Pacific-Basin Finance Journal

The Boundaries of Index Option Return Predictability: Evidence from Taiwan

--Manuscript Draft--

Manuscript Number:	PBFJ-D-26-01484
Article Type:	Research Paper
Section/Category:	Research Paper
Keywords:	Index Option Return Predictability, Delta-Hedged Returns, Variance Risk Premium, Put--Call Relative-value Wedge, Taiwan Market
Corresponding Author:	Xian-Ji Kuang, Ph.D. National Changhua University of Education Changhua, Changhua City TAIWAN
First Author:	Zong-Wei Yeh, Ph.D.
Order of Authors:	Zong-Wei Yeh, Ph.D.
	Pin-Chi Chen
	Xian-Ji Kuang, Ph.D.
	Tzu-Chieh Lin
Abstract:	Out-of-sample evidence on index option return predictability comes almost entirely from U.S. data. We provide, to our knowledge, the first systematic pseudo-real-time out-of-sample evaluation of monthly delta-hedged index-option return forecasts in an Asia-Pacific market, using recursive forecasts of dividend-inclusive returns on at-the-money Taiwan index (TXO) puts over 2002–2022; the mean put premium is about−0.71% per month. What predictability we find runs through the variancerisk premium, and it is episodic rather than steady. Realized stock variance (SVAR), the state variable that tracks the premium's physical leg, is the only one of 164 candidates whose full-sample slope survives the Holm correction; its out-of-sample R2 of 11.03% over the 2007–2022 evaluation window (Clark–West p = 0.008) is earned mostly in crisis months and largely disappears after 2015. In a frozen 2023–2025 holdout the decay continues: the put premium shrinks to−0.24% per month and the SVAR forecast falls behind the historical mean. Point estimates concentrate the signal in puts, though formal put–call tests cannot separate the two sides, and the incremental value of predictive timing is economically negligible. Entry-date implied variance fits the target return largely by construction, so we analyze it separately as a channel variable rather than include it in the exogenous-predictor comparisons. Same-strike puts are persistently richer than calls—a wedge that closely tracks the index-futures basis, with a near-unit slope and an economically negligible mean gap: a market-wide carry phenomenon consistent with downside-insurance demand, though alternative carry-based explanations remain open.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Cover Letter

Department of Finance
National Changhua University of Education
Changhua, Taiwan

July 17, 2026

The Editors
Pacific-Basin Finance Journal

Dear Editors,

We are pleased to submit our manuscript, "**The Boundaries of Index Option Return Predictability: Evidence from Taiwan**," for consideration at the *Pacific-Basin Finance Journal*.

Out-of-sample evidence on index option return predictability comes almost entirely from U.S. data. Using monthly delta-hedged returns on at-the-money Taiwan index (TXO) puts over 2002–2022, constructed from a dividend-inclusive self-financing profit-and-loss identity, we provide, to our knowledge, the first systematic pseudo-real-time out-of-sample evaluation of such forecasts in an Asia-Pacific market. Among 164 candidate predictors, realized stock variance is the only one whose full-sample slope survives the Holm correction; its unadjusted out-of-sample R^2 over 2007–2022 is 11.03%. The gain is episodic, falling to 0.22% after the five largest-gain months are removed. In a frozen 2023–2025 holdout, the SVAR forecast underperforms the recursive historical mean, while the continuous 2007–2025 evaluation remains positive but crisis-driven. The paper therefore maps four boundaries—predictor, contract, time, and complexity—so that the negative and qualifying results are as informative as the positive ones.

The manuscript addresses an empirical asset-pricing question in an Asia-Pacific derivatives market and thus aligns closely with the journal's focus. Three specific points of fit stand out. First, the setting is distinctly Pacific-Basin: TXO is one of the Asia-Pacific region's most actively traded index option markets, with high retail participation and distinctive institutional features, and our tests engage directly with the journal's Taiwan derivatives literature (Lee and Wang, 2016; Yang, Jhang, and Chang, 2016; Lo and Liu, 2025), complementing the recent put–call-ratio timing evidence of Lo and Liu with a forecast-evaluation framework. Second, the paper offers methodological transparency for boundary-condition evidence: an information-timing audit with an entry-aligned robustness check, a three-tier separation of full-sample and out-of-sample inference, a dependence-robust null for anomaly-spillover counts, and a frozen-design post-2022 holdout. Third, it contributes a cross-market validation of independent interest: a direct joint test against TAIFEX futures settlement prices shows that the same-strike put–call relative-value wedge moves essentially one-for-one with the index-futures basis (correlation 0.99)—the wedge is the option-market mirror of Taiwan's well-known futures backwardation.

This manuscript is original, has not been published previously, and is not under consideration elsewhere. All authors have approved the submission and declare no conflicts of interest. The option and firm-level data are licensed from the Taiwan Economic Journal (TEJ) and cannot be redistributed, but all other series come from public official sources (TWSE, TAIFEX, the Central Bank of the Republic of China, and the Directorate-General of Budget, Accounting and Statistics), and our complete replication code is available upon request for editorial review.

Thank you for considering our manuscript.

Sincerely,

Xian-Ji Kuang (on behalf of all authors)
Department of Finance, National Changhua University of Education
xjkuang@gm.ncue.edu.tw

Highlights

The Boundaries of Index Option Return Predictability: Evidence from Taiwan

- First systematic Asia-Pacific test of delta-hedged option return forecasts
- Realized variance alone survives full-sample multiple-testing correction
- Predictability is crisis-driven and largely disappears after 2015
- Realized variance underperforms the historical mean in the 2023–2025 holdout
- The put–call wedge closely mirrors the Taiwan index-futures basis

The Boundaries of Index Option Return Predictability: Evidence from Taiwan

Zong-Wei Yeh[a], Pin-Chi Chen[b], Xian-Ji Kuang[b,*], Tzu-Chieh Lin[b]

[a] Department of Money and Banking, National Chengchi University, Taipei, Taiwan

[b] Department of Finance, National Changhua University of Education, Changhua, Taiwan

Author e-mail addresses:

Zong-Wei Yeh 111352502@nccu.edu.tw

Pin-Chi Chen hawan10270827x1@gmail.com

Xian-Ji Kuang xjkuang@gm.ncue.edu.tw

Tzu-Chieh Lin m1367023@gm.ncue.edu.tw

*** Corresponding author.** Xian-Ji Kuang, Department of Finance, National Changhua University of Education, Changhua, Taiwan. E-mail: xjkuang@gm.ncue.edu.tw.

Declarations of interest: no potential conflict of interest was reported by the authors.

Declaration of generative AI and AI-assisted technologies in the writing process: during the preparation of this work, the authors used OpenAI ChatGPT and Anthropic Claude for language proofreading, programming assistance, and reproducibility checks. The authors independently reviewed and verified all text, code, analyses, and references and take full responsibility for the content of the article.

Manuscript (without Author Details) Click here to view linked References

The Boundaries of Index Option Return Predictability: Evidence from Taiwan

Abstract

Out-of-sample evidence on index option return predictability comes almost entirely from U.S. data. We provide, to our knowledge, the first systematic pseudo-real-time out-of-sample evaluation of monthly delta-hedged index-option return forecasts in an Asia-Pacific market, using recursive forecasts of dividend-inclusive returns on at-the-money Taiwan index (TXO) puts over 2002–2022; the mean put premium is about -0.71% per month. What predictability we find runs through the variance risk premium, and it is episodic rather than steady. Realized stock variance (SVAR), the state variable that tracks the premium's physical leg, is the only one of 164 candidates whose full-sample slope survives the Holm correction; its out-of-sample R^2 of 11.03% over the 2007–2022 evaluation window (Clark–West $p = 0.008$) is earned mostly in crisis months and largely disappears after 2015. In a frozen 2023–2025 holdout the decay continues: the put premium shrinks to -0.24% per month and the SVAR forecast falls behind the historical mean. Point estimates concentrate the signal in puts, though formal put–call tests cannot separate the two sides, and the incremental value of predictive timing is economically negligible. Entry-date implied variance fits the target return largely by construction, so we analyze it separately as a channel variable rather than include it in the exogenous-predictor comparisons. Same-strike puts are persistently richer than calls—a wedge that closely tracks the index-futures basis, with a near-unit slope and an economically negligible mean gap: a market-wide carry phenomenon consistent with downside-insurance demand, though alternative carry-based explanations remain open.

Keywords: Index Option Return Predictability, Delta-Hedged Returns, Variance Risk Premium, Put–Call Relative-value Wedge, Taiwan Market

1 Introduction

Index options perform an important risk-transfer function, and their prices reflect at once compensation for volatility risk, tail-risk premia, and hedging demand, which makes them a central setting for asset-pricing research. Coval and Shumway (2001) document that expected returns on U.S. index options are significantly negative, a regularity that frictionless pricing models cannot explain. Bakshi and Kapadia (2003) use the profit and loss of a delta-hedged portfolio as a measurement device and connect it to the variance risk premium (VRP): implied volatility in option markets is persistently higher than realized volatility, reflecting the systematic compensation investors demand for bearing volatility risk. If this compensation adjusts with the macroeconomic state, option returns become predictable in the time series. Bollerslev et al. (2009) show that the VRP predicts future equity excess returns, Drechsler and Yaron (2011) rationalize its time variation within a long-run-risk framework and link it to macroeconomic uncertainty, and Bollerslev et al. (2014) provide cross-country evidence on VRP predictability. Building on this line, Beckmeyer et al. (2023) study U.S. S&P 500 index options, systematically incorporate macroeconomic variables and anomaly factors, and establish an empirical framework in which index option returns are predictable out of sample (OOS).

The existing evidence, however, is concentrated in the United States. Taiwan provides a distinct institutional setting because retail investors play an unusually prominent role in its broader financial markets: Barber et al. (2009) use complete transaction records to document that individual investors sustain systematic and economically large losses over time. This makes TXO a useful setting for evaluating external validity without requiring retail participation to be the identified mechanism. If a realized-variance state variable such as SVAR predicts TXO returns, the U.S. evidence travels; if not, its external validity is limited. Cross-sectional anomaly signals let us ask, separately, whether demand-side forces transmit into index option prices.

Our aim is not only to record whether predictability exists but to delimit the range within which it holds. We take the monthly delta-hedged return on the ATM TXO put over January 2002 to December 2022 as the target variable (236 valid months after screening, 176 OOS months), and we reconstruct the target from a complete self-financing profit-and-loss identity that includes a dividend cash-flow term. The Taiwan Capitalization Weighted Stock Index is a price index; if the dividend income of the hedging spot leg is omitted, the profitability of the put writer is overstated. After the correction, the mean ATM put delta-hedged return rises from about -1.0% to about -0.71% per month, which serves as the benchmark for every magnitude statement in this paper.

We organize the findings by their evidentiary weight rather than by novelty. *The primary result is that predictability is sparse and episodic.* A single realized-variance state variable (SVAR) accounts for most of the full-sample OOS gain: of 164 candidates it is the only one whose full-sample slope survives the Holm–Bonferroni correction. Over the 2007–2022 recursive evaluation window, its unadjusted out-of-sample R^2 is 11.03% (Clark–West $p = 0.008$). That gain, however, is concentrated in stress periods: it falls to 5.35% once November 2008 is excluded and to 0.22% once the five largest-gain months are removed, and the 2015–2022 subsample delivers only 0.69%. Full-sample statistical evidence is strong, but economically and temporally concentrated: the signal survives Clark–West significance in every crisis-exclusion variant while shrinking sharply in economic magnitude.

Two additional findings sharpen this conclusion. Contract-level evidence is *put-*

concentrated: point-estimated OOS power is stronger on puts than on calls and rises toward the money, though a formal same-strike put–call test is not significant (Section 6). And the variance-pricing channel, measured by entry-date implied variance (IV^2), is largely mechanical because IV^2 is the entry pricing input of the contract being predicted; we therefore analyze its explanatory power separately as a channel/accounting variable rather than include it in the exogenous-predictor horse race (Section 5.3).

The evidence also reveals four boundaries, each with an explicit empirical criterion. The *predictor boundary* is the joint outcome of the full-sample Holm family (SVAR alone survives) and the OOS Clark–West family (no predictor survives the same correction, so we make no multiple-testing-corrected OOS claim; Section 4). The *contract boundary* is a common-sample moneyness grid with pairwise put–call tests (the smile gradient weakens sharply and the put–call difference is not formally separable; Section 6). The *time boundary* is the crisis-exclusion, leave-one-year-out, and fluctuation-test suite, together with a frozen post-2022 holdout and an extended 2002–2025 full-sample re-estimation that anchors it (Section 8). The *complexity boundary* is a same-information-set comparison of nonlinear alternatives, a forecast-encompassing test, and a calibrated power analysis (Section 8.5). Anomaly spillover and the available retail-participation proxies fall outside these boundaries as well (Section 8).

A separate price-based test provides complementary evidence. Same-strike puts are systematically richer than calls—a same-strike put–call relative-value wedge. A direct joint test against TAIFEX futures settlement prices shows this wedge closely tracks the Taiwan index-futures basis, with a correlation of 0.99 and a near-unit regression slope: it is the backwardation's option-market mirror. The wedge is independent of return prediction and is *consistent with* downside-insurance demand, though it does not uniquely identify it.

We make three contributions. First, to our knowledge we provide the first systematic pseudo-real-time out-of-sample evaluation of monthly delta-hedged index-option return forecasts in an Asia-Pacific market—filling a literature that focuses on the United States and complementing conditional option-writing evidence for Taiwan (Lo and Liu, 2025)—and we provide a cross-market validation of the price structure: the put–call relative-value wedge of delta-hedged P&L, which a direct test ties to the index-futures basis. Second, we use a mechanical decomposition to separate the "forecasting future variance" component of the IV^2 channel from its mechanical variance-risk-premium component, avoiding reading a definitional correlation as predictive power. Third, through linear diagnostics, a forecast-encompassing test, a wide hyperparameter search over the tree family, and a calibrated Monte Carlo power analysis, we delimit the complexity and sample-length boundaries of predictability, turning negative results into informative evidence.

The remainder of the paper is organized as follows. Section 2 reviews the related literature and institutional setting; Section 3 describes the data and target-variable construction; Section 4 presents the empirical framework; Section 5 reports the predictability results and decomposes the variance-pricing channel; Section 6 presents the contract-level evidence, from the moneyness grid to the put–call wedge and its futures-basis validation; Section 7 evaluates economic value; Section 8 collects the robustness evidence—time concentration, the frozen holdout, the retail tests, anomaly spillover, and the complexity boundary; Section 9 concludes.

2 Related Literature and Institutional Setting

2.1 Option Returns and Risk Compensation

The theory of option returns rests on the no-arbitrage pricing framework of Black and Scholes (1973) and Merton (1973). Under complete markets, continuous trading, and frictionless hedging, options can be dynamically replicated by the underlying and the risk-free asset and offer no systematic excess return in theory. Real markets, however, contain stochastic volatility, jump risk, discrete hedging, and trading frictions, so that even a delta-hedged option can bear higher-order risk and earn observable risk compensation. Coval and Shumway (2001) document that U.S. index option strategies still earn significantly negative returns after directional exposure is removed; Bakshi and Kapadia (2003) establish the theoretical link between delta-hedged P&L and the VRP, showing that when implied volatility is systematically higher than realized volatility, a delta-hedged long option earns a negative return over time while the writer is compensated for bearing volatility and jump risk. Carr and Wu (2009) and Kilic and Shaliastovich (2019) further characterize the structure of the VRP, the latter decomposing it into upside and downside components and relating the downside variance premium to compensation for downside protection on the put side.

At the cross-sectional level, Goyal and Saretto (2009) find that the gap between implied and historical volatility of individual equity options predicts future option returns; Cao and Han (2013) document the relation between idiosyncratic volatility and delta-hedged option returns; Zhan et al. (2022) use firm characteristics to systematically predict delta-hedged option returns, the closest cross-sectional counterpart to our study; Bali et al. (2023) characterize option return predictability with machine learning and large datasets, providing a methodological template; and Heston et al. (2023) document the long-run predictability of option momentum. Bali and Murray (2013) show that risk-neutral skewness predicts the returns of equity option portfolios, providing direct support for the link between skewness-related signals and option returns. For index options specifically, Chambers et al. (2014) note that pricing their expected returns remains challenging, which frames our focus on an index (rather than single-name) underlying.

Demand-side forces may also shape option returns. Garleanu et al. (2009) build a demand-based theory of option pricing, showing that when the option market exhibits a systematic demand imbalance, market makers cannot fully diversify their inventory and hedging costs and charge an additional premium, so prices deviate from the frictionless benchmark. Garleanu et al. (2009) supply the theory; its original evidence is from the U.S. market, and we cite it as the source of a mechanism hypothesis rather than as Taiwan evidence. Muravyev (2016) uses order flow to document the demand channel's effect on expected option returns; Han (2008) documents the effect of investor sentiment on the index option implied distribution; and Chordia et al. (2021) find that index option trading activity predicts market returns, linking the demand side to aggregate predictability.

2.2 Forecast Evaluation and Predictor Construction

The return-predictability literature was built first on equity markets (Campbell, 1987; Campbell and Shiller, 1988), and after Goyal and Welch (2008) raised a comprehensive challenge to out-of-sample validity, it turned toward out-of-sample evaluation and the R^2_{OOS} metric. Campbell and Thompson (2008) show that imposing economic restrictions improves OOS performance, and Rapach et al. (2010) propose forecast combination to

reduce model uncertainty. When the candidate set is large, the literature develops shrinkage, selection, and dimension-reduction methods: the LASSO of Tibshirani (1996), the elastic net of Zou and Hastie (2005), the three-pass partial least squares of Kelly and Pruitt (2015), and the targeted principal components of Bai and Ng (2008); Gu et al. (2020) systematically evaluate machine learning in empirical asset pricing.

In the context of using anomaly factors to predict *aggregate market returns*, Dong et al. (2022) is the flagship methodological parent: it extracts the predictable component of aggregate equity returns from a large set of cross-sectional anomalies, and it is the methodological blueprint for our question of whether anomaly factors translate into an aggregate time-series signal for index option returns; Chen and Velikov (2023) provide a method for estimating anomaly expected returns. We clarify that Dong et al. (2022) target *equity market* returns, whereas we transport the method to index option returns; the underlyings differ, and we cite it as a methodological paradigm rather than a conclusion analogy. On the direct framework for option return prediction, Beckmeyer et al. (2023) study U.S. index option delta-hedged returns, systematically test macroeconomic variables and anomaly factors, and compare six multivariate methods, which provides our closest forecasting benchmark. We adapt their forecasting architecture to TXO and extend it through a dividend-inclusive P&L identity, an implied-variance channel decomposition, contract-level tests, futures-basis validation, and nonlinear diagnostics. Because Taiwan lacks a standardized option-anomaly dataset, we do not include that class of factors.

2.3 Taiwan Market Structure and TXO Index Options

Taiwan provides an empirical setting distinct from the United States. Barber et al. (2009) use complete transaction records to document that individual investors sustain systematic losses over time, revealing pronounced retail behavior in the Taiwan market; Kumar (2009), Barberis and Huang (2008), and Mitton and Vorkink (2007) document retail preference for high-skewness (lottery-type) assets, providing a behavioral basis for skewness-related demand. Bryzgalova et al. (2023) use U.S. data to directly document a retail-dominated option market structure (the retail share once exceeded 60%), which provides an empirical anchor for a "high-retail-share option market." For Taiwan index options specifically, Lee and Wang (2016) (PBFJ) study the information content of TXO investor trading behavior, Yang et al. (2016) (PBFJ) examine the effects of sentiment, weather, and disaster on hedging performance, and Chang et al. (2009) and Chang et al. (2010) (JBF) document the information content of informed option investors and of option trading volume in Taiwan. Together these studies characterize Taiwan as a market with relatively high retail participation and distinctive local institutional features. In closely related work on the same market, Lo and Liu (2025) (PBFJ) show that a publicly observable sentiment signal—the option put–call ratio—carries market-timing value for index-option writing in Taiwan: conditioning buy-write, put-write, covered-combo, and delta-hedged call-writing strategies on the put–call ratio, and writing only when the ratio is low, raises their Sharpe and Sortino ratios and CAPM alphas over March 2007–December 2023 while sharply reducing trading frequency. Their design is a single-signal conditional trading rule evaluated by realized strategy performance. We differ in both question and method: we conduct a systematic pseudo-real-time out-of-sample evaluation of monthly delta-hedged index-option return forecasts—screening 164 candidate predictors under a multiple-testing hierarchy, an R^2_{OOS} metric against the prevailing-mean benchmark, and a frozen post-2022 holdout—and we take the ATM *put* delta-hedged return, rather than a call-writing timing

strategy, as the forecast target.

Retail structure serves as institutional background and motivation rather than an identified mechanism: TXO is a market with relatively high individual participation, and we do not claim that retail demand drives or blocks any predictability channel. As the negative identification tests in Section 8 show, both direct identifications of a retail channel (time-series state dependence and a retail/institutional within-group split) go unsupported, so any retail mechanism appears only as a candidate explanation consistent with the timing evidence and is not advanced as an identified causal claim.

2.4 Gap and Positioning

Taken together, the OOS evidence on index option return predictability comes mainly from the United States, and Asia-Pacific markets with different investor structures still lack a direct test; whether cross-sectional anomaly factors translate into an index option return signal outside the United States has likewise not been tested directly. We use Taiwan index options as the setting, provide pseudo-real-time OOS evidence on these two questions, and put the emphasis on *delimiting the boundaries of predictability*: the cross-sectional (moneyness) distribution of the signal, its time concentration, its complexity ceiling, and the statistical identifiability of anomaly spillover. This "predictability and its boundaries" framing makes negative and qualifying results equally informative.

3 Data and Return Construction

3.1 The TXO Market and Sample Screening

The Taiwan Futures Exchange (TAIFEX) launched options on the Taiwan Capitalization Weighted Stock Index (TXO) in December 2001, with formal trading from January 2002, making it one of the most actively traded index option markets in the Asia-Pacific region. Our sample spans January 2002 to December 2022 and includes major events—the 2003 SARS outbreak, the 2008 global financial crisis, the 2011 European debt crisis, and COVID-19 in 2020—which helps in observing the time-varying character of the risk premium. Raw daily data are from the Taiwan Economic Journal (TEJ).

To keep illiquidity, stale quotes, and abnormal prices from contaminating return measurement, we impose the following filters: (i) the bid is strictly positive; (ii) the ask exceeds the bid; (iii) implied volatility is not missing; (iv) open interest is strictly positive; (v) the relative bid–ask spread $(\text{Ask} - \text{Bid})/\text{Mid} \leq 50\%$; and (vi) prices satisfy no-arbitrage and convexity restrictions. We exclude weekly options and restrict moneyness $S_t/K \in [0.8, 1.2]$. Contract selection takes the third Thursday of each month as the entry date, and among contracts expiring on the third Wednesday of the following month with more than 20 calendar days to maturity, we select the put closest to the money; if no exactly at-the-money contract exists, we take one with $K \leq S_t$ to maintain consistency. After filtering, we exclude 16 months (1 for an insufficient valid-quote path length and 15 for valid-quote interruptions exceeding 7 calendar days within the holding period, 14 of which occur around the Lunar New Year), leaving 236 valid monthly observations, of which 176 fall in the OOS evaluation period (June 2007 to December 2022).[1]

[1]The dropped months do not introduce a detectable sample-selection bias: 15 of the 16 are January, when the Lunar New Year holiday interrupts the daily hedging path. Comparing dropped versus retained months on the contemporaneous market return and realized volatility, neither differs significantly (Welch

3.2 Dividend-Inclusive ATM Put Delta-Hedged Returns

We accumulate daily marked-to-market P&L over a one-month holding period. Let t be the entry date and $t+1$ the settlement date, with N daily hedging intervals over the holding period. Relative to a basic P&L identity that counts only the option and financing cost, we **add the dividend cash-flow term of the hedging spot leg**. The delta-hedged portfolio is a long option plus $(-\Delta)$ units of the spot: for a put, $\Delta < 0$ makes the spot position long, so it receives dividends, and the daily dividend term is positive (the opposite for a call). The accumulated daily hedging P&L path is

$$\Pi_{t+1} = \sum_{n=0}^{N-1} \left[(P_{t_{n+1}} - P_{t_n}) - \Delta_{t_n}(S_{t_{n+1}} - S_{t_n}) - r_{t_n} \tfrac{\delta\tau_n}{365}(P_{t_n} - \Delta_{t_n} S_{t_n}) - q_t \, \Delta_{t_n} S_{t_n} \tfrac{\delta\tau_n}{365} \right], \tag{1}$$

where P_{t_n} is the put mid-price, S_{t_n} is the weighted-index spot, Δ_{t_n} is the put delta recomputed daily under the Black–Scholes model (with dividend yield q_t: d_1 contains $r - q$ and $\Delta_{\text{put}} = -e^{-q_t T} N(-d_1)$), r_{t_n} is the annualized one-year time-deposit rate,[2] and q_t is the annualized dividend yield of the entry month (held fixed over the holding period). The penultimate term is the financing cost, and the last term is the dividend cash flow of the hedging spot leg. The monthly delta-hedged return is

$$R^{dh}_{t+1} = \frac{\Pi_{t+1}}{|P_t - \Delta_t S_t|}. \tag{2}$$

In a frictionless Black–Scholes world, equation (1) is identically zero; its departure from zero is the variance (and jump) risk premium realized by the delta-hedged put. Relative to the basic identity that ignores dividends, adding the dividend term yields a complete self-financing P&L identity.[3]

The dividend term is a one-sided positive item for the put and directly raises the mean return: including it moves the mean ATM put delta-hedged return from about -1.0% to -0.71% per month, so a dividend-omitting identity overstates the writer's profitability by about 0.3 percentage points; the core fact that the VRP is positive is unchanged, and -0.71% remains of the same magnitude as the U.S. ATM delta-hedged put premium (about -0.5 to -1% per month).

3.3 Predictors

Macroeconomic predictors (11). Following Goyal and Welch (2008) and Beckmeyer et al. (2023), these cover the dividend–price ratio (DP), earnings–price ratio (EP), book-

$p = 0.44$ and 0.81; Mann–Whitney $p = 0.32$ and 0.62). Market trading volume is not tested because the daily index file carries no volume field.

[2]We use the one-year time-deposit rate because our data set does not include a short-tenor Taiwan money-market rate (overnight/1M/3M); the choice has negligible effect on monthly excess returns, since the deposit rate is only about 1–2.5% annualized (well under 0.2% per month) against a put premium of about 0.7% per month, and it enters the financing term linearly without affecting the predictive slopes.

[3]Under a risk-neutral simulation of geometric Brownian motion with a continuous dividend, the per-unit-of-investment P&L in equation (1) has a mean of about zero (no double-counting of interest, no missing term); omitting the dividend term produces a systematic bias consistent with the theoretical missing term. The q series is constructed from the implied realized dividend yield of the total-return versus price index from 2003 on, preserving the ex-dividend seasonality (a July mid-year mean of about 22%, near zero in non-ex-dividend months); 2002 is backfilled by smoothing the market's trailing dividend yield (that year's q is only about 2.7–4.6%, with limited effect), and this measurement difference is disclosed in the robustness analysis.

to-market ratio (BM), dividend–earnings ratio (DE), dividend yield (DY), net stock issuance (NTIS), realized stock variance (SVAR), term spread (TMS), inflation (INFL), long-term government bond yield (LTBY), and the risk-free rate (RF); full definitions are in Appendix A. **Equity anomaly factors (153).** These are drawn from the Taiwan-listed sample of the Jensen et al. (2023) global factor database, using the monthly-rebalanced, value-weighted long–short decile portfolio return as the time-series signal. All 164 predictors are lagged one period uniformly, so the information vector $\mathbf{x}_t$ available before the entry date of month t predicts R^{dh}_{t+1}.

3.4 Descriptive Statistics

Table 1 reports descriptive statistics for the target variable and the macroeconomic predictors. The monthly ATM put delta-hedged return has a mean of -0.707%, a standard deviation of 1.698%, skewness of 0.054, and excess kurtosis of 3.999. Its distribution is decidedly left-heavy in outcome even though its mean is only mildly negative: only 23.7% of months are positive, the median is -0.640%, the interquartile range is $[-1.540\%, -0.045\%]$, and the tails run from a 5th percentile of -3.303% and a minimum of -8.576% (November 2008) to a 95th percentile of $+2.069\%$ and a maximum of $+6.534\%$—the writer collects a small premium in most months and pays out in occasional crashes. SVAR displays relatively low time-series persistence (AC(1) = 0.542), consistent with mean reversion in realized variance; the other macro variables are mostly highly persistent, which may mechanically raise in-sample fit yet need not deliver stable out-of-sample predictive power—this motivates our use of strict pseudo-real-time OOS evaluation. Figure 1 shows that the return series fluctuates in a narrow band around its -0.7% mean over most of the sample and takes large negative jumps in crises, indicating an intermittent signal.

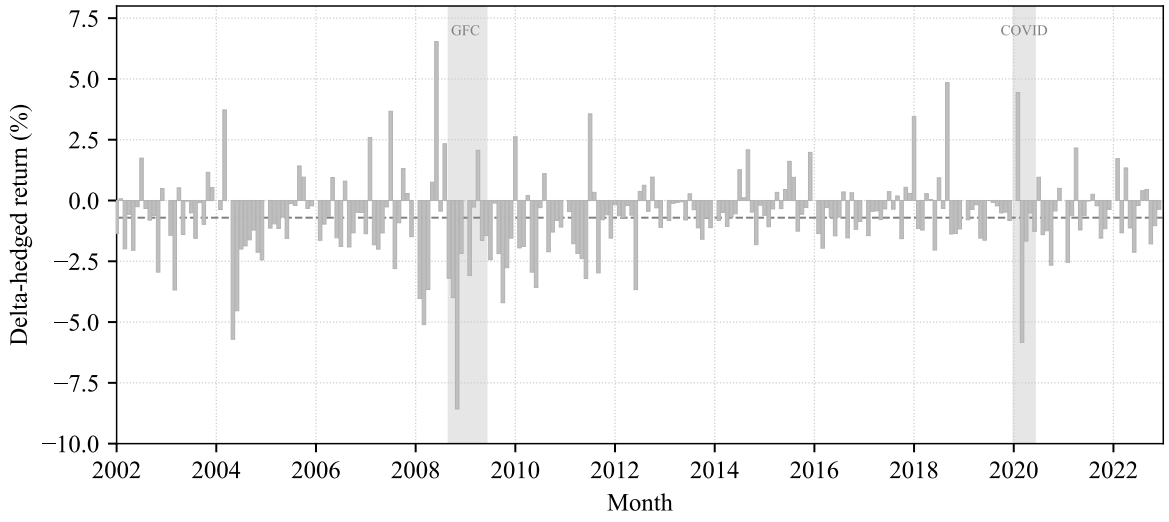


Figure 1: Time series of monthly ATM put delta-hedged returns, dividend-inclusive ($T = 236$)
Notes: Shaded bands mark the global financial crisis (2008/09–2009/06) and COVID-19 (2020/01–06). The horizontal dashed line is the full-sample mean (-0.707%).

Table 1: Descriptive statistics

	N	Mean	SD	10%	Q1	Q2	Q3	90%	Skew	Kurt	AC(1)
Panel A: Option returns (dividend-inclusive)											
R^{dh} (%)	236	−0.707	1.698	—	−1.540	−0.640	−0.045	—	0.054	3.999	0.073
Panel B: Macroeconomic predictors											
DP	236	4.217	1.087	3.105	3.545	4.195	4.573	5.325	2.400	10.137	0.924
EP	236	5.693	1.951	2.393	4.915	6.040	6.584	7.776	−0.438	0.636	0.948
BM	236	0.585	0.079	0.499	0.549	0.581	0.617	0.669	1.059	4.245	0.897
DE	236	0.875	0.526	0.551	0.618	0.714	0.879	1.412	2.986	9.514	0.913
DY	236	4.220	1.010	3.127	3.602	4.161	4.584	5.261	2.103	9.035	0.924
NTIS	236	2.936	2.314	0.560	1.167	2.399	3.730	6.688	0.939	−0.096	0.974
SVAR	236	0.299	0.332	0.073	0.101	0.179	0.349	0.712	2.641	8.513	0.542
TMS	236	0.153	0.501	−0.340	−0.191	−0.007	0.328	0.885	1.351	1.258	0.957
INFL	236	0.097	0.654	−0.743	−0.226	0.144	0.468	0.823	−0.109	0.771	−0.086
LTBY	236	1.552	0.776	0.665	1.038	1.440	1.965	2.655	0.855	0.741	0.984
RF	236	1.399	0.525	0.770	1.035	1.355	1.525	2.245	0.972	0.092	0.986

Notes: Panel A is the dividend-inclusive monthly ATM put delta-hedged return (see Section 3.2 and equation (1)), with a mean of −0.707% and a standard deviation of 1.698%. For Panel A, Q1/Q2/Q3 are $-1.540/-0.640/-0.045$; the 5th and 95th percentiles are −3.303 and +2.069, the minimum and maximum are −8.576 and +6.534, and the frequency of positive months is 23.7% (Skew/Kurt are bias-corrected, Kurt is the excess value). The 10%/90% quantile columns are shown for the macro variables only. Descriptive statistics for the macro variables use the full sample and are all lagged one period. AC(1) is the first-order autocorrelation.

4 Empirical Framework

4.1 Economic Motivation and the Leading Predictable Term

Following Bakshi and Kapadia (2003), the expected return of a delta-hedged long put position can be approximately decomposed, in continuous time, as

$$\mathbb{E}\big[R^{dh}_{t+1}\big] \approx -\tfrac{\tilde{\Gamma}_t}{2}\,\mathrm{VRP}_t - J_t + \varepsilon_{t+1}, \tag{3}$$

where $\tilde{\Gamma}_t \equiv \Gamma_t S_t^2/|P_t| > 0$ is the standardized dollar gamma, $\mathrm{VRP}_t \equiv \sigma^2_{IV,t} - \sigma^2_{RV,t}$, and J_t is the jump compensation. The expression is only heuristic, but it reveals the key pricing logic: variables related to volatility dynamics or to the time variation of the VRP should carry intertemporal information about R^{dh}_{t+1}. SVAR directly captures the realized-variance component of the VRP and is the most natural candidate predictor (Bollerslev et al., 2009; Drechsler and Yaron, 2011). The leading predictable term in equation (3) is linear in the VRP, which *motivates* a parsimonious linear variance-based benchmark; whether any nonlinear alternative reliably improves on that benchmark is an empirical question we take up in Section 8.5, and we do *not* claim that theory rules out nonlinearity.

4.2 A Three-Tiered Inference Scheme

Because we screen 164 predictors and then evaluate them out of sample, we keep three layers of inference strictly separate and label every claim by the layer on which it rests.

Tier 1: full-sample slope inference. We estimate $R^{dh}_{t+1} = \alpha + \beta\, x_{i,t} + \varepsilon_{t+1}$ ($i = 1, \ldots, 164$) by OLS over the full in-sample period, with Newey and West (1987) HAC

standard errors (4 lags).[4]

Tier 2: full-sample familywise correction. Because the 164 slopes are examined jointly, we apply the sequentially rejective Holm (1979) procedure to *the family of 164 full-sample Newey–West tests*, controlling the familywise error rate at 10%. SVAR is the *only* predictor that survives (Holm $p = 0.009$).

Tier 3: out-of-sample inference. We evaluate each predictor recursively (Section 4.3) and assess it with the one-sided Clark and West (2007) statistic. This OOS family is a *distinct* testing problem, and we ask explicitly whether any OOS finding survives multiple testing. Applying the same Holm and Benjamini–Hochberg (Benjamini and Hochberg, 1995) corrections to the 164 OOS Clark–West p-values, **no predictor survives**: although SVAR has the smallest raw OOS p-value (0.008, rank 1 of 164), its Holm-adjusted p is 1.00 and its Benjamini–Hochberg p is 0.775, and neither correction leaves any survivor at the 10% or 5% level. **We therefore make no claim of multiple-testing-corrected out-of-sample predictability.** The force of the OOS evidence comes instead from the concentration of the ex ante hypothesis set—SVAR is the unique Tier-2 survivor, and its OOS performance is the strongest in the (uncorrected) family—rather than from surviving an OOS family correction.

We add that SVAR is comparatively insulated from the small-sample predictive-regression bias that afflicts highly persistent regressors (the Stambaugh bias): its first-order autocorrelation is only 0.54, against 0.90–0.95 for the dividend–price, earnings–price, and related valuation ratios.

4.3 Out-of-Sample Evaluation and Inference

Our forecasts are **pseudo-real-time**: at each origin we re-estimate all parameters recursively on an expanding window using only information dated before that month's entry date, but the evaluation is not strictly real-time in that it uses the finally revised (rather than first-vintage) macro series; the timing audit of Section 4.4 verifies that no target-date information leaks into the predictors. Following Campbell and Thompson (2008), we measure predictive power relative to the prevailing-mean benchmark with $R^2_{\text{OOS}} = 1 - \sum(R^{dh}_{t+1} - \hat{R}^{dh}_{t+1})^2 / \sum(R^{dh}_{t+1} - \bar{R}^{dh}_{t})^2$, where $\bar{R}^{dh}_{t} = \frac{1}{t-1}\sum_{s<t} R^{dh}_{s}$ is the prevailing mean computed *only* from returns realized before the forecast origin t; we use an initial estimation window of 60 months and expand it month by month, with the evaluation period June 2007 to December 2022 (176 months). For models *nested* within the benchmark, we use the one-sided Clark and West (2007) (CW) MSPE-adjusted statistic. For *non-nested* comparisons (such as tree models against a linear benchmark), the CW test does not apply, and we use the one-sided Diebold–Mariano (DM) test (HAC bandwidth). Every CW/DM table states the loss-differential definition, the null, and whether the test is one- or two-sided.

4.4 Information Timing and the Pseudo-Real-Time Protocol

Because we describe the forecasts as pseudo-real-time, we audit the information timing explicitly against the code that builds each series. The option position is entered on the **third Thursday** of each month (deferred to the next trading day when that is

[4]The lag choice is not material: SVAR's Newey–West t is -3.84, -4.04, -4.38, and -4.26 at 1, 4, 6, and 12 lags, respectively—all highly significant—and we report the conservative middle value of 4 lags.

a holiday: 229 of 236 entries fall on a Thursday, 7 are deferred), so the entry day-of-month has a median of 18 and ranges over 15–30. The macro predictors—SVAR, the valuation ratios, NTIS, the term spread, and the rates—are measured over the *prior* calendar month ($M-1$) and dated at its month-end; because the code lags them one month, the measurement window ends a median of 18 days (range 15–30) *before* the entry date. There is thus no look-ahead, and the convention is if anything conservative, since it discards the roughly three weeks of information between the month-end and the entry date. The one predictor with a genuine release lag is inflation: Taiwan's CPI for month $M-1$ is published around the fifth-to-eighth business day of month M, still before the third-Thursday entry. Entry-date implied variance IV_t^2 is measured *on* the entry day and is therefore an entry condition (a channel/accounting variable), not an exogenous predictor, and $\mathrm{VRP}_t = \mathrm{IV}_t^2 - \mathrm{RV}_{M-1}$ combines two quantities both known at entry. Figure 2 summarizes the sequence.

As a robustness check we replace the calendar-month SVAR with a strict entry-aligned realized variance computed over the 21 trading days ending on the entry date (still using only pre-entry information); its OOS R^2 is *higher*, at 18.66% (CW $p = 0.005$, correlation 0.77 with the paper version), confirming that our monthly convention introduces no look-ahead and, if biased at all, understates predictability.

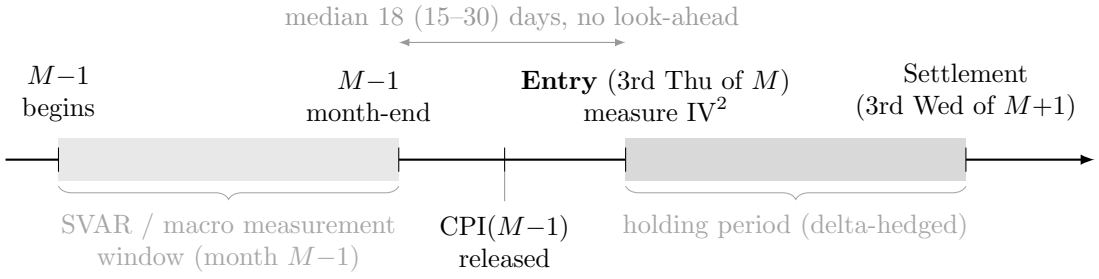


Figure 2: Information-timing sequence for the pseudo-real-time forecasts

Notes: Predictors dated at the $M-1$ month-end lead the entry date (the third Thursday of month M) by a median of 18 days (range 15–30). Inflation, the only predictor with a genuine publication lag, is released before entry. IV^2 is measured on the entry date and is treated as a channel/accounting variable rather than an exogenous predictor. The delta-hedged position is held from the entry date to the settlement date (the third Wednesday of month $M+1$).

4.5 Multivariate and Nonlinear Methods

We evaluate six multivariate methods, all estimated recursively on an expanding window using only information available up to the current period: the elastic net (ENet; Zou and Hastie, 2005), a selective forecast combination (Combine, retaining variables with in-sample $|t| > 3$ and equal-weighting them; Stock and Watson, 2004; Timmermann, 2006), the forecast average (P-Aver), hard- and soft-threshold targeted principal components (tPCA-h/tPCA-s), and the Kelly–Pruitt three-pass partial least squares (KP-PLS; Kelly and Pruitt, 2015). To delimit the complexity boundary, we also evaluate the random forest, gradient boosted trees (XGBoost), the generalized additive model (GAM), and kernel ridge regression (KRR), whose hyperparameters are selected within each window by time-series cross-validation, with inference by DM (see Section 8.5).

4.6 The Implied-Variance Channel and Its Mechanical Component

We also construct the entry-date ATM implied variance IV_t^2 (the squared, monthized average of the near-month ATM call and put implied volatilities) and define the **variance-risk-premium proxy** $\mathrm{VRP}_t = \mathrm{IV}_t^2 - \mathrm{RV}_t$, where RV_t is the prior calendar month's realized variance (constructed from the same realized-variance series as SVAR, correlation 0.9998). We call this a *proxy* deliberately: the theoretical variance risk premium is the risk-neutral expected variance minus the physical expected variance, whereas $\mathrm{IV}_t^2 - \mathrm{RV}_t$ is an implied-minus-lagged-realized variance spread in which lagged realized variance stands in for the physical expected variance. We define its nature up front: because IV_t^2 is the entry pricing input of *the very contract being predicted* (the target return's investment, delta, and entry price are all determined by the same IV), the Bakshi and Kapadia (2003) decomposition gives $R_{t+1}^{dh} \approx c\,(\mathrm{RV}_{t+1} - \mathrm{IV}_t^2)$ with $c > 0$, so the prediction of the return by IV_t^2 contains a mechanical component *by construction*. We therefore position IV^2 as a *direct measure of the variance-risk-premium channel* and characterize its nature through a mechanical decomposition in Section 5.3. Accordingly, IV^2 and VRP are excluded from the exogenous-predictor tables and from the paper's primary predictability claims; we retain them in the separate channel, economic-value, holdout, and complexity analyses, where they are labeled as channel variables.

4.7 Economic Value

We use a mean–variance active allocation, $w_t = \hat{R}_{t+1}^{dh}/(\gamma\hat{\sigma}_t^2)$, where $\hat{\sigma}_t^2$ is the rolling variance of the prior 24 months, weights are capped at $\pm 200\%$, and $\gamma = 10$. We measure performance by the annualized certainty equivalent return, $\mathrm{CER} = \hat{\mu}_{port} - \frac{\gamma}{2}\hat{\sigma}_{port}^2$, with the Sharpe ratio as a supplement. Transaction costs are accounted for on the correct notional (see Section 7): the bid–ask spread applies only to the *option premium* actually traded (about 4% of the position's investment notional), not to the entire investment including the hedging leg, and we also add the daily delta-rehedging cost of the futures leg.

5 Predictability and the Variance-Pricing Channel

5.1 Univariate: Predictability Is Highly Concentrated

Table 2 reports univariate results for the core predictors. Among the 11 macro variables, SVAR is the only one with both economic magnitude and statistical robustness: $\hat{\beta} = -1.735$, NW $t = -4.04$, $R^2_{\mathrm{IS}} = 12.00\%$, $R^2_{\mathrm{OOS}} = 11.03\%$, CW $p = 0.008$, and it is the *only* one of the 164 candidates that remains significant under the Holm–Bonferroni correction (adjusted $p = 0.009$). Its negative slope is consistent with the VRP mechanism: when market volatility rises, writers demand higher compensation, depressing the buyer's subsequent delta-hedged return. The result is robust to the hedging-delta convention: rebuilding the target with a Black–Scholes delta that omits the dividend yield leaves SVAR's OOS R^2 essentially unchanged at 11.77% (NW $t = -4.05$, CW $p = 0.010$) against 11.03% for the dividend-adjusted target, and the two target-return series correlate 0.955.

Among the 153 anomaly factors, idiosyncratic skewness (*iskew_capm_21d*) is the strongest in sample ($\hat{\beta} = -0.198$, $t = -3.19$, $R^2_{\mathrm{IS}} = 7.01\%$) and is significant under the Clark–West OOS test ($R^2_{\mathrm{OOS}} = 5.89\%$, CW $p < 0.01$). It is nonetheless a secondary

signal. It fails the Holm correction over the 164 candidates ($p = 0.234$); 62.9% of its OOS gain accrues in November 2008 alone, and the 2015–2022 subsample is negative (−0.70%, with the failing months concentrated in the ex-dividend season). Its transmission, moreover, runs through large-cap stocks that overlap heavily with the TAIEX constituents: the signal vanishes under equal weighting and is fully replicated by a size split (Section 8), so we read iskew as an index-level common-skewness state rather than as evidence about investor structure.

Table 2: Univariate predictability

Predictor	$\hat{\beta}$	NW t	R^2_{IS} (%)	R^2_{OOS} (%)	CW p	Holm p
SVAR (realized stock variance)	−1.735	−4.04	12.00	11.03	0.008	**0.009**
iskew_capm_21d (idiosyncratic skewness)	−0.198	−3.19	7.01	5.89	0.009	0.234

Notes: $T_{\text{IS}} = 236$, $T_{\text{OOS}} = 176$ (2007/06–2022/12), initial window 60 months. NW t uses 4 lags. Holm p is the familywise-corrected value over all 164 candidates; only SVAR remains significant at 10%. The table contains exogenous predictors only; the channel measures IV^2 and VRP, which are mechanically linked to the target, are analyzed separately in Section 5.3.

5.2 Multivariate: The Macro Set Is Stably Positive; Anomalies Only Dilute

Table 3 reports the complete OOS panel of the six multivariate methods over the three exogenous predictor sets. On the macro set (Panel A), all six methods have positive R^2_{OOS} (from 2.70% to 8.53%) with all CW $p < 0.06$; tPCA-h is highest (8.53%) and Combine second (8.36%), the latter mainly through its implicit sparsity: SVAR's in-sample $|t|$ is well above the selection threshold $T^* = 3$ and it is stably selected at every start point.[5] On the anomaly set (Panel B), **five of the six models are negative and one positive**—only Combine is positive (4.86%, CW $p = 0.074$, marginal at best)—so the anomaly set does not translate into stable OOS gains. The combined set (Panel C) is on par with the macro set (best: Combine 9.58%): the anomaly factors provide *no systematic or model-robust incremental improvement* beyond the macro set (a single specification, Combine, edges above its Macro-11 counterpart; the other five do not). The anomaly boundary is treated systematically in Section 8 with factor-strength and spillover tests; the channel-accounting specification that adds IV^2/VRP to the macro set is analyzed in Section 5.3.

5.3 The Implied-Variance Channel and Its Mechanical Component

We draw a sharp distinction between two classes of variables. **Class A** comprises *exogenous predictors*—SVAR, the other macro variables, and the anomaly factors—which are not mechanically embedded in the target return and can legitimately race against one another. **Class B** comprises *payoff-accounting / channel variables*—IV^2, the variance-risk-premium proxy, and the Macro-13 specification that contains them—which mechanically embed entry-price information and are therefore analyzed separately from the exogenous-predictor horse race. The paper's primary exogenous-predictor claims rest on Class A;

[5] Combine's OOS R^2 is insensitive to the inclusion threshold: it is 8.26%, 8.36%, and 8.36% at $|t| > 2.5$, 3.0, and 3.5.

Table 3: Multivariate out-of-sample predictability: three predictor sets × six models

	ENet	Combine	P-Aver	tPCA-h	tPCA-s	KP-PLS
Panel A: Macro variables (Macro-11, $p = 11$)						
R^2_{OOS} (%)	5.96	8.36	6.55	**8.53**	6.36	2.70
CW p	0.042	0.028	0.039	0.017	0.023	0.058
Panel B: Equity anomaly factors ($p = 153$)						
R^2_{OOS} (%)	−1.08	4.86	−1.24	−1.98	−2.22	−2.48
CW p	0.157	0.074	0.058	0.057	0.095	0.154
Panel C: Combined set ($p = 164$)						
R^2_{OOS} (%)	0.87	9.58	4.58	5.61	6.50	−0.64
CW p	0.002	0.029	0.016	0.012	0.030	0.128

Notes: R^2_{OOS} (%) and one-sided Clark–West p-values for the three exogenous predictor sets × six methods; evaluation period 2007/06–2022/12, initial window 60; the prevailing-mean benchmark has $R^2_{\text{OOS}} = 0$ by definition. (a) The CW test measures the nested-adjusted improvement in population MSPE; for heavily regularized models it can be significant even when the sample R^2_{OOS} is near zero (e.g., ENet in Panel C: 0.87%, $p = 0.002$), and should be read as "the forecast contains information beyond the prevailing mean" rather than "a sizable out-of-sample gain." (b) The panel is insensitive to the dividend treatment of the target except for tPCA-s in Panel B, whose sign flips under a dividend-omitting construction (−2.22% versus +2.60%). When $R^2_{\text{OOS}} < 0$, a low CW p is not interpreted as an economically meaningful improvement.

Class B is retained in the separate channel, economic-value, holdout, and complexity analyses. This section analyzes the Class-B measures.

In the univariate accounting regression, IV^2 attains an OOS R^2 of 27.63% ($\hat{\beta} = -1.970$, NW $t = -11.52$, $R^2_{\text{IS}} = 26.15\%$, CW $p = 0.012$) and the variance-risk-premium proxy 9.36% (CW $p = 0.045$), both higher than or comparable to SVAR's 11.03%; the near-equality of IV^2's in-sample and out-of-sample R^2 is itself a signature of a mechanical relation. In bivariate regressions, SVAR loses significance under SVAR+IV^2 ($t = -0.80$), and VRP is likewise insignificant under IV^2+VRP ($t = 0.77$); an encompassing test shows that IV^2 almost fully encompasses SVAR (H_0: IV^2 encompasses SVAR, $\lambda = 0.086$, $p = 0.113$, so SVAR's incremental information is insignificant).

Mechanical characterization (core). This "victory" *cannot* be read as IV^2 being a stronger predictor. Because the delta-hedged put return is theoretically $\approx c\,(\text{RV}_{t+1} - \text{IV}^2_t)$, the slope of the return on IV^2 splits into two opposing components: a *forecasting* component $+c\,\beta_f$ (where β_f is the predictive slope of IV^2 for future realized variance) and a *mechanical* component $-c$ (the definitional "paying implied variance" leg). With the measured $c = 3.70$ and $\beta_f = 0.53$, these are $+c\,\beta_f \approx +1.96$ and $-c = -3.70$. **Measured against the sum of the two components' absolute magnitudes (**$1.96+3.70 = 5.66$**), the mechanical leg alone is about 65.5**%; because the two components carry opposite signs, the mechanical leg exceeds the absolute net slope and pins its negative sign (net theoretical slope $\approx +1.96 - 3.70 = -1.74$). Put differently, if only the legitimate forecasting channel were present, IV^2 and the return would be *positively* related; the negative relation actually observed is driven by the definitional term. This decomposition is an approximation—the theoretical net slope of −1.74 differs from the directly estimated IV^2

slope of -1.970 by the combined approximation and estimation error—and the accounting regression $R^{dh} \sim (\mathrm{RV}_{t+1} - \mathrm{IV}^2)$ attains an R^2 of 53.9%, which is a *payoff-decomposition fit* rather than conventional out-of-sample forecastability. That SVAR (the prior month's exogenous realized variance) is absorbed by the contemporaneous, more direct IV^2 is expected and is not evidence that IV^2 has stronger predictive power.

We therefore position IV^2 as a **direct contract-level measure of the variance-pricing component embedded in the target payoff**: it records the strength of the variance-pricing channel driving delta-hedged put returns—it does not precisely quantify the theoretical VRP, since $\mathrm{IV}^2 - \mathrm{RV}$ is a proxy, c is an approximation, and the realized P&L also contains jump, discrete-hedging, and delta-model components—and most of its high R^2 is mechanical/definitional. Adding IV^2/VRP to the macro set (the Macro-13 channel-accounting specification) raises the six-model R^2_{OOS} across the board to 19.55%–26.73% (ENet 25.32%, Combine 25.37%, P-Aver 25.96%, tPCA-h 26.46%, tPCA-s 26.73%, KP-PLS 19.55%; all CW $p < 0.04$); this increase likewise inherits the mechanical channel above rather than being a genuine improvement in macro predictive power, and it must carry the same caveat. This characterization also explains the grid structure of Section 6 (predictability concentrates on the put side toward the money, exactly where the VRP/skew wedge is cleanest) and the near-linearity of Section 8.5 (the first-order payoff component is linear in IV^2).

Across targets (put-specific). Constructing an ATM call target by the same method, its delta-hedged return has a mean of -0.376% (the dividend term is negative for the call), and the three predictors all have negative R^2_{OOS} (SVAR -2.57%, iskew -0.17%, IV^2 -3.50%): the call delta-hedged return is not predictable. For the OTM95 put ($S/K \approx 0.95$), only IV^2 survives and is even stronger ($R^2_{\mathrm{OOS}} = 32.31\%$, CW $p = 0.036$). This is consistent with the economic intuition that the VRP/crash premium concentrates in downside protection (puts): for calls, implied and realized variance are close (the VRP is near zero) and the return is near zero; for puts, the volatility skew makes implied far higher than realized, so the return is significantly negative and IV^2's mechanical predictive power is strong.

6 Contract-Level Evidence and Futures-Basis Validation

This section presents the contract-level evidence in three steps: a moneyness grid of delta-hedged returns, its common-sample re-estimation with formal put–call tests, and the put–call relative-value wedge with its direct validation against the futures basis.

6.1 The Moneyness Grid

The grid upgrades a point conclusion to a structural picture. Using the same dividend-inclusive P&L identity, we construct six moneyness series (puts at $K/S \in \{0.90, 0.95, 1.00\}$ and calls at $K/S \in \{1.00, 1.05, 1.10\}$; coverage is in Appendix Table 14) and run a univariate OOS prediction on each. Table 4 and Figure 3 present the "predictability smile."

The grid results trace out a *put-concentrated* point pattern: (i) in point estimates predictability is stronger on the put side—the three call-side series all have negative or near-zero R^2_{OOS}; and (ii) it rises toward the money along the strike—SVAR's R^2_{OOS} rises from the deep out-of-the-money put (K090P, -0.68%) to the at-the-money put (K100P,

5.40%), and the channel measures trace the same shape amplified (IV^2 from 0.75% at K090P to 26.37% at K100P; see Section 5.3). This pattern is *not* organized by gamma exposure: directly measuring the standardized dollar gamma at entry shows that the call side has high gamma exposure (K105_CALL the largest) yet no predictive power, while the ATM put's gamma-to-investment ratio is not the highest—the gamma ordering does not match the R^2_{OOS} ordering. The pattern is *put-concentrated, rising toward the money along* $|\Delta|$. Its candidate mechanism requires two conditions at once: only puts carry the skew-premium wedge $IV^2 - \mathbb{E}[RV]$ (on the call side implied is roughly equal to realized), and near the money the noise is lower (deep out-of-the-money puts have high standard deviation and large skewness, which dilutes the signal). This is consistent with the mechanical decomposition of Section 5.3: the payoff-decomposition term is proportional to $(RV - IV^2)$, and the ATM put is where that wedge is cleanest. We emphasize, however, that Table 4 uses each series's own full sample, so the moneyness comparison is confounded by differing sample months. Below we re-estimate on a balanced common sample and subject the put–call gap to a formal test; the "smile" weakens sharply, and the put–call difference—though large in point estimate—is *not* statistically separable. We therefore report a put-concentrated point pattern and do not claim that predictability is confined to puts.

Table 4: Predictability grid: univariate R^2_{OOS} by moneyness series (%)

Predictor	K090P	K095P	K100P	K100C	K105C	K110C
SVAR	−0.68	1.90	**5.40**	−4.64	−2.32	−2.04
iskew	−0.96	0.32	5.26	−0.99	−0.61	−2.34
N (months)	204	234	229	208	220	126

Notes: KxxxP/C denotes the put/call series at $K/S = x.xx$. The expanding-window initial length is 60 months (48 for K110C). Each cell uses its *own* full sample, so the moneyness comparison is confounded by differing months (see the common-sample re-estimation in Table 5). In point estimates, predictability is put-concentrated (rising toward the money at K100P) while the call side (K100C–K110C) is uniformly negative or near zero; the pattern is organized along $|\Delta|$ rather than gamma exposure (see this section). The channel measures trace the same shape with larger magnitudes (IV^2_{ATM}: 0.75, 7.58, 26.37, −17.16, −4.96, 0.02; VRP: −0.22, 2.63, 11.87, −5.94, −1.24, 1.76; Section 5.3). K110_CALL has insufficient coverage ($N = 126$) and is listed here for comparison; the formal analysis downgrades it to the appendix.

6.2 Common-Sample Re-estimation and a Formal Put–Call Test

To purge the sample-composition confound, we form the balanced intersection of the five core cells (K090P, K095P, K100P, K100C, K105C; K110C is excluded, $N = 126$), which yields 159 common months (99 OOS after the initial window), and re-estimate the univariate SVAR OOS prediction with a common initial window. Table 5 contrasts the own-sample and common-sample R^2_{OOS} and reports each cell's characteristics. On the common sample the smile flattens dramatically: only the at-the-money put retains a positive point estimate (+0.99%, CW $p = 0.001$), the deep and out-of-the-money puts turn negative, and even the ATM call is marginally positive (+0.79%). The cell characteristics show why the deep out-of-the-money puts carry little signal: their return standard deviation is large (K090P 7.9% per month against 1.6% at the money) and their relative spreads are wide, diluting any wedge.

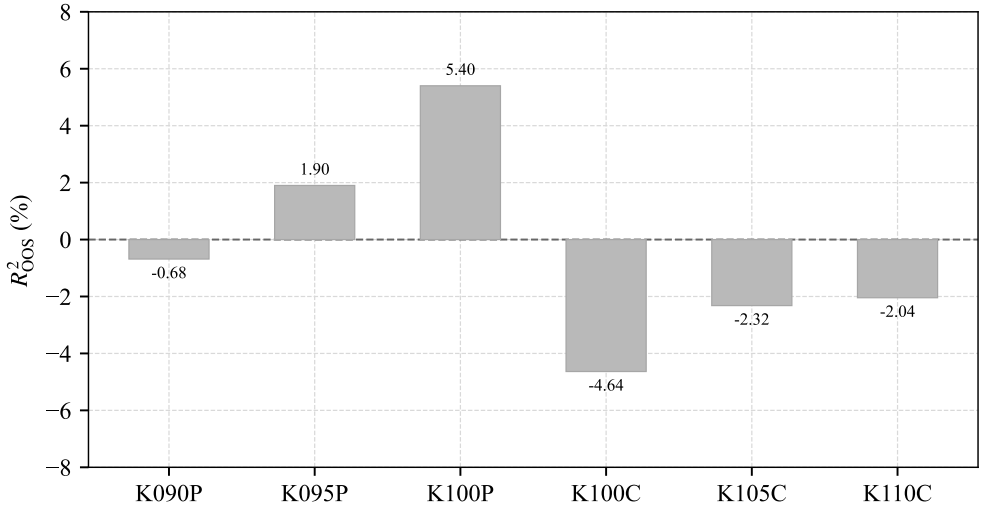


Figure 3: Predictability across the moneyness grid: R^2_{OOS} of SVAR

Notes: On the put side (K090P→K100P), SVAR's R^2_{OOS} rises monotonically toward the money, to 5.40%; on the call side (K100C→K110C) it is uniformly negative. The left three bars are puts and the right three are calls. Each series uses its own full sample (Table 4); the common-sample re-estimation in Table 5 flattens the profile substantially.

Table 5: Common-sample moneyness grid: SVAR R^2_{OOS} and cell characteristics

Cell	K/S	N months	mean $\lvert\Delta\rvert$	ret. SD (%)	rel. spread	own R^2_{OOS} (%)	common R^2_{OOS} (%)	common CW p
K090_PUT	0.90	204	0.064	7.91	0.066	−0.68	−1.83	0.638
K095_PUT	0.95	234	0.186	3.49	0.026	1.90	−1.34	0.106
K100_PUT	1.00	229	0.498	1.60	0.019	5.40	**0.99**	0.001
K100_CALL	1.00	208	0.488	1.42	0.016	−4.64	0.79	0.038
K105_CALL	1.05	220	0.135	2.55	0.046	−2.32	−1.23	0.418

Notes: The common sample is the balanced intersection of the five cells (159 months, 99 OOS; K110_CALL excluded, $N = 126$). "own" uses each cell's full sample; "common" uses the balanced sample with a common initial window. mean $\lvert\Delta\rvert$, return SD, and relative spread are common-sample cell characteristics. On the common sample only the ATM put keeps a positive point estimate.

We test the put–call difference formally on the two clean same-strike pairings that the grid's one-sided-OTM design permits (K100 and the ATM q-fixed series). Table 6 reports the OOS R^2 difference (put minus call), a stationary block bootstrap (Politis and Romano, 1994) 95% confidence interval (block ≈ 6, 2000 resamples), and a one-sided Diebold–Mariano test. At the same K100 strike the put–call gap is a trivial +0.20pp with a wide interval $[-5.71, +6.28]$ (DM one-sided $p = 0.478$); for the ATM pair the point gap is large (+16.20pp) but the interval $[-1.98, +47.21]$ contains zero (DM one-sided $p = 0.123$). **The smile gradient weakens sharply under the common sample, and the put–call difference—large in point estimate—is not statistically separable.** We therefore report a put-concentrated point pattern and explicitly *do not* claim that only puts are predictable.

Table 6: Formal same-strike put–call predictability difference

Pair	N_{OOS}	R^2_{OOS} put (%)	R^2_{OOS} call (%)	diff put−call (pp)	boot 95% CI (pp)	DM p
K100 same-strike	99	0.99	0.79	+0.20	[−5.71, +6.28]	0.478
ATM (q-fixed)	159	12.50	−3.70	+16.20	[−1.98, +47.21]	0.123

Notes: The loss differential is defined as $d_t = (\text{call SSE contribution}) - (\text{put SSE contribution})$, and the DM p is one-sided for $H_0 : R^2_{\text{put}} \leq R^2_{\text{call}}$ against put > call. Both confidence intervals contain zero, so neither difference is statistically separable despite the large ATM point gap. Clean same-strike pairings exist only at K100 and the ATM series (the grid is otherwise one-sided OTM).

6.3 Bridging the Main-Table ATM Put and the Grid ATM Put

The main-table ATM put ($R^2_{\text{OOS}} = 11.03\%$) and the grid K100 put (5.40%) differ because of *contract selection*, not a genuine moneyness effect. The two rules differ: the ATM series takes the nearest-the-money contract with $K \leq S$ and no band, whereas K100 requires $K/S \in [0.99, 1.01]$ and drops the entire month when no in-band strike exists—so K100 loses several high-variance months (including the crisis month October 2008). On the 227 months they share, the two monthly return series correlate 0.96 (means −0.66% and −0.62%), and their common-sample SVAR R^2_{OOS} are close, 7.71% versus 6.30% (a gap of only 1.4pp); once November 2008 is removed both fall to about 2–3% (Table 15, Internet Appendix E). The 5.6pp headline gap thus comes mainly from sample composition, and both selection rules are crisis-driven.

6.4 The Put–Call Wedge and Synthetic-Forward Basis

Beyond return predictability, a put–call parity test on delta-hedged P&L provides a class of demand-side price evidence *independent of* R^2_{OOS}. We keep four layers strictly separate: (i) an empirical price fact; (ii) an accounting interpretation; (iii) an economic interpretation consistent with the fact but not identified by it; and (iv) the alternative mechanisms the fact does not rule out.

(i) The price fact. For same-month, same-strike calls and puts we compute $(\pi_{\text{call}} - \pi_{\text{put}})/S_0$, which removes the investment-denominator effect and is zero under European put–call parity. All relative-value quantities in this paper follow a *call-minus-put* sign convention, so at the P&L level a *positive* value means the put is richer (put-rich). The wedge is systematically positive. Aggregating it across strikes with equal monthly weights—our headline aggregation, which we prefer because it avoids overweighting months with many

listed strikes—the monthly mean is +0.245% of the underlying (NW $t = 4.64$), and all five moneyness points are positive and mostly significant: **same-strike puts are systematically richer than calls**, by about 0.2–0.25% of the underlying per month.[6] As a cross-check immune to dividend measurement error, the entry-date *price-level* synthetic-forward basis, computed from the realized dividends over each contract's life, is −0.240% of S_0 ($t = -9.18$). The sign flips *by construction*: at the price level "puts richer than calls" is a synthetic-forward *discount*, so put-richness registers as a *negative* basis, whereas at the P&L level put-richness registers as a *positive* π difference. Both diagnostics point the same way; only the sign convention is reversed.

We subject this deviation to a series of adversarial robustness tests (Table 7 notes): (i) *dividend sensitivity*—the deviation's sensitivity to an error in annualized q is about 0.07pp per month per 1pp, and it would take a uniform underestimation of annualized q by +3pp over the whole sample to bring it to zero (economically implausible), and the deviation is strongest in non-ex-dividend months where the dividend flow is near zero (+0.287%, $t = 7.45$, where q error has nowhere to hide), so the realized-dividend version is fully immune to this concern; (ii) *interest-rate choice* affects it by < 0.07pp per 1pp; (iii) using only high-quality quotes with a two-sided relative spread $\leq 10\%$ still gives +0.178% ($t = 4.64$), and the deviation is unrelated to the spread/volume proxies; and (iv) *time pattern*—it is not yet significant over 2002–2008 (+0.087%), strengthens over 2009–2014 (+0.360%), and persists over 2015–2022 (+0.183%), the opposite of a story of declining market immaturity and consistent with persistent downside-hedging demand.

(ii) Accounting interpretation. For an index with tradable futures, "puts richer than calls" is mechanically a synthetic-forward discount—a carry/basis wedge that is **the option-market analogue of the persistent backwardation documented in the Taiwan index futures**, and *not* an arbitrage opportunity. Because the Taiwan index is not short-sale-constrained (the futures can be shorted), this is *not* a single-name, short-sale-constrained parity violation; that literature identifies with single stocks and shorting costs and has a different mechanism, which we do not analogize here.

(iii) Economic interpretation (consistent, not identified). The wedge is *consistent with* persistent downside-insurance demand: it strengthens rather than fades over the sample and integrates naturally with the put-concentrated predictability. But being consistent with that story does not *uniquely identify* it.

(iv) Unidentified alternatives. Several mechanisms could generate the same wedge and are not separated here: dealer balance-sheet and inventory constraints, funding costs, margin asymmetries between long and short option positions, the state of the futures basis, settlement conventions, institutional hedging demand, and market segmentation. Our incremental contribution is to quantify the wedge cleanly with the parity of delta-hedged P&L (about 0.2–0.25% of the underlying per month), to rule out dividend and liquidity explanations with the realized-dividend test, and—in (v) below—to tie the wedge directly to the index-futures basis; the wedge is independent of return prediction and serves as cross-market validation and economic interpretation of the price side of our narrative, not as a second independent class of demand-side evidence.

(v) A direct joint test against the futures basis. Using TAIFEX daily settlement prices

[6] The pooled (per-pair equal-weight) mean is +0.212% (NW $t = 5.10$) but overweights strike-rich months, so we report the monthly-aggregated value as the headline. The two aggregations are highly consistent: the monthly wedge correlates 0.92 with a fixed near-ATM series and 0.98 with the inner three strikes, and the number of listed strikes per month correlates only −0.10 with the wedge, so strike availability does not drive the result.

for the TX contract (2002–2025), we compute for every entry month the synthetic forward implied by the matched same-strike pairs, $F^{\text{syn}} = K + (C_{\text{mid}} - P_{\text{mid}})e^{r\tau}$, and compare its basis $(F^{\text{syn}} - S)/S$ with the same-expiry futures basis $(F^{\text{fut}} - S)/S$ over the 230 wedge months. The two series track each other almost one-for-one. Their means are -0.426% (NW $t = -7.0$) and -0.441% ($t = -7.5$); their correlation is 0.991 in levels and 0.987 in first differences; and regressing the synthetic basis on the futures basis gives a slope of 1.006 (we cannot reject a unit slope, $p = 0.59$), an intercept of $+0.017\%$ (we cannot reject zero), and $R^2 = 0.98$. The direct gap between the synthetic and observed forwards averages $+0.015\%$ of the underlying—about 3% of the basis itself and economically negligible, though marginally significant ($p = 0.06$). In the 2008–2009 crisis both bases widen in lockstep from roughly -0.4% to -1.35% while the gap stays near 0.01%, and a 2023–2025 extension repeats the pattern (correlation 0.992, slope 0.97). Because both series embed the same carry, the level correlation is partly mechanical; the identification comes jointly from the near-zero gap, the unit slope, and the first-difference correlation. The wedge is therefore not merely an analogue of the futures backwardation: it is that backwardation's option-market mirror, with no economically important independent component. This finding repositions the wedge's evidentiary role. It provides an option-market representation of the futures basis and confirms that the apparent relative-value difference is a market-wide carry phenomenon rather than an option-specific arbitrage violation; because the futures basis itself may reflect expected dividends, financing conditions, dealer balance-sheet costs, institutional positioning, and settlement effects as well as hedging demand, the wedge's economic origin remains unidentified, with downside-insurance demand one plausible contributor.

Table 7: Put–call parity deviations $(\pi_{\text{call}} - \pi_{\text{put}})/S_0$ (% of S_0)

	K090	K095	K100	K105	K110	Full sample
Deviation mean (%)	0.286	0.202	0.133	0.256	0.347	**0.245**
NW t	2.45	2.92	1.94	4.75	4.61	**4.64**
N (pairs)	60	113	201	142	48	564

Notes: π is delta-hedged P&L; the sign convention is call minus put, so a positive value means put-rich. The full-sample column is the *monthly-aggregated* deviation (+0.245%, NW $t = 4.64$); the pooled (per-pair equal-weight) mean is +0.212% (NW $t = 5.10$) but overweights strike-rich months and is reported only here. The per-strike t's are NW t's within each strike. The price-level synthetic-forward basis computed from realized dividends is -0.240% of S_0 ($t = -9.18$; the negative sign is the price-level counterpart of put-richness), significant in all three subperiods. Robustness: the non-ex-dividend-month deviation is +0.287% ($t = 7.45$); the high-quality-quote subsample is +0.178% ($t = 4.64$); the time pattern is +0.087 (ns)/+0.360/+0.183. A direct joint test against TAIFEX settlement prices shows the wedge closely tracks the index-futures basis with a near-unit slope (part (v) of this subsection); it is not an arbitrage violation.

7 Economic Value

Whether statistical predictability delivers practical benefit depends on whether it translates into allocation performance and survives transaction costs. Table 8 reports annualized CER and Sharpe ratios. The prevailing-mean benchmark has a **gross CER of** 8.21% **and a Sharpe ratio of** 1.34; under a dividend-omitting construction of the target these

figures would be 14.6% and 1.9, so the difference comes *almost entirely from the dividend correction* (the premium itself shrinks by about 6.4pp) rather than from transaction costs. The short-volatility strategy has positive economic value, but its magnitude is far below what dividend-omitting accounting suggests.

Transaction costs are accounted for on the correct notional. The bid–ask spread applies only to the option premium actually traded (about 4% of the position's investment), not to the entire investment including the hedging leg—misapplying the option's relative spread to the whole investment would overstate the cost by about 25 times. On this correct notional, the option-leg cost lowers the benchmark CER from 8.21% to 6.85% (-1.4pp); we further add the daily delta-rehedging *futures-leg* cost (the TX futures spread estimated at 1 tick, with a median of about 0.023% per month), giving a CER, net of quoted option spreads and futures rebalancing costs, of 6.39% (-0.5pp). This net CER is positive across all γ and weight caps ($\gamma = 3$: about 6.4–11.8%; $\gamma = 5$: 6.0–10.1%; $\gamma = 10$: 4.8–6.9%), of order about 5–8%. Throughout, the CER is expressed per unit of self-financing investment notional and does *not* net out exchange margin, regulatory capital, or market impact. The IV^2 channel-based timing strategy has the highest *point-estimated* net CER (6.88%, Sharpe 1.20); as we show below, this edge is not statistically distinguishable from the benchmark, and the other models roughly tie it.

To locate the source of the economic value, we decompose it into a nested ladder of four strategies (Table 9): (1) a *constant* short-volatility position (fixed premium and fixed variance from the initial 60-month window, so the weight never moves); (2) rolling-window volatility scaling with no return prediction; (3) the historical-mean-based volatility-scaled benchmark (a prevailing-mean forecast, updated recursively); and (4) predictive timing (SVAR and IV^2). Nearly all of the certainty-equivalent value comes from *unconditionally* bearing the short-volatility premium: the constant strategy already earns a net CER of 6.27%, which is 98.2% of the benchmark's and 91.2% of the best timing strategy's. Every increment up the ladder is small and statistically indistinguishable from zero: rolling-window volatility scaling adds +0.12pp (bootstrap 95% CI $[-2.26, +2.81]$), the prevailing-mean update adds -0.01pp $[-0.30, +0.31]$, SVAR timing adds -0.39pp $[-1.68, +1.12]$, and IV^2 timing adds +0.49pp $[-1.27, +2.19]$—**all four intervals contain zero**. **The incremental economic value of predictive timing is thus economically negligible**, and we record it as one of the paper's boundary results: statistical predictability does not translate into a reliable, cost-adjusted market-timing gain.

The IV^2 strategy's +0.49pp edge over the benchmark is a point estimate, *not* a statistical superiority: its stationary block bootstrap 95% CI is $[-1.27, +2.19]$ (right-tail $p = 0.28$), and the Sharpe-ratio difference of 0.04 has a CI that likewise contains zero ($[-0.15, +0.20]$), so we attach no economic meaning to it. Table 16 (Internet Appendix E) discloses the tail and leverage profile of the short-volatility strategies. All are short volatility, with average absolute weights near the $\pm 200\%$ cap; the cap binds in 72% of months for the benchmark but only 35% for IV^2 (which trades more actively, turnover 0.28 versus 0.04). Maximum drawdowns are -12.6% (benchmark) and -11.1% (IV^2), with fat, right-skewed monthly tails (ES_{95} of -6.5% and -5.3%). In this sample the crisis months accumulate *positive* returns—because entry-date implied volatility is very high and realized volatility comes in lower, the short-volatility premium is large (the benchmark earns +6.83% in November 2008 and +8.16% in March 2020; IV^2 earns +16.60% and +8.74%)—but this reflects the historical crashes and does not guarantee the future.

Table 8: Economic value: annualized CER and Sharpe ratio ($\gamma = 10$, $\pm 200\%$)

	Gross CER (%)	Net CER (option leg only, %)	Net CER (+ futures leg, %)	Net Sharpe
Benchmark (prevailing mean)	8.21	6.85	6.39	1.15
ENet	7.94	6.59	6.13	1.11
Combine	7.87	6.57	6.13	1.11
SVAR univariate	7.73	6.42	5.99	1.10
IV^2 channel-based timing	**8.41**	**7.26**	**6.88**	**1.20**

Notes: Gross CER is before costs; net CER (option leg only) includes the bid–ask spread applied to the premium; net CER (+ futures leg) additionally includes the daily delta-rehedging TX futures cost (median about 0.023% per month). The benchmark's gross Sharpe is 1.34. The benchmark CER is 8.21% against 14.6% under a dividend-omitting target, a difference due chiefly to the dividend correction rather than costs. IV^2 has the highest point-estimated net CER, but its difference from the benchmark is statistically insignificant (bootstrap 95% CI $[-1.27, +2.19]$pp).

Table 9: Decomposition of economic value: nested four-strategy ladder (net CER, $\gamma = 10$, $\pm 200\%$)

Strategy	Net CER (%)	Net Sharpe	Incremental CER (pp)	boot 95% CI (pp)
(1) Constant short volatility	6.27	1.12	—	—
(2) Rolling-window volatility scaling	6.40	1.16	+0.12 (vs 1)	[−2.26, +2.81]
(3) Historical-mean benchmark	6.39	1.15	−0.01 (vs 2)	[−0.30, +0.31]
(4a) SVAR predictive timing	5.99	1.10	−0.39 (vs 3)	[−1.68, +1.12]
(4b) IV^2 channel-based timing	6.88	1.20	+0.49 (vs 3)	[−1.27, +2.19]

Notes: The four strategies are nested (1) $\subset$ (2) $\subset$ (3) $\subset$ (4). Strategy (1) fixes the premium and variance at their initial-window values, so its weight is constant (causal, no look-ahead). Incremental CER is relative to the preceding strategy; the confidence intervals are stationary block bootstrap 95% intervals (block ≈ 6, 2000 resamples), and all four contain zero. The constant strategy already attains 98.2% of the benchmark's net CER and 91.2% of the best timing strategy's, so the incremental value of predictive timing is economically negligible.

8 Robustness, Holdout, and Model Complexity

8.1 Time Concentration and Predictive Instability

Crisis concentration. The headline predictability is statistically strong at the individual-test level but episodic and clustered in stress periods. Table 10 strips out crisis observations one layer at a time. SVAR's full-sample R^2_{OOS} of 11.03% falls to 5.35% once November 2008 is removed, to 3.51% once all of 2008 is removed, to 2.04% once the three largest-gain months are removed, and to just 0.22% once the five largest-gain months are removed; the Macro-11 Combine forecast behaves the same way (8.36% down to −0.29%). Yet the Clark–West statistic stays significant in every variant ($p < 0.001$–0.01): *statistical significance survives while the economic magnitude collapses.* A leave-one-year-out analysis tells the same story—dropping any single year leaves SVAR's R^2_{OOS} near 11% (median 11.14%, maximum 14.57% dropping 2020) *except* 2008, whose removal alone cuts it to the minimum of 3.51%.

Table 10: Crisis-exclusion suite: R^2_{OOS} collapses while CW significance survives

	SVAR R^2_{OOS} (%)	SVAR CW p	Combine R^2_{OOS} (%)	Combine CW p
Full sample	**11.03**	0.008	**8.36**	0.028
Ex November 2008	5.35	< 0.001	3.69	0.001
Ex all of 2008	3.51	< 0.001	3.92	< 0.001
Ex top-3 gain months	2.04	< 0.001	1.31	0.002
Ex top-5 gain months	0.22	< 0.001	−0.29	0.011

Notes: The one-sided Clark–West p tests the nested-adjusted MSPE improvement over the prevailing mean. Combine is the Macro-11 selective combination. Significance survives every exclusion, but the economic magnitude falls sharply, so we characterize the predictability as bounded and crisis-concentrated.

Time-varying predictive ability. A Giacomini–Rossi (Giacomini and Rossi, 2010) fluctuation statistic (CW-adjusted, window $m = 60$, $\mu = 0.34$) and the rolling-window estimates indicate pronounced time variation in SVAR's predictive ability: its fluctuation statistic peaks at 2.73 in May 2010 (the post-crisis recovery window) and troughs at 0.54 in the late sample, far below the peak. Figure 4 plots the rolling 60-month R^2_{OOS}: it stays near 15–25% through 2013–2018 and falls to around zero after 2019. Splitting at 2015, the early subsample (2007–2014) has $R^2_{\text{OOS}} = 16.53\%$ and the late subsample (2015–2022) only 0.69%. The CW test continues to detect adjusted predictive content in the late subsample ($p = 0.0016$), but the corresponding sample MSPE gain is economically negligible: the CW statistic tests an adjusted population-MSPE null for nested forecasts, so a near-zero sample R^2_{OOS} and statistical significance can coexist. Throughout the paper we read such cases as a divergence between statistical detection and economic magnitude, not as evidence of exploitable late-period predictability.

Table 11 reports robustness across subsamples, windows, and winsorization. Three core conclusions: (i) *crisis concentration*—the terminal values of the Goyal–Welch cumulative SSE difference are all positive but highly concentrated in the single month of November 2008 (SVAR 56.8%, iskew 62.9%, Combine 60.7%); (ii) *SVAR flattens to near zero in the late period*—SVAR's R^2_{OOS} is 16.53% in the early period (2007–2014) and *near zero* in the late period (2015–2022, +0.69%), so the qualitative "no late predictive power" holds; and (iii) *iskew is also early/crisis-concentrated*—iskew is 9.39% in the early period and −0.70% in the late period, turning negative under init120 (which misses the

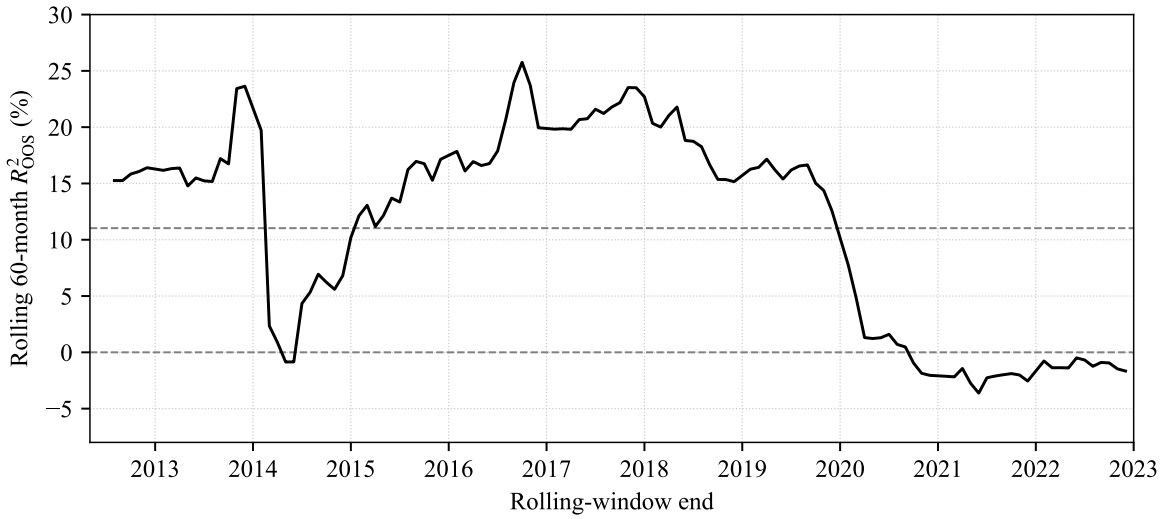


Figure 4: Rolling 60-month out-of-sample R^2 of the SVAR forecast

Notes: The solid line is SVAR's rolling 60-month R^2_{OOS} (%); the upper dashed line is the full-sample value (11.03%) and the lower dashed line is zero. Predictive power is high through the mid-2010s and falls to around zero after 2019, confirming that predictability is concentrated in the earlier part of the sample.

crisis, −6.25%). Thus the OOS gains of both SVAR and iskew concentrate in 2007–2014 (including the crisis) and neither has predictive power after 2015. The results are robust to winsorization (1%/99%) and to Campbell–Thompson non-negativity truncation (only a small decline).

Table 11: Robustness: subsamples, windows, and truncation (R^2_{OOS}, %)

	Early 07–14	Late 15–22	init120	rolling60	winsor 99	CT non-neg
SVAR	16.53	0.69	4.56	10.56	9.97	11.03
iskew	9.39	−0.70	−6.25	3.94	5.71	6.69
Macro-Combine	10.46	4.42	7.15	5.76	7.78	8.36

Notes: The full-period R^2_{OOS} are SVAR 11.03, iskew 5.89, and Combine 8.36. The November 2008 single-month share of the cumulative SSE difference is SVAR 56.8%, iskew 62.9%, and Combine 60.7%. SVAR is +0.69% in the late period ("near zero") and iskew is −0.70% (early/crisis-concentrated). init120 misses the 2008 crisis.

Anatomy of the crisis gains. Month by month, the gains concentrate in October–November 2008—high prior-month realized variance drove a sharply negative SVAR forecast that the market validated—while March 2020 is the mirror image, a tail event the model *missed*, so the predictive power is concentrated and does not span the tails (Table 17, Internet Appendix E).

8.2 A Frozen Post-2022 Holdout

The recursive forecasts of Sections 5–7 use only information dated before each origin. As an additional validation, we freeze every variable definition, filter, model class, and tuning rule at its end-of-2022 specification and extend the raw TXO data through January 2026; regression coefficients and Combine selections continue to update recursively within

the frozen rules as the window expands. Table 12 reports both the 2023–2025 holdout and the continuous 2007–2025 evaluation. Of the 36 candidate months (January 2023 to December 2025), 35 are buildable; the only exclusion is January 2025, removed by the frozen holding-path filter through the same Lunar-New-Year mechanism that removes fifteen Januaries from the estimation sample. Predictors come from the same official sources as the estimation sample, except that the 2025 aggregate valuation inputs are built from TEJ's market-level monthly series; this source choice is immaterial to the holdout (Internet Appendix E).

The holdout put premium is -0.24% per month ($t = -1.21$), roughly a third of the full-sample -0.71% and not distinguishable from zero, and the holdout return path is markedly less fat-tailed (excess kurtosis -0.2, against 4.0 in sample). SVAR underperforms the prevailing mean, with a holdout R^2_{OOS} of -10.7% that deepens to -19.1% when its two best months are removed—consistent with the late-subsample evidence. IV^2 delivers $+13.6\%$, the Macro-11 Combine $+21.7\%$ (Clark–West one-sided $p = 0.004$), and tPCA-h $+9.2\%$. These positive estimates warrant caution for three reasons. The sample is short (35 months). The gains are concentrated and model-dependent: the largest single month (April 2024 for the univariate models, March 2025 for the multivariate ones) accounts for 13–17% of each model's positive gain, and dropping the two largest months leaves Combine at $+16.9\%$ but cuts tPCA-h to $+1.0\%$. And part of the measured gain is level-mechanical: the recursive historical mean remains anchored to the deeper premia of the 2008–2011 era (-0.68%) while the realized holdout mean is -0.24%, so a forecast that shrinks toward zero beats the benchmark before any conditional timing enters. Stress events do occur within the window—the August 2024 volatility spike and the April 2025 tariff shock, the latter the largest SVAR reading in the holdout—but neither generates 2008-style forecast gains, and the December 2025 rebound month ($+0.76\%$ realized against negative forecasts) penalizes the variance-based models most.

Three diagnostics clarify the source of the Combine result. *Selection.* SVAR is selected at every one of the 35 holdout origins (its in-sample $|t| \approx 5.7$ keeps it far above the entry threshold), the book-to-market ratio at every origin, and the dividend–price ratio at 46% of origins; no other predictor enters. Yet SVAR alone delivers -10.7% in the same window, while univariate BM and DP deliver $+13.5\%$ and $+14.6\%$: the combination's holdout gain is carried by the valuation ratios, not by the variance channel that drives the full-sample result. *Level versus timing.* The average Combine forecast is -0.39%, between the benchmark's -0.68% and the realized -0.24%; decomposing the MSPE gain, 52% comes from correcting the benchmark's level bias and 48% from tracking month-to-month variation. Against benchmarks free of the anchoring problem the gain shrinks but does not vanish: $+13.9\%$ versus a zero forecast, $+16.2\%$ versus a rolling 60-month mean, and $+12.4\%$ versus a post-2015 expanding mean. *Demeaned content.* Within the holdout, the Combine forecast correlates 0.36 with realized returns, its directional hit rate is 60%, and the demeaned R^2_{OOS} is 11.8%. Taken together, the diagnostics attribute roughly half of the Combine gain to adaptive level correction and the remainder to conditional timing that runs through valuation ratios rather than SVAR—a channel rotation relative to the full-sample evidence—and on 35 months short-sample luck cannot be excluded. A longer post-2022 sample is needed to determine whether the timing component persists.

Panel B of Table 12 reports the extended full-sample results obtained by carrying the recursive evaluation through 2025. Over 2002–2025 (271 months) the put premium is -0.65% per month (NW $t = -6.37$), and SVAR's full-sample slope remains significant ($\hat{\beta} = -1.470$, NW $t = -3.76$, $R^2_{\text{IS}} = 10.58\%$). Because the forecasts form one continuous

recursive sequence, the extended evaluation window 2007:06–2025:12 (211 months) is the main window carried forward: SVAR's R^2_{OOS} is 9.09% (CW $p = 0.004$), the IV^2 channel's 26.38%, Combine's 9.55%, and tPCA-h's 8.60%. The extended estimates are modestly smaller than their 2002–2022 counterparts but qualitatively unchanged, and they locate the source of the decline: the largest single-month contribution to every model's extended gain remains November 2008 (dropping it cuts SVAR to 3.76%), while the post-2022 segment is where the SVAR forecast loses to the prevailing mean.

Table 12: Frozen post-2022 holdout and the extended evaluation window

Model	N	R^2_{OOS} (%)	CW t	CW p (one-sided)	Evaluation window
Panel A: Frozen holdout					
SVAR (external predictor)	35	−10.71	1.60	0.059	2023:01–2025:12
IV^2 (channel variable)	35	13.62	1.83	0.038	2023:01–2025:12
Macro-11 Combine	35	21.67	2.79	0.004	2023:01–2025:12
Macro-11 tPCA-h	35	9.25	2.22	0.017	2023:01–2025:12
Panel B: Extended evaluation window (main window carried forward)					
SVAR (external predictor)	211	9.09	2.72	0.004	2007:06–2025:12
IV^2 (channel variable)	211	26.38	2.32	0.011	2007:06–2025:12
Macro-11 Combine	211	9.55	2.30	0.011	2007:06–2025:12
Macro-11 tPCA-h	211	8.60	2.51	0.006	2007:06–2025:12

Notes: All specifications are frozen at their end-of-2022 form; forecasts are recursive with no reselection or retuning, and the recursive prevailing-mean benchmark runs continuously from 2002 in both panels. CW p-values are one-sided with HAC standard errors (Andrews automatic bandwidth: 4 lags at $N = 35$, 5 at $N = 211$; fixing 4 lags throughout changes no conclusion). The 35 holdout months are January 2023 through December 2025, excluding only January 2025 (frozen Lunar-New-Year path filter); the 211 months of Panel B are the 176 main-window evaluation months plus these 35 (Januaries without an eligible entry are absent from both segments). The 2025 aggregate valuation inputs are built from TEJ's market-level monthly series; rebuilding the entire 2023–2025 segment from the TEJ source alone changes no R^2_{OOS} by more than 0.05 percentage points. In Panel A, dropping each model's two largest-gain months leaves SVAR at −19.1%, IV^2 at +8.1%, Combine at +16.9%, and tPCA-h at +1.0%; in Panel B, the largest single-month contribution to every model's gain is November 2008, dropping which leaves SVAR at 3.76%, IV^2 at 18.61%, Combine at 5.47%, and tPCA-h at 2.49%. The holdout mean put premium is −0.240% per month; the extended 2002–2025 mean is −0.647% (NW $t = -6.37$, 271 months) against −0.707% in the 2002–2022 sample.

8.3 Negative Identification Tests: The Retail Channel

Taiwan's institutional distinctiveness in this literature rests on its high retail participation, so we test the retail-demand channel directly, with two identifications; *neither supports it* (full detail in Table 18, Internet Appendix E). *(i) Time-series state dependence.* Splitting high/low retail states by the expanding-window median of the (one-period-lagged) individual-investor trading share, SVAR's grouped R^2_{OOS} shows an apparent gradient (high retail 20.19% versus low retail 2.37%), but the high-retail months fall almost entirely in the early sample, and **after removing the 9 financial-crisis months the gradient disappears entirely and even reverses** (SVAR 2.97%/3.65%)—the state dependence is a repackaging of crisis concentration, not evidence for a retail channel. *(ii) Cross-sectional within-group split.* Splitting stocks into retail-heavy and institution-heavy terciles by firm-level ownership and building the iskew decile value-weighted long–short factor *within*

each group, the result runs *in the opposite direction*: the institution-heavy group has the larger absolute slope and in-sample R^2 (holdings-based 3.1% versus 0.5%), the difference between the two groups is insignificant in a monthly block bootstrap ($t = 0.75$–1.13 for $\Delta\beta$), and a 2,000-draw random-grouping placebo places each statistic in the central region (0.11–0.87; none in the 5% tails). The institution-heavy signal moreover vanishes under *equal weighting* and is fully replicated by the *large-cap* tercile—iskew's transmission to TXO concentrates in large-cap stocks that overlap heavily with the TAIEX, unrelated to investor structure. We conclude that the retail-participation proxies currently available cannot identify a retail-driven predictability channel. This is a failure to identify rather than a rejection of the channel itself, and retail structure accordingly remains institutional background throughout the paper; iskew's transmission is better described as an index-level common-skewness state carried by index-constituent overlap (Section 5). Sharper identification would require order-flow or position-level data beyond these proxies.

8.4 Anomaly Spillover: Null-Consistency

The cross-sectional strength of Taiwan's 153 anomaly factors is clearly weak on its own: by Newey–West t, only $16/153 = 10.5\%$ have $|t| > 1.96$, about a fifth of the Jensen et al. (2023) U.S. benchmark (a 55.6% replication rate for $|t| > 1.96$ on raw returns), and no one of the 13 themes reaches a 50% significant share (the highest, Value, is 44%). The direction, however, remains consistent (80.4% of factors have the correct sign, the cross-factor average annualized long–short return is about +2%, and Value/Momentum are robust), so the extreme "no signal at all" reading does not hold.

As for spillover into the index option: of all 153 factors, $10/153 = 6.5\%$ show nominal spillover into TXO ($R^2_{\text{OOS}} > 0$ and CW $p < 0.10$), close to the 10% nominal false-positive rate of the CW test. Independence is the wrong null here, though, because the 153 factors are strongly cross-correlated. We therefore build the null distribution of the discovery count by *stationary block bootstrap*: holding the factor panel fixed to preserve its cross-sectional correlation and resampling the target series (block ≈ 6, 2000 resamples, seed 42) to break any true target–factor link, we recount nominal discoveries under each draw. The null count has mean 9.45, standard deviation 8.37, and a 95th percentile of 27—far wider than the binomial benchmark under independence. The observed 10 discoveries fall at the 65.5th percentile of this null (right-tail $p = 0.344$), so the count is **consistent with chance under the nominal threshold**.[7] There is a theoretical reason to take the test seriously: individual-stock anomaly signals could transmit to index options through the aggregation logic of Dong et al. (2022), whereby the systematic component of cross-sectional mispricing maps into an index-level risk-premium state, so the null test has a genuine theoretical counterpart rather than being a mechanical multiple-comparison exercise.

This is the paper's anomaly boundary: no individual factor's nominal spillover can be separated from a false positive, though the directional consistency documented above cautions against reading the Taiwan cross-section as pure noise.

[7] Under the dividend-omitting target the count is 12/153, at the 75th percentile of its null (right-tail $p = 0.25$); the small difference comes entirely from the dividend correction, not from the method. A complementary conditional check gives the same reading: restricting to the 16 factors that are cross-sectionally significant in Taiwan, only 2 spill over, statistically indistinguishable from the full set (Fisher $p = 0.36$).

8.5 The Complexity Boundary: No Gains from the Nonlinear Methods Considered

This subsection delimits the complexity ceiling of predictability: *none of the nonlinear specifications we test provides reliable incremental information beyond the corresponding parsimonious linear variance specification.* Three pieces of evidence support this—linear diagnostics, a forecast-encompassing test, and a calibrated power analysis. The claim concerns the methods and sample at hand, not nonlinearity in general.

Motivation for a linear benchmark. From Section 5.3, under the Bakshi–Kapadia decomposition the leading predictable term of the delta-hedged return is $-c\,\mathrm{IV}^2$, which *motivates* a linear variance-based benchmark. This does not rule out nonlinearity: state-dependent gamma, a nonlinear map from predictors to expected realized variance, jumps, vol-of-vol, skew dynamics, and nonlinearity in the return denominator could all generate higher-order terms. We therefore treat the linear benchmark as a hypothesis to be tested, not as a theoretical ceiling, and we do not claim a universal upper bound on complexity.

Empirical linear diagnostics. For IV^2 and SVAR, the RESET, polynomial, spline, Hansen threshold (Hansen, 2000), and crisis-dummy interaction tests all fail to reject linearity (minimum $p = 0.16$), and the OOS increment of any nonlinear extension is negative (the polynomial is still -43pp after truncated extrapolation, not merely an extrapolation blow-up). Table 13 shows that the tree family (random forest, XGBoost) is uniformly mediocre on non-variance-channel variable sets: by the DM test (*non-nested; CW is not usable*), only the RF on Macro-13 (with IV^2/VRP) is significant ($R^2_{\mathrm{OOS}} = 12.35\%$, DM $p = 0.009$; 19.12% under the wide hyperparameter search), but its forecast correlates 0.65 with the IV^2 univariate forecast, and a leave-channel-out retraining shows the gain flows through the variance channel: removing IV^2/VRP and retraining collapses the RF from 12.35% to 3.51%, the XGBoost from 12.16% to 4.79%, and the wide-search RF from 19.12% to -0.92%—at least seven-tenths of the tree gain rides on the mechanical variance channel,[8] and it is still *below* the 27.63% of the IV^2 linear univariate. A matched-information-set Diebold–Mariano test makes the point directly: no tree beats the corresponding linear specification on the same predictor set (one-sided $p = 0.94$–0.99)—the tree is only a noisier repackaging of the variance channel, with no incremental nonlinear alpha; on Anomaly-153 both are negative. The intermediate-complexity methods (GAM 18.13%, KRR 15.40%) are likewise below linear. Across the complexity spectrum, from OLS to GAM/KRR to trees, no nonlinear alternative we consider improves on the linear benchmark.

Forecast encompassing. A direct encompassing test confirms that the nonlinear methods carry no information beyond the linear model. Using the Harvey–Leybourne–Newbold (Harvey et al., 1997) statistic, the forward test that the linear SVAR forecast encompasses the shallow random forest (Macro-11, $R^2_{\mathrm{OOS}} = 1.96\%$) fails to reject ($t = 0.05$, $p = 0.963$), while the reverse—that the random forest encompasses the linear model—is strongly rejected ($t = 3.06$, $p = 0.002$); the random-forest signal correlates only 0.005 with the linear model's out-of-sample residual. The nonlinear method thus adds nothing the linear variance channel does not already contain.

Calibrated power analysis. An empirical null at $T = 236$ could simply reflect low power, so we measure the power of the tree methods directly with a Monte Carlo

[8] We rely on this retraining comparison rather than impurity- or permutation-based feature importance, whose attribution is unreliable under correlated inputs.

calibrated to our data; Appendix C details the data-generating process, the signal-strength calibration, the error distribution, the number of simulations, the tree tuning, and the hit definition. For *interaction*-type nonlinearity at only 25% of the linear signal strength, the random forest detects the signal in about 73–78% of simulations at $T = 236$, so the uniform empirical absence is unlikely to reflect low power alone, though the remaining type-II error probability of 22–27% means absence is not proof. For *extreme-threshold* types (a high-volatility regime with a five-fold slope amplification), a flexible random forest has about 77% power, while the real data show no gain across the hyperparameter landscape, so this type is likewise unsupported. The *mild-threshold* type (slope amplification $k \lesssim 2$) sits beyond the power of every tool at this sample length; the Hansen test's failure to reject linearity ($p \geq 0.16$) is the only evidence we can offer there, and we flag it as an open limitation. Within the range these tools can identify, the data are consistent with a linear variance channel.

Table 13: Out-of-sample performance of nonlinear methods (DM inference, non-nested)

Method / variable set	R^2_{OOS} (%)	DM p	Note
RF | Macro-11	3.51	0.101	< linear Combine (8.36)
RF | Anomaly-153	−0.09	0.519	sample-length boundary
RF | Macro-13 (+IV^2/VRP)	12.35	0.009	inherits variance channel; corr. 0.65 with IV^2 forecast
RF | Macro-13 (wide search)	19.12	0.023	still < linear IV^2 (27.63)
GAM | Macro-13	18.13	0.082	< linear
KRR | Macro-13	15.40	0.067	< linear
Linear channel benchmark: IV^2 univariate R^2_{OOS} = 27.63%; Macro-13 tPCA-s 26.73%			

Notes: Every DM p-value in this table compares the listed model with the *recursive prevailing mean*: the loss differential is $d_t = (r_t - \bar{r}_t)^2 - (r_t - \hat{r}_t)^2$, tested one-sided ($H_0$: $\mathbb{E}[d_t] \leq 0$) with HAC standard errors; these are non-nested comparisons, so Clark–West is not used, and significance against the prevailing mean says nothing about superiority or inferiority relative to IV^2 or any other model. Matched-information-set comparisons are reported in the text: with $d_t = (r_t - \hat{r}_t^{\text{lin}})^2 - (r_t - \hat{r}_t^{\text{tree}})^2$, no tree beats the linear specification on the same predictor set (one-sided $p = 0.94$–0.99), and the only DM-significant cell (Macro-13) inherits its gain from the mechanical variance channel of IV^2. The calibrated power analysis gives trees 73–78% power against moderate interaction-type nonlinearity at $T = 236$, so the uniform empirical null is unlikely to reflect low power alone; the mild-threshold type ($k \lesssim 2$) remains an acknowledged power gap.

9 Conclusion

We ask three questions of the Taiwan index option market: whether the monthly delta-hedged return on the ATM put is predictable out of sample, whether any predictability is stable, and whether it carries economic value. Under a pseudo-real-time expanding-window design, with a target built from a dividend-inclusive P&L identity (a mean put premium of about −0.71% per month over 2002–2022), the answers are a narrow yes, a no, and a no at the margin.

Predictability is real but sparse. Of 164 candidate predictors, SVAR alone survives the full-sample Holm slope correction, and its out-of-sample R^2 over the 2007–2022 evaluation window is 11.03%; no predictor survives the analogous correction applied to the out-of-sample Clark–West family, so we make no multiple-testing-corrected OOS claim. Point estimates concentrate the signal in at-the-money puts, though the put–call difference is not formally separable. Across the nonlinear alternatives we consider, none

reliably improves on the linear variance benchmark, and Taiwan's anomaly factors do not translate into an index-option signal distinguishable from chance.

The signal is not stable. Excluding November 2008 cuts SVAR's out-of-sample R^2 to 5.35%, excluding the five largest-gain months to 0.22%, and the 2015–2022 subsample delivers 0.69%. In the frozen 2023–2025 holdout the put premium shrinks to −0.24% per month and the SVAR forecast falls behind the historical mean; over the extended 2002–2025 sample the premium is −0.65% and SVAR's extended-window R^2_{OOS} of 9.09% remains significant but is carried by the crisis years. The one positive holdout signal, the combination forecast, runs through valuation ratios rather than the variance channel, on a window too short to distinguish durable content from luck; it is the finding most worth re-examining as the post-2022 sample lengthens.

Economic value at the margin of prediction is negligible. A constant short-volatility position attains 98% of the benchmark's certainty-equivalent return, and every predictive increment up the strategy ladder has a bootstrap confidence interval containing zero: what value exists comes from unconditionally bearing the variance premium, not from timing it.

Two findings stand somewhat apart. Entry-date implied variance explains the target return largely by construction—the definitional leg accounts for about two-thirds of the decomposed slope magnitudes—so its apparent forecasting supremacy measures the variance-pricing channel rather than predictive skill. And same-strike puts are persistently richer than calls by about 0.2–0.25% of the underlying per month, a wedge that a direct test ties to the Taiwan index-futures backwardation (correlation 0.99, near-unit slope): the backwardation's option-market mirror, consistent with downside-insurance demand though not uniquely attributable to it. Retail structure, Taiwan's most salient institutional feature, remains background, as neither of our identifications supports a retail channel; order-flow or position-level data may yet sharpen that test. To our knowledge these results constitute the first systematic pseudo-real-time out-of-sample evaluation of monthly delta-hedged index-option return forecasts in an Asia-Pacific market, and their practical content is a map of where predictability ends.

Declaration of interest

No potential conflict of interest was reported by the authors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI ChatGPT and Anthropic Claude for language proofreading, programming assistance, and reproducibility checks. The authors independently reviewed and verified all text, code, analyses, and references and take full responsibility for the content of the article.

References

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*(2), 304–317.

Bakshi, G. and N. Kapadia (2003). Delta-hedged gains and the negative market volatility risk premium. *Review of Financial Studies 16*(2), 527–566.

Bali, T. G., H. Beckmeyer, M. Moerke, and F. Weigert (2023). Option return predictability with machine learning and big data. *Review of Financial Studies 36*(9), 3548–3602.

Bali, T. G. and S. Murray (2013). Does risk-neutral skewness predict the cross-section of equity option portfolio returns? *Journal of Financial and Quantitative Analysis 48*(4), 1145–1171.

Barber, B. M., Y.-T. Lee, Y.-J. Liu, and T. Odean (2009). Just how much do individual investors lose by trading? *Review of Financial Studies 22*(2), 609–632.

Barberis, N. and M. Huang (2008). Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review 98*(5), 2066–2100.

Beckmeyer, H., G. Tong, and G. Zhou (2023). Expected index option return: What can we learn from macro and anomalies? SSRN Working Paper 4674294.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 289–300.

Bergstra, J. and Y. Bengio (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research 13*, 281–305.

Black, F. and M. Scholes (1973). The pricing of options and corporate liabilities. *Journal of Political Economy 81*(3), 637–654.

Bollerslev, T., J. Marrone, L. Xu, and H. Zhou (2014). Stock return predictability and variance risk premia: Statistical inference and international evidence. *Journal of Financial and Quantitative Analysis 49*(3), 633–661.

Bollerslev, T., G. Tauchen, and H. Zhou (2009). Expected stock returns and variance risk premia. *Review of Financial Studies 22*(11), 4463–4492.

Bryzgalova, S., A. Pavlova, and T. Sikorskaya (2023). Retail trading in options and the rise of the big three wholesalers. *Journal of Finance 78*(6), 3465–3514.

Campbell, J. Y. (1987). Stock returns and the term structure. *Journal of Financial Economics 18*(2), 373–399.

Campbell, J. Y. and R. J. Shiller (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies 1*(3), 195–228.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies 21*(4), 1509–1531.

Cao, J. and B. Han (2013). Cross-section of option returns and idiosyncratic stock volatility. *Journal of Financial Economics 108*(1), 231–249.

Carr, P. and L. Wu (2009). Variance risk premiums. *Review of Financial Studies 22*(3), 1311–1341.

Chambers, D. R., M. Foy, J. Liebner, and Q. Lu (2014). Index option returns: Still puzzling. *Review of Financial Studies 27*(6), 1915–1928.

Chang, C.-C., P.-F. Hsieh, and H.-N. Lai (2009). Do informed option investors predict stock returns? evidence from the Taiwan Stock Exchange. *Journal of Banking and Finance 33*(4), 757–764.

Chang, C.-C., P.-F. Hsieh, and Y.-H. Wang (2010). Information content of options trading volume for future volatility: Evidence from the Taiwan options market. *Journal of Banking and Finance 34*(1), 174–183.

Chen, A. Y. and M. Velikov (2023). Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis 58*(3), 968–1004.

Chordia, T., A. Kurov, D. Muravyev, and A. Subrahmanyam (2021). Index option trading activity and market returns. *Management Science 67*(3), 1758–1778.

Clark, T. E. and K. D. West (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics 138*(1), 291–311.

Coval, J. D. and T. Shumway (2001). Expected option returns. *Journal of Finance 56*(3), 983–1009.

Dong, X., Y. Li, D. E. Rapach, and G. Zhou (2022). Anomalies and the expected market return. *Journal of Finance 77*(1), 639–681.

Drechsler, I. and A. Yaron (2011). What's vol got to do with it. *Review of Financial Studies 24*(1), 1–45.

Garleanu, N., L. H. Pedersen, and A. M. Poteshman (2009). Demand-based option pricing. *Review of Financial Studies 22*(10), 4259–4299.

Giacomini, R. and B. Rossi (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics 25*(4), 595–620.

Goyal, A. and A. Saretto (2009). Cross-section of option returns and volatility. *Journal of Financial Economics 94*(2), 310–326.

Goyal, A. and I. Welch (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*(4), 1455–1508.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies 33*(5), 2223–2273.

Han, B. (2008). Investor sentiment and option prices. *Review of Financial Studies 21*(1), 387–414.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica 68*(3), 575–603.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting 13*(2), 281–291.

Heston, S. L., C. G. Jones, M. Khorram, S. Li, and H. Mo (2023). Option momentum. *Journal of Finance 78*(6), 3141–3192.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*(2), 65–70.

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (2023). Is there a replication crisis in finance? *Journal of Finance 78*(5), 2465–2518.

Kelly, B. and S. Pruitt (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics 186*(2), 294–316.

Kilic, M. and I. Shaliastovich (2019). Good and bad variance premia and expected returns. *Management Science 65*(6), 2522–2544.

Kumar, A. (2009). Who gambles in the stock market? *Journal of Finance 64*(4), 1889–1933.

Lee, Y.-H. and D. K. Wang (2016). Information content of investor trading behavior: Evidence from Taiwan index options market. *Pacific-Basin Finance Journal 38*, 149–160.

Lo, C.-L. and W.-R. Liu (2025). Low risk, high return: Improving option writing performance with put-call ratios in Taiwan. *Pacific-Basin Finance Journal 90*, 102687.

Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science 4*(1), 141–183.

Mitton, T. and K. Vorkink (2007). Equilibrium underdiversification and the preference for skewness. *Review of Financial Studies 20*(4), 1255–1288.

Muravyev, D. (2016). Order flow and expected option returns. *Journal of Finance 71*(2), 673–708.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical Association 89*(428), 1303–1313.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies 23*(2), 821–862.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B 58*(1), 267–288.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 135–196. Elsevier.

Yang, C.-Y., L.-J. Jhang, and C.-C. Chang (2016). Do investor sentiment, weather and catastrophe effects improve hedging performance? Evidence from the Taiwan options market. *Pacific-Basin Finance Journal 37*, 35–51.

Zhan, X., B. Han, J. Cao, and Q. Tong (2022). Option return predictability. *Review of Financial Studies 35*(3), 1394–1442.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B 67*(2), 301–320.

Appendix

A Variable Definitions

Timing. $x_{i,t}$ is the most recent observation of predictor i available before the entry date of month t; the target R^{dh}_{t+1} is the delta-hedged return realized from the third Thursday of month t to the third Wednesday of the following month. Macro variables with release delays use the latest data made public as of the entry date.
Macroeconomic predictors. DP: trailing twelve-month cumulative dividend yield divided by the month-end index. EP: the inverse of the price–earnings ratio. BM: the inverse of the price-to-book ratio. DE: DP/EP. DY: dividends paid in the current month divided by the prior month's index. NTIS: trailing twelve-month net equity issuance as a share of market capitalization. SVAR: the sum of squared daily log returns of the weighted index in the immediately preceding completed calendar month (the predictor for entry month M is measured over $M-1$; see Section 4.4). TMS: the ten-year government bond yield minus the one-year time-deposit rate. INFL: the month-over-month growth of the CPI (released promptly, needing no extra lag). LTBY: the secondary-market yield of the ten-year government bond. RF: the one-year time-deposit rate (monthly). Sources: TWSE, TEJ, the Central Bank, and the Directorate-General of Budget, Accounting and Statistics.
Equity anomaly factors. The 153 factors are from the Taiwan sample of the Jensen et al. (2023) global factor database, as value-weighted long–short decile spreads rebalanced monthly; accounting variables use a six-month reporting lag and price-based variables use the month-end close. The *iskew_capm_21d* cited in the text is the idiosyncratic skewness of daily stock returns relative to the CAPM over the past 21 trading days.

B Model Estimation Details

At each forecast origin $t \in \{60, \ldots, 235\}$: (i) assemble the in-sample data; (ii) compute the expanding-window prevailing mean as the benchmark; and (iii) for each model, standardize by the training-set mean/standard deviation, estimate parameters using only in-sample data, and generate the out-of-sample forecast. The ENet's (λ, ρ) is selected within each window by five-fold time-series cross-validation ($\lambda \in [10^{-4}, 10^{2}]$ log-uniform, $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$, 30 random draws; Bergstra and Bengio, 2012). Combine/P-Aver/tPCA select on in-sample $|t| > T^{*} = 3$ (a conservative setting that reduces misselection in short samples). The tree models' hyperparameters are selected within each window by three-fold time-series cross-validation (`TimeSeriesSplit`), with inference by DM; to delimit the complexity boundary, we also use a fixed-hyperparameter landscape (an a priori fixed grid, each combination run over the full expanding-window OOS independently) to assess "whether the right combination could have been selected ex ante." Imposing economically motivated restrictions—Campbell–Thompson non-negativity truncation and sign restrictions on the predictive slope—leaves the SVAR result essentially unchanged (Table 11, CT column), so the finding does not depend on ignoring economic priors.

The wide hyperparameter search ranges over the following grids. **Random forest**: number of trees $\{200, 500, 1000\}$, maximum depth $\{2, 3, 4, 6, \text{None}\}$, minimum leaf size $\{1, 5, 10, 20\}$, and maximum features $\{\sqrt{\cdot}, 0.3, 0.5, 1.0\}$. **XGBoost**: number of trees

$\{100, 300, 600\}$, maximum depth $\{1, 2, 3, 4, 6\}$, learning rate $\{0.01, 0.03, 0.1\}$, subsample $\{0.5, 0.8, 1.0\}$, column subsample $\{0.3, 0.6, 1.0\}$, L_2 penalty $\{1, 10, 100\}$, and minimum child weight $\{1, 5, 20\}$. In the cross-validation stage the tree count is proxied by $\min(n, 300)$ for cost, and the selected combination is refit with the full tree count.

C Calibrated Monte Carlo Power Analysis

The power analysis in Section 8.5 uses a data-generating process (DGP) calibrated to our sample. The predictor x_t (standing in for IV^2) follows an AR(1) whose innovations are drawn by residual bootstrap to preserve right skewness; z_t (standing in for SVAR) is generated as $z_t = c_0 + c_1 x_t + v_t$ with the empirical correlation (≈ 0.62) and a residual bootstrap for v_t; and the target is $y_t = a + b\,x_t + e_t$, with a and $b \approx -1.97$ taken from the univariate regression and e_t drawn by residual bootstrap so that the in-sample R^2 ($\approx 26\%$) and the residual skewness and kurtosis match the data. Ten AR(1) noise features, independent of y, are added to the feature set.

We run two experiments. *Experiment 1* (linear DGP, $T = 236$) uses the features x plus ten noise columns and compares OLS (all features) with the random forest and XGBoost on an expanding-window OOS protocol (initial window 60, refit every 12 months, identical for all three models), establishing the size baseline—the probability that a tree beats OLS when the DGP is linear. *Experiment 2* (nonlinear DGP power surface) adds either (a) a *threshold*: in the high-x regime (top 20%) the slope is amplified by $k \in \{1.5, 2, 3, 5\}$; or (b) an *interaction*: a term $\gamma\,\tilde{x}\,\tilde{z}$ with $\text{var}(\text{interaction})/\text{var}(b\,x) = w \in \{0.25, 0.5, 1, 2\}$; over sample lengths $T \in \{236, 500, 1000, 2000, 5000\}$, under a 70/30 time-series split with a single fit. The reported metric (the "hit rate") is Pr(tree OOS MSPE $<$ OLS OOS MSPE), i.e., the probability that the tree beats the linear model. Each cell uses 500 simulations with seed 42 plus the cell index. The tree hyperparameters are fixed at values near the median of the wide-search landscape (random forest: 200 trees, depth 4, minimum leaf 10, max features 0.3; XGBoost: 300 trees, depth 2, learning rate 0.03, subsample 0.8, column subsample 0.6, L_2 penalty 10), a simplification that if anything favors the trees. These details support the Section 8.5 conclusion that low power alone is unlikely to explain the uniform empirical absence of moderate interaction-type or extreme-threshold nonlinearity, while the mild-threshold type ($k \lesssim 2$) remains a power gap we cannot close at this sample length.

D Coverage and Supplementary Tables

Table 14 reports the coverage of the moneyness grid (the number of months with a selectable contract at entry and a complete holding-period path). The deep-in-the-money side (puts with $K/S \geq 1.05$, calls with $K/S \leq 0.95$) has insufficient coverage, consistent with the default of "puts for $K/S < 1$, calls for $K/S > 1$"; the only downgraded cell is K110_CALL (126 months, since 10% out-of-the-money contracts were often not listed in low-index years), which the formal analysis downgrades to the appendix.

Table 14: Moneyness-grid coverage (path_ok months, out of 252 entry months)

K/S	Put	Call	Use
0.90	**204**	70	Put main series
0.95	**234**	121	Put main series
1.00	**229**	**208**	Put + call both main
1.05	145	**220**	Call main series
1.10	69	126	Call downgraded to appendix (< 150)

Notes: path_ok is the number of months with a selectable contract at entry and a complete daily hedging path over the holding period. Bold marks the main series used (≥ 150).

E Internet Appendix: Supplementary Tables

This Internet Appendix collects the diagnostic tables referenced in the main text. Every result's conclusion and headline numbers are stated in the body; the tables below carry the supporting detail and are presented in order of first reference: the ATM/K100 bridge (Table 15), the tail and leverage profile of the short-volatility strategies (Table 16), the crisis-month anatomy (Table 17), and the within-group retail split (Table 18).

E.1 Bridging the Main-Table ATM Put and the Grid ATM Put

Table 15 decomposes the headline 11.03%-versus-5.40% gap between the main-table ATM put and the grid K100 put into its contract-selection component (Section 6.3).

Table 15: Bridging the ATM and grid-K100 put series (SVAR R^2_{OOS})

Series	Sample	N	R^2_{OOS} (%)	CW p
ATM put	full	236	11.03	0.008
K100 put	full	229	5.40	0.006
ATM put	common months	227	7.71	0.006
K100 put	common months	227	6.30	0.008
ATM put	common, ex 2008-11	226	3.16	0.004
K100 put	common, ex 2008-11	226	2.15	0.013

Notes: On the 227 common months the two put series correlate 0.96; their SVAR R^2_{OOS} differ by only 1.4pp and both collapse toward 2–3% once the November 2008 crisis month is removed. The headline 11.03%-vs-5.40% gap is therefore driven by sample composition (the K100 band drops high-variance months), and predictability is crisis-driven under both rules.

E.2 Tail and Leverage Profile of the Short-Volatility Strategies

Table 16 reports the tail and leverage profile underlying the economic-value analysis of Section 7.

Table 16: Tail and leverage profile of the short-volatility strategies

Strategy	avg $\lvert w\rvert$	cap-hit freq	turnover	worst month (%)	MaxDD (%)	ES_{95} (%)	skew	excess kurt
Benchmark	1.78	0.716	0.036	−9.86	−12.64	−6.48	−0.94	3.35
SVAR univariate	1.62	0.551	0.259	−9.86	−12.75	−6.26	+0.32	6.96
IV^2 univariate	1.50	0.352	0.278	−7.99	−11.12	−5.35	+0.91	6.64

Notes: "cap-hit freq" is the fraction of months at the ±200% weight cap; ES_{95} is the monthly 95% expected shortfall; MaxDD is the maximum drawdown. The CER and these statistics are computed per unit of self-financing investment notional and do *not* net out exchange margin, regulatory capital, or market impact. Crisis-month net returns are positive for all strategies (benchmark +6.83% in 2008-11 and +8.16% in 2020-03; IV^2 +16.60% and +8.74%).

E.3 Crisis-Month Anatomy

Table 17 gives the month-by-month mechanics of the crisis gains discussed in Section 8.

Table 17: Crisis-month anatomy: entry conditions, SVAR forecast, and realized return

Month	Entry date	Entry IV (ann.)	Prev-month RV	Current-month RV	SVAR forecast (%)	Actual return (%)
2008-10	2008-10-16	0.544	0.0190	0.0203	−3.08	−3.99
2008-11	2008-11-20	0.664	0.0203	0.0119	−3.49	−8.58
2008-12	2008-12-18	0.412	0.0119	0.0117	−2.63	−2.18
2011-08	2011-08-18	0.263	0.0018	0.0105	−0.54	+0.33
2015-08	2015-08-20	0.177	0.0023	0.0069	−0.61	+0.95
2020-03	2020-03-19	0.605	0.0016	0.0200	−0.43	−5.84

Notes: The 2008 gains come from high prior-month realized variance producing a strongly negative SVAR forecast that the market validated. March 2020 is a tail event the model missed (low prior-month RV, small forecast, large realized loss), illustrating that the predictive power is concentrated rather than a comprehensive tail hedge.

E.4 Within-Group Retail Split

Table 18 reports the cross-sectional within-group retail identification test of Section 8.

Table 18: Within-group iskew split predicting TXO put returns (main target)

Group factor	$\hat{\beta}$	NW t	R^2_{IS} (%)	R^2_{OOS} (%)	CW p
Trading-based H_{retail} (high retail)	−0.033	−1.75	2.1	−0.1	0.085
Trading-based L_{inst} (institution-heavy)	−0.064	−1.37	2.8	−0.7	0.097
Ownership-based H_{retail}	−0.026	−1.32	0.5	−1.6	0.281
Ownership-based L_{inst}	−0.073	−1.79	3.1	−1.9	0.144
$\Delta\beta$ (retail − institution, trading)	+0.031	0.75	—	—	—
$\Delta\beta$ (retail − institution, ownership)	+0.047	1.00	—	—	—
JKP official iskew (same-window baseline)	−0.235	−2.91	8.7	−1.0	0.029

Notes: The within-group iskew decile value-weighted long–short factor (low iskew − high iskew, same JKP direction) predicts the TXO ATM put delta-hedged return. In-sample uses NW(4), $n \approx 163$–168; OOS uses an expanding window with an initial length of 48 months (from about 2012). The institution-heavy group's point estimate is if anything stronger, the difference is insignificant, and the placebo test confirms it as random-grouping noise. The JKP official full-pool iskew is still significant in sample over the same window ($t = -2.91$), showing that the sample shortening alone does not destroy the signal.

E.5 Holdout Data-Source Sensitivity

The 2025 aggregate valuation inputs (DP, EP, BM, DE, DY, NTIS) used in the frozen post-2022 holdout (Table 12, Section 8.2) are built from TEJ's market-level monthly series, which track the official TWSE series closely over 2023–2024; rebuilding the entire 2023–2025 segment from the TEJ source alone shifts no holdout R^2_{OOS} by more than 0.05 percentage points, so the source choice is immaterial to any conclusion.



Click here to access/download

LaTeX Source File

latex_source_anon.zip