $\mathrm{t}_\nu(\mu, \sigma)$ on the last W months of median-forecast errors; center at $f_t^{\mathrm{med}} + \hat{\mu}$ with half-width from $\hat{\sigma}, \hat{\nu}$; $W \in \{12, 24, 36, 60, 120\}$.

Calibration is summarized by empirical coverage at each L and by the mean-absolute gap (MAG) between empirical and nominal coverages (smaller is better).

Table 3: Empirical coverage versus nominal targets for Student-t bands centerd at $\hat{\mu}$. The final column reports the mean-absolute gap (MAG).

Method	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**MAG**
Full-TS-t 120m	0.425	0.534	0.623	0.733	0.822	0.918	0.066
Full-TS-t 12m	0.449	0.547	0.657	0.728	0.862	0.921	0.047
Full-TS-t 24m	0.475	0.566	0.674	0.748	0.893	0.942	0.025
Full-TS-t 36m	0.470	0.565	0.643	0.770	0.887	0.943	0.029
Full-TS-t 60m	0.427	0.539	0.670	0.733	0.888	0.937	0.043
Full-XS-t	0.233	0.289	0.368	0.425	0.496	0.586	0.342

Table 3 reports results on the full panel.

- *Cross-section alone under-covers materially.* FULL-XS-T delivers 0.586 empirical coverage at the 95% band and 0.233 at the 50% band, yielding a large MAG of 0.342. The shortfall is broad-based across all levels, indicating that disagreement snapshots do not by themselves encode a stable predictive density.

- *Rolling time-series fits calibrate well, with a clear sweet spot at 24 months.* FULL-TS-T$_{24\text{m}}$ attains the lowest MAG (0.025) and tracks targets closely across the grid (e.g., 95% → 0.942, 90% → 0.893, 50% → 0.475). Windows that are too short (12m, MAG = 0.047) are overly reactive, while very long windows (60–120m, MAG = 0.043–0.066) are sluggish and under-adjust through regime shifts. The 36m window is competitive (MAG = 0.029) but marginally less aligned than 24m.

Implications

1. *Spread is informative but not sufficient.* The XS-T experiment shows that using the cross-section to *construct* intervals leads to systematic under-coverage. We therefore do not build bands directly from the cross-section.

2. *Adopt a* **24***-month rolling window as the baseline interval engine.* The 24-month window provides the best accuracy–stability trade-off for NFP. It also provides more dynamic adjustment to time-varying volatility relative to longer rolling windows. Thus, we deploy a 24-month rolling window for prediction band estimation for all methods

across our distributional engines.

3. *Use spread only as a modifier.* Given its directional signal in extremes, we incorporate cross-sectional disagreement as a *crisis multiplier* applied to the well-calibrated TS-$\text{T}_{24\text{m}}$ bands rather than as a stand-alone density.

4 Methods

In this section, we present an overview of our forecasting models, including ensemble methods and distributional engines tailored for macroeconomic time series.

4.1 Point and Directional Forecasts

Contiguity filter. All point–forecast engines in this study apply a *contiguity filter* to the economist panel before constructing weights or aggregating forecasts. The filter requires that an economist must have submitted non–missing forecasts for each of the W most recent releases in the chosen look–back window to be eligible for inclusion. This rule serves two purposes. First, it screens out sporadic forecasters whose intermittent submissions can inject high–variance noise into the aggregation, especially if they happen to be correct in a single outlier month and are overweighted by naive inverse–error schemes. Second, it stabilizes the composition of the forecast pool, ensuring that performance statistics used for weighting (e.g., mean squared error, log–likelihood) are based on comparable forecast histories rather than irregular or incomplete records. In practice, the contiguity filter reduces weight volatility and anchors aggregation to forecasters with demonstrated consistency.

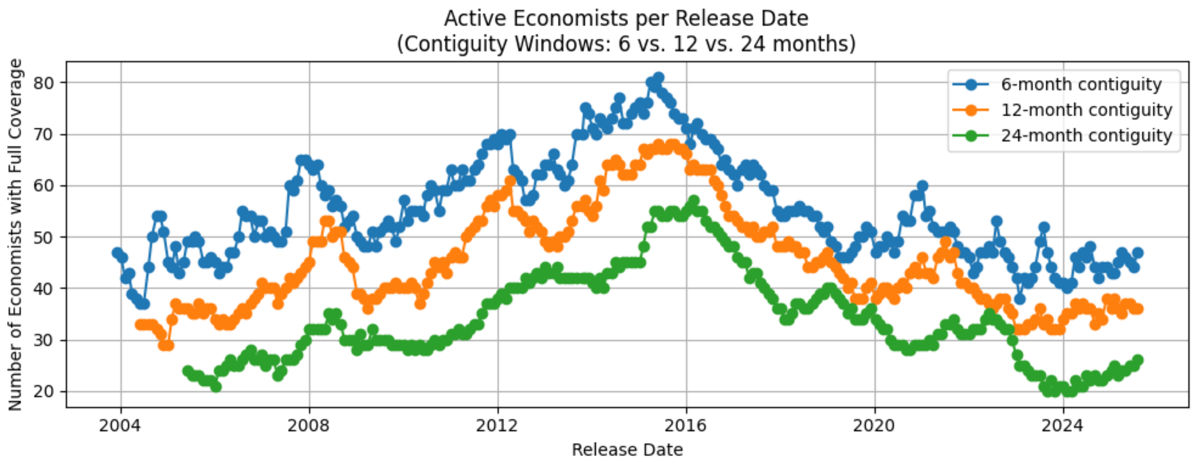

Figure 5: Active Economists per Release Date with Applied Contiguity Filters

The contiguity filter is capped at 12 months to preserve a sufficient number of forecasters in each cross-sectional sample across the full backtesting horizon.

Stratified Regimes. To gauge robustness of our point and directional forecasts, we evaluate point and directional accuracy within six macro regimes. Cut dates are anchored to known breaks and to visible shifts in loss variance and forecast disagreement (see Fig. 1): the **Global Financial Crisis (GFC)** and the **COVID shock** serve as explicit stress tests, while adjacent expansion phases probe stability in low-volatility backdrops. The regimes are:

- **Pre-GFC expansion:** 2003-12 to 2007-12
- **GFC:** 2008-01 to 2009-12
- **Early expansion:** 2010-01 to 2014-12
- **Late expansion:** 2015-01 to 2019-12
- **COVID shock:** 2020-01 to 2022-12
- **Post-COVID normalization:** 2023-01 to 2025-07-03

GFC and COVID windows are where squared-error volatility and cross-sectional disagreement spike, making naive weighting schemes brittle; expansions test whether gains persist when distributions are tight and drifting slowly. These strata are used only for reporting and diagnostics; all models are refit in a rolling, out-of-sample protocol with no look-ahead.

4.1.1 Static Inverse Error

Intuitively, the static inverse error forecast adopts the approach of weighing forecasters that have been accurate in the recent months more heavily while filtering out sporadic forecast noise with a contiguity filter. We present a more detailed and mathematical breakdown below.

Let t index data releases and let $\mathcal{E}_t(W) \subseteq \mathcal{E}$ be the set of economists who supplied non–missing forecasts in each of the W most recent releases.

Error history. For $i \in \mathcal{E}_t(W)$ define the point forecast errors

$$e_{i,t-k} = f_{i,t-k} - y_{t-k}, \qquad k = 1, \dots, W,$$

where $f_{i,\tau}$ is the submitted forecast and y_τ is the realised print.

Weight rules (re-estimated each t)

$$s_{i,t} = \begin{cases} 1 & \text{Equal weight} \\ \left(\text{MAE}_{i,t} + \lambda\right)^{-1} & \text{Inverse absolute error} \\ \left(\text{MSE}_{i,t} + \lambda\right)^{-1} & \text{Inverse squared error} \end{cases}$$

with $\text{MAE}_{i,t} = W^{-1} \sum_{k=1}^{W} |e_{i,t-k}|$, $\text{MSE}_{i,t} = W^{-1} \sum_{k=1}^{W} e_{i,t-k}^2$, and $\lambda = 10^{-6}$. Normalized weights are $w_{i,t} = s_{i,t} / \sum_{j \in \mathcal{E}_t(W)} s_{j,t}$. The lambda term is supplied for numerical stability.

Smart consensus forecast

$$\hat{y}_t^{\text{smart}} = \sum_{i \in \mathcal{E}_t(W)} w_{i,t} \, f_{i,t}.$$

The procedure is repeated across $W \in \{3, 6, 12\}$ and all weighting rules, yielding a specification grid evaluated walk-forward out-of-sample.

4.1.2 Exponentially Weighted Moving Average (EWMA)

EWMA adapts forecaster weights in real time by decaying the influence of older forecast errors. Relative to the static inverse–error scheme, the procedure introduces a temporal decay hyperparameter that emphasizes recent performance.

Temporal decay. Fix a rolling window length $W \in \{3, 6, 12\}$ (months) and a decay factor $\rho \in \{0.75, 0.80, \ldots, 0.95\}$. Define

$$\psi_k(\rho, W) = \frac{\rho^{W-k}}{\sum_{\ell=1}^{W} \rho^{W-\ell}}, \qquad k = 1, \ldots, W,$$

so that ψ_1 applies to the most-recent error and $\sum_k \psi_k = 1$. As $\rho \to 1$ the weights flatten, recovering the static window as a limiting case.

Error aggregation. For economist i with a complete W-month history let $e_{i,t-k} = f_{i,t-k} - y_{t-k}$ denote the k-step-old error. EWMA forms an exponentially weighted score

$$S_{i,t}^{(g)} = \sum_{k=1}^{W} \psi_k(\rho, W)\, g\big(e_{i,t-k}\big), \quad g(x) \in \{\, |x|,\ x^2 \,\},$$

corresponding to mean-absolute or mean-squared loss.

Weight rules. With a numerical ridge $\lambda = 10^{-6}$ the raw scores are converted to weights

$$s_{i,t} = \begin{cases} 1, & \text{Equal weight,} \\ \big(S_{i,t}^{(|\cdot|)} + \lambda\big)^{-1}, & \text{Inverse-MAE,} \\ \big(S_{i,t}^{(\cdot^2)} + \lambda\big)^{-1}, & \text{Inverse-MSE,} \end{cases} \qquad w_{i,t} = \frac{s_{i,t}}{\sum_{j \in \mathcal{E}_t(W)} s_{j,t}}.$$

Smart consensus forecast. The EWMA point prediction for release t is then

$$\hat{y}_t^{\text{EWMA}} = \sum_{i \in \mathcal{E}_t(W)} w_{i,t}\, f_{i,t},$$

where $\mathcal{E}_t(W)$ is the set of economists with non-missing submissions in all W look-back months.

Hyperparameter tuning. A walk-forward grid search traverses $\rho \in \{0.75, 0.80, \ldots, 0.95\}$ for each window $W \in \{3, 6, 12\}$ and weighting rule; each (W, ρ, rule) defines a distinct specification.

4.1.3 Soft Bayesian Model Averaging (soft-BMA)

soft-BMA builds a heavy-tailed likelihood for each economist's recent errors and converts the resulting log-evidence into a *soft-max* weight. While it stops short of full BMA, using rolling-window plug-in likelihoods rather than posterior model probabilities (hence *soft-BMA*), the procedure retains the Bayesian spirit of BMA while allowing weights to evolve smoothly with incoming data.

1. **Construct an error panel:** Fix a look-back horizon W months and degrees-of-freedom parameter ν. For each active economist i collect the centerd forecast errors $e_{i,t-k} = f_{i,t-k} - y_{t-k},\ k = 1, \ldots, W$.

2. **Fit Student-t error models:** Estimate the scale $\hat{\sigma}_{i,t} = \sqrt{\frac{1}{W-1}\sum_k e^2_{i,t-k}}$ and compute the cumulative log-likelihood

$$\ell_{i,t}(\nu) \;=\; \sum_{k=1}^{W} \log \mathrm{t}_\nu\big(e_{i,t-k};\, 0, \hat{\sigma}_{i,t}\big),$$

   where $\mathrm{t}_\nu(\cdot; 0, \sigma)$ denotes the Student-t density with ν degrees of freedom, zero mean, and scale σ.

3. **Convert evidence to weights (soft-max):** Define $w_{i,t} \;=\; \dfrac{\exp\big\{\ell_{i,t}(\nu)\big\}}{\sum_{j\in\mathcal{E}_t(W)} \exp\big\{\ell_{j,t}(\nu)\big\}}$, thereby favouring economists whose recent errors are more probable under their own Student-t fit.

4. **Form the crowd forecast:** Align the weight vector with current submissions and predict

$$\hat{y}_t^{\text{soft-BMA}} \;=\; \sum_{i\in\mathcal{E}_t(W)} w_{i,t}\, f_{i,t},$$

   alongside the directional flag $\mathbf{1}\{\hat{y}_t^{\text{soft-BMA}} > \text{median}_t\}$.

5. **Hyperparameter grid:** A walk-forward search spans $W \in \{3, 6, 12\}$ and $\nu \in \{3, 5, 10, 25\}$, generating an out-of-sample record per (W, ν) specification.

The soft-BMA procedure was selected as a result of initial data exploration observing the distributional pattern of median forecast errors (3.2). It generalizes static inverse-error weighting by translating likelihoods—not point errors—into weights, thereby accommodating heteroskedasticity and fat-tailed periods. Unlike MWU, which multiplicatively aggregates *all* past errors, soft-BMA restricts memory to a finite window but modulates the influence of extreme observations through the Student-t tail parameter ν.

4.1.4 Multiplicative Weights Update (MWU)

MWU is an online learning algorithm that treats each economist as an *expert* and adaptively reallocates probability mass toward forecasters that minimize squared error in real time. We adapt the classical scheme with a persistent global expert pool, probation and drop rules, sleep–tracking, and projection onto a capped simplex to reflect operational realities.

1. **Persistent global pool and probation:** The algorithm maintains a single global pool of experts across the full sample. An economist is eligible to enter the pool only after passing a 12-month *contiguity* screen (i.e. uninterrupted forecasts for the prior 12 releases). Entry is triggered the month after the probation window completes.

2. **Newcomer allocation and incumbents:** New entrants receive a fixed *newcomer share* $\alpha_{\text{new}} = 0.10$ split equally among them, with incumbent weights scaled down proportionally to preserve the unit sum. All weights are projected onto the capped simplex $\{w_i \in [\omega_{\min}, \omega_{\max}],\ \sum_i w_i = 1\}$ with $\omega_{\max} = 0.50$ and $\omega_{\min} = 10^{-3}$.

3. **Forecast construction:** For release t, let $\mathcal{A}_t$ denote experts with a live submission and no more than $S_{\max} = 2$ consecutive misses. If $|\mathcal{A}_t| \geq 10$ (minimum active experts), we re-project the active sub-portfolio to the capped simplex and form the smart consensus

$$\hat{y}_t^{\text{MWU}} = \sum_{i \in \mathcal{A}_t} w_{i,t} f_{i,t}.$$

   A directional signal $\mathbf{1}\{\hat{y}_t^{\text{MWU}} > \text{median}_t\}$ is logged for evaluation. Active-month weights (post-projection) are snapshotted for later diagnostics.

4. **Loss evaluation and multiplicative update:** Upon realisation of y_t, penalised squared-error losses $\ell_{i,t} = (f_{i,t} - y_t)^2 + \lambda$ with $\lambda = 10^{-6}$ are computed for active

experts, and weights updated via

$$w_{i,t+1} \;\propto\; w_{i,t} \exp(-\eta\, \ell_{i,t}), \qquad \eta \in [0.001, 0.020].$$

The global pool is then re-projected to the capped simplex to enforce bounds and unit sum.

5. **Sleep and expulsion logic:** Absences increment a *sleep counter* $c_{i,t}$, reset on submission. Experts are permanently dropped if they exceed $S_{\max} = 2$ consecutive misses or accumulate more than $M_{\max} = 6$ misses in any rolling 12-month window.

6. **Hyperparameter grid and evaluation:** A walk-forward grid over η is run separately on COVID-only and full-history panels. Out-of-sample diagnostics include RMSE, hit rate, Binomial and Pesaran–Timmermann p–values, and Diebold–Mariano tests, with regime-wise breakdowns to gauge robustness.

MWU differs from static inverse-error and EWMA schemes by compounding past losses multiplicatively, yielding a *long-memory* weight vector that adapts smoothly over time. Relative to soft-BMA, MWU operates in expert space rather than error-likelihood space, providing a complementary blend of adaptivity and interpretability, while the projection, cap, and drop mechanisms prevent domination by any single forecaster while still allowing the model the appropriately bias predictions toward recent winners among the expert pool.

4.1.5 Robust Ensemble for Directional Forecasting

Motivation. Ensemble methods are a standard remedy for model fragility. Take for example the case study of decision tree learning algorithms in machine learning. A single decision tree can be highly sensitive to minor perturbations in the training data; aggregating many trees (bagging, random forests) reduces correlation across errors and yields a predictor that is both lower variance and more stable to small changes. We adopt the same logic for macro forecast combination: instead of committing to one "best" specification, we average a small, diverse set of top performers so that idiosyncratic misspecification risk washes out while common signal is amplified.

Candidate pool (robust winner / HR / RMSE). For each base model we form a compact pool of specifications using out-of-sample (OOS) diagnostics on the full panel: (i) the *RMSE winner* (lowest OOS RMSE vs. the crowd median), (ii) the *Hit-Rate winner* (highest directional hit rate relative to the median), and (iii) a *robust winner* defined as the lowest-RMSE spec among those that simultaneously pass directional robustness screens (Diebold–Mariano $p < 0.10$ and Pesaran–Timmermann $p < 0.10$). Pool members must have live OOS histories. From this pool we build small directional forecast ensembles.

Ensemble construction. At release t, let $\hat{y}_t^{(j)}$ denote the OOS smart forecast from pool member $j \in \mathcal{J}$, with consensus median f_t^{med}. For a given ensemble $\mathcal{S} \subset \mathcal{J}$ of size k (we consider $k \in \{3, 5\}$), the ensemble directional forecast is produced by a strict majority vote relative to the median:

$$\hat{d}_t^{\text{ens}} = \mathbf{1}\left\{\sum_{j \in \mathcal{S}} \mathbf{1}\{\hat{y}_t^{(j)} > f_t^{\text{med}}\} > \tfrac{k}{2}\right\},$$

so the ensemble calls a *beat* when strictly more than half of the members sit above the median; it calls a miss otherwise.

Four horizon-anchored signals. To balance stability and adaptivity, we compute four *horizon-anchored* ensemble signals by re-evaluating candidate combinations on progressively shorter realized windows that end at the last observed print:

$$\mathcal{W} \in \{\text{Full history, last 12 months, last 6 months, last 3 months}\}.$$

For each window $\mathcal{W}$, we:

1. restrict each member's OOS series to $\mathcal{W}$;
2. enumerate all k-member combinations from the pool;
3. select the combination $\mathcal{S}^{\star}_{\mathcal{W}}$ that maximizes directional hit rate over $\mathcal{W}$ (primary criterion);
4. report the ensemble's RMSE vs. the median, the exact binomial p–value for hit rate (null = 50%), the Pesaran–Timmermann p–value to adjust for base-rate effects, and

an AC score

$$\mathrm{AC} \;=\; \big(1-\widehat{\mathrm{HR}}\big) \;+\; \lambda\,\sigma_{\mathrm{blocks}}(\widehat{\mathrm{HR}}),$$

where σ_{blocks} is the standard deviation of block-level hit rates across regimes. Lower AC indicates better accuracy and stability.

This yields four parallel signals—$\mathrm{Ens}_{\mathrm{Full}}, \mathrm{Ens}_{12\mathrm{m}}, \mathrm{Ens}_{6\mathrm{m}}, \mathrm{Ens}_{3\mathrm{m}}$—each optimized for its lookback.

Comparing signal horizons. The *Full-history* signal is the most stable and is preferred when the process is stationary and structural breaks are unlikely. The *12-month* signal offers a balanced bias–variance trade-off and typically tracks evolving seasonals and slow-moving shifts without overreacting. The *6-month* signal is more reactive and can surface regime changes sooner at the cost of higher variance. The *3-month* signal is the most adaptive but also the noisiest; it is informative as an early-warning overlay, not as a sole driver. In practice, we publish all four; downstream users can privilege stability or dynamic reactivity as the trading context requires. We recommend defaulting to the 6-month signal for most purposes as it delivers the highest historical hit rate.

4.2 Distributional Engines

Accurate point estimates alone are often insufficient for trading and risk management; a full predictive density can often be useful to gauge tail risk, size positions, and price optionality around macro prints. Accordingly, we deploy four complementary distributional engines: (i) a **Student-*t*** error model, (ii) a **Gaussian Mixture Model** (GMM) to capture latent regimes, (iii) a ***t*-GARCH(1,1)** filter for time-varying conditional volatility, and (iv) **Bayesian Model Averaging** (BMA) over Normal and Student-t specifications. Each engine ingests the historical error stream of the crowd median and outputs calibrated two-sided prediction bands for the upcoming release. We evaluate candidates on coverage *accuracy*—the mean absolute gap between empirical and nominal hit rates (i.e. for a 95% prediction band, how close historical empirical coverage is to the theoretical 95% nominal value)—and *consistency*, the variation of that gap across distinct market regimes.

Crisis Multiplier (shared across engines). Let s_t be the cross-sectional forecast spread and $\tilde{s}$ the median of its trailing 24-month history. We define

$$m_t = \left(\frac{s_t}{\tilde{s}}\right)^{\beta_t}, \qquad \beta_t = \begin{cases} 0, & s_t \le \mathrm{Pct}_{95}\{s_{t-k}\}_{k=1}^{24}, \\ 0.80, & s_t > \mathrm{Pct}_{95}\{s_{t-k}\}_{k=1}^{24}. \end{cases}$$

All interval engines below apply m_t multiplicatively to their baseline half-widths (see §3.3 for motivation).

Regime construction for distributional evaluation. Unlike the point–forecast analysis, which stratifies performance into shorter macro–regimes to examine directional and level accuracy, the distributional evaluation uses a coarser set of four long–horizon strata. This is intentional: coverage analysis requires many realized observations to obtain stable estimates of empirical inclusion rates at each nominal level. For point/directional evaluation we estimate a low-dimensional target (mean and sign), which stabilizes with relatively few observations, whereas coverage assessment interrogates the full predictive CDF—especially tail quantiles—whose binomial standard errors require materially larger T. Partitioning too finely would produce high–variance coverage estimates and potentially misleading inferences. The four strata are defined to be approximately equal in length to preserve enough data within each block to assess calibration while still allowing us to observe structural shifts.

4.2.1 Student-*t* Predictive Density

We model forecast errors $e_t = f_t^{\mathrm{med}} - y_t$ with a Student-$t(\nu, \mu, \sigma)$ distribution, estimated each month on the most recent 24 monthly errors. Let $\hat{\nu}, \hat{\mu}, \hat{\sigma}$ denote the maximum-likelihood parameters for that window. In the implementation, $\hat{\mu}$ is ignored so that every interval is centered directly on the median point forecast. For a nominal coverage level $L \in \{50, 60, 70, 80, 90, 95\}\%$, the half-width is

$$h_{L,t} \;=\; t_{1-\alpha/2,\hat{\nu}}\,\hat{\sigma} \times \underbrace{\left(\tfrac{s_t}{\text{median } s}\right)^{\beta}}_{\text{crisis multiplier}}, \qquad \alpha = 1 - \tfrac{L}{100},$$

where s_t is the cross-sectional forecast spread at t. The exponent β is set to a base value ($\beta = 0$) unless the current spread is above the 95^{th} percentile of its own 24-month history, in which case we apply the crisis multiplier m_t ($\beta = 0.80$) from 4.2.

Two specifications are considered: (i) no crisis adjustment, and (ii) with crisis adjustment. A walk-forward evaluation on the full-sample panel selects the variant that minimizes the mean-absolute coverage gap, which is then applied to the live-month forecast.

4.2.2 Gaussian Mixture Model (GMM) Predictive Density

We model forecast errors $e_t = f_t^{\text{med}} - y_t$ using a Gaussian mixture model (GMM) estimated each month on the most recent 24 monthly errors. For $k \in \{1, 2, 3, 4\}$ mixture components, we fit a full-covariance GMM by maximum likelihood and select the number of components that minimizes the Bayesian Information Criterion (BIC). This allows the error distribution to flexibly capture skewness, kurtosis, and potential multi-modality in the historical error process.

Given the fitted GMM, we simulate $N = 100{,}000$ draws $\{\tilde{e}^{(n)}\}_{n=1}^{N}$ from the mixture. For a nominal coverage level $L \in \{50, 60, 70, 80, 90, 95\}\%$, the lower and upper error quantiles are taken from the empirical simulation distribution:

$$q_L^{\text{lo}},\, q_L^{\text{hi}} \;=\; \text{Quantile}\left(\tilde{e}^{(n)},\, \frac{1-L}{2}\right),\, \text{Quantile}\left(\tilde{e}^{(n)},\, 1 - \frac{1-L}{2}\right).$$

The corresponding prediction interval is then

$$\left[\, f_t^{\text{med}} - m_t \cdot q_L^{\text{hi}},\; f_t^{\text{med}} - m_t \cdot q_L^{\text{lo}} \,\right],$$

where m_t is the crisis multiplier defined in 4.2.

Two specifications are considered: (i) no crisis adjustment ($\beta = 0$ always) and (ii) with crisis adjustment as above. A walk-forward evaluation on the full-sample panel selects the variant that minimises the mean-absolute coverage gap, which is then applied to generate the live-month forecast intervals.

4.2.3 *t*-GARCH Predictive Density

We model forecast errors $e_t = f_t^{\text{med}} - y_t$ with a constant-mean GARCH(1,1) process and Student-t innovations, re-estimated each month on the most recent 24 monthly errors. Let $\hat{\nu}$ denote the degrees-of-freedom of the Student-t innovations and $\hat{\sigma}$ the one-step-ahead conditional volatility from the filter. Forecast errors are rescaled by a constant factor κ (e.g., $\kappa = 100$ to map tens of thousands to hundreds of jobs) to improve numerical stability; outputs are transformed back to original units.

For a nominal coverage level $L \in \{50, 60, 70, 80, 90, 95\}\%$, the half-width is

$$h_{L,t} \; = \; t_{1-\alpha/2,\hat{\nu}} \; \hat{\sigma} \times \underbrace{\left(\tfrac{s_t}{\text{median}\, s}\right)^{\beta}}_{\text{crisis multiplier}}, \qquad \alpha = 1 - \tfrac{L}{100},$$

where s_t is the cross-sectional forecast spread at t. The exponent β is set to a base value ($\beta = 0$) unless the current spread is above the 95^{th} percentile of its own 24-month history, in which case we apply the crisis multiplier from §4.2.

Centering. Prediction intervals are *centered at the consensus median* f_t^{med} (no μ-shift). This keeps the interval engine focused on *dispersion* rather than level, and avoids double-counting any bias relative to the median.

We evaluate two specifications: (*i*) no crisis adjustment and (*ii*) with crisis adjustment.

4.2.4 Bayesian Model Averaging (BMA) Predictive Density

We model forecast errors $e_t = f_t^{\text{med}} - y_t$ each month using Bayesian model averaging (BMA) over two candidate error distributions, estimated on the most recent 24 monthly errors. The candidates are: (*i*) a Normal distribution $N(\mu, \sigma^2)$, fitted by closed-form maximum likelihood, and (*ii*) a Student-t distribution $\mathrm{t}_\nu(\mu, \sigma)$, fitted by numerical maximum likelihood with $\nu > 2$ constrained. The Student-t candidate allows for fat-tailed errors in periods of elevated volatility.

Let BIC_k denote the Bayesian Information Criterion for model k, and $\text{BIC}_{\min} = \min_k \text{BIC}_k$. We define *Occam weights*

$$w_k \; = \; \frac{\exp\left[-\frac{1}{2}\left(\text{BIC}_k - \text{BIC}_{\min}\right)\right]}{\sum_\ell \exp\left[-\frac{1}{2}\left(\text{BIC}_\ell - \text{BIC}_{\min}\right)\right]},$$

so that better-fitting models receive larger posterior weights.

We then simulate $N = 100{,}000$ synthetic errors by: 1. randomly selecting a model according to $\{w_k\}$, and 2. drawing an error from the selected distribution with its fitted parameters.

For each nominal coverage level $L \in \{50, 60, 70, 80, 90, 95\}\%$, the lower and upper quantiles of the simulated error distribution define the empirical error bounds $(q_L^{\mathrm{lo}}, q_L^{\mathrm{hi}})$. Prediction intervals are formed by centring these bounds on the median forecast f_t^{med} and optionally applying a *crisis multiplier* previously defined in 4.2.

Two specifications are considered: (i) no crisis adjustment ($\beta = 0$ always) and (ii) with crisis adjustment as above. A walk-forward evaluation on the full-sample panel selects the variant that minimizes the mean-absolute coverage gap, which is then applied to generate the live-month forecast intervals.

4.3 Evaluation Protocol

Formal test statements are collected in Appendix A.1; for completeness, we summarize the operational definitions here so results can be read at a glance.

Back-test protocol: At each release t, models are refit using information available through $t-1$ and then produce: a point forecast $\hat{y}_t$, a directional call relative to the consensus median f_t^{med}, and predictive intervals at levels $L \in \{50, 60, 70, 80, 90, 95\}\%$. Relative tests take the consensus median as the benchmark. The result of this rolling evaluation protocol are out-of-sample results with no lookahead bias.

(1) Point and directional evaluation

- **Point loss and RMSE:** Define squared loss $\ell_t = (\hat{y}_t - y_t)^2$. Report RMSE $= \sqrt{\frac{1}{T}\sum_t \ell_t}$ for the *smart* model and for the *median* benchmark.
- **Relative accuracy (DM):** Compare each model to the median using the Diebold–Mariano test on $d_t = \ell_t^{\mathrm{model}} - \ell_t^{\mathrm{median}}$, with Newey–West long-run variance. We report the DM statistic and two-sided p–value; negative $E[d_t]$ favors the model.
- **Directional skill:** We forecast the *direction of the surprise relative to the consensus*

median. Let the directional target be $d_t^{\text{act}} = \text{sign}(y_t - f_t^{\text{med}})$ and the model's directional forecast be $d_t^{\text{mod}} = \text{sign}(\hat{y}_t - f_t^{\text{med}})$. A *hit* occurs when $d_t^{\text{mod}} = d_t^{\text{act}} \neq 0$ (ties at the median are not scored and virtually never happen). We report the hit rate $\widehat{\text{HR}} = T^{-1} \sum_t \mathbf{1}\{d_t^{\text{mod}} = d_t^{\text{act}} \neq 0\}$, an exact binomial p-value under a 50% null, and the Pesaran–Timmermann statistic (with p-value) to account for base-rate effects.

- **Accuracy x consistency score across regimes:** To assess forecast performance across regimes, we summarize accuracy and regime stability with

$$\text{AC}_{\text{point}} = (1 - \widehat{\text{HR}}) + \lambda\, \sigma_{\text{block}},$$

where σ_{block} is the standard deviation of block-level (e.g., pre-COVID / COVID / post–COVID) hit rates and λ controls the accuracy–stability trade-off (default $\lambda = 1.0$, equally weight accuracy and consistency). Lower is better. We also display RMSE and DM alongside AC to keep level performance visible.

(2) Distributional (interval) evaluation

- **Coverage by level.** For a nominal level L, let the ex-ante interval be $I_{L,t} = [\ell_{L,t}, u_{L,t}]$. Empirical coverage is the proportion of test releases whose realization lands inside the band:

$$\widehat{C}_L \;=\; \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{\ell_{L,t} \le y_t \le u_{L,t}\},$$

with interval endpoints treated as inclusive and any missing y_t excluded from T.

- **Calibration summary:** Report the vector $\{\widehat{C}_L\}$ and the mean absolute coverage gap

$$\text{MAG} = \frac{1}{|\mathcal{L}|} \sum_{L \in \mathcal{L}} \bigl|\widehat{C}_L - L\bigr|,$$

where $\mathcal{L} = \{50, 60, 70, 80, 90, 95\}\%$. Smaller MAG indicates better average alignment to targets.

- **Selection score (densities).** We balance calibration and regime stability with

$$\text{AC}_{\text{dist}} = \text{MAG} + \lambda\, \sigma_{\text{block}},$$

where σ_{block} is the standard deviation of block-level coverage gaps across L. Lower is better. Where applicable, we report results with and without the spread-based crisis multiplier to verify that gating improves coverage in high-disagreement months without degrading tranquil periods.

How to read our backtest tables (quick guide): DM< 0 with small p favours the model over the median; higher $\widehat{\text{HR}}$ with small binomial/PT p indicates genuine directional skill; coverage rows close to nominal and low MAG indicate well-calibrated intervals; in all cases, a lower AC score marks the *accuracy x consistency* champion once stability is priced in.

5 Results

5.1 Point–Forecast Performance

We present primary findings from our point/directional forecast models in this section. Full backtest results for individual methods are available at §A.2.

5.1.1 Inverse–Error

Inverse–error schemes reweight economists according to recent performance within a strict contiguity screen. At each release t and window $W \in \{3, 6, 12\}$, weights are proportional to inverse loss (MAE or MSE) over the last W realized months; we also test an equal–weight baseline on the same contiguous panel. This construction is the simplest form of adaptive ensembling: it is transparent, fast to update, and provides a clean stress test of whether the crowd can be improved with light–touch learning.

COVID–filtered vs. full history. On the COVID–filtered panel, inverse–MSE with a 12-month window attains the lowest RMSE and passes the Diebold–Mariano test at conventional levels, while 6-month equal–weight delivers the highest hit rate with strong binomial and PT support. When the full history is reinstated, equal–weight with a 6-month window jointly attains the lowest RMSE and the highest hit rate; however, DM evidence weakens as the COVID spikes inflate loss variance, eroding level advantages even when directional calls remain above 55%.

Production choice and rationale. For reporting on the full sample we emphasize the *equal–weight, 6-month* specification. It balances adaptivity and stability: (i) short enough to track slow drifts in forecaster quality; (ii) long enough to avoid the noise we observe in 3-month variants; and (iii) robust across pre-GFC, GFC, late-cycle, COVID, and post–COVID subperiods. We retain the 12-month inverse–MSE model as a COVID-filtered benchmark where RMSE gains are clearest.

Regime diagnostics (selected spec: Full panel, `equal_weight`, $W = 6$) Table 4 summarizes performance by macro regime. Several features are notable. First, directional accuracy is consistently above 55% in tranquil regimes and rises to nearly 80% in the GFC, indicating that the ensemble tends to be on the right side of large moves. Second, RMSE differences are small in pre-COVID expansions but remain in favor of the smart consensus during COVID despite both series exploding in scale. Third, short trailing windows illustrate the classic bias–variance trade-off: very recent slices can look excellent (or poor) by chance; we therefore prefer the full-regime profile when judging robustness.

Table 4: Stratified diagnostics for the selected inverse–error specification on the full panel (`equal_weight`, $W = 6$). Metrics by regime: RMSE of smart and median forecasts, directional hit rate (vs. the median), and Diebold–Mariano p–value (smart vs. median).

Regime	**RMSE_smart**	**RMSE_median**	**HitRate**	**DM_p**
2003-12 to 2007-12 (pre-GFC)	80.074	80.345	0.551	0.670
2008-01 to 2009-12 (GFC)	73.908	76.637	0.792	0.100
2010-01 to 2014-12 (early expansion)	60.267	61.597	0.583	0.059
2015-01 to 2019-12 (late expansion)	62.551	62.883	0.550	0.395
2020-01 to 2022-12 (COVID)	1655.331	1722.868	0.611	0.284
2023-01 to 2025-07-03 (post–COVID)	91.877	93.334	0.516	0.336

Takeaways

- *Keep it simple under regime uncertainty.* Equal weighting across contiguous forecasters is hard to beat in the full history, where large shocks destabilize inverse–error weights.
- *Directional alpha is durable.* Even when RMSE improvements blur in crisis tails, the ensemble's sign relative to the median remains informative and statistically supported.

- *Window length is a tuning knob, not a free lunch.* Very short windows add variance; very long windows dull responsiveness. A 6–12 month horizon provides the best bias–variance trade-off here.

5.1.2 EWMA

EWMA adds controlled recency to the inverse–error idea: within a W-month window, older errors are exponentially down-weighted by a decay factor ρ, and the resulting scores feed the same inverse-MAE/MSE (or equal-weight) rules. As $\rho \to 1$, the scheme approaches the static window.

COVID–filtered panel. The sharpest level gains appear with a *longer* window and *slower* decay: the lowest RMSE is delivered by `ewma_w12_d0.95_minverse_mse` (69.81 vs. 70.91 for the median). Directional skill is strongest for `ewma_w6_d0.75_mequal_weight`, with a hit rate around 0.58 and statistically significant binomial and PT p–values. Taken together, these results suggest that (i) modest recency helps, but (ii) aggressive reweighting by recent squared error is not strictly necessary to achieve stable directional improvements—simple equal–weighting within a 6-month window is competitive and robust.

Full panel (including COVID). Heavy–tail months erode the level advantage of error–weighted variants. The best overall specification by both RMSE and hit rate is `ewma_w6_d0.75_mequal_weight` (619.6 vs. 644.7 for the median; HR $\approx$ 0.585). Despite the sizeable RMSE reduction, Diebold–Mariano p–values are generally not significant on the full sample because crisis months inflate variance. The robust winner for this panel (`ewma_w3_d0.75_mequal_weight`) reinforces the theme that light recency and small windows can be preferable when the error process is punctuated by rare, extreme shocks.

What the regime breakdown shows. Table 5 reports regime diagnostics for the full–panel winner by accuracy (`ewma_w6_d0.75_mequal_weight`). Relative to the crowd median, EWMA is neutral in tranquil periods (pre-GFC, late expansion), improves in stress (GFC, COVID) and early–recovery phases, and is roughly even post–COVID. This pattern is consistent with EWMA's design: it adapts enough to benefit when distributions broaden, yet remains simple enough (equal–weight, short window) to avoid chasing transitory idiosyncrasies.

Key takeaways.

- **Directional skill persists across panels.** Hit rates cluster in the mid-50s and rise in stress regimes; PT tests confirm dependence beyond chance on the COVID panel.
- **Level gains are regime-dependent.** On the COVID-filtered panel, w = 12, $\rho = 0.95$ inverse–MSE offers the best RMSE; on the full panel, simple equal–weighting with w = 6 dominates.
- **Decay is a second-order choice.** Within a given window, changing ρ from 0.75 to 0.95 nudges performance rather than overturning it. Window length (3–12 months) and whether we use error–weights vs. equal–weights matter more.
- **Operational guidance.** For live use, we favor w = 6 equal–weight for directional signaling (stable, low-variance), and w = 12, $\rho = 0.95$ inverse–MSE as an optional level overlay in non-crisis regimes (COVID-filtered evidence).

Table 5: Regime diagnostics for EWMA (full panel winner by accuracy: `ewma_w6_d0.75_mequal_weight`).

Regime	**RMSE_smart**	**RMSE_median**	**HitRate**	**DM_p**
2003–12 to 2007–12 (pre-GFC)	80.074	80.345	0.551	0.670
2008–01 to 2009–12 (GFC)	73.908	76.637	0.792	0.100
2010–01 to 2014–12 (early expansion)	60.267	61.597	0.583	0.059
2015–01 to 2019–12 (late expansion)	62.551	62.883	0.550	0.395
2020–01 to 2022–12 (COVID)	1655.331	1722.868	0.611	0.284
2023–01 to 2025–07–03 (post–COVID)	91.877	93.334	0.516	0.336

5.1.3 soft-BMA

soft-BMA converts fit into weights by scoring each economist's last W errors under a Student-t_ν model and mapping log-likelihoods through a soft-max. Heavier tails (ν small) damp outliers, and the weighted average of live submissions yields the point forecast.

Key patterns

- **COVID-filtered panel.** Twelve-month windows coupled with heavy tails dominate. The lowest RMSE arises at W=12, ν=3 (smart 67.7 vs median 70.9), and the highest hit rates cluster at W=12, $\nu \in \{10, 25\}$ (HR $\approx$ 0.58) with small binomial and PT p–values ($p = .018$–$.036$), and DM $p = .060$–$.067$. Three- and six-month windows are competitive on direction but less compelling on level.

- **Full panel (with COVID).** Level RMSE deteriorates materially for all specs (reflecting the extreme COVID miss), producing no "robust winner" under our 10% DM/PT gate. Directional accuracy remains resilient: HRs in the 0.56–0.58 range frequently attain exact-binomial $p < 0.05$ even when RMSE is worse than the median. In short, soft-BMA carries a stable directional edge, but its level advantage is eroded by crisis-scale errors that heavy tails alone do not neutralize.
- **Role of heavy tails and window length.** Moving from ν=25 toward ν=3 systematically helps in COVID-filtered tests (more protection against occasional large errors). Window length matters more than ν: W=12 emerges as the most reliable horizon for both RMSE and HR.

Selected specification and regime breakdown. For interpretability across regimes we display the W=12, ν=3 model (it is the COVID-panel RMSE leader and the full-panel HR leader). The table shows that soft-BMA is comparable or better than the median in tranquil expansions, suffers a level penalty during COVID, and recovers strongly post–COVID with borderline-significant DM in levels.

Table 6: Regime breakdown for `soft_bma_w12_nu3` (full panel).

Regime	**RMSE_smart**	**RMSE_median**	**HitRate**	**DM_p**
2004-06 to 2007-12 (pre-GFC)	71.137	71.584	0.581	0.759
2008-01 to 2009-12 (GFC)	77.873	76.637	0.542	0.665
2010-01 to 2014-12 (early-expansion)	60.551	61.597	0.567	0.328
2015-01 to 2019-12 (late-expansion)	62.452	62.883	0.567	0.777
2020-01 to 2022-12 (COVID)	2219.418	1722.868	0.556	0.283
2023-01 to 2025-07-03 (post–COVID)	76.355	93.334	0.645	0.066

Takeaways

- *Directionally valuable.* soft-BMA retains a consistent directional edge, particularly with W=12 and small ν, even when level RMSE parity versus the median cannot be guaranteed in crisis periods.

- *Level sensitivity to crises.* Heavy tails cushion but do not neutralize COVID-scale errors; this explains the absence of a full-panel robust winner (DM/PT $p \geq 0.10$).
- *Practical placement.* We treat soft-BMA as a strong directional component in the robust ensemble and rely on distributional engines (with crisis multipliers) for calibrated uncertainty in levels.

5.1.4 Multiplicative Weights Update

MWU treats each economist as an "expert" and updates weights multiplicatively with recent loss: $w_{i,t+1} \propto w_{i,t} \exp(-\eta\, \ell_{i,t})$, where $\ell_{i,t}$ is squared-error and η is a learning rate. Each month we form the smart forecast as the weighted average of live submissions, with weights projected onto a capped simplex to avoid dominance.

COVID-filtered panel. Across step sizes $\eta \in [0.001, 0.019]$, MWU consistently reduces level error relative to the crowd median (all entries have $\text{RMSE}_{\text{smart}} < \text{RMSE}_{\text{median}}$). The best RMSE occurs around $\eta = 0.005$, while the highest directional hit rate is at $\eta = 0.015$. However, directional skill remains modest (HR $\approx$ 0.48–0.54) and neither the exact binomial nor PT tests deliver strong significance; no configuration passes our robustness gate (both DM and PT < 0.10).

Full panel (with COVID months). When the crisis months are included, MWU's level performance deteriorates: for all η, $\text{RMSE}_{\text{smart}} > \text{RMSE}_{\text{median}}$. Directional accuracy hovers near coin-flip (HR = 0.47–0.54). DM p–values frequently indicate *worse* squared-error than the median at moderate/large η (e.g., $\eta = 0.007$–0.011), consistent with the algorithm overweighting experts that themselves became unstable during the COVID shock. Smaller η attenuates this variance but does not overturn the median.

Regime diagnostics. Table 7 reports performance for the configuration with the highest full-sample hit rate ($\eta = 0.015$). MWU improves upon the median pre-GFC (lower RMSE; HR = 0.63) but underperforms in the long expansions and especially in the post–COVID period, where the DM test flags statistically worse loss relative to the median. The pattern suggests that MWU's compounding memory, even with our caps and sleep/expulsion rules, does not reweight quickly enough after large distributional breaks.

Table 7: Stratified performance for MWU (`mwu_eta0.015`) on the full panel.

Regime	RMSE_smart	RMSE_median	HitRate	DM_p
2004–2007 (pre-GFC)	68.747	71.584	0.628	0.422
2008–2009 (GFC)	84.079	76.637	0.542	0.493
2010–2014 (early expansion)	65.827	61.597	0.533	0.229
2015–2019 (late expansion)	67.394	62.883	0.533	0.263
2023–2025 (post–COVID)	100.989	93.334	0.419	**0.004**

Takeaways. (i) MWU exhibits *directional persistence* in tranquil periods but its level errors are fragile to crisis-era volatility, with post–COVID degradation that is statistically detectable. (ii) Tuning η trades off variance and adaptivity but does not produce a robust full-sample winner under our DM/PT gate. (iii) Operationally, we therefore retain MWU as a *diversifying voter* within the robust ensemble—useful for directional tie-breaks and as a hedge against misspecified inverse-error/EWMA weights—rather than as a standalone champion. Full backtests and winners tables appear in the Appendix.

5.1.5 Cross Point–Forecast Signal Comparison

Across the four families—*Inverse–Error*, *EWMA*, *soft-BMA*, and *MWU*—two regularities anchor the evidence. First, **directional skill persists**: hit rates reliably sit in the mid-50s and strengthen in stress regimes (e.g., GFC), indicating that all families tend to get the sign right relative to the consensus median and have demonstrated edge. Second, **level RMSE gains are regime-dependent**: once COVID months are included, Diebold–Mariano evidence weakens and level advantages become fragile because variance explodes.

Family-wise patterns

- **Inverse–Error.** Simple specs dominate in unstable regimes. Equal-weight with a 6-month window delivers the most consistent full-sample profile (stable hit rates, no overreaction to regime breaks). Error-weighted variants (inverse-MSE/MAE) look best on the COVID-filtered panel—especially at 12 months—where tails are muted and past accuracy is more informative.

- **EWMA.** Light recency helps without being essential. On the full sample, equal-weight with a short window (6 months) again leads on both RMSE and hit rate; within the COVID-filtered panel, a longer window (12 months) with slow decay and inverse-MSE attains the cleanest level gains. Changing the decay factor moves the needle modestly relative to window length and the choice between equal vs. error weights.
- **soft-BMA.** Heavy-tailed likelihoods convert recent fit into *soft* weights, yielding a durable directional edge (often statistically supported) with a 12-month window and small ν. However, crisis-scale errors erode level RMSE on the full sample; heavy tails cushion but do not neutralize COVID outliers.
- **MWU.** The long-memory multiplicative update adapts quickly in principle, but in our macro panel it is *most sensitive* to distributional breaks. It can post good pre-GFC direction but loses level footing in long expansions and post–COVID, where compounding can overweight stale “winners.”

Regime view. In the GFC, equal-weight variants of Inverse–Error/EWMA show the largest directional lift (HR approaching 0.8 for some slices), consistent with diversified voting when individual experts wobble. In tranquil expansions, all families hover near modest positive direction with small RMSE differences versus the median. During COVID, every method’s level error inflates; equal-weight short-window designs are least fragile, soft-BMA retains sign information, and MWU degrades the most. Post–COVID, soft-BMA and the simple averages recover directionally, while level metrics converge toward the median with only small separations.

Implications for production There is *no single always-best* champion. Instead, a *cluster of simple, short-horizon averages* (6-month equal-weight across Inverse–Error/EWMA) delivers the best accuracy–stability trade-off on the full history; soft-BMA contributes a complementary directional signal; MWU acts as a diversifying voter rather than a standalone leader. These findings motivate the *robust ensemble* in the next section: rather than commit to any single learner, we aggregate a small set of top, regime-complementary specifications and score them by Accuracy×Consistency, producing a single, resilient signal that inherits the strengths and hedges the weaknesses of its constituents.

Economist Weights as an Analysis Tool. Beyond aggregate performance metrics, an additional diagnostic tool is to inspect the individual economists that each model is most heavily weighting at the latest forecast snapshot. For each model family, we aggregate weights across all live specifications and panels, normalize them to sum to one, and rank economists by their resulting model-specific weights. We then form an equal-model blend (25% weight per family) to identify the overall top contributors to the current month's point-forecast signal. This ranking is informative in two complementary scenarios: first, when the ensemble is performing well, a "hot" economist appearing near the top across multiple families can offer qualitative insight into the directional bias or level call driving the model; second, when model performance deteriorates, the list can be checked for overweighted economists with a history of underperformance, prompting targeted review or temporary down-weighting. The full ranked tables for the August 2025 NFP print are reported in Appendix A.3, providing transparency into the composition of the live forecast signal.

5.2 Robust Ensemble Performance

5.2.1 In–Sample Search Results

Our in–sample search combines candidate specifications drawn from the point–forecast families in §4.1 by selecting, for each model and panel, the lowest–RMSE specification, the highest–hit–rate specification, and any "robust" winner (both Diebold–Mariano and Pesaran–Timmermann $p < 0.10$). This naturally produces a small, high–quality pool. We then exhaustively evaluate all $k \in \{3, 5\}$–member combinations under different evaluation windows: the full history, trailing 12, 6, and 3 months. Combinations are scored by hit rate (direction vs. median), with the Accuracy × Consistency (AC) score serving as a tie–breaker.

Table 8 reports stratified diagnostics for the best k in each window, excluding trailing windows from the breakdown. The **FULL**–window winner (k=5) achieves a 58.7% hit rate over 218 months with strong exact–binomial and PT significance ($p < 0.02$), and an AC–score of 0.48, reflecting both accuracy and stability. Performance is strongest in the GFC (hit rate ≈ 0.71) and early expansions, with robust gains over the median in COVID and post–COVID periods. The **T12M** and **T6M** winners produce exactly 50% hit rates in their respective short windows, unsurprising given the small sample sizes, and their stratified profiles indicate mixed performance across regimes. The **T3M** winner reaches 66.7% in its narrow evaluation band, but with only three observations, offering little statistical reliability.

Table 8: Stratified performance for in–sample robust ensembles (best k per window). Trailing windows omitted.

Window	**Spec ID**	**Regime**	**HitRate**	**Binom_p**	**PT_p**	**AC λ=1.0**
FULL	F1	pre-GFC	0.558	0.542	0.432	0.477
		GFC	0.708	0.064	0.054	
		early-expansion	0.600	0.155	0.121	
		late-expansion	0.550	0.519	0.405	
		COVID	0.581	0.473	0.283	
		post–COVID	0.542	0.839	0.562	
T12M	T12–1	pre-GFC	0.558	0.542	0.432	0.540
		GFC	0.583	0.541	0.728	
		early-expansion	0.550	0.519	0.466	
		late-expansion	0.583	0.245	0.165	
		COVID	0.528	0.868	0.877	
		post–COVID	0.645	0.150	0.796	
T6M	T6–1	pre-GFC	0.512	1.000	0.807	0.594
		GFC	0.750	0.023	0.015	
		early-expansion	0.600	0.155	0.121	
		late-expansion	0.567	0.366	0.273	
		COVID	0.611	0.243	0.196	
		post–COVID	0.484	1.000	0.959	
T3M	T3–1	pre-GFC	0.512	1.000	0.807	0.427
		GFC	0.750	0.023	0.015	
		early-expansion	0.600	0.155	0.121	
		late-expansion	0.567	0.366	0.273	
		COVID	0.611	0.243	0.196	
		post–COVID	0.484	1.000	0.959	

Spec Legend: F1 = (`ewma_w6_d0.75_mequal_weight`, `inv_err_w12_minverse_mse`, `inv_err_w6_mequal_weight`, `mwu_eta0.001`, `soft_bma_w12_nu10`)

T12–1 = (`ewma_w12_d0.95_minverse_mse`, `soft_bma_w12_nu10`, `soft_bma_w12_nu3`)

T6–1 = (`ewma_w12_d0.95_minverse_mse`, `ewma_w3_d0.75_mequal_weight`, `ewma_w6_d0.75_mequal_weight`)

T3–1 = (`ewma_w12_d0.95_minverse_mse`, `ewma_w3_d0.75_mequal_weight`, `ewma_w6_d0.75_mequal_weight`)

Overall, the in–sample results confirm that blending diverse model families delivers a persistent directional edge, especially in stress regimes (GFC, COVID) and early recoveries. However, shorter evaluation windows (T3M, T6M) are inherently more volatile, with wide swings in hit rates that reflect small–n sensitivity rather than genuine robustness. The FULL–history $k = 5$ blend remains the most compelling candidate for live deployment from an in–sample perspective unless one has a strong reason to favor the dynamic nature of shorter-horizon signals.

5.2.2 Dynamic Evaluation

The in–sample procedure above benefits from a look–ahead bias: ensemble specifications are chosen using the full history, then evaluated on that same history. To assess true live–feasibility, we re–run the selection process in a rolling, time–anchored fashion. At each month t, only data available up to $t - 1$ are used to (i) identify the candidate pool, (ii) select the best k–spec ensemble for each evaluation horizon (T3, T6, T12), and (iii) generate a directional signal for t. This produces an honest out–of–sample sequence of predictions.

We compare these dynamic ensembles to a **baseline** computed as the average of individual economist hit rates within each regime, where each economist's hit rate is calculated relative to the consensus median direction. This baseline captures the intrinsic "signal strength" of the economist panel without any ensembling.

Table 9 reports stratified diagnostics for each dynamic ensemble and the baseline, excluding trailing windows. All three dynamic horizons substantially outperform the baseline's overall hit rate (49.3%) in both accuracy and AC–score, with T6 achieving the highest overall hit rate (57.6%) and best AC (0.477). Performance is particularly strong in the GFC and post–COVID recovery, with T12 also excelling early expansion. T3, while competitive on average, shows greater variability, confirming that very short windows tend to overfit transient patterns. This, however, can prove to be valuable in times of unusually high volatility.

Table 9: Stratified diagnostics for dynamic robust ensembles and baseline.

Regime	**T3 HitRate**	**T6 HitRate**	**T12 HitRate**	**Baseline HitRate**
2004–2007 (pre-GFC)	0.500	0.475	0.525	0.516
2008–2009 (GFC)	0.667	0.625	0.583	0.526
2010–2014 (early-expansion)	0.567	0.600	0.600	0.497
2015–2019 (late-expansion)	0.583	0.583	0.567	0.469
2020–2022 (COVID)	0.639	0.600	0.560	0.499
2023–2025 (post–COVID)	0.516	0.581	0.452	0.475

Figure 6 plots the rolling hit rates for each horizon's winning ensemble over the evaluation period. T3 is visibly the noisiest series, with sharp month–to–month swings reflecting its susceptibility to small–sample variance. T6 is more stable, with smoother transitions and fewer abrupt reversals, while T12 offers the most consistent profile over long regimes but reacts more slowly to regime shifts. These patterns align with the AC–scores: T6's combination of high mean accuracy and low volatility makes it the most balanced performer, while T3 is better suited for opportunistic, high–beta directional calls, and T12 for slow–moving macro backdrops.

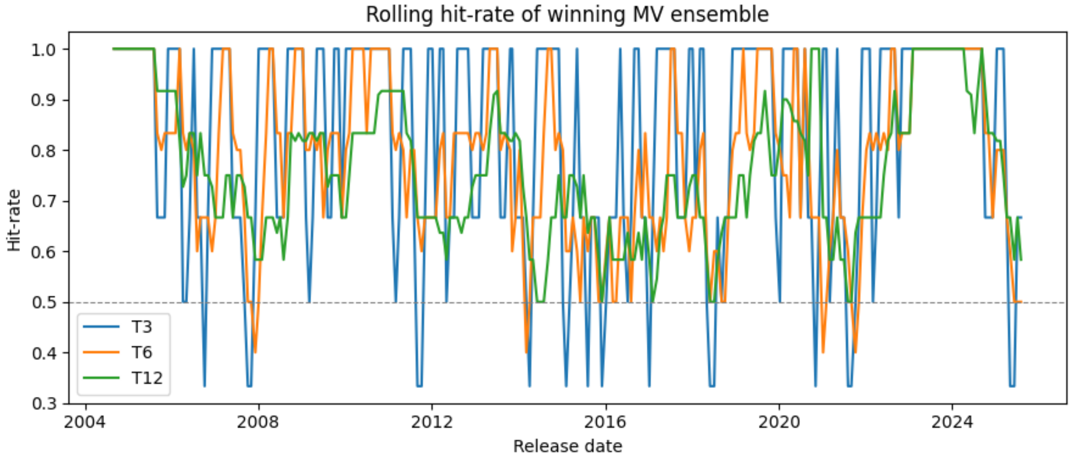

Figure 6: Rolling hit rate of the winning dynamic majority–vote ensemble for each horizon. The dashed line denotes the 50% no–skill level.

In sum, the dynamic evaluation confirms that the robust–ensemble framework materially improves on the baseline consensus direction, with the T6 ensemble offering the most attractive balance of accuracy and consistency for live deployment.

5.3 Distributional Performance

Methods and interval–construction differences. We evaluate four interval–forecasting engines:

- **Student–*t***: Fits a rolling 24–month Student–t error distribution to the residuals of the point forecast. Crisis–adjusted variants scale intervals in high–volatility regimes.
- **GARCH(1,1)–t**: Models conditional volatility dynamics directly from residuals via a GARCH(1,1) process with t–distributed innovations, optionally applying crisis multipliers.
- **Gaussian Mixture (GMM)**: Fits a two–component Gaussian mixture to the rolling 24–month residual set, with or without crisis scaling. Captures multi–modal error structures.
- **Bayesian Model Averaging (BMA)**: Averages predictive distributions from candidate engines weighted by recent likelihood, allowing for heavy–tailed members. The best–performing BMA variant here is *without* crisis adjustment.

These approaches differ in how they capture distributional shape (single–parametric tail vs. mixture), dynamics (static rolling fit vs. conditional volatility), and crisis–period scaling.

Summary of back–test results. Table 10 consolidates the best–performing variant of each family on the *Full* panel by mean absolute coverage gap (AvgAbsGap) and Accuracy–Consistency score (AC–Score, lower is better).

Table 10: Best–performing distributional models on the *Full* panel.

Method	**Best Tag**	**AvgAbsGap**	**AC–Score**
Gaussian–Mixture	Roll24_CrisisAdj	0.0114	**0.0173**
Student–t	Roll24_CrisisAdj	0.0116	0.0221
BMA	Roll24_NoAdj	**0.0101**	0.0371
GARCH(1,1)–t	GARCH_CrisisAdj	0.0189	0.082

Detailed observations by method.

Gaussian–Mixture. Crisis–adjusted GMM achieves the lowest AC–Score (0.0173) and the second–lowest AvgAbsGap (0.0114). Coverage is stable across strata, with only mild over–coverage in the most volatile blocks. The mixture form appears to capture asymmetric tail risks without producing excessive width in tranquil periods.

Student–*t*. Crisis–adjusted Student–t matches GMM in AvgAbsGap (0.0116) and performs slightly worse in AC–Score (0.0221). Its parametric simplicity yields well–behaved intervals, though tails are somewhat too narrow in post–COVID volatility, even with crisis multipliers.

BMA. The best–performing BMA variant is *without* crisis adjustment, yielding the lowest AvgAbsGap overall (0.0101) but a weaker AC–Score (0.0371). This reflects periods of excellent calibration offset by sharp degradation in specific regimes (notably 2015–2020), which inflates the consistency penalty.

GARCH(1,1)–t. While GARCH with crisis adjustment improves on its unadjusted counterpart in both AvgAbsGap and AC–Score, it remains the weakest performer overall (0.0189, 0.082). Its intervals are noticeably tighter than other engines, which aids in certain low–volatility phases but produces under–coverage in expansionary periods and post–COVID.

Regime–level patterns. Across methods, crisis–adjusted variants tend to improve calibration in the COVID block but can slightly overshoot in earlier periods. The BMA no–adj variant, while leading on average gap, shows significant regime dependence—excellent in blocks 2 and 4, weaker in block 3—suggesting sensitivity to shifts in the underlying point–forecast error process. GMM maintains the most even performance across all blocks, while Student–t is nearly as stable but more prone to under–coverage in extreme volatility. GARCH displays the strongest regime dependence, benefiting from its dynamic volatility adaptation in short–lived stress but struggling to match nominal targets in sustained expansions.

For NFP interval forecasting, *Gaussian Mixture with crisis adjustment* emerges as the most reliable all–rounder—its combination of low AvgAbsGap and the best AC–Score indicates both accurate calibration and stability across regimes. *Student–*t is a close second, offering a simpler alternative with competitive coverage. *BMA* is valuable when the goal is to minimize average gap, but its regime sensitivity suggests complementing it with a more stable engine in production. *GARCH* provides the tightest intervals and fastest adaptation but at a calibration cost, making it better suited for risk–seeking or directional–trading contexts rather than probability–calibrated forecasting. These findings indicate no single engine dominates all objectives; a composite or regime–switching approach could leverage each method's strengths.

5.3.1 Interpreting Multiple Densities

When the live forecasting system produces multiple competing predictive densities for the same NFP release—for example, from Gaussian–Mixture, Student–t, BMA, and GARCH(1,1)–t models—it is important to interpret them in light of their historical performance characteristics and methodological strengths.

1. Start with historical calibration metrics. The most direct way to assess which density to lean on is through its empirical–coverage record:

- **AvgAbsGap:** Lower values indicate that nominal coverage levels match realised frequencies more closely; this speaks to calibration.
- **AC–Score:** Balances calibration accuracy with stability across regimes. A low AC–Score signals both good fit and robustness.

For NFP, the Gaussian–Mixture and Student–t engines have the tightest calibration (AvgAbsGap $\approx$ 0.011–0.012) and strongest AC–Scores. This implies that, all else equal, their interval widths and shapes are most trustworthy as direct probabilistic statements.

2. Recognise regime–specific strengths. Some engines have profiles that vary meaningfully by macro regime:

- **Gaussian–Mixture:** Most stable across blocks; particularly strong in early and late expansions.
- **Student–*t*:** Heavy tails can better accommodate crisis–era outliers; tends to produce wider central intervals in volatile conditions.

- **BMA:** When not crisis–adjusted, can be the most sharply calibrated overall (lowest AvgAbsGap), but with more variability across regimes.
- **GARCH:** Narrower bands in calm periods; most reactive to volatility spikes; suitable when market–implied vol is a key conditioning input.

In practice, if the upcoming release is in a regime with elevated macro uncertainty (e.g., post–shock recovery), it is prudent to weigh Student–*t* or crisis–adjusted variants more heavily.

3. Examine shape and tail behaviour. Even for equally well–calibrated methods, the tails can differ markedly:

- **Gaussian–Mixture**: Can produce asymmetric or multi–modal densities if component means diverge, reflecting genuine disagreement in the expert pool.
- **Student–*t*** and **BMA**: Heavier tails; interpret tail quantiles as more generous allowances for extreme surprises.
- **GARCH**: Often produces the narrowest tail estimates in calm regimes, but may overshoot in crisis–adjusted form.

If decision–making is tail–sensitive, heavier–tailed densities deserve greater weight.

4. Avoid over–reliance on a single engine. No single method dominates all metrics in all regimes. A prudent approach is:

1. Identify the historically most reliable density by AC–Score.
2. Adjust subjective weight based on current regime characteristics.
3. Cross–check agreement in central coverage bands (e.g., 50–70%) across engines.
4. Investigate outliers in tail probabilities; if one method diverges sharply without regime justification, treat with caution.

5. Operational guidance for NFP.

- **Baseline:** Gaussian–Mixture (`Roll24_CrisisAdj`) for primary probabilistic guidance; it has the best AC–Score and stable calibration. Use student-t as a fallback.

- **Sharpest bands:** Use BMA (`Roll24_NoAdj`) if calibrated tightness is preferred in calm regimes, but cross–check against heavier–tailed alternatives.
- **Vol–responsive overlay:** Use GARCH(1,1)–t (`CrisisAdj`) when market–implied volatility or recent macro releases suggest rapidly changing risk.

In live use, we recommend reporting a range of intervals from two complementary engines, allowing the user to internalize both the central tendency and the plausible extremes.

6 Conclusion

Our unified forecasting framework demonstrates that even in a high–volatility, regime–shifting environment such as U.S. Nonfarm Payrolls, carefully designed ensemble methods can deliver a persistent directional edge over the crowd median and produce well–calibrated probabilistic forecasts. On the point–forecast side, *simple, short–horizon equal–weight averages* across contiguous forecasters—particularly 6–month equal–weight variants from the inverse–error and EWMA families—provide the most stable accuracy–consistency trade–off on the full history, while soft–BMA adds complementary directional strength in calmer regimes. Multiplicative–weights updates contribute as a diversifying voter but are not stand–alone leaders under our robustness criteria. On the distributional side, the *Gaussian–Mixture model with crisis adjustment* emerges as the most reliable all–rounder for interval calibration, followed closely by crisis–adjusted Student–t. BMA delivers the smallest average coverage gap but with greater regime sensitivity, while GARCH's adaptivity comes at the cost of under–coverage in long expansions.

Operational Guidance

- *Directional-signal baseline:* Use the T6 robust ensemble as default; T3 (reactive overlay) and T12 (stable backdrop) as context-dependent complements.
- *Density publication:* Primary = Gaussian–Mixture (Roll24, crisis-adjusted); secondary cross-check = Student–t (Roll24, crisis-adjusted). Tail-sensitive decisions can upweight heavier-tailed engines; calm-market options work may prefer the tighter BMA bands (cross-check stability first).
- *Regime awareness:* Use stratified diagnostics to match the current environment to

historical blocks where specific specs excelled.

Generalizability Roadmap

Although our empirical evaluation is anchored to NFP, the pipeline is designed to be indicator–agnostic. Extending the framework involves:

1. *Schema harmonization:* Ingest and standardize historical forecast vintages for the target indicator (e.g., CPI, Retail Sales, ISM).
2. *Error–process characterization:* Replicate exploratory diagnostics (distribution fit, spread–error elasticity) to select appropriate distributional families and crisis–gating logic.
3. *Model re–tuning:* Re–run walk–forward hyperparameter searches for point–forecast learners to accommodate indicator–specific forecast dispersion and volatility patterns.
4. *Density calibration:* Back–test all distributional engines with indicator–specific residuals to determine the best–performing primary/secondary densities and adjust crisis multipliers after running the corresponding regressions.
5. *Regime definition:* Adapt macro–regime partitions to the indicator's sensitivity

By following this process, the ensemble/density framework can be ported to a range of macroeconomic releases, yielding a library of consistent, cross–indicator probabilistic forecasts suitable for portfolio–level aggregation and macro–risk monitoring.

Limitations

Several constraints qualify our findings. Economist forecasts are made *ex–ante* using initially released NFP figures, while evaluation uses the latest revised values; revisions can materially alter measured errors in ways forecasters could not anticipate. In addition, the contiguity filter stabilizes weights but excludes intermittent forecasters, potentially biasing the panel toward established institutions.

A Appendix

A.1 Statistical Tests and Definitions

This section provides brief definitions of the statistical tests and metrics used to evaluate forecast performance in our pipeline.

- **Root Mean Squared Error (RMSE):** Measures the square root of the mean of squared forecast errors, providing an aggregate measure of point forecast accuracy in the same units as the target variable.
- **Hit Rate (HR):** The proportion of forecasts for which the predicted direction of change (relative to the consensus median) matches the actual realised direction.
- **Binomial Test:** A nonparametric test of whether the observed hit rate differs significantly from the null hypothesis of a 50% success probability. Useful for detecting directional skill.
- **Pesaran–Timmermann Test (PT):** A statistical test for directional accuracy that accounts for potential biases in the unconditional distribution of actual and predicted directions. It tests whether forecasts and outcomes are positively dependent beyond chance.
- **Diebold–Mariano Test (DM):** Compares the predictive accuracy of two competing forecasts (here, the "smart" model vs. the consensus median) by testing whether the mean loss differential is statistically different from zero, accounting for serial correlation.
- **Mean Absolute Coverage Gap (MAG):** For prediction intervals, the mean absolute deviation between empirical coverage and the nominal target coverage level, averaged across all levels tested.
- **Accuracy × Consistency Score (AC):** A composite measure of performance across regimes. For *point/directional forecasts*,

  $$\mathrm{AC} = (1 - \mathrm{HitRate}) + \lambda \cdot \sigma_{\text{block}},$$

  where σ_{block} is the standard deviation of block-level hit rates. For *distributional fore-*

casts, since Accuracy $= 1 - \text{MAG}$,

$$\text{AC} = \text{MAG} + \lambda \cdot \sigma_{\text{block}},$$

where σ_{block} is the standard deviation of block-level coverage gaps.

A.2 Comprehensive Backtest Results

Specification identifier convention. For compactness, each backtest entry is labeled with a `spec_id` that encodes the key hyperparameters of the forecast specification in a single string. The naming convention follows the template:

`{family}_w{W}_d{`ρ`}_m{rule}`

where:

- `family` denotes the model family, e.g., `inv_err` (inverse–error), `ewma` (exponentially weighted moving average), `soft_bma` (soft Bayesian model averaging), or `mwu` (multiplicative weights update).
- `w{W}` is the trailing window length in months used to compute performance statistics (e.g., `w6` means a 6-month window).
- `d{`ρ`}` is the temporal decay factor for EWMA families only; it is omitted for static-weight families (e.g., `d0.85` means $\rho = 0.85$).
- `rule` specifies the weighting scheme applied to individual forecasters: `equal_weight`, `inverse_mae`, or `inverse_mse`.
- For MWU, the spec identifier instead uses `{family}_eta{`η`}` where η is the learning rate.
- For soft–BMA, the identifier takes the form `soft_bma_w{W}_nu{`ν`}`, where ν is the fixed degrees–of–freedom parameter in the Student–t likelihood.

Example: The specification `ewma_w6_d0.85_minverse_mse` corresponds to an exponentially weighted moving average (`ewma`) model with a 6-month look-back window (`w6`), a temporal decay factor of $\rho = 0.85$ (`d0.85`), and inverse–mean–squared–error weighting of forecasters (`minverse_mse`).

A.2.1 Inverse–Error Backtests: Full Tables

This appendix reports the complete backtests for the inverse–error family across all window lengths ($W \in \{3, 6, 12\}$) and weighting rules (inverse–MSE, inverse–MAE, equal–weight). We list, for each specification, the RMSE of the smart forecast and the crowd median, directional hit rate and its exact binomial p–value, the Pesaran–Timmermann p–value (`PT_p`), and the Diebold–Mariano p–value (`DM_p`). Summary "winners" for each panel are provided below the full tables (candidate specifications for robust ensemble).

Table 11: Inverse–error backtests on the COVID–filtered panel. Abbreviations: W=window; RMSE_s=RMSE(smart); RMSE_m=RMSE(median); HR=hit rate.

spec_id	**W**	**method**	**RMSE_s**	**RMSE_m**	**HR**	**Binom_p**	**PT_p**	**DM_p**
`inv_err_w3_minverse_mse`	3	inverse_mse	72.256	73.173	0.529	0.426	0.372	0.139
`inv_err_w3_minverse_mae`	3	inverse_mae	72.238	73.173	0.533	0.353	0.313	0.036
`inv_err_w3_mequal_weight`	3	equal_weight	72.316	73.173	0.555	0.111	0.095	0.008
`inv_err_w6_minverse_mse`	6	inverse_mse	71.746	72.955	0.571	0.038	0.029	0.020
`inv_err_w6_minverse_mae`	6	inverse_mae	71.865	72.955	0.558	0.095	0.075	0.013
`inv_err_w6_mequal_weight`	6	equal_weight	71.951	72.955	0.580	0.019	0.015	0.013
`inv_err_w12_minverse_mse`	12	inverse_mse	69.782	70.913	0.564	0.067	0.054	0.028
`inv_err_w12_minverse_mae`	12	inverse_mae	69.923	70.913	0.550	0.155	0.126	0.045
`inv_err_w12_mequal_weight`	12	equal_weight	70.096	70.913	0.555	0.119	0.100	0.077

Table 12: COVID–filtered panel winners (inverse–error family).

Category	**Specification**	
Lowest RMSE	`inv_err_w12_minverse_mse`	(window = 12, method = inverse_mse)
Highest HitRate	`inv_err_w6_mequal_weight`	(window = 6, method = equal_weight)
Robust Winner	`inv_err_w12_minverse_mse`	

Table 13: Inverse–error backtests on the full panel. Abbreviations: W=window; RMSE_s=RMSE(smart); RMSE_m=RMSE(median); HR=hit rate.

spec_id	**W**	**method**	**RMSE_s**	**RMSE_m**	**HR**	**Binom_p**	**PT_p**	**DM_p**
`inv_err_w3_minverse_mse`	3	inverse_mse	664.827	641.034	0.551	0.109	0.104	0.331
`inv_err_w3_minverse_mae`	3	inverse_mae	641.926	641.034	0.551	0.109	0.101	0.837
`inv_err_w3_mequal_weight`	3	equal_weight	635.159	641.034	0.559	0.064	0.058	0.085
`inv_err_w6_minverse_mse`	6	inverse_mse	666.750	644.653	0.569	0.030	0.028	0.303
`inv_err_w6_minverse_mae`	6	inverse_mae	635.931	644.653	0.565	0.040	0.037	0.264
`inv_err_w6_mequal_weight`	6	equal_weight	619.565	644.653	0.585	0.008	0.007	0.283
`inv_err_w12_minverse_mse`	12	inverse_mse	679.432	651.933	0.571	0.028	0.027	0.368
`inv_err_w12_minverse_mae`	12	inverse_mae	643.409	651.933	0.559	0.069	0.068	0.111
`inv_err_w12_mequal_weight`	12	equal_weight	631.802	651.933	0.555	0.090	0.085	0.207

Table 14: Full panel winners (inverse–error family).

Category	**Specification**
Lowest RMSE & Highest HitRate	`inv_err_w6_mequal_weight`
Robust Winner	`inv_err_w3_mequal_weight`

A.2.2 EWMA Backtests: Full Tables

This appendix reports the complete EWMA backtests across all window lengths $W \in \{3, 6, 12\}$, decay factors $\rho \in \{0.75, 0.85, 0.95\}$, and weighting rules (inverse–MSE, inverse–MAE, equal–weight). For each specification we list the RMSE of the smart forecast and the crowd median, the directional hit rate and its exact binomial p–value, the Pesaran–Timmermann p–value (`PT_p`), and the Diebold–Mariano p–value (`DM_p`). Summary "winners" for each panel are provided below the full tables (candidate specifications for robust ensemble).

Table 15: EWMA backtests on the COVID–filtered panel. Abbreviations: RMSE_s=RMSE(smart); RMSE_m=RMSE(median); HR=hit rate.

spec_id	**method**	**RMSE_s**	**RMSE_m**	**HR**	**Binom_p**	**PT_p**	**DM_p**
ewma_w3_d0.75_minverse_mse	inverse_mse	72.218	73.173	0.542	0.232	0.199	0.126
ewma_w3_d0.75_minverse_mae	inverse_mae	72.213	73.173	0.533	0.353	0.311	0.035
ewma_w3_d0.75_mequal_weight	equal_weight	72.316	73.173	0.555	0.111	0.095	0.008
ewma_w3_d0.85_minverse_mse	inverse_mse	72.236	73.173	0.537	0.288	0.246	0.132
ewma_w3_d0.85_minverse_mae	inverse_mae	72.225	73.173	0.529	0.426	0.377	0.035
ewma_w3_d0.85_mequal_weight	equal_weight	72.316	73.173	0.555	0.111	0.095	0.008
ewma_w3_d0.95_minverse_mse	inverse_mse	72.250	73.173	0.529	0.426	0.372	0.136
ewma_w3_d0.95_minverse_mae	inverse_mae	72.234	73.173	0.537	0.288	0.253	0.036
ewma_w3_d0.95_mequal_weight	equal_weight	72.316	73.173	0.555	0.111	0.095	0.008
ewma_w6_d0.75_minverse_mse	inverse_mse	71.762	72.955	0.567	0.052	0.039	0.032
ewma_w6_d0.75_minverse_mae	inverse_mae	71.866	72.955	0.549	0.160	0.135	0.021
ewma_w6_d0.75_mequal_weight	equal_weight	71.951	72.955	0.580	0.019	0.015	0.013
ewma_w6_d0.85_minverse_mse	inverse_mse	71.760	72.955	0.580	0.019	0.013	0.027
ewma_w6_d0.85_minverse_mae	inverse_mae	71.869	72.955	0.549	0.160	0.133	0.018
ewma_w6_d0.85_mequal_weight	equal_weight	71.951	72.955	0.580	0.019	0.015	0.013
ewma_w6_d0.95_minverse_mse	inverse_mse	71.752	72.955	0.576	0.027	0.019	0.022
ewma_w6_d0.95_minverse_mae	inverse_mae	71.867	72.955	0.562	0.071	0.056	0.015
ewma_w6_d0.95_mequal_weight	equal_weight	71.951	72.955	0.580	0.019	0.015	0.013
ewma_w12_d0.75_minverse_mse	inverse_mse	69.895	70.913	0.537	0.310	0.262	0.070
ewma_w12_d0.75_minverse_mae	inverse_mae	69.968	70.913	0.528	0.456	0.400	0.069
ewma_w12_d0.75_mequal_weight	equal_weight	70.096	70.913	0.555	0.119	0.100	0.077
ewma_w12_d0.85_minverse_mse	inverse_mse	69.860	70.913	0.555	0.119	0.096	0.050
ewma_w12_d0.85_minverse_mae	inverse_mae	69.953	70.913	0.537	0.310	0.265	0.059
ewma_w12_d0.85_mequal_weight	equal_weight	70.096	70.913	0.555	0.119	0.100	0.077
ewma_w12_d0.95_minverse_mse	inverse_mse	69.809	70.913	0.560	0.090	0.071	0.034
ewma_w12_d0.95_minverse_mae	inverse_mae	69.933	70.913	0.546	0.198	0.163	0.049
ewma_w12_d0.95_mequal_weight	equal_weight	70.096	70.913	0.555	0.119	0.100	0.077

Table 16: COVID–filtered panel winners (EWMA family).

Category	**Specification**
Lowest RMSE	`ewma_w12_d0.95_minverse_mse` (window = 12, decay = 0.95, method = inverse_ms
Highest HitRate	`ewma_w6_d0.75_mequal_weight` (window = 6, decay = 0.75, method = equal_weigh
Robust Winner	`ewma_w12_d0.95_minverse_mse`

Table 17: EWMA backtests on the full panel. Abbreviations: RMSE_s=RMSE(smart); RMSE_m=RMSE(median); HR=hit rate.

spec_id	**method**	**RMSE_s**	**RMSE_m**	**HR**	**Binom_p**	**PT_p**	**DM_p**
`ewma_w3_d0.75_minverse_mse`	inverse_mse	666.253	641.034	0.563	0.048	0.045	0.329
`ewma_w3_d0.75_minverse_mae`	inverse_mae	643.615	641.034	0.551	0.109	0.102	0.655
`ewma_w3_d0.75_mequal_weight`	equal_weight	635.159	641.034	0.559	0.064	0.058	0.085
`ewma_w3_d0.85_minverse_mse`	inverse_mse	665.679	641.034	0.559	0.064	0.061	0.330
`ewma_w3_d0.85_minverse_mae`	inverse_mae	642.891	641.034	0.548	0.139	0.132	0.718
`ewma_w3_d0.85_mequal_weight`	equal_weight	635.159	641.034	0.559	0.064	0.058	0.085
`ewma_w3_d0.95_minverse_mse`	inverse_mse	665.110	641.034	0.551	0.109	0.104	0.331
`ewma_w3_d0.95_minverse_mae`	inverse_mae	642.233	641.034	0.555	0.084	0.078	0.794
`ewma_w3_d0.95_mequal_weight`	equal_weight	635.159	641.034	0.559	0.064	0.058	0.085
`ewma_w6_d0.75_minverse_mse`	inverse_mse	668.305	644.653	0.573	0.022	0.020	0.306
`ewma_w6_d0.75_minverse_mae`	inverse_mae	638.035	644.653	0.562	0.054	0.050	0.263
`ewma_w6_d0.75_mequal_weight`	equal_weight	619.565	644.653	0.585	0.008	0.007	0.283
`ewma_w6_d0.85_minverse_mse`	inverse_mse	667.697	644.653	0.585	0.008	0.007	0.304
`ewma_w6_d0.85_minverse_mae`	inverse_mae	637.168	644.653	0.562	0.054	0.050	0.264
`ewma_w6_d0.85_mequal_weight`	equal_weight	619.565	644.653	0.585	0.008	0.007	0.283
`ewma_w6_d0.95_minverse_mse`	inverse_mse	667.071	644.653	0.573	0.022	0.020	0.303
`ewma_w6_d0.95_minverse_mae`	inverse_mae	636.334	644.653	0.573	0.022	0.020	0.264
`ewma_w6_d0.95_mequal_weight`	equal_weight	619.565	644.653	0.585	0.008	0.007	0.283
`ewma_w12_d0.75_minverse_mse`	inverse_mse	683.621	651.933	0.543	0.188	0.187	0.362
`ewma_w12_d0.75_minverse_mae`	inverse_mae	649.546	651.933	0.531	0.347	0.347	0.541
`ewma_w12_d0.75_mequal_weight`	equal_weight	631.802	651.933	0.555	0.090	0.085	0.207
`ewma_w12_d0.85_minverse_mse`	inverse_mse	682.431	651.933	0.559	0.069	0.067	0.364
`ewma_w12_d0.85_minverse_mae`	inverse_mae	647.630	651.933	0.539	0.233	0.233	0.214
`ewma_w12_d0.85_mequal_weight`	equal_weight	631.802	651.933	0.555	0.090	0.085	0.207
`ewma_w12_d0.95_minverse_mse`	inverse_mse	680.687	651.933	0.567	0.038	0.037	0.366
`ewma_w12_d0.95_minverse_mae`	inverse_mae	645.029	651.933	0.555	0.090	0.089	0.107
`ewma_w12_d0.95_mequal_weight`	equal_weight	631.802	651.933	0.555	0.090	0.085	0.207

Table 18: Full panel winners (EWMA family).

Category	**Specification**
Lowest RMSE & Highest HitRate	`ewma_w6_d0.75_mequal_weight`
Robust Winner	`ewma_w3_d0.75_mequal_weight`

A.2.3 Soft-BMA Backtests: Full Tables

This appendix reports the complete soft-BMA backtests across all window lengths $W \in \{3, 6, 12\}$ and tail parameters $\nu \in \{3, 5, 10, 25\}$. For each specification we list the RMSE of the smart forecast and the crowd median, the directional hit rate (HR), the exact binomial p–value, the Pesaran–Timmermann p–value (`PT_p`), and the Diebold–Mariano p–value (`DM_p`). To keep the layout compact we omit auxiliary columns; the window and ν are encoded in `spec_id`. Summary "winners" for each panel are provided below the full tables.

Table 19: soft-BMA backtests on the COVID-filtered panel. Abbreviations: RMSE_s=RMSE(smart); RMSE_m=RMSE(median); HR=hit rate.

spec_id	**RMSE_s**	**RMSE_m**	**HR**	**Binom_p**	**PT_p**	**DM_p**
`soft_bma_w3_nu3`	72.549	73.173	0.520	0.596	0.538	0.449
`soft_bma_w3_nu5`	72.556	73.173	0.533	0.353	0.308	0.455
`soft_bma_w3_nu10`	72.563	73.173	0.533	0.353	0.308	0.461
`soft_bma_w3_nu25`	72.569	73.173	0.524	0.507	0.449	0.466
`soft_bma_w6_nu3`	71.667	72.955	0.558	0.095	0.076	0.185
`soft_bma_w6_nu5`	71.663	72.955	0.562	0.071	0.057	0.185
`soft_bma_w6_nu10`	71.664	72.955	0.562	0.071	0.057	0.186
`soft_bma_w6_nu25`	71.668	72.955	0.562	0.071	0.057	0.188
`soft_bma_w12_nu3`	67.656	70.913	0.578	0.025	0.020	0.067
`soft_bma_w12_nu5`	67.679	70.913	0.573	0.036	0.028	0.064
`soft_bma_w12_nu10`	67.719	70.913	0.583	0.018	0.013	0.062
`soft_bma_w12_nu25`	67.771	70.913	0.583	0.018	0.014	0.060

Table 20: COVID-filtered panel winners (soft-BMA).

Category	**Specification**
Lowest RMSE	`soft_bma_w12_nu3` (window = 12, $\nu = 3$)
Highest HitRate	`soft_bma_w12_nu10` (window = 12, $\nu = 10$)
Robust Winner	`soft_bma_w12_nu3`

Table 21: soft-BMA backtests on the full panel. Abbreviations: RMSE_s=RMSE(smart); RMSE_m=RMSE(median); HR=hit rate.

spec_id	**RMSE_s**	**RMSE_m**	**HR**	**Binom_p**	**PT_p**	**DM_p**
`soft_bma_w3_nu3`	686.489	641.034	0.525	0.459	0.441	0.296
`soft_bma_w3_nu5`	687.010	641.034	0.536	0.267	0.255	0.295
`soft_bma_w3_nu10`	687.472	641.034	0.536	0.267	0.255	0.293
`soft_bma_w3_nu25`	687.793	641.034	0.529	0.388	0.376	0.292
`soft_bma_w6_nu3`	761.342	644.653	0.558	0.072	0.067	0.279
`soft_bma_w6_nu5`	762.128	644.653	0.558	0.072	0.067	0.278
`soft_bma_w6_nu10`	762.680	644.653	0.558	0.072	0.067	0.277
`soft_bma_w6_nu25`	762.940	644.653	0.558	0.072	0.067	0.275
`soft_bma_w12_nu3`	837.900	651.933	0.575	0.020	0.019	0.285
`soft_bma_w12_nu5`	841.044	651.933	0.567	0.038	0.036	0.284
`soft_bma_w12_nu10`	844.073	651.933	0.571	0.028	0.027	0.284
`soft_bma_w12_nu25`	846.392	651.933	0.571	0.028	0.027	0.282

Table 22: Full panel winners (soft-BMA).

Category	**Specification**
Lowest RMSE	`soft_bma_w3_nu3`
Highest HitRate	`soft_bma_w12_nu3`
Robust Winner	— (no specification meets the DM/PT < 0.10 gate)

A.2.4 MWU Backtests: Full Tables

This appendix reports the complete backtests for the multiplicative weights update (MWU) family across step sizes $\eta \in \{0.001, \dots, 0.019\}$. For each specification we list the RMSE of the smart forecast and the crowd median, directional hit rate with exact binomial p–value, Pesaran–Timmermann p–value (`PT_p`), and Diebold–Mariano p–value (`DM_p`). Summary "winners" are provided below the full tables.

Table 23: MWU backtests on the COVID–filtered panel. Abbreviations: RMSE_s=RMSE(smart); RMSE_m=RMSE(median); HR=hit rate.

spec_id	**RMSE_s**	**RMSE_m**	**HR**	**Binom_p**	**PT_p**	**DM_p**
`mwu_eta0.001`	476.387	651.933	0.524	0.490	0.717	0.307
`mwu_eta0.003`	595.166	651.933	0.539	0.233	0.277	0.293
`mwu_eta0.005`	435.725	651.933	0.528	0.415	0.942	0.309
`mwu_eta0.007`	631.401	651.933	0.488	0.754	0.469	0.284
`mwu_eta0.009`	631.507	651.933	0.480	0.572	0.469	0.287
`mwu_eta0.011`	435.769	651.933	0.496	0.950	0.772	0.309
`mwu_eta0.013`	435.990	651.933	0.531	0.347	0.942	0.309
`mwu_eta0.015`	436.015	651.933	0.543	0.188	0.828	0.309
`mwu_eta0.017`	436.000	651.933	0.535	0.286	0.828	0.309
`mwu_eta0.019`	436.032	651.933	0.531	0.347	0.942	0.309

Table 24: COVID–filtered panel winners (MWU family).

Category	**Specification**
Lowest RMSE	`mwu_eta0.005`
Highest HitRate	`mwu_eta0.015`
Robust Winner	None (DM_p & PT_p ≥ 0.10)

Table 25: MWU backtests on the full panel. Abbreviations: RMSE_s=RMSE(smart); RMSE_m=RMSE(median); HR=hit rate.

spec_id	**RMSE_s**	**RMSE_m**	**HR**	**Binom_p**	**PT_p**	**DM_p**
mwu_eta0.001	72.519	70.913	0.509	0.839	0.755	0.220
mwu_eta0.003	76.251	70.913	0.523	0.542	0.696	0.032
mwu_eta0.005	77.652	70.913	0.500	1.000	0.348	0.003
mwu_eta0.007	80.261	70.913	0.472	0.456	0.211	0.001
mwu_eta0.009	79.540	70.913	0.459	0.250	0.159	0.002
mwu_eta0.011	75.939	70.913	0.477	0.542	0.211	0.015
mwu_eta0.013	74.683	70.913	0.523	0.542	0.532	0.070
mwu_eta0.015	74.856	70.913	0.537	0.310	0.639	0.052
mwu_eta0.017	74.860	70.913	0.528	0.456	0.639	0.052
mwu_eta0.019	75.069	70.913	0.523	0.542	0.532	0.040

Table 26: Full panel winners (MWU family).

Category	**Specification**
Lowest RMSE	mwu_eta0.001
Highest HitRate	mwu_eta0.015
Robust Winner	None (DM_p & PT_p ≥ 0.10)

A.2.5 Distributional Methods: Full Tables

This appendix reports complete empirical–coverage back–tests for all distributional–forecasting engines evaluated on the *Full* panel. For each method, we present empirical coverage at nominal levels $(50\%, 60\%, 70\%, 80\%, 90\%, 95\%)$, mean absolute coverage gap (AvgAbsGap; lower is better), and the Accuracy–Consistency (AC) score computed across four long–horizon strata. Crisis–adjusted (`CrisisAdj`) variants scale predictive intervals during high–volatility regimes.

Table 27: Gaussian–Mixture empirical coverage results (*Full* panel).

Spec	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**AvgAbsGap**
Roll24_CrisisAdj	0.506	0.626	0.716	0.798	0.893	0.938	0.0114
Roll24_NoAdj	0.490	0.609	0.700	0.774	0.872	0.922	0.0170

Gaussian–Mixture. Best spec: `Roll24_CrisisAdj`, AC–Score = 0.0173.

Table 28: Stratified AvgAbsGap by block for `GMM Roll24_CrisisAdj` (*Full* panel).

Block	**Start**	**End**	**AvgAbsGap**
1	2005-06-03	2010-06-04	0.030
2	2010-07-02	2015-07-02	0.017
3	2015-08-07	2020-08-07	0.025
4	2020-09-04	2025-08-01	0.028
All	2003-06-06	2025-08-01	0.011

Table 29: Student–t empirical coverage results (*Full* panel).

Spec	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**AvgAbsGap**
Roll24_CrisisAdj	0.504	0.603	0.682	0.822	0.917	0.955	0.0116
Roll24_NoAdj	0.492	0.574	0.657	0.802	0.901	0.938	0.0150

Student–*t*. Best spec: `Roll24_CrisisAdj`, AC–Score = 0.022.

Table 30: Stratified AvgAbsGap by block for `Student-`t `Roll24_CrisisAdj` (*Full* panel).

Block	**Start**	**End**	**AvgAbsGap**
1	2005-06-03	2010-06-04	0.023
2	2010-07-02	2015-07-02	0.016
3	2015-08-07	2020-08-07	0.034
4	2020-09-04	2025-08-01	0.044
All	2003-06-06	2025-08-01	0.009

Table 31: BMA empirical coverage results (*Full* panel).

Spec	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**AvgAbsGap**
Roll24_CrisisAdj	0.521	0.628	0.723	0.810	0.909	0.950	0.0150
Roll24_NoAdj	0.504	0.612	0.702	0.785	0.897	0.934	**0.0101**

BMA. Best spec: `Roll24_NoAdj`, AC–Score = 0.0371.

Table 32: Mean–absolute coverage gap by block for `BMA Roll24_NoAdj` (*Full* panel).

Block	**Start**	**End**	**AvgAbsGap**
1	2005-06-03	2010-06-04	0.028
2	2010-07-02	2015-07-02	0.010
3	2015-08-07	2020-07-02	0.075
4	2020-08-07	2025-07-03	0.053
All	2003-06-06	2025-07-03	0.009

Table 33: GARCH(1,1)–t empirical coverage results (*Full* panel).

Spec	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**AvgAbsGap**
GARCH_CrisisAdj	0.554	0.657	0.723	0.818	0.909	0.955	0.0276
GARCH_NoAdj	0.533	0.645	0.707	0.802	0.888	0.934	**0.0189**

GARCH(1,1)–*t*. Best spec: `GARCH_NoAdj`, AC–Score = 0.0823.

Table 34: Stratified AvgAbsGap by block for `GARCH_NoAdj` (*Full* panel).

Block	**Start**	**End**	**AvgAbsGap**
1	2005-06-03	2010-06-04	0.014
2	2010-07-02	2015-07-02	0.018
3	2015-08-07	2020-07-02	0.086
4	2020-08-07	2025-07-03	0.147
All	2003-06-06	2025-07-03	0.019

Table 35: Consolidated best–spec performance across distributional methods (*Full* panel).

Method	**Best Tag**	**AvgAbsGap**	**AC–Score**
Gaussian–Mixture	Roll24_CrisisAdj	0.0114	**0.0173**
Student–t	Roll24_CrisisAdj	0.0116	0.0221
BMA	Roll24_NoAdj	**0.0101**	0.0371
GARCH(1,1)–t	GARCH_NoAdj	0.0189	0.0823

A.3 Top Weighted Economists for August 2025 Print

This appendix reports the top ten most heavily weighted economists for each model family in the live forecast snapshot for the August 2025 NFP release, as well as the top ten in the equal–model blend (25% per family).

Methodology. At the live evaluation date, we compute weights for each economist within each model family by:

1. Filtering the snapshot to the most recent month.
2. Averaging weights across all live specifications and panels for that family.
3. Normalising so weights sum to 1 within each family.

For the equal–model blend, each family's weight vector is scaled to a target share of 25% and then summed across families to yield the final blended weights. These represent each economist's proportional influence on the aggregate signal.

How to read this table: Higher model weights indicate that the economist's forecasts currently have greater influence on the model family's point forecast. A high rank across multiple families signals an economist whose recent accuracy patterns have been broadly rewarded. Conversely, low or absent weights indicate that the economist has either been inactive or received little weight based on recent performance.

	Economist	**Model Weight**
Inverse–Error	David P Kelly	3.90%
	Seiji Katsurahata	3.11%
	Ashworth/Dales	2.88%
	Rhys Herbert	2.77%
	Derek Holt	2.60%
	Jason M Schenker	2.58%
	Joe Brusuelas/Tuan Nguyen	2.50%
	Michael R Englund	2.49%
	Russell T Price	2.45%
	Michael E Feroli	2.42%

	Economist	**Model Weight**
EWMA	David P Kelly	3.99%
	Seiji Katsurahata	3.19%
	Ashworth/Dales	3.05%
	Rhys Herbert	3.03%
	Jason M Schenker	2.69%
	Derek Holt	2.68%
	Michael R Englund	2.63%
	Joe Brusuelas/Tuan Nguyen	2.51%
	Michael E Feroli	2.50%
	Oscar Munoz	2.43%

	Economist	**Model Weight**
soft-BMA	David P Kelly	10.91%
	Seiji Katsurahata	7.43%
	Russell T Price	6.10%
	David H Sloan	4.97%
	Derek Holt	4.69%
	Richard F Moody	4.65%
	Ashworth/Dales	4.35%
	Joe Brusuelas/Tuan Nguyen	4.32%
	Rhys Herbert	3.38%
	James Egelhof	3.19%

	Economist	**Model Weight**
MWU	Yongxin Chen	6.92%
	Joe Brusuelas/Tuan Nguyen	6.75%
	Andreas Busch	2.06%
	Andrew Zatlin	2.06%
	Avery Shenfeld	2.06%
	Besch/Luetje	2.06%
	Brett Ryan	2.06%
	Christophe Barraud	2.06%
	Christopher Hodge	2.06%
	Crandall/Jordan	2.06%

	Economist	Weight
Equal–Model Blend	David P Kelly	5.21%
	Joe Brusuelas/Tuan Nguyen	4.02%
	Seiji Katsurahata	3.95%
	Yongxin Chen	3.39%
	Russell T Price	3.24%
	Ashworth/Dales	3.08%
	Derek Holt	3.01%
	David H Sloan	2.83%
	Rhys Herbert	2.81%
	Richard F Moody	2.71%