

What Forecasts the US Macroeconomy Out of Sample?

Specification, Shocks and Real-Time Evidence

Nicholas Hong

Nanyang Technological University, Singapore
nhong001@e.ntu.edu.sg

Abstract—This paper asks one question of US quarterly data: what forecasts output growth, inflation, the policy rate and the real exchange rate out of sample once look-ahead leakage, overfitting and multiple comparisons are controlled? The answer is built in three stages. Stage 1 runs a theory-pruned block forward search with hard validity gates. It selects a four-variable core with oil as an exogenous regressor, rejects a household net worth block that lowers raw out-of-sample error but fails parsimony and whiteness, and finds that no estimator, whether OLS, Minnesota Bayesian VAR, ridge, LASSO or elastic-net, is separable from a univariate autoregression by the Model Confidence Set; full-sample Hodrick-Prescott filtering of the policy rate revises the endpoint by 66 percent of the cycle's own standard deviation. Stage 2 asks whether constructed exogenous shocks and Bayesian shrinkage help and whether densities are calibrated. A naively coded Minnesota prior leaves the posterior unshrunk (largest companion eigenvalue 1.068) while dummy observations restore stability (0.96); stacking all shocks inflates out-of-sample error by an order of magnitude; the only significant benchmark gain under full-sample shock standardization disappears in real time; and a stochastic-volatility BVAR evaluated over 48 expanding-window origins produces sharper, better calibrated densities than random-walk and AR(1) benchmarks for output and inflation. Stage 3 takes the shock idea to scale across 1,632 specification cells and 1,014 models. Shocks buy about a 12 percent one-quarter GDP RMSE gain that survives a real-time ALFRED replication almost intact, but only 4.0 percent of specifications beat the core VAR at the 5 percent level, which is what chance delivers; externally identified oil and monetary shocks do not beat cheap growth-rate proxies; and shock-model densities are overconfident until future-shock uncertainty is priced, after which they are no sharper than the pure VAR. Parsimony and leakage-free evaluation dominate estimator sophistication, reduced-form forecasting does not reward structural identification, and the honest product is a set of near-equivalent models rather than a champion.

Index Terms—Vector autoregression, Bayesian VAR, exogenous shocks, out-of-sample evaluation, Model Confidence Set, density forecasting, real-time data, look-ahead bias.

I. INTRODUCTION

Applied macroeconomic forecasting with vector autoregressions (VARs) faces a familiar tension. The information set is large, the sample is short, and the analyst is free to choose variables, transformations, lag orders, estimators, constructed regressors and evaluation windows. Each choice can be made to look good in a backtest. The question here is narrow: on US quarterly data, what actually forecasts real GDP growth, inflation, the policy rate and the real dollar out of sample,

once the three ways a backtest can flatter itself, look-ahead leakage, overfitting and multiple comparisons, are controlled at the same time?

The answer is assembled in three stages that build on one another. Stage 1 settles the specification and estimator question. It runs a block-level forward selection over economically motivated groups of variables, gates every candidate through hard validity filters, and then holds the winning specification fixed while varying the estimator across OLS, Minnesota Bayesian VAR [2], [3] and three regularized variants, summarizing the comparison with the Model Confidence Set (MCS) of Hansen, Lunde and Nason [5]. It also quantifies two construction choices that are common in practice and damaging in a forecasting context: full-sample Hodrick-Prescott (HP) filtering [10] and year-over-year overlapping transforms.

Stage 2 asks whether the next natural step, augmenting a Bayesian VAR with constructed exogenous shocks, buys anything, and extends the evaluation from point forecasts to full predictive densities. It documents a coding trap in the Minnesota prior, shows that stacking all available shocks is catastrophic in a conventional design, measures what look-ahead in standardized shock construction does to apparent skill, and then evaluates a stochastic-volatility BVAR against naive benchmarks with proper scoring rules [9] over an expanding window with enough origins to have power. A five-economy G7 extension checks which of those conclusions are systematic.

Stage 3 takes the shock idea to scale. Rather than defend one specification, it enumerates a full grid of endogenous sets, shock subsets, shock treatments and lag rules, 1,632 cells and 1,014 unique models, gates each through the same kind of validity filters, and reads the shape of the surface. It then subjects the headline result to three stress tests that the first two stages could not: a real-time replication on archival ALFRED vintages, a head-to-head between cheap growth-rate proxies and externally identified oil [21] and monetary [22] shocks, and a density evaluation that prices the uncertainty of the future shock path.

Read together, the stages give a consistent and modest answer. A parsimonious core with one exogenous oil regressor cannot be separated from a univariate autoregression at these sample sizes. Constructed shocks help short-horizon output forecasts by about a tenth, a gain concentrated in the pandemic

quarters that survives real-time data and sits inside a large set of statistically equivalent models. Shocks processed toward structural identification forecast worse, and shocks do not earn their keep in the predictive distribution. Bayesian shrinkage matters when the alternative is an explosive OLS system, and a stochastic-volatility BVAR produces calibrated densities for output and inflation, but interest rates remain close to a no-change forecast everywhere. The paper reports these results, including the nulls, as they came out.

II. DATA

All series are quarterly US observations from public sources. The macroeconomic panel is drawn from the Federal Reserve Economic Data (FRED) interface, utilization-adjusted total factor productivity from the Federal Reserve Bank of San Francisco [13], the Holston-Laubach-Williams natural rate from the Federal Reserve Bank of New York [14], and archival data vintages from ALFRED. The externally identified shocks used in Stage 3 come from the authors' public releases: the oil supply news shock of Känzig [21] and the FOMC surprises of Bauer and Swanson [22]. Stage 2's shock inputs are constructed from public series (WTI crude, the S&P 100, personal consumption expenditure, private fixed investment, domestic nonfinancial debt and the cyclically adjusted primary balance). The G7 extension uses comparable OECD and Bank for International Settlements series.

The three stages were run on different samples, and the differences are deliberate. Stage 1 uses 1994Q2 to 2026Q1, 128 quarters, with the start bound by the availability of the real broad dollar index; its year-over-year variants begin in 1995Q1 and span 125 quarters. Stage 2 uses 2000Q1 to 2025Q1, 101 quarters, the common span of its six constructed shock inputs, and its G7 panel runs 2000 to 2025 for every economy except Japan (2002 to 2021, for source data reasons). Stage 3 uses 1990Q1 to 2026Q1, 145 quarters, the longest span over which the three-variable core and the four growth-rate shocks are all observed, and its real-time replication rebuilds the panel at each origin from the ALFRED vintage current roughly 45 days after the quarter closed. The stages also differ in how the core variables enter. Stages 1 and 3 use annualized quarterly log changes, $400 \Delta \ln x_t$, for flows and prices, because overlapping four-quarter differences inject moving-average residual autocorrelation of order three that a short-lag VAR cannot whiten; Stage 1 demonstrates this directly. Stage 2 inherited year-over-year growth rates for GDP, CPI and the real effective exchange rate from the applied convention it set out to audit, and estimates in levels after Johansen testing. The policy rate enters in levels in Stages 1 and 2 and as a stance gap, the real federal funds rate minus the one-sided natural rate, in Stage 3 (Table I).

Two construction details matter later. In Stage 3 the broad dollar splices the discontinued broad index onto its successor at January 2006 with a level match, and impulse dummies for 2020Q2 and 2020Q3 enter every model. In Stage 2 the shocks are z -scored AR(1) residuals on quarterly log growth for the four price-type series and z -scored first differences for credit

TABLE I
VARIABLE SETS BY STAGE. GROWTH RATES ARE $400 \Delta \ln$ UNLESS MARKED YOY.

Stage	Endogenous	Exogenous or candidate blocks
1	GDP (GDPC1), core PCE (PCEPILFE), fed funds (level), real broad USD (RBUSBS); 1994Q2–2026Q1	Oil (WTISPLC), TFP, investment, consumption, fiscal, exports, household net worth, house prices
2	GDP YoY, CPI YoY, REER YoY, 10y yield, policy rate (levels); 2000Q1–2025Q1	Six z -scored shocks: oil, equity wealth, consumption, investment, credit impulse, fiscal impulse
3	GDP, core PCE, stance $(i_t - \pi_t^{4q}) - r_t^*$; add-ons: dollar, wages, payrolls, term spread, permits; 1990Q1–2026Q1	Real oil, investment (GPDIC1), real net worth (TNWBSHNO), fiscal (GCEC1); Känzig oil news; Bauer-Swanson FOMC surprises

and fiscal; how those moments are estimated is itself one of the results.

III. STAGE 1: WHICH SPECIFICATION AND WHICH ESTIMATOR

A. Design

The out-of-sample design is an expanding window with the first forecast origin at 55 percent of the sample. At each origin every model is estimated on data up to that origin only and iterated forward to horizons $h \in \{1, 4\}$. For a target y the metric is relative root mean squared error against a no-change random walk,

$$\text{relRMSE}(y, h) = \frac{\text{RMSE}_{\text{model}}(y, h)}{\text{RMSE}_{\text{RW}}(y, h)}, \quad (1)$$

with a univariate AR whose order is chosen by BIC as a second benchmark. Targets on different scales are never averaged in raw RMSE; cross-model ranking uses the mean of relative RMSE across targets. Exogenous regressors are projected forward by their own AR(BIC) for multistep evaluation, so no future value enters a forecast.

Selection is greedy at the level of economic blocks rather than individual series. Starting from a core of output, inflation, the policy rate and the exchange rate, the block that most improves mean relative RMSE at $h = 1$ is added, and the search stops when no block improves it. A variable that is plausibly exogenous for forecasting, such as oil, is tested as an exogenous regressor rather than as a new equation. A specification is eligible only if it passes four hard filters on the full-sample fit: stability, all companion eigenvalues inside the unit circle; residual whiteness, a portmanteau p -value above 0.01; parsimony, at least three effective observations per parameter; and conditioning of the regressor moment matrix. A Johansen trace test on the core levels returns rank zero at 5 percent, so the models are estimated on stationary transforms rather than as an error-correction system.

Pairwise accuracy is tested with Diebold-Mariano [6] for non-nested comparisons and Clark-West [7] for nested ones. To guard against selection across many models the MCS [5]

TABLE II
STAGE 1 SPECIFICATION SEARCH, $h = 1$ (REL RMSE VS. RW)

Specification	mean	T/k	stable	white	pass
CORE	1.04	3.5	yes	yes	yes
CORE+OIL(x) (<i>sel.</i>)	0.97	3.2	yes	yes	yes
CORE+TFP(x)	1.06	3.2	yes	yes	yes
CORE+WEALTH	0.95	2.3	yes	yes	no
CORE+WEALTH+OIL(x)	0.91	2.1	yes	no (0.04)	no

returns the set of models not statistically separable from the best at a given confidence level, which is the formal answer to the data-snooping concern.

B. The specification search

Block forward selection from the core adds oil as an exogenous regressor and stops. The resulting specification, CORE+OIL(x), attains mean relative RMSE of 0.97 at $h = 1$ and 0.88 at $h = 4$.¹ Per target at $h = 1$ it records 0.84 for output, 1.07 for inflation, 1.13 for the policy rate and 0.83 for the exchange rate. Clark-West tests against the random walk are significant for the exchange rate ($p < 0.001$) and the policy rate ($p = 0.013$), marginal for output ($p = 0.081$) and insignificant for inflation ($p = 0.176$).

The validity filters do consequential work. Before the gates are applied, the greedy path runs from the bare core at 1.04 to a core plus household net worth block at 0.95 and then, with oil, to 0.91, the best raw score in the run. Neither net worth specification is admissible. The two-block form fails parsimony with 2.3 observations per parameter, and the three-block form fails both parsimony (2.1) and residual whiteness ($p = 0.04$). Utilization-adjusted TFP as an exogenous regressor is admissible and has clean residuals, but at 1.06 it is worse than the bare core, so it is eligible and never selected. Table II summarizes the search and Fig. 2 shows the relative RMSE surface. The rejected 0.91 is the case that shows the filters overruling raw out-of-sample error for a good reason. At 90 percent the MCS places the winner in the same set as the univariate AR and the bare core for every target, and excludes the random walk only for the exchange rate.

C. Rate representation and HP look-ahead

The policy rate is borderline non-stationary, which motivates a cyclical, stance-based representation. The usual choice is the deviation from an HP trend. The HP filter is two-sided, so a trend computed on the full sample uses future data to set the value at each date, and in a recursive exercise this leaks information into exactly the recent quarters on which forecasts most depend. Comparing a full-sample two-sided HP cycle of the fed funds rate with a real-time one-sided cycle, defined as the last value of an expanding HP filter anchored on the rate history from 1954, the two series correlate at only 0.52. The mean absolute revision is 0.86 percentage points, and over the

¹The repository README reports these as 0.9679 and 0.8786; the paper rounds them.

final eight quarters it is 0.71 percentage points, which is 66 percent of the cyclical component's own standard deviation.² Fig. 1 shows the divergence together with the Hamilton regression filter [11] as a one-sided alternative, consistent with the real-time output-gap unreliability documented by Orphanides and van Norden [12].

Testing the rate as a level against a leakage-free one-sided gap inside the same block search, on the three targets common to both, the cores are close: mean relative RMSE of 0.915 for the level and 0.927 for the gap at $h = 1$. The gap improves only the rate's own forecast, at 0.75 of the random walk against 1.13 for the level, and the MCS does not separate the two. A rate gap is defensible, and Stage 3 adopts one, but it must be constructed one-sided within each training window.

D. The estimator horse race

Holding CORE+OIL(x) fixed with two lags, the estimator is varied across OLS, Minnesota BVAR at tight and loose settings, and ridge, LASSO and elastic-net VAR. The Minnesota prior is implemented by dummy observations [3], so the point forecast is the posterior mean, and the tight-to-loose limit was verified to recover OLS to machine precision. Shrinkage penalties are chosen by blocked time-series cross-validation inside each training window. A wide information set of ten endogenous variables plus two exogenous regressors, on which OLS is deliberately overparameterized, tests whether shrinkage plus more data can beat the parsimonious baseline.

Table III reports the outcome. The univariate AR is the single best forecaster on average at $h = 1$ with mean relative RMSE of 0.85. The small-system estimators cluster just behind, between 0.88 and 0.90. Every wide-information model is worse and the overparameterized OLS is worst, at 1.05 and 1.15. The tight BVAR delivers the best single output forecast at 0.69, the Minnesota prior giving up accuracy on the other targets for accuracy on output. The decisive result is the joint MCS: at 90 percent it contains every estimator and the AR benchmark on all four targets, and the random walk on every target except the exchange rate. The data cannot separate the models, and the failure of the wide set vindicates the filters: the variables OLS could not hold carry little incremental quarterly signal.

E. Why year-over-year transforms are excluded

Replacing the core flow transforms with year-over-year percent changes and rerunning the search produces two effects. Predictability migrates in horizon: a four-quarter overlapping difference barely moves at $h = 1$, where a no-change forecast is nearly unbeatable, so output loses to the random walk at $h = 1$ (1.04) and beats it at $h = 4$ (0.82), while under annualized quarterly changes it wins at both (0.90 and 0.72). Inflation, close to a unit root under the overlapping transform with an augmented Dickey-Fuller p of 0.81, loses at both horizons. More consequentially, the transform injects moving-average structure of order three into the residuals, and with all

²The README records 66.3 percent in the frozen reference run and 65.8 percent on a later FRED vintage; the difference is data revision.

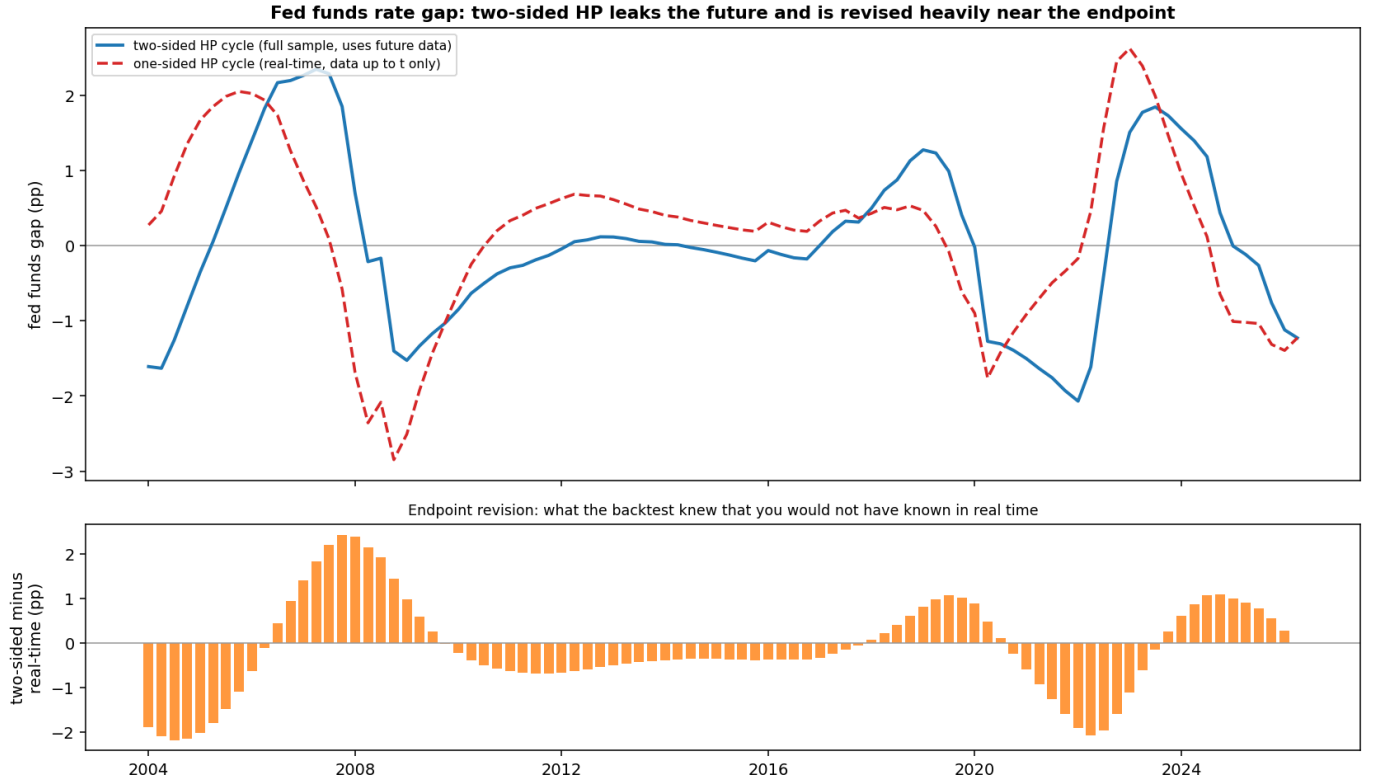


Fig. 1. Stage 1. Fed funds rate gap. Top: the full-sample two-sided HP cycle diverges from the real-time one-sided cycle, most visibly near the sample end. Bottom: the endpoint revision, the difference between what a full-sample backtest encodes and what was knowable in real time.

TABLE III
STAGE 1 ESTIMATOR COMPARISON, MEAN RELRMSE VS. RW

Model	set	$h=1$	$h=4$	in 90% MCS
AR(BIC)	—	0.85	0.83	yes
Ridge	small	0.88	0.88	yes
BVAR-Loose	small	0.88	0.89	yes
OLS	small	0.89	0.90	yes
BVAR-Tight	small	0.89	0.88	yes
Elastic-net	small	0.90	0.90	yes
LASSO	small	0.90	0.90	yes
BVAR-Tight	wide	0.93	0.88	yes
Ridge	wide	0.96	0.93	yes
Random walk	—	1.00	1.00	all but FX
LASSO	wide	1.01	1.12	yes
Elastic-net	wide	1.02	0.99	yes
OLS (overfit)	wide	1.05	1.15	yes

core flows in year-over-year form every specification fails the whiteness filter. This is why Stage 3 never uses overlapping transforms, and why Stage 2, which does, is read with its Johansen and stability diagnostics in view.

IV. STAGE 2: DO CONSTRUCTED SHOCKS HELP, AND ARE DENSITIES CALIBRATED?

A. The model and a coding trap

Stage 2 estimates the five-variable system of Table I in levels with p lags and exogenous regressors x_t ,

$$y_t = c + \sum_{j=1}^p B_j y_{t-j} + \Gamma x_t + u_t, \quad u_t \sim \mathcal{N}(0, \Sigma), \quad (2)$$

under a natural-conjugate Minnesota prior [2] that centers each variable's own first lag at a value near unity for persistent levels and below unity for growth rates, with prior variance on the coefficient of lag ℓ of variable j in equation i proportional to $(\lambda/\ell)^2(\sigma_i/\sigma_j)^2$. The prior is imposed through dummy observations [3], [4]: artificial rows are appended so that OLS on the stacked system reproduces the conjugate posterior, with a finite scale-tied prior on the exogenous coefficients and a sum-of-coefficients prior that damps near-unit-root behavior. Estimation in levels is justified by the Johansen trace test, which returns full cointegration rank ($r = 5$) for the system, meaning every linear combination is stationary, so an error-correction model collapses to a level VAR. Augmented Dickey-Fuller and KPSS tests classify GDP growth and the policy rate as $I(0)$, inflation and the 10-year yield as $I(1)$, and REER growth as ambiguous.

The first finding is about implementation rather than choice. On these data the OLS VAR is explosive, with a largest

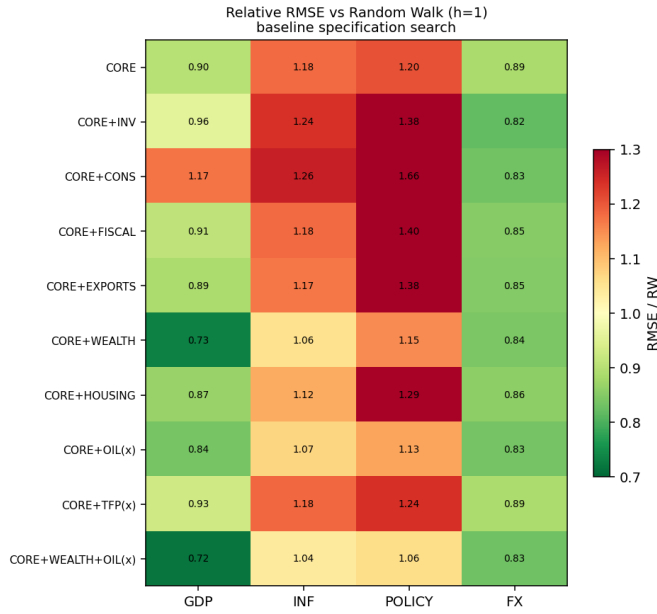


Fig. 2. Stage 1. Relative RMSE versus a random walk at $h = 1$ across specifications (rows) and targets (columns). Values below one beat the random walk. The oil-augmented core beats the benchmark on output and the exchange rate; endogenous block additions raise the ratio and fail the validity filters.

companion eigenvalue of 1.068, and a Minnesota prior coded with the precision and variance conventions inverted leaves the posterior effectively unshrunk, so multistep forecasts diverge. The dummy-observation prior with $\lambda = 0.05$ and sum-of-coefficients tightness $\mu = 0.3$ brings the largest eigenvalue to 0.96, with 78 percent of posterior draws stable and bounded forecasts. Tighter priors than textbook defaults are needed because the data are highly persistent. The Stage 1 BVAR followed the same rule: impose the prior through dummy observations and monitor the companion eigenvalue.

B. Shock stacking and look-ahead

The second finding reverses a common instinct. Over a static one-step window of 13 quarters, 2022Q1 to 2025Q1, with roughly 83 training quarters, a no-shock baseline is compared with the single best shock per equation and with all six shocks together. Stacking all shocks is dramatically worse than using none: GDP RMSE rises from 0.60 to 4.59 and REER from 2.82 to 8.28. The best single shock helps modestly, and oil is the most consistently useful, winning three of five equations. The result depends on the design, in which the shocks enter with full lag structure in a heavily parameterized level system; Stage 3 shows that a two-lag cap reverses its sign.

The third finding concerns leakage in shock construction. Standardizing the shocks with full-sample mean and variance, and fitting the AR(1) on the full sample, leaks post-sample information into the evaluation period. Two leak-free variants are built: a train-frozen version in which the AR and standardization parameters are estimated only through 2021Q4, and a rolling real-time version using a trailing five-year window. The leakage is modest in RMSE terms but

systematically flattering, lowering GDP RMSE from 0.526 (frozen) and 0.618 (rolling) to 0.478. Decisively, the only statistically significant benchmark improvement in the static window under leaky shocks, REER beating the random walk at $p = 0.026$, vanishes once shocks are built in real time ($p = 0.14$ frozen, $p = 0.18$ rolling). Apparent skill that does not survive leak-free construction is not skill.

C. Evaluation window and density forecasts

The static 13-quarter window cannot discriminate models: with leak-free shocks no Diebold-Mariano test rejects equal accuracy at 5 percent. The evaluation is therefore moved to an expanding window from 2013Q1 with 48 origins, re-estimating every model each quarter and scoring the full predictive density. Point accuracy is summarized by RMSE and tested by Diebold-Mariano with the Harvey-Leybourne-Newbold small-sample correction [6], [8]; densities are scored by the log predictive score and the continuous ranked probability score (CRPS) [9], with the empirical coverage of the 90 percent interval. The BVAR predictive density is the posterior predictive ensemble; benchmark densities are Gaussian with variance estimated on the training window.

To accommodate time-varying variance the BVAR carries a univariate stochastic-volatility (SV) model on the relevant equation, estimated with the Kim-Shephard-Chib mixture sampler [15], [16]. A diagnostic motivates the variable-specific treatment: residuals are 3.9 times their normal magnitude for GDP growth over 2020Q2 to 2021Q2 but only 1.6 times for inflation and essentially unchanged for the exchange rate and rates, so the pandemic variance spike is output-specific rather than common. A single common volatility factor [17], which was implemented, dilutes it and proves unidentifiable on a 95-quarter sample. The univariate SV path on GDP residuals recovers elevated volatility at both the global financial crisis (about 1.6 times the calm level) and the pandemic (about 2.7 times) with persistence $\phi \approx 0.85$. Pandemic indicator variables [18] are a pragmatic equivalent, and Stage 3 uses them.

Table IV and Fig. 3 report the outcome. The SV-BVAR's predictive densities are sharper and better calibrated than both benchmarks for GDP growth and inflation on CRPS and log score, and tie or beat the long rate on log score; coverage lies in 0.85 to 0.92 against a 0.90 target. With 48 origins the Diebold-Mariano tests acquire power: GDP beats the random walk at $p = 0.068$ and inflation beats AR(1) at $p = 0.051$, against no significance at all with 13 origins. The interest-rate block remains hard to beat, since a persistent rate is well approximated by a no-change forecast.

Extending the evaluation to horizons $h \in \{1, 2, 4, 8\}$ and adding an unrestricted OLS VAR and a no-shock BVAR isolates the contribution of each component. Shrinkage is decisive at every horizon: for GDP the unrestricted VAR has CRPS of 2.27 against 1.42 for the BVAR at $h = 4$, so the explosive OLS dynamics translate directly into poor forecasts. The constructed shocks, by contrast, do not help: the full BVAR and the no-shock BVAR are essentially indistinguishable across variables and horizons. Because the shocks

TABLE IV
STAGE 2 EXPANDING-WINDOW ONE-STEP EVALUATION, 48 ORIGINS
(2013Q1–2025Q1). RMSE AND CRPS: LOWER IS BETTER; LOG SCORE
(LS): HIGHER IS BETTER; COV: 90% INTERVAL COVERAGE.

Variable	Model	RMSE	CRPS	LS	Cov
GDP growth	SV-BVAR	1.785	0.685	−2.78	0.92
	RW	2.532	1.068	−3.33	0.88
	AR(1)	2.311	0.948	−3.30	0.90
Inflation	SV-BVAR	0.379	0.183	−0.49	0.90
	RW	0.407	0.201	−0.61	0.83
	AR(1)	0.439	0.211	−0.78	0.85
REER	SV-BVAR	4.321	2.117	−2.69	0.85
	RW	3.465	1.996	−2.68	0.90
	AR(1)	3.349	1.932	−2.64	0.94
10y yield	SV-BVAR	0.448	0.215	−0.44	0.88
	RW	0.383	0.216	−0.50	0.94
	AR(1)	0.386	0.217	−0.50	0.92
Policy rate	SV-BVAR	0.535	0.258	−0.57	0.88
	RW	0.443	0.229	−0.65	0.92
	AR(1)	0.460	0.236	−0.68	0.92

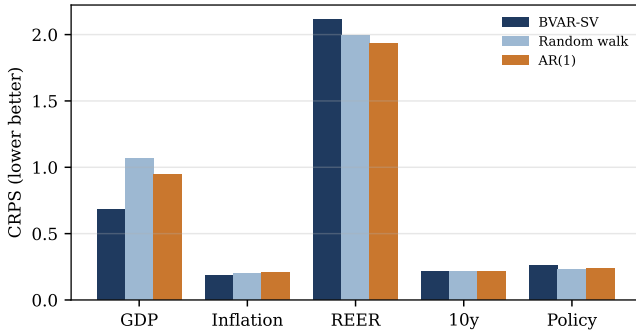


Fig. 3. Stage 2. Mean CRPS by variable over 48 expanding-window origins (lower is better). The SV-BVAR improves on both naive benchmarks for GDP growth and inflation; the rate variables are close to a no-change forecast.

are observed only contemporaneously, they aid in-sample fit but cannot improve multistep prediction unless the shocks themselves are forecast.

D. Identification and the G7 extension

Local projections [19] to a one-standard-deviation oil surprise show inflation peaking near 0.20 percentage points after two quarters and the policy rate climbing to roughly 0.37 after two years; the early positive output response shows that an unconditional oil surprise conflates supply and demand [20]. Used as instruments in a proxy-SVAR, the constructed shocks are mostly weak (first-stage $F < 10$); only oil is a defensible, if moderate, instrument. Stage 3 returns to this point with properly identified shocks.

Replicating the evaluation across the United States, United Kingdom, Canada, Germany and Japan, with the oil shock as the only exogenous regressor and an identical pipeline per

economy, shows which findings are systematic. The BVAR beats the unrestricted VAR in all five economies for every variable, beats the random walk for output growth in all five, and beats AR(1) in all five for inflation and the policy rate. It rarely beats the no-shock BVAR (one of five economies for both GDP and inflation), the exception being the exchange rate (four of five), the variable most directly tied to oil. Pooling one-step loss differentials across countries restores statistical power that no single short sample has: the BVAR significantly outperforms the random walk for output growth ($p = 0.016$) and the policy rate ($p < 0.001$), while inflation, the exchange rate and the long yield are not separable from no change. The 90 percent MCS eliminates the unrestricted VAR for every variable except GDP, and for the policy rate only the two BVAR variants survive. Shrinkage is the one feature that systematically earns its place; constructed shocks improve fit but not forecasts.

V. STAGE 3: THE SHOCK IDEA AT SCALE

A. Design

Stage 2 compared one shock-augmented BVAR with one no-shock BVAR. Stage 3 asks the same question across the whole specification space. Each model is estimated by equation-by-equation OLS as

$$Y_t = c + \sum_{i=1}^p A_i Y_{t-i} + B_0 X_t + B_1 X_{t-1} + \Gamma D_t + \varepsilon_t, \quad (3)$$

where Y_t is the chosen endogenous vector ($n \in \{3, 4, 5\}$) built around the three-variable core of GDP growth, core PCE inflation and the policy stance, X_t is the chosen shock vector under one of three treatments, and D_t holds the pandemic dummies. Shocks enter with lags zero and one only, which caps the parameter cost of each additional shock at two coefficients per equation. The design crosses 16 endogenous sets with 12 shock subsets drawn from real oil, private investment, real household net worth and government spending growth, three shock treatments (raw, purged of contemporaneous GDP, purged of the full endogenous vector) and three information-criterion lag rules over $p \in \{1, \dots, 4\}$, for 1,632 specification cells and 1,014 unique models. Both purges are refit on each training window, so the derived regressors are leak-free by construction; the full purge approximates treating the shock as an external instrument.

Lag order is chosen on the initial training window, 1990Q1 to 2009Q4, and held fixed across origins. Every model must pass the same kind of gates as Stage 1: largest companion eigenvalue below one, minimum Ljung-Box whiteness p -value above 0.01 across equations at eight lags, and at least three observations per parameter. The gates keep 85.6 percent of specifications. Forecasts are iterated from 65 expanding-window origins, 2009Q4 through 2025Q4, at horizons $h \in \{1, 4, 8\}$. Future shocks follow either an unconditional path, set to their training-window mean, which is the honest forecast, or the realized path, a scenario exercise measuring the value of knowing the shock path. Metrics are RMSE, an out-of-sample

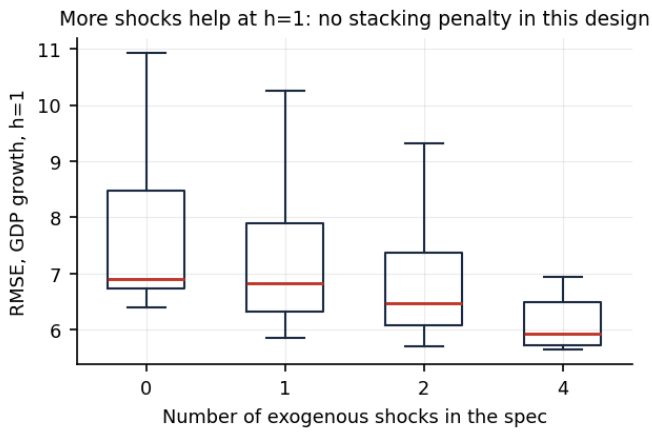


Fig. 4. Stage 3. Distribution of $h = 1$ GDP RMSE across validity-gated specifications by number of included shocks. Medians in red.

R^2 against a univariate AR(BIC) benchmark, and a Diebold-Mariano test on squared-error loss against the core three-variable VAR under the same lag rule with HAC variance at $h - 1$ lags, all reported for the full window and excluding target quarters from 2020Q1 to 2021Q1.

B. What the surface shows

Every winner on the validity-gated leaderboard includes shocks. The best GDP specification at $h = 1$, the core plus the term spread with all four raw shocks at $p = 2$, has RMSE 5.65 against 6.40 for the best shockless VAR, a reduction of about 12 percent, and an out-of-sample R^2 of 0.33 against the AR benchmark on the full window that falls to 0.10 once the pandemic quarters are excluded. Inflation and the dollar are barely forecastable (R^2 of 0.07 and 0.04 at $h = 1$). The stance loses to its own AR at $h = 1$ (-0.11), where persistence dominates, but earns R^2 of 0.40 at $h = 8$ with investment and net worth shocks: the system-based forecast of where the rate gap mean-reverts to is the single largest gain in the study, consistent with Stage 1's finding that a gap representation helps most for the rate's own forecast.

Shock stacking does not overfit in this design. Median $h = 1$ GDP RMSE across gated specifications falls monotonically in the number of included shocks: 6.91 with none, 6.82 with one, 6.47 with two and 5.92 with all four (Fig. 4). This is the opposite of the Stage 2 result, and the reconciliation is parameter cost. At two coefficients per equation per shock, with the pandemic dummies absorbing the outlier quarters, each additional shock behaves like an additional leading indicator; with deep exogenous lag structures or leaky construction, the same stacking destroys performance. Neither result should be extrapolated to the other design.

Residualization is an identification tool rather than a forecasting tool. Across shocked specifications, median $h = 1$ RMSE is lowest for raw shocks for GDP (6.24 against 6.68 GDP-purged and 6.75 fully purged), inflation (1.00 against 1.01 and 1.03) and the dollar (9.90 against 9.95 and 10.29). The stance is the exception, where both purges beat raw

(0.55 against 0.56). Investment and net worth growth carry contemporaneous business-cycle information that predicts next quarter's activity, and projecting it out removes the signal a forecaster wants; the policy rate responds to that systematic component, so removing it isolates the innovation that still carries news.

Lag rules matter less than the persistence of the target. On median $h = 1$ RMSE, BIC beats AIC for GDP (6.46 against 6.63), inflation (0.99 against 1.07) and the dollar (9.66 against 10.34); AIC wins only for the stance (0.555 against 0.558). Knowing the future shock path is worth little: conditioning on realized shocks improves median GDP RMSE by 0.24 points at $h = 1$ for raw shocks and roughly 0.3 points at $h = 8$, against an unconditional RMSE of about 6.

C. What chance delivers

Across all validity-gated specifications, targets and horizons, only 4.0 percent of Diebold-Mariano tests against the core VAR reject at the 5 percent level, which is indistinguishable from the false-positive rate under the null. The best GDP specification's p -value against the core is 0.38 on the full window. With 65 evaluation points and squared-error loss dominated by a handful of quarters, the ordering among the top specifications is not statistically meaningful. The honest summary is a set of near-equivalent models spanned by oil, net worth and one real-activity shock in raw form, the conclusion the MCS delivered in Stages 1 and 2, reached here by enumeration.

D. Real-time replication

Final-vintage results assume an end-of-quarter information set on revised data, which flatters any model that leans on the national accounts. The core comparison is therefore rerun on ALFRED vintages, rebuilding the panel at each origin from the vintage current roughly 45 days after the quarter closed, just after the advance GDP print. This carries both the publication lag (net worth trails the output block by a quarter) and the pre-revision values of every national-accounts series. GDP, core PCE, investment, government spending and net worth are pulled at their contemporaneous vintages; the federal funds rate, oil and CPI are held at final values because they are not meaningfully revised, and the natural rate at its final one-sided estimate for want of a real-time series. Forecasts are scored against final actuals.

Table V and Fig. 5 report the outcome for GDP. The absolute RMSE of every specification rises by one to six percent out of vintage, an almost parallel shift, but the relative advantage of the all-four-shock system over the pure core VAR is essentially unchanged: 16.1 percent falls to 15.9 percent on the full window and 8.1 to 8.0 percent excluding the pandemic, with the shocked system still earning a real-time out-of-sample R^2 of 0.42 against the AR benchmark.³ The one fragile choice is residualization. Purging against contemporaneous GDP means projecting on a figure that is itself heavily revised,

³The 16.1 percent is measured against the pure three-variable core (6.74); the 12 percent of the leaderboard is measured against the best shockless VAR with add-ons (6.40). Both figures appear in the Stage 3 source and README.

TABLE V

STAGE 3 GDP FORECAST RMSE, FINAL VINTAGE AGAINST REAL-TIME ALFRED VINTAGES, CORE SPECIFICATIONS, $h = 1$, UNCONDITIONAL PATH. ADVANTAGE IS THE RMSE REDUCTION OF THE ALL-FOUR-SHOCK SYSTEM RELATIVE TO THE PURE CORE VAR.

Model	Full window		Ex COVID	
	Final	Real-time	Final	Real-time
Pure VAR	6.74	7.10	1.90	1.93
All four, raw	5.66	5.97	1.74	1.78
<i>Advantage</i>	<i>16.1%</i>	<i>15.9%</i>	<i>8.1%</i>	<i>8.0%</i>
All four, GDP-purged	5.70	7.68	—	—

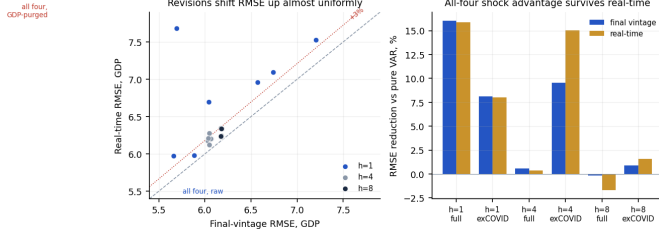


Fig. 5. Stage 3. Left: real-time against final-vintage GDP RMSE across specifications and horizons; points sit just above the 45-degree line except the GDP-purged system. Right: the all-four-shock advantage over the pure VAR, final vintage against real-time.

so the GDP-purged system degrades by 35 percent out of vintage against 5 percent for the raw system, turning a near tie on final data into a clear loss. Real-time data therefore sharpen both practical conclusions: shock augmentation helps GDP at short horizons, and raw shocks beat purged ones, now also because purging imports revision noise.

E. External identification against cheap proxies

The shocks used so far are growth-rate proxies, and the natural objection is that a properly identified structural shock would forecast better. The oil block allows a direct test. Känzig’s oil supply news shock [21], identified from high-frequency movements in oil futures around OPEC announcements, is available monthly from 1975 and summed to quarters, so it spans the whole evaluation window. Three exogenous regressors are compared inside the core VARX at $p = 2$: the study’s own real-oil-price growth proxy, which is leak-free and persistent; Känzig’s surprise instrument, leak-free but zero outside announcement quarters; and Känzig’s news shock recovered from his full-sample SVAR-IV, which is dense but embeds sample-wide information the other two do not, a look-ahead that stacks the deck for identification.

It changes nothing. On the full window at $h = 1$ the raw proxy cuts inflation RMSE by 7.0 percent against 4.4 percent for the news shock and 2.5 percent for the surprise, and it is the only oil regressor that helps GDP at all, while the look-ahead-advantaged news shock is the worst of the three for GDP (6.91 against 6.79 for the pure VAR and 6.66 for the proxy). Across all twelve target, horizon and window cells the raw proxy beats the pure VAR in eight and the news shock

TABLE VI

STAGE 3 PREDICTIVE DENSITY AT $h = 1$, FULL WINDOW: 90 PERCENT INTERVAL COVERAGE AND MEAN LOG SCORE, UNDER THE NAIVE PLUG-IN DENSITY AND THE FAIR DENSITY THAT PRICES FUTURE-SHOCK UNCERTAINTY.

Target	Model	90% coverage		Log score	
		Naive	Fair	Naive	Fair
GDP	Pure VAR	0.88	0.88	−6.49	−6.49
GDP	All four	0.68	0.91	−13.01	−4.78
Inflation	Pure VAR	0.82	0.82	−2.27	−2.27
Inflation	All four	0.74	0.75	−2.45	−2.19

in three, and no Diebold-Mariano test rejects at 5 percent. Identification orthogonalizes away the persistent, predictable component of oil-price movements, which is exactly what a point forecaster wants.

The monetary margin is the natural place for a counterexample, because the stance variable is the weakest core proxy and a high-frequency surprise moves even at the zero lower bound. Bauer and Swanson’s FOMC surprises [22], added over their 2009 to 2023 coverage, do not provide one. The raw surprise helps core inflation at the four-quarter horizon by about one percent (1.22 against 1.23 for the core alone), the horizon its transmission lag predicts, but the gain is insignificant and absent for GDP, where the identified shocks degrade forecasts. The orthogonalized surprise is worse than the raw one in eleven of twelve cells. The canonical identified fiscal series, Ramey’s military news [23], ends in 2015 and is near-flat after 1990, so the fiscal margin cannot be tested the same way.

Combining the proxy and identified forecasts, by equal weights or by leak-free inverse-MSE weights, is 0.2 to 0.8 percent worse than keeping the better single model in every margin and target group. In a forecast-encompassing regression the proxy forecast encompasses the identified one in 22 of 24 cells; the two exceptions, core inflation at eight quarters on the monetary margin, resolve to a 0.4 percent edge no test separates from zero. Identification and prediction pull in opposite directions, and for the forecasting task the cheap proxy is the better tool rather than a compromise.

F. Density forecasts and future-shock risk

Every Stage 3 result so far scores the conditional mean. A shock-augmented VARX absorbs in-sample variance, so its plug-in residual covariance is smaller and its predictive density tighter than the pure VAR’s. Table VI evaluates that density for the core VAR against the all-four-shock model, and Fig. 6 shows the mechanism.

Under the naive plug-in density, which holds future shocks at their training mean, the sharpening is a trap. The all-four-shock model’s nominal 90 percent GDP interval covers only 68 percent of outcomes against 88 for the pure VAR, and the single pandemic quarter, landing far outside the too-narrow density, drives its log score to −13.01. The cause is mechanical: holding future shocks at their mean ignores the uncertainty of the future shock path. Propagating that uncertainty through

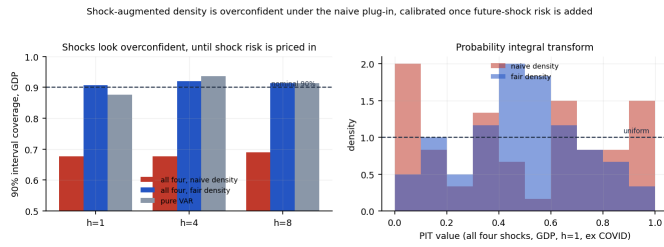


Fig. 6. Stage 3. Left: 90 percent GDP interval coverage for the all-four-shock model under the naive and fair densities against the pure VAR; the naive density is badly overconfident and the fair density restores it. Right: the probability integral transform for the same model at $h = 1$; the naive PIT piles up in the tails, the fair PIT is close to uniform.

the same moving-average structure restores GDP coverage to 91 percent and the log score to -4.78 , but it widens the intervals, because the in-sample variance the shocks explain is offset almost exactly by the cost of having to forecast the shocks. Under the fair density the shock model is no sharper than the pure VAR, every log-score difference is insignificant, and the one residual edge, GDP at one quarter on the full window, is the same pandemic-era mean advantage the point forecasts already show.

Even the fair density leaves the pure VAR a little undercovered, most visibly for long-horizon inflation. A Bayesian predictive separates the two candidate causes. Simulating from a Normal-inverse-Wishart posterior centered at OLS, so that coefficients and covariance are drawn rather than fixed, lifts the pure VAR's 90 percent coverage from 0.88 to 0.91 for GDP and from 0.80 to 0.85 for inflation at one quarter,⁴ closing the gap and over-covering at long horizons. Resampling training residuals in place of Gaussian forward shocks does not help, because the pandemic quarters are already absorbed by the dummies. The under-coverage is estimation noise rather than misspecification, and because parameter uncertainty widens every model's intervals alike, it leaves the cross-model verdict unchanged. Stage 2's SV-BVAR, a full posterior predictive, reached the same calibration from the other direction, covering 0.90 to 0.92 for output and inflation without any shock contribution.

VI. WHAT THE THREE STAGES SAY TOGETHER

A. Headline results

Table VII collects the headline numbers. Three agreements run through them.

First, parsimony and leakage-free evaluation dominate estimator sophistication in short macroeconomic samples. Stage 1 could not separate any estimator from a univariate autoregression, and its filters rejected the best raw score in the run. Stage 2 found that the one feature that systematically matters is shrinkage against an explosive OLS system, not the shocks. Stage 3 found 1,014 models whose ordering is

statistically indistinguishable from noise. Every stage located a source of flattering leakage, the HP endpoint, full-sample shock standardization, revised national accounts, a full-sample identified shock, a plug-in density, and in each case removing it shrank or erased an apparent result.

Second, reduced-form forecasting does not reward structural identification. Stage 2's constructed shocks aided in-sample fit but not forecasts. Stage 3 made the reason explicit: purging the shocks against the endogenous block costs accuracy for output, prices and the dollar, externally identified oil and monetary shocks are encompassed by cheap growth-rate proxies, and the more a series is processed toward clean identification the worse it forecasts. The exception, the policy rate, is the one variable whose systematic component is itself a response to the others, so stripping it leaves news.

Third, the honest product is a set, not a champion. Stage 1 reports an MCS, Stage 2 a panel MCS, Stage 3 a false-positive rate, and all three say that the leaderboard winner is one member of a large equivalence class.

Where the stages appear to disagree, the disagreement is informative. Stage 2's catastrophic stacking and Stage 3's monotone stacking gain differ because exogenous lag depth and the treatment of 2020 differ, not the shocks. Stage 2's SV-BVAR beating naive benchmarks and Stage 1's inability to separate a BVAR from an AR differ because Stage 2 scored densities over persistent levels, where shrinkage against an explosive system and volatility modeling both have work to do, while Stage 1 scored point forecasts of stationary transforms, where a correctly specified AR already captures most of what there is. Stage 3's 12 percent GDP gain and Stage 2's null on shocks differ in the same way: a two-lag exogenous block on annualized growth with pandemic dummies is a leading-indicator design, and a full-lag block in a level system is not.

B. Limitations

Each stage works with 100 to 145 observations and 48 to 65 evaluation origins, which is why so few comparisons separate statistically and why the pandemic quarters dominate squared-error loss; every Stage 3 metric is therefore reported with and without them. The stages use different samples and transforms, so cross-stage numbers are compared qualitatively rather than pooled. Stage 2's endogenous block uses year-over-year growth rates, which Stage 1 shows fail whiteness, and relies on final-vintage data; its G7 panel uses a single exogenous shock. Stage 3's real-time replication holds the financial series and the natural rate at final values, and the Holston-Laubach-Williams rate, while one-sided, is re-estimated by its authors over the full sample, so the stance variable inherits mild look-ahead through parameter revisions. The external-identification check covers oil and monetary shocks but not fiscal ones. The headline Stage 3 density figures are plug-in, and the Bayesian check leaves a mild shortfall in long-horizon inflation intervals that a linear Gaussian VAR cannot close. None of this argues against Bayesian, shrinkage or structural methods where larger cross-sections or higher frequency make their advantages material; it says that on a short national

⁴The source paper's plug-in table reports 0.82 for the same pure-VAR inflation interval; its Bayesian paragraph reports 0.80. Both are reproduced as stated.

TABLE VII
HEADLINE RESULTS ACROSS THE THREE STAGES. ALL FIGURES ARE TAKEN FROM THE STAGE PAPERS AND REPOSITORY READMEs.

Question	Stage	Finding
Which specification survives a gated search?	1	Core plus oil as exogenous regressor: relRMSE 0.97 ($h=1$), 0.88 ($h=4$); net worth block at 0.91 raw rejected ($T/k = 2.1$, whiteness $p = 0.04$).
Does any estimator beat a univariate AR?	1	No. AR(BIC) 0.85; small-system OLS, BVAR, ridge, LASSO, elastic-net 0.88 to 0.90; 90% MCS contains all estimators on all four targets.
How much does full-sample HP filtering leak?	1	Endpoint revision 0.71 pp over the last eight quarters, 66% of the cycle's standard deviation; two-sided and one-sided cycles correlate at 0.52.
Do year-over-year transforms work as model variables?	1	No. MA(3) residuals; every specification fails whiteness.
Is the Minnesota prior shrinking?	2	Naive coding: largest eigenvalue 1.068; dummy observations: 0.96, 78% of draws stable.
Does stacking all shocks help?	2, 3	Depends on parameter cost. Stage 2 (full lags, levels): GDP RMSE 0.60 to 4.59. Stage 3 (two lags, dummies): median GDP RMSE 6.91 to 5.92.
Does look-ahead in shock construction matter?	2	The only significant benchmark win (REER vs. RW, $p = 0.026$) vanishes leak-free ($p = 0.14, 0.18$).
Are SV-BVAR densities calibrated?	2	Yes for output and inflation: coverage 0.90 to 0.92, best CRPS and log score; GDP vs. RW $p = 0.068$, inflation vs. AR(1) $p = 0.051$ over 48 origins.
Do constructed shocks help the BVAR forecast?	2	No. Full and no-shock BVAR indistinguishable at every horizon; G7 full BVAR beats no-shock in 1/5 economies for GDP and inflation.
Is shrinkage systematic?	2	Yes. BVAR beats OLS VAR in 5/5 economies for every variable; panel DM vs. RW $p = 0.016$ (GDP), $p < 0.001$ (policy rate).
How much do shocks buy at scale?	3	Best GDP $h=1$: RMSE 5.65 vs. 6.40 (about 12%); R_{oos}^2 0.33 full window, 0.10 ex-COVID; stance R_{oos}^2 0.40 at $h=8$.
Is the ranking statistically meaningful?	3	No. 4.0% of DM tests reject at 5%; best GDP model $p = 0.38$ against the core.
Does the gain survive real-time data?	3	Yes. All-four-shock advantage 16.1% to 15.9% (full), 8.1% to 8.0% (ex-COVID); GDP-purged degrades 35%.
Do identified shocks beat proxies?	3	No. Proxy encompasses identified in 22/24 cells; blends 0.2 to 0.8% worse; raw proxy cuts inflation RMSE 7.0% vs. 4.4% (news).
Are shock-model densities honest?	3	Only when future-shock risk is priced: naive GDP coverage 0.68 and log score -13.01 , fair 0.91 and -4.78 ; then no sharper than the pure VAR.

quarterly panel the parsimonious baseline is hard to beat and easy to fool oneself about.

VII. CONCLUSION

On US quarterly data, and once look-ahead leakage, overfitting and multiple comparisons are controlled together, what forecasts the macroeconomy out of sample is less than the applied toolkit suggests. A four-variable core with oil as an exogenous regressor is the honest baseline, and nothing estimated on it, Bayesian or regularized, narrow or wide, is separable from a univariate autoregression. Bayesian shrinkage earns its place when the alternative is an explosive OLS system, and a stochastic-volatility BVAR delivers calibrated densities for output and inflation, but interest rates stay close to a no-change forecast in every design and every country examined. Cheap growth-rate shocks entering with two lags improve short-horizon GDP forecasts by about a tenth, a gain that is concentrated in the pandemic, survives real-time data almost intact, and sits inside a set of a thousand statistically equivalent models. Processing those shocks toward structural identification makes them worse forecasters, and their apparent sharpness in the predictive distribution is an

artifact of pretending the future shock path is known. The useful outputs are the gates, the leak-free constructions, the confidence sets and the reported nulls, which together quantify how little the added sophistication buys at this sample size.

VIII. REPRODUCIBILITY

The three stages are implemented as three unchanged pipelines in the repository, and this paper introduces no new computation: every figure is read from the directories below by relative path and every number is taken from the stage papers or their READMEs. All data are public and keyless.

- *Stage 1* (repository root). `src/reproduce.py` is a single-seed driver (seed 20260710) that pulls FRED and FRBSF data and regenerates every Stage 1 number. Executed outputs are in `results/`: `reproduce_results.json`, `phase2_summary.json` and the `var_horserace_*` tables. Typeset figures are in `papers/var_selection_horserace/figs/`.
- *Stage 2* (`studies/bvar_shocks_density/`). Modules `shock_builder.py`, `var_advanced.py`, `sv_univariate.py`,

eval_multi.py and panel.py in src/. Cited tables in results/: eval_multi_us.csv, rolling_density_eval.csv, panel_metrics.csv, gdp_sv_path.csv. Figures in figures/. DATA.md lists the public series behind the six shocks.

- *Stage 3* (studies/varx_permutation_study/). src/fetch_data.py then run_study.py reproduces the full walk-forward (1,014 models, 63,918 result rows, about 35 seconds). The extensions run_realtime.py, varx_identified.py, varx_monetary.py, varx_combine.py, varx_density.py and varx_bayes_density.py run in seconds against cached data. Results in results/: paper_stats.json, rt_stats.json, id_stats.json, mp_stats.json, density_stats.json, bayes_density_stats.json and the matching *_compare.parquet tables. Figures in figures/.

REFERENCES

- [1] C. A. Sims, “Macroeconomics and reality,” *Econometrica*, vol. 48, no. 1, pp. 1–48, 1980.
- [2] R. B. Litterman, “Forecasting with Bayesian vector autoregressions: five years of experience,” *J. Bus. Econ. Stat.*, vol. 4, no. 1, pp. 25–38, 1986.
- [3] M. Bańbura, D. Giannone, and L. Reichlin, “Large Bayesian vector auto regressions,” *J. Appl. Econometrics*, vol. 25, no. 1, pp. 71–92, 2010.
- [4] C. A. Sims and T. Zha, “Bayesian methods for dynamic multivariate models,” *Int. Econ. Rev.*, vol. 39, no. 4, pp. 949–968, 1998.
- [5] P. R. Hansen, A. Lunde, and J. M. Nason, “The model confidence set,” *Econometrica*, vol. 79, no. 2, pp. 453–497, 2011.
- [6] F. X. Diebold and R. S. Mariano, “Comparing predictive accuracy,” *J. Bus. Econ. Stat.*, vol. 13, no. 3, pp. 253–263, 1995.
- [7] T. E. Clark and K. D. West, “Approximately normal tests for equal predictive accuracy in nested models,” *J. Econometrics*, vol. 138, no. 1, pp. 291–311, 2007.
- [8] D. Harvey, S. Leybourne, and P. Newbold, “Testing the equality of prediction mean squared errors,” *Int. J. Forecasting*, vol. 13, no. 2, pp. 281–291, 1997.
- [9] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” *J. Amer. Statist. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.
- [10] R. J. Hodrick and E. C. Prescott, “Postwar U.S. business cycles: an empirical investigation,” *J. Money Credit Bank.*, vol. 29, no. 1, pp. 1–16, 1997.
- [11] J. D. Hamilton, “Why you should never use the Hodrick-Prescott filter,” *Rev. Econ. Stat.*, vol. 100, no. 5, pp. 831–843, 2018.
- [12] A. Orphanides and S. van Norden, “The unreliability of output-gap estimates in real time,” *Rev. Econ. Stat.*, vol. 84, no. 4, pp. 569–583, 2002.
- [13] J. Fernald, “A quarterly, utilization-adjusted series on total factor productivity,” Federal Reserve Bank of San Francisco, Working Paper 2012-19, 2014.
- [14] K. Holston, T. Laubach, and J. C. Williams, “Measuring the natural rate of interest after COVID-19,” Federal Reserve Bank of New York Staff Reports, no. 1063, 2023.
- [15] S. Kim, N. Shephard, and S. Chib, “Stochastic volatility: likelihood inference and comparison with ARCH models,” *Rev. Econ. Stud.*, vol. 65, no. 3, pp. 361–393, 1998.
- [16] G. E. Primiceri, “Time varying structural vector autoregressions and monetary policy,” *Rev. Econ. Stud.*, vol. 72, no. 3, pp. 821–852, 2005.
- [17] A. Carriero, T. E. Clark, and M. Marcellino, “Common drifting volatility in large Bayesian VARs,” *J. Bus. Econ. Stat.*, vol. 34, no. 3, pp. 375–390, 2016.
- [18] M. Lenza and G. E. Primiceri, “How to estimate a vector autoregression after March 2020,” *J. Appl. Econometrics*, vol. 37, no. 4, pp. 688–699, 2022.
- [19] Ò. Jordà, “Estimation and inference of impulse responses by local projections,” *Amer. Econ. Rev.*, vol. 95, no. 1, pp. 161–182, 2005.
- [20] L. Kilian, “Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market,” *Amer. Econ. Rev.*, vol. 99, no. 3, pp. 1053–1069, 2009.
- [21] D. R. Känzig, “The macroeconomic effects of oil supply news: evidence from OPEC announcements,” *Amer. Econ. Rev.*, vol. 111, no. 4, pp. 1092–1125, 2021.
- [22] M. D. Bauer and E. T. Swanson, “A reassessment of monetary policy surprises and high-frequency identification,” *NBER Macroeconomics Annual*, vol. 37, pp. 87–155, 2023.
- [23] V. A. Ramey, “Identifying government spending shocks: it’s all in the timing,” *Quart. J. Econ.*, vol. 126, no. 1, pp. 1–50, 2011.