When Does Synthetic Data Improve Forecast Combination?

An Information-Channel Decomposition

Hyun Hak Kim

Summary

We develop an empirical framework for determining whether simulation-trained forecast combiners benefit from outside information. The framework distinguishes information imported by a simulator from regularization induced by synthetic training. In a calendar-cutoff-clean U.S. inflation application, cross-country pooling improves prediction of episode amplitude, while tested financial features add little information about episode onset. Across 16 prespecified arm–horizon comparisons, no outside-information arm beats the own-data control after familywise adjustment; 12 are equivalent within a 2% RMSFE margin and four remain inconclusive. Thus severity information need not translate into a detectable unconditional point-forecast gain.

Keywords: forecast combination; synthetic data; prior-fitted networks; equal-weights puzzle; inflation forecasting.
JEL: C32, C45, C52, C53, E31, E37.

Correspondence: Hyun Hak Kim, Department of Economics, Kookmin University, 77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Republic of Korea.
Email: hyunhak.kim@kookmin.ac.kr. ORCID: 0000-0002-4909-500X.

Acknowledgments: The author thanks Norman R. Swanson for comments that motivated the information-channel framing of the paper.

1 Introduction

More than fifty years after Bates and Granger (1969), the central puzzle of forecast combination is still that the simplest rule wins. Weights estimated to minimize combined error rarely beat the equal-weighted average out of sample, across data sets, loss functions, and estimation schemes (Stock and Watson, 2004; Timmermann, 2006; Claeskens et al., 2016). The standard reading is an estimation story: estimated weights add variance that their in-sample optimality does not repay (Claeskens et al., 2016). We take a different vantage. Suppose the estimation problem were solved — suppose a combiner could be handed, for free, an oracle's worth of weight-fitting. What would it then need to *know* to beat equal weights, and where could that knowledge come from? Posed this way, the puzzle becomes a question about information, and it is the question a recent and fast-growing practice forces upon us.

That practice is synthetic data. Generative models trained to reproduce a macroeconomic panel are increasingly used to augment forecasting samples, on the implicit premise that more history — even simulated history — yields better learned decision rules. The premise is ill-posed until one asks what the simulator knows. A generator estimated on the forecaster's own dataset produces histories that are randomized functions of that sample. Conditional on the sample, those histories add no information about the future. They can still alter finite-sample performance through regularization, model restriction, and inductive bias; information neutrality is not performance equivalence to equal weights or to any particular classical method. Information *beyond* the main sample must come from outside sources such as high-frequency financial data, foreign experience, or distant history, but its presence is necessary rather than sufficient for a forecast gain. Section 2 formalizes the conditional-information result and a hierarchical example in which cross-country episodes can reduce uncertainty about episode severity.

We apply this framework to U.S. inflation. A common pipeline is run in five generator-

mixture arms built from own data, daily financial data, a 111-country inflation panel, pre-1960 history, or all three outside sets. We separately ask whether daily finance adds evidence about episode onset and whether cross-country experience improves prediction of episode severity and duration. A neural combiner, the prior-fitted forecast combiner, is trained on 200 replicates per generator, in which the recursive forecasting exercise and pool re-estimation are re-enacted, and is then deployed without real-data training on six U.S. inflation measures. The calculation separates 1985–2004 design from a common 2005–2024 evaluation, fixes generator weights independently of diagnostic outcomes, and uses ten training seeds per arm and horizon. It is calendar-cutoff clean under a fixed revised vintage, not a real-time-vintage experiment.

The evidence establishes a deliberately narrow pattern. Calendar-rolling cross-country pooling improves predictive density for episode amplitude, while the locked daily-financial specification adds little incremental evidence about episode onset. None of the sixteen outside-arm–horizon contrasts shows a familywise-supported point-forecast gain over A0. Twelve are equivalent within a two-percent RMSFE margin; four long-horizon cells remain inconclusive. We interpret the severity–onset distinction as a mechanism consistent with these facts: knowing how severe an episode may be need not improve an unconditional point forecast if the observable state contains little incremental information about whether the episode will start. This is not a general impossibility claim.

Our main contribution of this paper is as follows. First, we formalize a conditional-information boundary: a corpus simulated from the forecaster's own sample is a randomized sample-dependent procedure and adds no information conditional on that sample (Proposition 1). This does not preclude finite-sample gains or losses from regularization and inductive bias. A hierarchical episode model then shows how cross-country observations can reduce posterior uncertainty about episode severity under explicit exchangeability assumptions (Proposition 2). Then we introduce the *prior-fitted forecast combiner* (PFFC), a permutation-equivariant attention network trained entirely on sim-

ulated forecasting histories — replicate corpora in which the full pseudo-out-of-sample exercise, including pool re-estimation, is re-enacted — and deployed zero-shot on real data. The design extends the prior-fitted-network mechanism from prediction to the combination-weight functional, with a gated head that nests equal weights. Third, we organize the empirical exercise as an information-channel decomposition: five arms share one architecture and forecasting pool while their generator mixtures draw on three candidate outside information sets — daily financial data, a 111-country inflation panel, and pre-1960 history. Channel diagnostics are reported separately from point-forecast performance so that information content is not inferred from a forecast ranking alone. Finally, we use the distinction between episode severity and duration (hereafter, episode geometry) and episode onset to interpret why a channel may help the former without producing a detectable point-forecast gain. This is a stylized mechanism consistent with the diagnostics, not a claim that severity information can never affect a point forecast.

The paper joins three strands. The first is the forecast-combination puzzle and the econometrics of weight estimation, from Bates and Granger (1969) and Granger and Ramanathan (1984) through the equal-weights evidence of Stock and Watson (2004), the survey of Timmermann (2006), and the variance-based explanation of Claeskens et al. (2016) and its structural- and instability-based predecessors (Diebold and Pauly, 1987; Smith and Wallis, 2009; Elliott and Liao, 2026); we add an information audit to the usual estimation-error account. Recent JAE evidence provides two demanding benchmarks for that audit. Disaggregated commodity information can improve both point and density inflation forecasts (Garratt and Petrella, 2022), while the relative success of machine-learning inflation forecasts changes across countries and the pandemic period, with simple combinations remaining competitive (Naghi et al., 2024). Hierarchical pooling across related time series can also produce sizable gains when similarity and dependence are modeled directly (Müller and Watson, 2026). Our matched-arm design asks the prior question for simulation training: which information was imported, and did it reach the

final forecast loss?

The second strand is simulation-trained, or prior-fitted, inference (Müller et al., 2022; Nagler, 2023), including its recent applications to zero-shot time-series prediction (Dooley et al., 2023; Taga et al., 2025; Ansari et al., 2024; Woo et al., 2024; Carriero et al., 2026); we extend the mechanism from the prediction functional itself to the combination-weight functional and equip it with the information-theoretic discipline of Section 2 — these zero-shot forecasters answer "what will happen," trained on synthetic sequences, while the PFFC answers "whom to trust," trained on synthetic *track records*, and Proposition 1 applies to the information content of either use whenever the training corpus carries no outside information. The third is the use of outside and cross-country information in inflation forecasting — the global inflation factor of Ciccarelli and Mojon (2010), the cross-country database of Ha et al. (2023), the long-run panel of Jordà et al. (2017), the distinction between mean and uncertainty objectives in density combination (Knüppel and Krüger, 2022), and the durable finding that domestic activity variables struggle to beat naive inflation benchmarks (Atkeson and Ohanian, 2001). Our contribution to this strand is to make "what does the simulator know?" the identifying question for synthetic-data value, and to test the answer channel by channel.

The remainder proceeds as follows. Section 2 develops the theory and its five testable hypotheses; Section 3 describes the panel and the three outside sets; Section 4 lays out the experiment; Section 5 presents the implementation record, empirical results, and their interpretation; and Section 6 concludes.

2 An information-channel theory of synthetic data

This section formalizes a single question: *through what channel could synthetic data improve a forecast combination?* Synthetic histories drawn from generators estimated on the forecaster's own dataset are randomized functionals of that dataset and add no in-

formation conditional on it (Proposition 1). They may nevertheless change finite-sample performance through regularization and inductive bias. Information beyond the sample requires an outside source, but outside information improves a point forecast only if it changes decision-relevant conditional moments. Proposition 2 gives one hierarchical example in which additional cross-country episodes reduce uncertainty about episode severity. Section 2.4 translates these boundaries into empirical questions rather than performance equalities implied by the theory.

2.1 Setting

A true data-generating process φ_0 produces the monthly macro panel; the realized *main dataset* is $\mathcal{H} = \{y_t, X_t\}_{t=1}^{T}$, with $\mathcal{H}_{\text{tr}} \subset \mathcal{H}$ a fixed training window. *Outside information* $\mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3)$ denotes data that are neither contained in nor deterministic functions of $\mathcal{H}$: in the application, daily financial series, a cross-country inflation panel, and pre-sample historical series (Section 3).

A pool of M base forecasting models produces the forecast vector $f_{t,h} \in \mathbb{R}^M$ for the target y_{t+h} at each origin t of a recursive pseudo-out-of-sample protocol. The pool uses $\mathcal{H}$ only and is held fixed across all experimental arms, so every information effect in the experiment flows through the combination step — this exclusion restriction is what makes the arm comparisons interpretable. The *forecasting state* $s_t \in \mathcal{S}$ collects real-time-observable features: rolling relative accuracy, signed errors, forecast dispersion, regime proxies, and — in information-augmented arms — $\mathcal{Z}_1$-derived daily-frequency measures. A *combination rule* is a measurable map $w : \mathcal{S} \to \Delta^{M-1}$; the combined forecast is $w(s_t)' f_{t,h}$. Risk under a DGP φ is $R_\varphi(w) = \mathbb{E}_\varphi\big[(y_{t+h} - w(s_t)' f_{t,h})^2\big]$, and for a distribution π over DGPs the Bayes risk is $R_\pi(w) = \mathbb{E}_{\varphi\sim\pi} R_\varphi(w)$.

2.2 Synthetic training as prior fitting

A generator ensemble $\{G_1, \ldots, G_K\}$ with estimated parameters $\hat{\varphi}_k = g_k(\mathcal{H}_{\text{tr}}, \mathcal{Z})$ and mixture weights ρ induces a *design prior* π over DGPs. Replicate histories $\mathcal{H}^{(b)} \sim \pi$ are simulated at monthly frequency, jointly with the $\mathcal{Z}_1$-type state features where the generator models them. Re-enacting the full pseudo-out-of-sample exercise — pool estimation, forecasting, and state construction, by the same code path as on real data — yields training tuples $\big(s_t^{(b)}, f_t^{(b)}, y_{t+h}^{(b)}, \mu_{t+h|t}^{(b)}\big)$, where $\mu_{t+h|t}$ is the DGP's true conditional mean, observable only in simulation. The prior-fitted forecast combiner (PFFC) is a network $w_\theta : \mathcal{S} \to \Delta^{M-1}$ trained on m such tuples; θ^* denotes the population optimum under π and $\hat{\theta}$ its Monte-Carlo estimate. The network output is gated,

$$w_\theta(s) \;=\; \lambda_\theta(s)\,\operatorname{softmax}\big(g_\theta(s)\big) \;+\; \big(1-\lambda_\theta(s)\big)\,\tfrac{1}{M}\mathbf{1}, \qquad \lambda_\theta(s) \in [0,1], \tag{1}$$

so equal weights (EW) are contained in the rule class at $\lambda_\theta \equiv 0$.

The object the training procedure approximates is the state-conditional Bayes rule:

Definition 1 (State-conditional Bayes weights)**.**

$$w^*(s) = \arg\min_{w \in \Delta^{M-1}} \mathbb{E}_\pi\big[(y_{t+h} - w' f_{t,h})^2 \,\big|\, s_t = s\big]. \tag{2}$$

Lemma 1 (L^2 characterization)**.** *Let $A(s) = \mathbb{E}_\pi[f_{t,h} f_{t,h}' \mid s]$ be positive definite and $c(s) = \mathbb{E}_\pi[f_{t,h}\, y_{t+h} \mid s]$. The unconstrained minimizer of Definition 1 is $A(s)^{-1}c(s)$, and $w^*(s)$ is its projection onto Δ^{M-1} under the $A(s)$-norm; it exists and is unique.*

Two features of this construction matter for everything that follows. First, w^* depends on the primitives only through the conditional moments $A(\cdot)$ and $c(\cdot)$ — a fact Remark 2 exploits. Second, the rule is a function of the *state*, not of calendar time or of the identity of the target series; this is what makes zero-shot deployment across inflation measures meaningful (Hypothesis 5).

2.3 The information dichotomy

Write $\mathcal{D}$ for the full synthetic corpus (all simulated tuples) and U for the simulation randomness — the stream of seeds, which is independent of $(\varphi_0, \mathcal{H}, \mathcal{Z})$ by construction.

Proposition 1 (Information neutrality of own-data synthesis)**.** *Suppose $\mathcal{Z} = \emptyset$, so $\hat{\varphi} = g(\mathcal{H}_{\text{tr}})$ and $\mathcal{D} = G(\hat{\varphi}, U)$. Then:*

(a) No additional information. *For any random element ψ — in particular φ_0 itself, or any post-training realization of the data — $I(\mathcal{D}; \psi \,|\, \mathcal{H}_{\text{tr}}) = 0$. Consequently the trained rule $w_{\hat{\theta}}$ is a randomized statistic of $\mathcal{H}_{\text{tr}}$. Any performance difference relative to another $\mathcal{H}_{\text{tr}}$-measurable procedure reflects the chosen functional class, regularization, optimization, or averaging over U, not information about ψ beyond $\mathcal{H}_{\text{tr}}$.*

(b) Bootstrap-family membership. *The class of combination rules reachable by "estimate $\hat{\varphi}$ on $\mathcal{H}_{\text{tr}}$, simulate, fit a functional" is a subclass of randomized functionals of $\mathcal{H}_{\text{tr}}$. It is therefore bootstrap-like in its information content, although its finite-sample behavior need not equal that of a particular classical bootstrap estimator. For block-bootstrap generators the identity is literal.*

(c) Rare-state honesty. *Regime-conditional oversampling manufactures rare-state* samples*, not rare-state* information*: the simulated tail dynamics are the parametric extrapolation of the same few realized episodes in $\mathcal{H}_{\text{tr}}$, and the identification of $w^*(s)$ on rare states remains bounded by what $\mathcal{H}_{\text{tr}}$ contains.*

Proposition 1 is an information statement, not an estimator- equivalence theorem. An own-data arm can outperform or underperform equal weights and classical own-data methods because those procedures impose different restrictions and incur different estimation error. The control arm is useful because it measures what simulation training does

without an outside channel; it does not turn non-rejection against a selected classical benchmark into a test of the mutual-information result.

Proposition 2 (The information channel)**.** *Let episode geometry (amplitude, duration) in unit c be governed by $\theta_c = \bar{\theta} + \eta_c$ with $\eta_c \sim N(0, \tau^2)$ i.i.d. across units, and let each recorded episode deliver an observation $x_{c,i} = \theta_c + \varepsilon_{c,i}$, $\varepsilon_{c,i} \sim N(0, \sigma^2)$. The target unit contributes n_0 episodes; the outside panel $\mathcal{Z}_2$ contributes C additional units with n_c episodes each, $n_e = \sum_{c=1}^{C} n_c$. Then, with variance components known:*

(a) Posterior contraction in the pooled episode count. *The posterior variance of the target-unit geometry parameter θ_0 satisfies*

$$V(C) \;=\; \left[\, \tfrac{n_0}{\sigma^2} \;+\; \tfrac{1}{\tau^2 + V_{\bar{\theta}}(C)} \,\right]^{-1}, \qquad V_{\bar{\theta}}(C) \;=\; \left[\, \sum_{c=1}^{C} \left(\tau^2 + \sigma^2/n_c\right)^{-1} \right]^{-1}, \tag{3}$$

which is strictly decreasing in C, strictly below the own-data-only value σ^2/n_0 for any $C \geq 1$, and converges as $C \to \infty$ to the oracle-prior limit $\left[n_0/\sigma^2 + 1/\tau^2\right]^{-1}$. The gain is bounded by the heterogeneity floor $1/\tau^2$: pooling helps exactly to the extent that units are exchangeable.

(b) Risk transmission under regularity conditions. *If the map from the geometry block to the Bayes weights is Lipschitz on the rare-state set and the transfer bias is controlled, the corresponding component of an excess-risk bound is bounded by a constant times $V(C)$ plus squared bias. Importing $\mathcal{Z}_2$ can therefore tighten this component; posterior contraction alone does not guarantee a realized forecast gain.*

(c) Synthetic data as the carrier. *The prior-fitting mechanism requires only the ability to sample from π; hence arbitrarily heterogeneous outside information — daily frequencies, foreign panels, archival series that no combination regression could ingest directly — is compressed into $\hat{\varphi}$ and delivered to the combiner as simulated tuples on a common monthly frame.*

The empirical counterpart of the exchangeability caveat in (a) is the cross-country severity diagnostic of Section 4.2, which is evaluated by calendar rolling origin so that other countries' future episodes cannot enter the prior for a given target episode.

Three remarks complete the theory. The first two concern what the corpus teaches the combiner; the third locates the equal-weights puzzle inside the rule class.

Remark 1 (Rao–Blackwellization of training targets). In simulation the DGP's conditional mean $\mu_{t+h|t}$ is observable. For any rule class, the population realized-y and μ-target objectives differ by the constant $\mathbb{E}[\mathrm{Var}(y_{t+h} \mid \mathcal{F}_t, \varphi)]$ and therefore have the same population minimizers. Replacing y by μ removes conditional outcome noise and can reduce the sampling variance of empirical losses and gradients. This variance-reduction statement requires a common target definition across generators; a mixture that uses μ for some generators and y for others is not covered by the comparison.

Remark 2 (Geometry versus timing). Decompose the target as $y_{t+h} = \mu_0(s_t) + D_{t,h}\, A_{t,h} + e_{t+h}$, where $D_{t,h} = \mathbf{1}\{\text{an episode is active in } (t, t+h]\}$ and $A_{t,h}$ is the episode's contribution, with geometry block θ_g governing the law of A and timing block θ_τ governing $p(s) = \mathbb{P}(D_{t,h} = 1 \mid s_t = s)$. Then

$$\mathbb{E}[y_{t+h} \mid s] \;=\; \mu_0(s) \;+\; p(s)\, a(s; \theta_g), \qquad a(s; \theta_g) = \mathbb{E}[A_{t,h} \mid s, D_{t,h} = 1]. \tag{4}$$

Information about θ_g moves the conditional *mean* through its product with $p(s)$. A sufficient condition for no change in combination weights is stronger than "unpredictable timing": suppose (i) $p(s) = \bar{p}$, (ii) $a(s; \theta_g) = a(\theta_g)$ is also state-invariant, and (iii) the target and every pool forecast are translation-equivariant, so a change in geometry knowledge produces the same constant shift in all of them. Sum-to-one combined errors are then invariant and the Bayes weights do not change. Geometry may still alter predictive variances and tails. If severity is state-dependent, pool forecasts are not translation-equivariant, or onset probabilities vary with observable state, severity information can

affect a point forecast. The severity–onset distinction is therefore a stylized transmission mechanism, not a general impossibility theorem.

Remark 3 (Insurance and the nested puzzle). Because EW lies in the rule class (1) at $\lambda_\theta \equiv 0$, a globally optimized population rule cannot have greater design-prior risk than EW. Strict improvement is possible when the Bayes weights differ from EW on a set of positive probability and the network represents that difference. Under exchangeability of pool members' joint performance, symmetry gives $w^* = \frac{1}{M}\mathbf{1}$. The gate itself is not identified by this result: equal scores in the softmax reproduce equal weights for any value of λ_θ. Gate paths are therefore descriptive diagnostics, not direct tests of information neutrality.

2.4 Testable implications

The framework generates five empirical questions. Their current operationalizations and the amendments made during analysis are documented in Sections 4, 5 and 5.1.2; they should not be read as an untouched preregistration.

H 1 (Own-data information benchmark). The own-data PFFC adds no information conditional on the training sample. Its comparison with equal weights and classical own-data weighting quantifies finite-sample regularization and estimator efficiency; neither outperformance nor non-rejection by itself proves or falsifies the conditional mutual-information statement.

H 2 (Gains require a validated channel). An information-augmented arm has an interpretable outside-information claim only if its channel shows incremental evidence in the diagnostic matched to that claim. Passing a diagnostic is not sufficient for a point-forecast gain: the information must also change decision-relevant conditional moments and survive transfer through the generator and combiner. The primary empirical contrast is each information arm against A0.

H 3 (Channel decomposition). Outside information can enter generator-side (the training distribution embodies $\mathcal{Z}$-informed dynamics) or state-side (s_t includes $\mathcal{Z}_1$ features at deployment). The two channels are separately ablatable; their contributions are identified by the with/without-state-features grid of Section 4.3.

H 4 (Gate behavior). The learned gate and the effective distance of the final weights from EW describe how strongly the network uses its adaptive component. Because the gate parameter is not separately identified when softmax scores are equal, no particular gate level is required by Remark 3.

H 5 (Zero-shot cross-measure transfer). Because w_θ conditions only on track-record and state features — never on the identity of the target series — a trained combiner can be deployed unchanged on another inflation measure's pool. Whether its relative performance is stable across measures is an empirical transportability question, not a consequence of target-agnostic inputs.

3 Data

The experiment holds one object fixed — the base forecasting pool, which sees only the main U.S. panel — and varies a second: the information a generator may draw on. The identifying logic of Section 2 therefore requires a clean separation between the *main dataset* $\mathcal{H}$, on which every pool member and every generator's own-data component is estimated, and the three *outside information sets* $\mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3)$, which by construction are neither contained in nor deterministic functions of $\mathcal{H}$. This section describes each and states the principle we follow when a series does not cover the full sample: availability is data, never to be imputed away.

3.1 The main panel and six inflation measures

The main dataset is a monthly U.S. macroeconomic panel spanning 1960M1 through 2025M12, built from the May 2026 fixed revised vintage of FRED-MD (McCracken and Ng, 2016) with the standard transformation codes. The eight-variable system on which the generators and the base pool operate — CPI inflation, industrial production, unemployment, the federal funds rate, the ten-year Treasury yield, producer prices, M2, and the Baa–Aaa credit spread — is fully observed from 1960M3 (the first month at which every differenced transform is available) and is frozen at 1960M3–1984M12 for the own-data generator components. The reported final exercise uses 2005M1–2024M12 forecast origins and is pseudo-out-of-sample with respect to those components, but it is not a real-time-vintage evaluation. G-Z2 applies a separate 2004M12 cutoff to foreign information and a 1984M12 cutoff to its U.S. local component.

We evaluate six inflation measures rather than one, both to test the zero-shot transfer of Hypothesis 5 and to guard against a measure-specific artifact driving any conclusion: headline and core CPI, headline and core PCE, the Dallas Fed trimmed-mean PCE, and the Cleveland Fed median CPI. The first four are observed over the full window; the trimmed-mean and median measures begin in 1977M2 and 1983M1 respectively. All six have sufficient pre-evaluation history and are evaluated on the same 240 origins from 2005M1 through 2024M12. For each measure we construct the one-, three-, six-, and twelve-month cumulative targets that define the four forecast horizons.

A methodological point that recurs below originates here. When a series does not reach the sample boundary, naïve interpolation silently fabricates observations at the tails; we instead restore every value outside a series' true first- and last-observed dates to missing, and verify the rebuilt panel against the frozen archive to floating-point tolerance. The same no-fabrication discipline governs the outside sets.

3.2 Daily financial data ($\mathcal{Z}_1$)

The first outside set is high-frequency financial information — the canonical example when one asks whether synthetic data could ever help a monthly forecaster (Andreou et al., 2013; Ghysels et al., 2006): daily Treasury yields across maturities (from 1962), Moody's Aaa and Baa corporate yields (from 1983 and 1986), the CBOE volatility index (from 1990), WTI crude (from 1986), the broad dollar index (from 2006), and TIPS-implied inflation breakevens (from 2003) (Gürkaynak et al., 2010). From the daily series we compute, for each month, thirty-two aggregate features — realized volatilities, within-month changes in the yield-curve slope and credit spread, drawdowns, and jump counts — that are candidates both for the generator's regime-transition dynamics and for the deployment-time state vector (the two channels of Hypothesis 3). Fourteen of these are available from 1962, covering the generator estimation window; the remainder inherit their series' start dates. Every feature carries an explicit availability mask, and the combiner receives the mask alongside the value, so that the absence of, say, a market-implied breakeven before 2003 is information the network conditions on rather than a zero it misreads as a signal.

3.3 The cross-country episode library ($\mathcal{Z}_2$)

The second outside set operationalizes the pooled-episode-count mechanism of Proposition 2. From the World Bank global inflation database (Ha et al., 2023) we retain 111 economies with sufficient monthly CPI history, and detect surge and disinflation episodes in each country's year-on-year inflation series by a zigzag turning-point rule (minimum swing three percentage points, minimum duration six months). This yields 1,529 episodes; restricting to the eighteen advanced economies that anchor the hierarchical generator gives 122 surge episodes. Against this the United States contributes only a handful of modern surge episodes — the gap between n_e^{US} and n_e that Proposition 2(a) says the pooling must exploit. For each episode we record its amplitude, duration, and normalized shape, and

we compute a Ciccarelli–Mojon-style global inflation factor (Ciccarelli and Mojon, 2010) as the first principal component of the standardized cross-country panel.

The generator library is rebuilt after imposing the information cutoff: foreign observations end in 2004M12 and U.S. local episode statistics end in 1984M12. Episodes beginning in 2005–2024 are physically separated and used only for the retrospective calendar-rolling severity diagnostic; they do not enter generator estimation or select arm weights.

3.4 Long-run history ($\mathcal{Z}_3$)

The third outside set reaches back before the main sample. We use the U.S. consumer price index from 1913 to 1959 at monthly frequency — 552 observations of a regime distribution (mean 2.4%, standard deviation 6.6% in year-on-year terms) that the post-war estimation window never sees — and the eighteen-economy annual macro-financial panel of Jordà et al. (2017) from 1870. Because that panel contains genuine currency collapses (Weimar Germany foremost) whose moments would otherwise dominate any pooled statistic, we report both the raw and a hyperinflation-trimmed version of the pre-1960 regime-frequency and tail statistics, and use the trimmed version to calibrate the long-history generator's priors. These statistics enter only as priors on regime frequency, persistence, and tail thickness; unlike $\mathcal{Z}_1$ and $\mathcal{Z}_2$, $\mathcal{Z}_3$ has no real-time channel and so no incremental-predictive-content test applies to it (Section 4.2).

4 Experimental design

The revision separates design information from evaluation. Generator parameters and U.S. local information are frozen at 1984M12; foreign information used by G-Z2 ends at 2004M12. The 1985M1–2004M12 interval is used for design and tuning, and the untouched forecast evaluation contains the 240 origins from 2005M1 through 2024M12. All calculations use a fixed revised vintage, so the exercise is calendar-cutoff clean but is not

Table 1: Information sets and calendar chronology

Information set	Frozen generator cutoff	Role after cutoff	Chronology status
Main U.S. panel	1984M12 for own-data generators	1985–2004 design; 2005–2024 forecast evaluation	Calendar-cutoff clean; fixed revised vintage
Z1 daily finance	1984M12 for generator estimation	Origin-t state values use calendar data through month t	Calendar-clean; historical retrieval vintages unavailable
Z2 foreign CPI	2004M12	Later episodes excluded from G-Z2 and used only in the severity holdout	Cutoff-clean reconstruction from a later fixed vintage
Z2 U.S. local component	1984M12	No evaluation-period U.S. episode enters generator estimation	Observation dates end before design and evaluation
Z2 severity holdout	2005M1–2024M12 excluded from generator	Retrospective calendar-rolling diagnostic only	Physically separated; never selects arm mass
Z3 long history	1959M12	Prior on regime frequency, persistence, and tails	All observation dates predate the main panel

Notes: The study follows Route B: a retrospective calendar-cutoff-clean exercise under fixed revised vintages, not a real-time-vintage exercise. Origin t is defined after observation-month t macro releases.

a real-time-vintage study. The full information-set audit is reported in Table 1.

4.1 Generators, arms, and replicate corpus

Five generators estimated only on the U.S. training sample form A0: a regime-switching VAR, a Minnesota-prior BVAR, a dynamic factor model, a stationary block bootstrap (Politis and Romano, 1994), and a copula VAR. A1 adds the daily-financial generator G-Z1; A2 adds the cross-country episode generator G-Z2; A3 adds the pre-1960 generator G-Z3; and A_{full} contains all eight. Generator weights are fixed before final evaluation and normalized within each arm. The diagnostics below neither choose those weights nor read final forecast losses.

For each generator we simulate $B = 200$ histories of 792 months. The same five-model pool used on real data is re-estimated recursively on every history and produces four-horizon forecasts at 480 origins. Replicates 0–159 train the combiner, 160–179 select training checkpoints, and 180–199 form an untouched synthetic test. The half-corpus stability gate that triggered the extension from 100 to 200 replicates is recorded in the

Table 2: Outside-information diagnostics: onset and conditional severity

Channel	Estimand	Metric	Estimate	95% CI	p_+
Z1 daily finance	Onset	Delta AUC	-0.021	[-0.043, 0.007]	0.893
Z1 daily finance	Onset	Brier improvement	-0.002	[-0.012, 0.012]	0.478
Z2 cross-country	Severity: amplitude	Delta log score	0.045	[0.017, 0.100]	0.000
Z2 cross-country	Severity: amplitude	CRPS improvement	0.026	[0.013, 0.046]	0.000
Z2 cross-country	Severity: duration	Delta log score	0.050	[-0.075, 0.138]	0.249
Z2 cross-country	Severity: duration	CRPS improvement	0.019	[-0.022, 0.044]	0.213

Notes: Positive values favor the augmented-information specification. Onset coefficients are frozen before 2005; severity uses calendar rolling origins. p_+ is one-sided.

replication registry.

4.2 Incremental-information diagnostics

The diagnostics distinguish two estimands. The onset diagnostic asks whether daily-financial variables improve the probability that a U.S. inflation episode begins within twelve months. Its coefficients are tuned on the design period with a twelve-month label embargo and frozen before 2005. The severity diagnostic asks whether cross-country histories improve the predictive distribution of log amplitude and log duration conditional on an episode having begun. It is evaluated by calendar rolling origin: only episodes ending before a target episode starts enter its predictive distribution. These are different prediction problems, not alternative measures of one latent quantity.

4.3 Combiner training and common-calendar evaluation

The prior-fitted forecast combiner is a permutation-equivariant attention network over the five pool-member tokens with 40,370 parameters and a gated head that nests equal weights. The primary specification uses realized outcomes for every generator. It is trained independently for five arms, four horizons, and ten deterministic seeds, giving 200 primary fits. Common-μ, no-gate, no-state, high-dose, and matched-placebo variants are restricted to the pre-specified $h = 12$ robustness grid.

On real data, all neural and classical methods use the same five-model pool and ex-

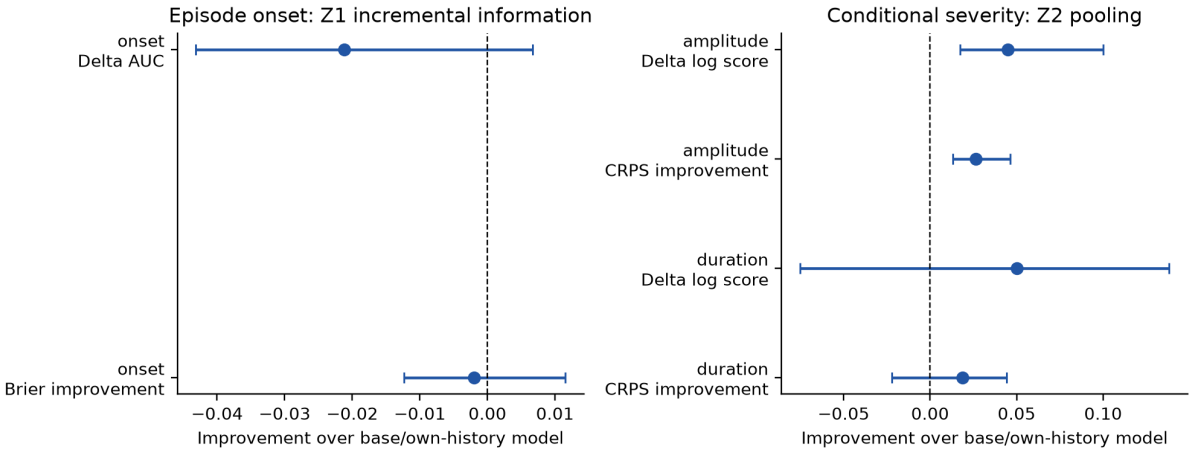

Figure 1: Episode onset and conditional severity are distinct information targets. Points are incremental predictive-score estimates and bars are 95% calendar-bootstrap intervals. Daily-financial variables do not improve the locked onset prediction, while cross-country pooling improves amplitude density prediction; duration estimates are less precise.

actly the same 2005M1–2024M12 origin calendar for six inflation measures. The classical battery contains equal weights, robust location rules, inverse-loss rules, selection, BMA, DMA, complete-subset regression, and rolling unconstrained, constrained, and ridge regressions. Every adaptive benchmark passes a future-target perturbation test: at origin t and horizon h, only forecast errors from origins i satisfying $i+h \leq t$ may enter its weights.

The primary effects are the four outside arms relative to A0 at each horizon. Squared losses are normalized by the equal-weight MSE within each measure and the six measures receive equal mass. A stationary calendar bootstrap applies the same date blocks jointly to every measure, horizon, and arm. The primary expected block length is 24 months, with 12 and 36 months as sensitivity checks. We use 4,999 draws and max-t adjustment over all sixteen arm–horizon contrasts. A 90% simultaneous interval inside $[0.98, 1.02]$ establishes equivalence at the two-percent RMSFE margin; a one-percent margin is reported as sensitivity.

5 Empirical Results

5.1 Implementation and analysis record

5.1.1 Computation and reproducibility

The final corpus contains 200 replicates for each of ten generators. The primary PFFC grid contains 200 completed fits; the targeted $h = 12$ grid contains 180 fits. Frozen deployment files store ten seed forecasts, their weights and gate values, and an equal-seed forecast ensemble at every origin. The real pool, classical benchmark battery, loss tensor, bootstrap inference, and result ledger each have a machine-readable registry. The final six-measure calendar is verified independently at every stage.

5.1.2 Analysis record

The revision responds to five issues found in the exploratory analysis. First, onset labels are now embargoed and final coefficients are frozen. Second, conditional information neutrality is no longer tested as cell-by-cell performance equality with equal weights. Third, information arms are compared directly with A0 using simultaneous inference rather than by selecting the best arm ex post. Fourth, duplicated country series in the source workbook are detected automatically; the U.S. series is replaced by the independently retrieved CPI series and the two corrupted countries without an independent replacement are excluded. Fifth, G-Z2 is reconstructed with foreign observations ending in 2004 and U.S. local statistics ending in 1984. Diagnostics are evidence about information content only and cannot change arm mass or the final claim rule.

The claim rule was fixed before R6 inference. A positive severity result combined with no onset or point-forecast gain supports the severity–onset interpretation. If all primary cells satisfy the two-percent equivalence margin, an economically bounded null may be stated; otherwise the wording is limited to failure to detect a point-forecast gain. The

Table 3: Primary outside-information arms relative to the own-data control

Arm	h	Rel. RMSFE	90% simultaneous CI	Loss benefit	FWER p_+
A1	1	1.006	[0.997, 1.015]	-0.011	1.000
A1	3	1.002	[0.988, 1.016]	-0.004	1.000
A1	6	0.996	[0.978, 1.015]	+0.007	0.942
A1	12	1.001	[0.956, 1.048]	-0.001	1.000
A2	1	1.003	[0.995, 1.011]	-0.006	1.000
A2	3	0.998	[0.995, 1.000]	+0.004	0.151
A2	6	1.000	[0.997, 1.004]	-0.001	1.000
A2	12	0.996	[0.984, 1.009]	+0.007	0.891
A3	1	1.000	[0.998, 1.002]	-0.001	1.000
A3	3	0.999	[0.995, 1.003]	+0.002	0.917
A3	6	0.999	[0.994, 1.003]	+0.003	0.908
A3	12	0.998	[0.992, 1.003]	+0.005	0.769
A_full	1	1.005	[0.997, 1.012]	-0.009	1.000
A_full	3	1.004	[0.991, 1.019]	-0.009	1.000
A_full	6	1.006	[0.985, 1.029]	-0.013	1.000
A_full	12	1.015	[0.977, 1.054]	-0.030	1.000

Notes: Six measures receive equal mass. Positive loss benefit favors the outside arm. Confidence intervals and one-sided p-values use 4,999 common-date stationary-bootstrap draws and max-t adjustment across 16 arm–horizon contrasts.

latter branch applies here because four of sixteen simultaneous intervals extend beyond the two-percent band.

5.2 Outside information: severity versus onset

Table 2 and Figure 1 show the central diagnostic separation. Adding the daily-financial block lowers onset AUC by -0.021 (95% interval $[-0.043, 0.007]$) and does not improve the Brier score. In contrast, cross-country pooling improves the calendar-rolling log score for episode amplitude by 0.045 ($[0.017, 0.100]$), with a positive CRPS improvement as well. Duration estimates are positive but their calendar-block intervals include zero. Thus the evidence supports a predictive channel for amplitude conditional on episode onset, not a broad claim that every aspect of episode geometry is predictable.

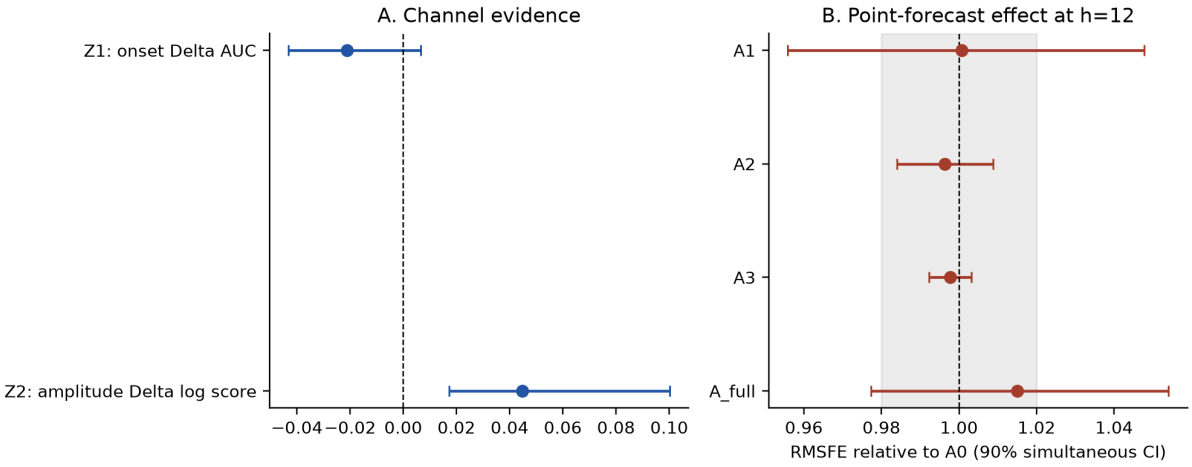

Figure 2: Channel evidence and realized point-forecast effects. Panel A reports the locked incremental onset and amplitude-severity diagnostics. Panel B reports $h = 12$ RMSFE ratios relative to A0 with 90% simultaneous intervals; shading marks the $\pm 2\%$ equivalence band. The panels use different estimands and scales and are not a regression of gain on a diagnostic score.

5.3 Primary forecast contrasts

Measure-specific simultaneous intervals for all 96 arm–horizon–measure contrasts are reported as Tables S1–S4 in the Supporting Information.

No outside arm is superior to A0 after familywise adjustment (0 of 16 contrasts). Point ratios are close to one, but precision differs across cells. The 90% simultaneous interval lies entirely inside the $\pm 2\%$ RMSFE band in 12 of sixteen cells. A1 at $h = 6, 12$ and $\mathrm{A}_{\mathrm{full}}$ at $h = 6, 12$ remain inconclusive: their intervals are too wide to establish either superiority or two-percent equivalence. The same classification holds for 12 cells with a 36-month expected block and for 11 cells with a 12-month block. We therefore report failure to detect a point-forecast gain, not universal practical equivalence.

The seed distribution reinforces that caution. At $h = 12$, respectively 7, 6, 6, and 5 of ten seeds favor A1, A2, A3, and $\mathrm{A}_{\mathrm{full}}$ over their matched A0 seed. A result that depends on selecting one seed would consequently be unstable; the table and inference use the pre-specified equal-seed forecast ensemble.

Table 4: Common-calendar benchmark comparison

Family	Method	$h = 1$	$h = 3$	$h = 6$	$h = 12$
classical	BG60	0.999	0.995	0.992	0.989
classical	CSR3	1.001	0.979	0.976	0.976
classical	DMSFE	0.998	0.994	0.991	0.988
classical	EW	1.000	1.000	1.000	1.000
classical	WinsorMean	0.995	0.997	0.993	0.990
primary	A0	0.998	0.995	0.996	1.004
primary	A1	1.003	0.996	0.992	1.005
primary	A2	1.001	0.992	0.996	1.000
primary	A3	0.998	0.993	0.995	1.002
primary	A_full	1.002	0.999	1.002	1.019
simulator_free	simulator_free	1.031	1.042	1.024	1.000

Notes: RMSFE is relative to equal weights after normalizing within measure and giving all six measures equal mass. Every entry uses 2005M1–2024M12.

5.4 Classical and simulator-free benchmarks

Several classical combinations modestly improve on equal weights. At $h = 12$, complete-subset regression has relative RMSFE 0.976, discounted MSFE 0.988, rolling Bates–Granger 0.989, and the winsorized mean 0.990. The simulator-free nonlinear combiner is essentially tied with equal weights at $h = 12$ and worse at shorter horizons. The PFFC arms occupy the same broad neighborhood as equal weights and the robust classical rules; the model-confidence set retains a large collection at most horizons. These comparisons show that the outside-arm null is not an artifact of an empty or deliberately weak benchmark battery.

5.5 Targeted robustness and mechanism diagnostics

No targeted perturbation yields a stable reversal of the primary result. Removing A1 state features worsens its $h = 12$ relative RMSFE from 1.005 to 1.025, but the matched no-link placebo is also worse (1.033); this pattern does not identify a beneficial timing channel. The A2 own-history placebo and high-dose variant remain near equal weights. Common-μ, no-gate, and high-dose results vary by arm without producing a coherent

Table 5: Targeted robustness checks at $h = 12$

Category	Specification	Rel. RMSFE vs EW	Ratio vs primary arm
common-mu target	A0_common_mu	1.001	0.997
common-mu target	A1_common_mu	1.040	1.035
common-mu target	A2_common_mu	1.008	1.008
common-mu target	A3_common_mu	0.998	0.996
common-mu target	A_full_common_mu	1.016	0.997
high-dose	A1_high_dose	1.019	1.014
high-dose	A2_high_dose	1.000	1.000
high-dose	A3_high_dose	1.003	1.002
high-dose	A_full_high_dose	1.015	0.996
matched placebo	P_gz1_no_link	1.033	1.028
matched placebo	P_gz2_own	0.995	0.995
no-gate	A0_no_gate	0.994	0.990
no-gate	A1_no_gate	1.014	1.009
no-gate	A2_no_gate	1.014	1.014
no-gate	A3_no_gate	1.009	1.007
no-gate	A_full_no_gate	1.020	1.001
no-state	A1_no_state	1.025	1.020
no-state	A_full_no_state	1.036	1.017
Seed stability: number of seeds (out of 10) beating matched A0 at $h = 12$			
seed distribution	A1	7/10	[0.972, 1.012]
seed distribution	A2	6/10	[0.971, 1.029]
seed distribution	A3	6/10	[0.968, 1.045]
seed distribution	A_full	5/10	[0.983, 1.032]

outside- information ordering.

The learned weights help explain why a large gate value is not itself an information result. Across ensemble deployments, normalized entropy is generally high and effective distance from equal weights is small for A0, A2, and A3. A1 and A_{full} move farther from equal weights, yet do not earn a forecast gain. Gate level, distance, entropy, and turnover are therefore reported separately in the ledger.

The 2020–2024 calculations are descriptive because they contain only 60 origins. They do not determine the superiority conclusion. Cell-level DM–HLN statistics, Holm-adjusted across 96 information-arm cells, and the four-horizon MCS survivor lists are documented in the replication ledger and the Supporting Information file map.

5.6 From predictive information to forecast value

The results establish two empirically distinct facts. First, cross-country histories improve locked, calendar-rolling density prediction of inflation episode amplitude. Second, im-

porting outside channels into the synthetic training distribution does not improve point forecasts relative to the matched own-information arm. Their conjunction is the main empirical contribution: predictive content in an outside channel and its incremental value for a particular forecast decision are separate objects that can be measured in a common design.

The severity–onset decomposition explains how the two findings fit together. For outside information to change unconditional point-combination risk, it must affect the forecast-origin probability of entering the relevant episode or alter which pool member is expected to perform best. Information about amplitude conditional on an episode occurring need not supply either signal. The observed configuration—a positive amplitude diagnostic, limited incremental onset evidence, and no familywise-supported gain over A0—is the empirical counterpart of the transmission logic in Remark 2. It locates the relevant link between information content and target loss rather than treating forecast accuracy as a generic test of whether outside data are informative.

The matched own-information arm is essential to that interpretation. A0 holds the architecture, synthetic-training protocol, forecast pool, and seed ensemble fixed while removing only the outside channel. The comparison therefore separates information imported by the simulator from regularization and inductive bias created by simulation training itself. The classical and simulator-free benchmarks provide an additional check: improvements from own-data rules are visible when they occur, so the direct outside-arm result does not depend on using equal weights as the sole or deliberately weak reference. Joint inference and the pre-specified equivalence margin then distinguish lack of superiority from economically bounded similarity.

For applied work, this framework yields a portable evaluation protocol. Researchers should name the information carried by a simulator, freeze its calendar cutoff before evaluation, validate the proposed channel with an estimand aligned to that information, and compare the resulting decision rule with a matched own-information control. This

sequence turns simulation-based forecasting from an undifferentiated model comparison into a test of where predictive information enters and whether it reaches the target loss.

5.6.1 Further research

The decomposition also makes the next empirical steps specific. Because severity information is naturally aligned with density, tail, and episode- conditional objectives, future work can evaluate those losses using proper scores and quantile or expected-shortfall criteria. This direction is consistent with evidence that outside commodity information can improve inflation-risk densities and that mean-optimal and variance-optimal combination weights need not coincide (Garratt and Petrella, 2022; Knüppel and Krüger, 2022). Richer embargoed onset signals, including market expectations and text-based news measures, can test whether a stronger timing channel completes the transmission from severity to point accuracy. Real-time-vintage reconstruction can replace the fixed-vintage pseudo-out-of-sample design, while broader forecast pools and pre-specified multi-country evaluation can increase cross-model variation and long-horizon precision. Finally, an explicit simulator-to-real transport audit—with held-out channel calibration and matched transport placebos—can distinguish attenuation in the generator from attenuation in the forecast combiner.

6 Conclusion

We provide a disciplined empirical framework for determining whether simulation-trained forecast combiners benefit from outside information. Synthetic histories generated only from the forecaster's sample add no information conditional on that sample, although they may change finite-sample performance through regularization. Outside data can add information, but a forecast gain requires that information to be relevant to the target loss at the forecast origin.

Applied to U.S. inflation, cross-country pooling improves the predictive density for episode amplitude, while the tested financial variables add limited incremental information about episode onset. Across four outside arms and four horizons, none outperforms the own-data control after familywise adjustment. Twelve contrasts are equivalent within a two-percent RMSFE margin; four long-horizon contrasts remain too imprecise to establish either superiority or equivalence. The appropriate conclusion is therefore that we fail to detect a point-forecast gain, not that all outside information is economically irrelevant.

The severity–onset distinction reconciles these findings without turning them into an impossibility theorem. A channel can predict how severe an episode will be once it occurs yet fail to improve unconditional point combination when it gives little guidance about when that episode will start or which model will forecast it best. The broader lesson is to treat simulation-based training as an information-delivery mechanism: identify the imported information, validate its channel, and test whether it matters for the final decision loss on an untouched common sample.

References

Andreou, E., Ghysels, E., and Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, 31(2):240–251. https://doi.org/10.1080/07350015.2013.767199.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, Y. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=gerNCVqqtR.

Atkeson, A. and Ohanian, L. E. (2001). Are Phillips curves useful for forecasting inflation?

Federal Reserve Bank of Minneapolis Quarterly Review, 25(1):2–11. https://doi.org/10.21034/qr.2511.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20(4):451–468. https://doi.org/10.1057/jors.1969.103.

Carriero, A., Pettenuzzo, D., and Shekhar, S. (2026). MACROCAST: A vintage-consistent time series foundation model for real-time macroeconomic forecasting. Working paper, arXiv. https://doi.org/10.48550/arXiv.2606.28670.

Ciccarelli, M. and Mojon, B. (2010). Global inflation. *Review of Economics and Statistics*, 92(3):524–535. https://doi.org/10.1162/REST_a_00008.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762. https://doi.org/10.1016/j.ijforecast.2015.12.005.

Diebold, F. X. and Pauly, P. (1987). Structural change and the combination of forecasts. *Journal of Forecasting*, 6(1):21–40. https://doi.org/10.1002/for.3980060103.

Dooley, S., Khurana, G. S., Mohapatra, C., Naidu, S. V., and White, C. (2023). ForecastPFN: Synthetically-trained zero-shot forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2403–2426. https://doi.org/10.52202/075280-0112.

Elliott, G. and Liao, J. (2026). Combining forecasts—on why averaging beats optimal linear weights. *Journal of Business & Economic Statistics*. Advance online publication. https://doi.org/10.1080/07350015.2026.2661980.

Garratt, A. and Petrella, I. (2022). Commodity prices and inflation risk. *Journal of Applied Econometrics*, 37(2):392–414. https://doi.org/10.1002/jae.2868.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1–2):59–95. https://doi.org/10.1016/j.jeconom.2005.01.004.

Granger, C. W. J. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204. https://doi.org/10.1002/for.3980030207.

Gürkaynak, R. S., Sack, B., and Wright, J. H. (2010). The TIPS yield curve and inflation compensation. *American Economic Journal: Macroeconomics*, 2(1):70–92. https://doi.org/10.1257/mac.2.1.70.

Ha, J., Kose, M. A., and Ohnsorge, F. (2023). One-stop source: A global database of inflation. *Journal of International Money and Finance*, 137:102896. https://doi.org/10.1016/j.jimonfin.2023.102896.

Jordà, Ò., Schularick, M., and Taylor, A. M. (2017). Macrofinancial history and the new business cycle facts. *NBER Macroeconomics Annual*, 31:213–263. https://doi.org/10.1086/690241.

Knüppel, M. and Krüger, F. (2022). Forecast uncertainty, disagreement, and the linear pool. *Journal of Applied Econometrics*, 37:23–41. https://doi.org/10.1002/jae.2834.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589. https://doi.org/10.1080/07350015.2015.1086655.

Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., and Hutter, F. (2022). Transformers can do Bayesian inference. In *International Conference on Learning Representations*. https://openreview.net/forum?id=KSugKcbNf9.

Müller, U. K. and Watson, M. W. (2026). Forecasting related time series. *Journal of Applied Econometrics*, 41(4):481–498. https://doi.org/10.1002/jae.70050.

Naghi, A. A., O'Neill, E., and Zaharieva, M. D. (2024). The benefits of forecasting inflation with machine learning: New evidence. *Journal of Applied Econometrics*, 39(7):1321–1331. https://doi.org/10.1002/jae.3088.

Nagler, T. (2023). Statistical foundations of prior-data fitted networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25660–25676. PMLR. https://proceedings.mlr.press/v202/nagler23a.html.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313. https://doi.org/10.1080/01621459.1994.10476870.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355. https://doi.org/10.1111/j.1468-0084.2008.00541.x.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430. https://doi.org/10.1002/for.928.

Taga, E., Ildiz, M. E., and Oymak, S. (2025). TimePFN: Effective multivariate time series forecasting with synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20761–20769. https://doi.org/10.1609/aaai.v39i19.34288.

Timmermann, A. (2006). Forecast combinations. In Elliott, G., Granger, C. W. J., and

Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 135–196. Elsevier. https://doi.org/10.1016/S1574-0706(05)01004-9.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. (2024). Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 53140–53164. PMLR. https://proceedings.mlr.press/v235/woo24a.html.

Funding: This work was supported by the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2025S1A5C3A0201187111).

Conflict of Interest: The author declares no conflict of interest.

Data Availability Statement: The replication package accompanying the submission contains the code, frozen configurations, derived data, hashes, and retrieval instructions needed to reproduce the reported results. Third-party source files that cannot be redistributed are identified by source URL and SHA-256 hash. Upon acceptance, all redistributable data and specialized programs will be deposited in the Journal of Applied Econometrics Data Archive.

A Proofs

A.1 Proof of Lemma 1

The map $w \mapsto \mathbb{E}_\pi[(y_{t+h} - w' f_{t,h})^2 \mid s] = w' A(s) w - 2w' c(s) + \mathbb{E}_\pi[y_{t+h}^2 \mid s]$ is a quadratic with Hessian $2A(s) \succ 0$, hence strictly convex, and Δ^{M-1} is convex and compact; a unique minimizer exists. Completing the square, the objective equals $(w - u(s))' A(s)(w - u(s)) +$ const with $u(s) = A(s)^{-1} c(s)$, so the constrained minimizer is the projection of $u(s)$ onto Δ^{M-1} in the $A(s)$-norm, characterized by the KKT conditions of this projection. □

A.2 Proof of Proposition 1

(a) By construction $\mathcal{D} = G(g(\mathcal{H}_{\text{tr}}), U)$ with $U \perp\!\!\!\perp (\psi, \mathcal{H})$, so $\mathcal{D}$ is $\sigma(\mathcal{H}_{\text{tr}}, U)$-measurable. By the chain rule for conditional mutual information and monotonicity under measurable maps (data-processing inequality),

$$I(\mathcal{D};\, \psi \,|\, \mathcal{H}_{\text{tr}}) \;\le\; I((\mathcal{H}_{\text{tr}}, U);\, \psi \,|\, \mathcal{H}_{\text{tr}}) \;=\; I(U;\, \psi \,|\, \mathcal{H}_{\text{tr}}) \;=\; 0, \tag{5}$$

the last equality because

$$U \perp\!\!\!\perp (\psi, \mathcal{H}_{\text{tr}}). \tag{6}$$

Since mutual information is nonnegative, $I(\mathcal{D};\, \psi \,|\, \mathcal{H}_{\text{tr}}) = 0$.

For the risk statement, note that the fitted rule is $w_{\hat{\theta}} = \mathcal{A}(\mathcal{D}, U')$ for a training algorithm $\mathcal{A}$ and auxiliary randomization U' (initialization, batching), hence $w_{\hat{\theta}} = \tilde{\mathcal{A}}(\mathcal{H}_{\text{tr}}, \tilde{U})$: a *randomized* $\mathcal{H}_{\text{tr}}$-measurable procedure. For any convex loss, Jensen's inequality applied conditionally on $(\mathcal{H}_{\text{tr}}, s_t)$ gives that the derandomized rule $\bar{w}(\cdot) = \mathbb{E}[\, w_{\hat{\theta}}(\cdot) \mid \mathcal{H}_{\text{tr}}\,]$ (a non-randomized $\mathcal{H}_{\text{tr}}$-measurable rule; the expectation preserves the simplex by convexity) satisfies $R_\varphi(\bar{w}) \le \mathbb{E}_{\tilde{U}}\, R_\varphi(w_{\hat{\theta}})$ for every φ. Hence the procedure's risk is bounded below by that of some non-randomized $\mathcal{H}_{\text{tr}}$-measurable rule, and any improvement over a given $\mathcal{H}_{\text{tr}}$-measurable benchmark can originate only in the choice of the functional class and the averaging over $\tilde{U}$ — regularization — never in information about ψ beyond $\mathcal{H}_{\text{tr}}$.

(b) The synthetic-training pipeline of (a) is a randomized measurable functional of $\mathcal{H}_{\text{tr}}$ and hence belongs to the broad information class of generalized bootstrap procedures. This is a membership statement, not a claim that a given generator–trainer architecture can represent every randomized functional of the sample or that its risk equals that of a particular classical bootstrap estimator. For generators that resample blocks of $\mathcal{H}_{\text{tr}}$ the identification is literal: the simulated histories *are* block-bootstrap pseudo-samples.

(c) Conditional on $\hat{\varphi}$, the law of simulated tuples on any rare-state set is a deterministic functional of $\hat{\varphi}$; with $\mathcal{Z} = \emptyset$, the rare-regime components of $\hat{\varphi}$ are estimated from the finitely many realized episodes in $\mathcal{H}_{\text{tr}}$. Oversampling changes the *measure* assigned to rare states in the training corpus, not the conditional law of tuples given the rare state, which remains the parametric extrapolation of those episodes. Identification of w^* on rare states is therefore bounded by the information in $\mathcal{H}_{\text{tr}}$, as claimed. □

A.3 Proof of Proposition 2

(a) Marginalizing the unit effects, the within-unit episode means satisfy $\bar{x}_c \sim N(\bar{\theta},\, \tau^2 + \sigma^2/n_c)$, independent across c, and $(\bar{x}_1, \ldots, \bar{x}_C)$ is sufficient for $\bar{\theta}$ in the outside panel. Under the flat prior on $\bar{\theta}$,

$$\bar{\theta} \mid \mathcal{Z}_2 \;\sim\; N\big(\hat{\bar{\theta}},\, V_{\bar{\theta}}(C)\big), \qquad V_{\bar{\theta}}(C) = \Big[\textstyle\sum_{c=1}^{C} (\tau^2 + \sigma^2/n_c)^{-1}\Big]^{-1}, \tag{7}$$

by standard Gaussian conjugacy. The target-unit parameter satisfies $\theta_0 = \bar{\theta} + \eta_0$ with $\eta_0 \sim N(0, \tau^2)$ independent of $\mathcal{Z}_2$, hence $\theta_0 \mid \mathcal{Z}_2 \sim N(\hat{\bar{\theta}},\, \tau^2 + V_{\bar{\theta}}(C))$. Combining with the own-unit likelihood $\bar{x}_0 \mid \theta_0 \sim N(\theta_0, \sigma^2/n_0)$ by conjugacy again yields the posterior precision $n_0/\sigma^2 + (\tau^2 + V_{\bar{\theta}}(C))^{-1}$, which is Equation (3).

Monotonicity: each additional unit adds the positive quantity $(\tau^2 + \sigma^2/n_{C+1})^{-1}$ to the precision of $\bar{\theta}$, so $V_{\bar{\theta}}$ strictly decreases in C, hence $(\tau^2 + V_{\bar{\theta}})^{-1}$ strictly increases and $V(C)$ strictly decreases. At $C = 0$ the flat prior on $\bar{\theta}$ renders the prior on θ_0 flat, giving $V(0) = \sigma^2/n_0$; for $C \geq 1$, $V(C) < \sigma^2/n_0$ because the added precision term is positive. As $C \to \infty$, $V_{\bar{\theta}}(C) \to 0$ and $V(C) \to [n_0/\sigma^2 + 1/\tau^2]^{-1}$: the posterior attainable if $\bar{\theta}$ were known — the oracle-prior limit. Since the added precision never exceeds $1/\tau^2$, the total gain is bounded by the heterogeneity floor, as claimed. With a balanced panel ($n_c \equiv \bar{m}$), $V_{\bar{\theta}}(C) = (\tau^2 + \sigma^2/\bar{m})/C$, displaying the $1/C$ (equivalently $\bar{m}/n_e$) rate.

(b) Let the transfer bound of the training procedure be

$$R_{\varphi_0}(w_{\theta^*}) - R_{\varphi_0}(w_0^*) \le C_1 \, \mathbb{E}_{\varphi_0} \|w_{\theta^*}(s_t) - w_0^*(s_t)\|^2, \tag{8}$$

where w_0^* is the Bayes rule under the truth. Partition $\mathcal{S}$ into the rare-episode set $\mathcal{S}_r$ and its complement, and suppose the map from the geometry block θ_g to the Bayes weights is Lipschitz on $\mathcal{S}_r$ with constant L (finite by Lemma 1 when $A(\cdot)$ is uniformly positive definite and moments are bounded). Then

$$\mathbb{E}_{\varphi_0}\big[\|w_{\theta^*}(s_t) - w_0^*(s_t)\|^2 \mathbf{1}\{s_t \in \mathcal{S}_r\}\big] \;\le\; L^2 \left(V(C) + b^2\right) \mathbb{P}(s_t \in \mathcal{S}_r), \tag{9}$$

where $V(C)$ is the posterior spread of the geometry block under the design prior and b the (heterogeneity) bias of its posterior mean. The rare-state component of the bound is thus proportional to $V(C)$ up to the squared bias controlled by hierarchical shrinkage — the bias–variance tradeoff at the episode level stated in the text. □

A.4 Argument for Remark 1

Condition on $(\varphi, \mathcal{F}_t)$; then $\mu_{t+h|t} = \mathbb{E}[y_{t+h} \mid \mathcal{F}_t, \varphi]$ and $w(s_t)' f_{t,h}$ is $\mathcal{F}_t$-measurable. Hence

$$\mathbb{E}\big[(y_{t+h} - w' f_{t,h})^2\big] = \mathbb{E}\big[(\mu_{t+h|t} - w' f_{t,h})^2\big] + \mathbb{E}\big[\operatorname{Var}(y_{t+h} \mid \mathcal{F}_t, \varphi)\big], \tag{10}$$

the cross term vanishing by iterated expectations. The second term is free of w (and of θ), so both population objectives have identical minimizers. The realized-y objective carries the additional conditional outcome noise $(y - \mu)$; removing it can reduce the sampling variance of empirical losses and gradients, although the size of that variance reduction depends on the rule and optimization path. □

A.5 Argument for Remark 2

$\mathbb{E}[y_{t+h} \mid s] = \mu_0(s) + \mathbb{P}(D_{t,h} = 1 \mid s)\,\mathbb{E}[A_{t,h} \mid s, D_{t,h} = 1] = \mu_0(s) + p(s)\,a(s;\theta_g)$ by the tower property. Suppose $p(s) = \bar{p}$ and $a(s;\theta_g) = a(\theta_g)$ on the support of s (timing unpredictable; geometry-given-occurrence not state-dependent). A change of geometry knowledge $\theta_g \to \theta_g'$ then shifts the conditional mean by the state-constant amount $\bar{p}\,[a(\theta_g) - a(\theta_g')]$. If pool members are translation-equivariant — adding a constant to the target shifts every member's forecast by that constant, as holds for linear predictors with intercepts under re-estimation — the same constant is added to y and to every entry of $f_{t,h}$; since combination weights satisfy $w'\mathbf{1} = 1$, the combined error $y - w'f_{t,h}$ is invariant to this common shift for *every* rule w. The squared-loss objective, hence A, c, and the Bayes weights, are unchanged. What the geometry block continues to govern is the conditional law of $D\,A$ *around* its mean — predictive variance and tails — establishing that density-relevant effects may remain. If $\partial p(s)/\partial s \neq 0$ on a set of positive measure, the shift $p(s)\Delta a$ can be state-dependent and can propagate to $c(s)$. Whether it changes the simplex projection and improves risk depends on the joint forecast moments; a timing signal is a possible transmission channel, not a sufficient condition for gain. □

A.6 Argument for Remark 3

EW is attained at $\lambda_\theta \equiv 0$ in Equation (1), so a globally optimized population rule in the class weakly improves on it under the design prior. Under exchangeability of pool members' joint performance, symmetry gives $w^* = \frac{1}{M}\mathbf{1}$. More generally, strict improvement is possible whenever the Bayes weights differ from EW on a set of positive probability and that difference is representable by the network; the differing weights need not vary with the state. Finally, λ_θ^* is not identified when the softmax scores are equal because every gate value then produces EW. The observable gate path therefore has no unique population value implied by exchangeability. □