

7504478-what-substitutes-for-a-stale-mark-breadth-implied-volatility-and-the-price-of-be

SUMMARY: The article examines the use of breadth (correlations with other markets) and implied volatility from the options market as substitutes for delayed market data in forecasting. The study uses S&P 500 realised variance as a case study, withholding the forecaster's own data for varying periods and measuring the performance of other sources. The findings indicate that a current foreign cross-section returns 71% [59, 81] of what a delay of eleven weeks would have cost, and that an options market on the asset itself makes the cross-section redundant. The study also finds that the substitution rate can be measured before any purchases are made, and that the effect of extra staleness on housing prices is significant.

METHOD: The study uses S&P 500 realised variance as a case study, with the forecaster's own data withheld for varying periods (δ days). Other sources are left untouched, and the performance of each is measured. The experiment is conducted where the truth is observable, using a housing-price index panel with similar characteristics to the private equity portfolio described in the introduction. The results are measured on the same set of days as a previous section.

KEY FINDINGS:

- A current foreign cross-section returns 71% [59, 81] of what a delay of eleven weeks would have cost.
- An options market on the asset itself makes the cross-section redundant, but only on the asset itself.
- The substitution rate can be measured before any purchases are made.
- Extra staleness on housing prices gives 67.8% [45, 92] on the next month's return and 19.0% [6, 31] on a backward-looking variance.

IMPLICATION FOR TRADING: The findings suggest that using breadth and implied volatility from the options market can be a useful substitute for delayed market data. Practitioners can use the substitution rate to determine which sources to use and how much they can rely on them. The study's results can inform trading strategies by providing insights into the effectiveness of different substitutes for delayed market data.

What Substitutes for a Stale Mark? Breadth, Implied Volatility, and the Price of Being Late

Sunny Alex Vellankaran

Independent

22 September 2026

Corresponding author. Sunny Alex Vellankaran, Independent. Email sunnyalex1234@gmail.com.

Abstract

A forecaster holding a stale mark must buy something in place of timeliness, with no accepted way to price that trade. Handicapping an S&P 500 realised-variance forecaster by δ days, other sources current, we define the substitution rate $R(\delta)$, the share of delay damage a substitute repairs, and treat its inference as the weak-identification problem it is. A current foreign cross-section returns 71% [59, 81] of what eleven weeks of staleness takes. In both own-index option cases examined, the compressed cross-section adds no detectable information once own-index implied volatility is included; a chain on another asset does not do the same. Twenty house price indices carry it to a smoothed market with no listed options chain of its own.

Keywords: information delay; substitution rate; volatility forecasting; ragged edge; weak identification.

JEL codes: C53, C58, G17, G12, C22.

1. Introduction

A private equity portfolio is marked once a quarter. Direct property is appraised on a cycle, and the resulting series is smoothed in ways that have been documented for thirty years (Geltner et al. 2003). Hedge fund returns arrive with a serial correlation that understates their true volatility (Getmansky et al. 2004). In each case someone must produce a risk number today from marks that are weeks or months old, while public markets have gone on trading in plain sight.

The instinct is to reach for those markets. The question is which of them, and how far it gets you. Two candidates present themselves. The first is breadth: other markets are correlated with yours, and some of them have already closed today. The second is the options market, which reports a forward-looking risk number continuously and without any delay at all. These are different purchases, and a practitioner choosing between a wider data feed and an options subscription would like to know which one buys back more of what staleness took.

We cannot observe the truth for a private portfolio, so we run the experiment where the truth is observable. We take a forecaster of S&P 500 realised variance, withhold their own data for δ days, leave the other sources untouched, and measure what each gives back. For foreign equities the asymmetry costs nothing to construct: Tokyo closes at 06:00 UTC, Hong Kong at 08:00, Mumbai at 10:00, London and Frankfurt at 16:30, New York at 21:00. On any given day the rest of the world has already finished trading when New York does.

An experiment run where the truth is observable is still a proxy for the case that motivates it, and a paper that stopped there would be asking its reader to extrapolate from liquid equities to appraised assets and to trust the extrapolation. The last step is therefore to narrow that extrapolation as far as public data allows, with a housing-index panel that has appraisal-like persistence and infrequent

marking. Twenty metropolitan house price indices, fourteen of them published monthly since January 1987 and the rest beginning later, carry the properties the first paragraph describes, though they remain a published index and not a book: each monthly figure is a three-month moving average of closings, no options chain on any of the twenty is currently listed and no daily own-market implied-volatility index exists for them, so the substitute Section 7 races against is unavailable here; CME does list futures on the composite and on ten of the metro indices, which is a derivative but not the one that race needs, and all twenty are published on the same day, so a stale mark sits inside a fresh cross-section exactly as the design requires. The same measurement is made there, on the same estimator, and it is the result the paper is built to deliver.

Four findings carry it. First, breadth substitutes for timeliness and the amount can be priced: at eleven weeks of staleness a current foreign cross-section returns 71% [59, 81] of what the delay took, measured on the implementable target, the original model, the stronger control, the compressed specifications and the state-space alternative all landing between 68% and 75%, and 81% [74, 87] in this sample's stressed markets where the estimate is tight. Read as a clock rather than a share, a fifty-five-day-old mark plus today's foreign close forecasts as well as a mark 4.6 days old [2.5, 8.4]. Second, an options market on the asset itself makes the cross-section redundant, but only on the asset itself: where the only chain available is written on another index the two candidates finish three points apart rather than thirty-one, and a compressed cross-section still adds in thirteen of twenty-four cells. Third, which of those worlds a holder is in can be measured before anything is bought, because the target's own correlation with the cross-section tracks the substitution rate at +0.98 across eight markets, and where that correlation is small the cross-section repairs nothing and charges for the attempt. Fourth, on the house price panel one quarterly appraisal cycle of extra staleness gives 67.8% [45, 92] on the next month's return and 19.0% [6, 31] on a backward-looking variance, positive on fourteen metros of fourteen and thirteen of fourteen.

Those figures are not competing estimates of one quantity and it is worth saying so before a reader tries to reconcile them. Each answers a different question, and Table 1 puts the question beside the number so that the paper's own hierarchy is visible in one place rather than assembled from six sections.

Table 1. The substitution rate at eleven weeks under each specification the paper regards as defensible, with the question each one answers. All are R^2 -scale. The first five are on the same 4,862 test days as Section 5; the stressed row is the 1,618 of those days in the upper volatility tercile; and the last is a different experiment on a different asset class at a different delay, 310 test months, included because it is the result the paper is built to deliver. No row here uses the 4,525-day window of Section 7, which is shortened by VDAX-NEW availability and carries the options race alone. Intervals are the full-ratio bootstrap of Section 6. Two rows carry none, because their evidence is a paired difference rather than a marginal interval: the stronger control is 4.1 points below the paper's own, interval $[-9.7, -0.6]$ on that difference (Section S4), and the state-space filter is 3.8 points above it (Section S20).

specification	R(55)	the question it answers
implementable target, original block	71% [59, 81]	what a forecaster could have acted on
strongest domestic control	68%	is it an artefact of a weak benchmark?
equal-weighted cross-section	74.3% [64.4, 82.7]	what a practitioner should actually run
one real-time factor	75.2% [65.4, 83.5]	where in the block the recovery lives
state-space alternative	75.4%	does a better estimator change it?
stressed markets only	81% [74, 87]	when is the mechanism worth having?
<i>house prices, one quarter, non-overlapping</i>	57%	<i>does it hold on a smoothed index of an illiquid market?</i>

Read down the second column and the paper's own reading is the obvious one: nothing here turns on which row is called the headline, because the span is narrow relative to every interval in it. The 71% is the lower choice, and lower for two separable reasons: it is the implementable target, and it is the original specification and the implementable target; the 74% to 75% band is what the compression of Section 6.1 recommends and is what a practitioner should expect; and the paper quotes the first while recommending the second, which is a gap worth naming rather than smoothing over.

The contribution is four-fold, and the first of the four is a measurement problem rather than a finding. What a stale mark costs, and what a substitute returns, are both differences in forecast skill, so the quantity a buyer actually wants is their ratio. This paper defines that ratio as $R(\delta)$, gives it a real-time design in which the training window stops $\delta + h$ days before the forecast origin so that no label used to fit the model had resolved after the stale mark was current, and treats its inference as the weak-identification problem it is: the denominator is what delay costs, it approaches zero wherever delay costs little, and a percentile bootstrap of a ratio reports finite endpoints whether or not they mean anything. The intervals reported here are a moving-block bootstrap of the whole ratio, and every one of them is then re-derived by inverting a HAC t-test after Fieller (1954), a construction free to return a bounded interval, two half-lines or the whole line. Section 4.6 reports what that returned: an ordinary interval at every delay, within 1.6 points of the bootstrap's throughout. The measure and that treatment of it are what the rest of the paper reuses, across loss functions, markets, candidate substitutes and one asset class outside equities.

The other three follow from applying it. The rate prices the trade between timeliness and breadth in a unit-free way, so that a wider data feed and an options subscription can be compared on one ruler rather than argued about. The scope condition on that rate is identified in the examined markets rather than assumed: it depends on whether an option chain exists on the asset itself, and on a coupling statistic a practitioner can compute from data already in hand. And the illiquid case, which this literature has treated as the motivating example and left unmeasured, is brought within one step of measurement, on an asset class with no intraday range, no option chain and a mark that is smoothed before anyone touches it. One step, not none: a published index is not a book.

The paper proceeds as follows. Section 2 places it against what is already known. Sections 3 and 4 give the data and the design, including the estimator, the loss functions and the definition of $R(\delta)$. Section S26 states the specification the measurements support. Sections 5 to 7 measure what breadth recovers, where it stops, and how it fares against implied volatility. Section 8 carries the measurement to house prices. Section 9 tests the result and bounds where it holds, with the working in the Internet Appendix. Section S30 says what a practitioner should do with it, Section 10 what the paper cannot say, and Section 11 what would test it next.

2. Related work

Four findings underpin this exercise, and none of them is ours.

Foreign equity volatility forecasts domestic volatility. Engle, Ito and Lin (1990) rejected the idea that volatility is a purely local phenomenon, and the spillover literature since has measured the connections in considerable detail (Corsi 2009; Diebold and Yilmaz 2014). Korkusuz augments a HAR model of S&P 500 volatility with a European block and improves on it (2025). Buncic and Gisler (2016) run the mirror experiment, forecasting seventeen foreign markets from US volatility, and report out-of-sample R^2 between 4.6% and 14.5%.

Non-synchronous closing times define which observations a forecaster may legitimately use. Zhao and Qiao say so explicitly and build an attention model whose mask prevents a closed exchange from acting as a source (Zhao and Qiao 2026). Jayawardena and co-authors forecast Tokyo volatility from markets that trade while Tokyo sleeps (2020).

A live market can price a stale one. Goetzmann, Ivković and Rouwenhorst showed that same-day US returns predict the next day's international fund NAVs strongly enough to trade on (Goetzmann et al. 2001): the mechanism of this paper stripped to its simplest form.

Implied volatility carries much of the cross-market gain. This is the finding that shapes the design below. Both Korkusuz (2025) and Buncic and Gisler (2016) report that a large share of what looks like a cross-market effect is attributable to implied volatility, the VIX in particular. Their work does not settle the matter for a *delayed* forecaster, since in both cases the domestic information set is current, but it makes the comparison unavoidable: any claim that foreign markets substitute for timeliness has to be set against an options market that never goes stale.

A second literature has the same problem and a different name for it. Macroeconomic nowcasting is built around exactly the asymmetry this paper constructs on purpose: series arrive on different schedules, so the most recent rows of a panel are present for some variables and missing for others. Wallis called it the ragged edge (1986), and the modern treatment is a dynamic factor model with a Kalman filter that handles the missing rows in the measurement equation rather than by imputation, which is how Giannone, Reichlin and Small measure the value of each release as it lands (2008) and how Doz, Giannone and Reichlin estimate the factor in two steps (2011). The mixed-frequency alternative regresses a low-frequency target directly on high-frequency data with a parsimonious lag polynomial, which is MIDAS (Ghysels et al. 2007), and the two approaches have been compared on precisely the ragged-edge problem (Marcellino and Schumacher 2010). A third strand insists on the vintages: Croushore and Stark's real-time data set exists because a forecast evaluated on revised data is not the forecast anyone could have made (Croushore and Stark 2001). And forecast combination is the standing answer to not knowing which of several models to trust (Timmermann 2006).

The connection is close enough to be worth stating precisely, because the differences are what leave room for this paper. The apparatus is shared: our design is a ragged edge with one variable's edge pushed back by hand, our Kalman comparison in Section 10 is a nowcasting model, and our compressed peer block is a factor. What is not shared is the question. That literature asks how accurately the latent current value can be estimated, and answers in the units of the target. We ask what share of a *known* loss a *named purchasable* source repairs, and answer in a unit-free ratio that lets a data feed and an options subscription be compared with each other. Nor is the ragged edge in that work a variable: it is whatever the release calendar happens to deliver, so the delay is a fact about the data rather than a parameter to be swept. Sweeping it is what produces a delay curve, a ceiling and a rate.

The stale-price literature in finance is the other half of the inheritance and is older. Scholes and Williams (1977) and Lo and MacKinlay (1990) showed what non-synchronous trading does to measured covariances, which is why a cross-section of closes taken at different hours cannot be treated as simultaneous; Getmansky, Lo and Makarov (2004) and Geltner and co-authors (2003) showed what smoothed or infrequent marking does to measured risk. Closest of all in spirit is Brown, Ghysels and Gredil (2023), who nowcast private equity net asset values from public market information between reporting dates: the same act on the same kind of asset, reported as a level rather than as a price paid for timeliness. Section 8's house price panel is the object that literature and this one share, and the substitution rate is what this paper adds to it.

What none of this work does is treat the delay as a variable. In every case additional data is added to an information set that is already current, and the question is whether it helps. Ours is the complementary question: when domestic information is not current, how much of the damage does each candidate undo, and how does that fraction move as the damage grows? That is a rate of substitution between things a practitioner can actually buy, and we could not find it measured.

3. Data

The data are daily open, high, low and close for eight equity indices: S&P 500 (SPX), Nikkei 225 (N225), S&P/ASX 200 (AXJO), Hang Seng (HSI), Nifty 50 (NSEI), FTSE 100 (FTSE), DAX (DAX) and Bovespa (BVSP), from January 2000 to September 2026 as raw data, of which the out-of-sample evaluation uses 4,862 days once the burn-in, training and validation blocks of Section 4.3 are spent, together with three implied-volatility indices: CBOE VIX, its nine-day counterpart VIX9D, and VDAX-NEW. The two CBOE series are Cboe's own published daily histories; the eight indices and VDAX-NEW are retail end-of-day exports from Investing.com, downloaded in September 2026. The repository lists every file, the identifier it maps to, and the column dictionary, and pins the environment to NumPy and pandas alone. Section S14 of the Internet Appendix adds four further index series, for Pakistan, Indonesia, Sri Lanka and Malaysia, from a different vendor and as one file each rather than two. They are kept in their own folder, they enter one subsection, and no other number in this paper is computed from them. Each retail series came in two exports, because the source caps a download at 5,000 rows, which leaves nine joins to be checked rather than trusted. Eight of the nine overlap by 120 to 251 trading days, and across every overlapping row the two exports agree exactly: no mismatches, largest absolute difference zero. The ninth is the S&P itself, where the exports abut without overlapping and no cross-check is possible. We report it as unverified and fall back to two weaker tests: the join is contiguous, Friday 9 March 2018 to Monday 12 March

2018 with no session missing or duplicated, and the return across it is -0.0013 , or 0.11 standard deviations of the series' own daily returns, so there is no level shift where the exports meet.

A foreign close is admissible only when its timestamp falls before the target's. Markets closing at or after the target therefore contribute their previous session. The rule generalises without special cases, which matters for Section S9 of the Internet Appendix: a Tokyo forecaster gets every other market a day late, because Tokyo closes first. The rule is implemented from a table of standard-time closes, and whether that simplification is safe is a question about real exchange clocks rather than about the design, so Section S35 of the Internet Appendix settles it with one: every close is rebuilt in UTC from its exchange's own local closing time and its own timezone, daylight-saving transitions included, on all 7,021 dates. Across every target, every peer and every date there are **zero** days on which the true clocks would have reordered a block, and the narrowest margin the rule relies on anywhere is two hours. The audit also shows the implementation is stricter than the rule it implements, which is the direction that costs information rather than validity, and Section S35 puts a number on that too. It also disciplines the implied-volatility block, and the discipline is awkward in a way worth setting out precisely. VDAX-NEW satisfies the rule outright, since Frankfurt finishes hours before New York. VIX does not. Its printed daily close is stamped 21:15 UTC, fifteen minutes after the 21:00 S&P cash close, so that print is inadmissible. VIX is computed continuously from live SPX option quotes, so a value of it did exist at 21:00 and would be the right observation to use, but a retail end-of-day export carries no intraday snapshot and we do not have it. What we can run is a pair of timing specifications on either side of the one we want: the previous day's close, which is strictly admissible and discards a full day of information, and the same-day close, which contains fifteen minutes we are not entitled to. Section 7 leads with the lagged arm, because it is the one that obeys our own rule, and reports the same-day arm as a sensitivity check. The 21:00 observation is temporally between the two, and we are careful not to infer from that that its forecasting performance would be numerically between their results. Timestamps order; skill need not.

4. Methodology

The object of this section is one number and the conditions under which it means anything. A forecaster loses skill when their own mark goes stale, and a substitute returns some of it; the quantity a buyer wants is the second as a share of the first, which is the ratio $R(\delta)$ of equation (9). Everything before that equation exists to make the two skills comparable: one volatility estimator, one decision variable, one real-time training window that stops $\delta + h$ days before the forecast origin, one family of losses. Section 4.6 is what stops the ratio being read further than the data supports, since its denominator is what delay costs and approaches zero wherever delay costs little. A reader interested only in the measure can read 4.5 and 4.6 and return.

4.1 The volatility estimator

Volatility is the Yang-Zhang estimator (2000) over rolling five-day windows, which uses the whole daily range and so sees more of each session than the close alone. Write O_s , H_s , L_s and C_s for the open, high, low and close of session s , and

$$o_s = \log(O_s / C_{s-1}), \quad c_s = \log(C_s / O_s), \quad u_s = \log(H_s / O_s), \quad d_s = \log(L_s / O_s)$$

for the overnight, open-to-close, high and low log returns. Over the n sessions ending at t the estimator is

$$\sigma^2_{YZ,t} = \sigma^2_o + k \sigma^2_c + (1 - k) \sigma^2_{RS}, \quad k = 0.34 / (1.34 + (n+1)/(n-1)), \quad (1)$$

where σ^2_o and σ^2_c are the sample variances of o_s and c_s over those n sessions, σ^2_{RS} is the mean over them of the Rogers-Satchell term $u_s(u_s - c_s) + d_s(d_s - c_s)$ (Rogers and Satchell 1991), and $n = 5$ throughout. The three components measure the overnight gap, the drift within the session, and the range net of that drift; k is the weight for which the sum is minimum-variance and drift-independent under a Brownian price (Yang and Zhang 2000). All variances use the $n - 1$ divisor.

4.2 The decision variable and the forecast target

Let i index the eight series and t the target's trading day. The decision variable each model sees is the log of that series' variance against its own recent level,

$$a_{i,t} = \log(\sigma^2_{YZ,i,t} / M_{i,t}), \quad M_{i,t} = \text{median}(\sigma^2_{YZ,i,s} : t - 251 \leq s \leq t), \quad (2)$$

and the quantity to be forecast, with horizon $h = 5$ trading days and target market denoted SPX, is

$$y_t = a_{\text{SPX}, t+h}. \quad (3)$$

So $M_{i,t}$ is the trailing 252-day median, and the implied-volatility series enter the same way, as the log of the level over its own trailing median. Dividing by a series' own median puts Tokyo, São Paulo and the VIX on one scale and keeps each series' units out of the regression. The target at $t + h$ carries a median whose window includes the five days between origin and outcome, which a forecaster at t does not know; no feature is affected, since every one is dated $t - \delta$ or earlier for the domestic block and t for the others. Section S16 of the Internet Appendix runs the whole experiment under the implementable alternative, which normalises the outcome at the origin. **One definition per place, stated once:** unless a caption says otherwise, the equity-result tables in this paper report the target-dated rate and every headline sentence reports the implementable origin-dated one. Table 1 is the labelled exception, since it collects the headline specifications and therefore carries the origin-dated figure, and its housing row is a different experiment on a different asset class. The two differ by at most a point at every delay, and Table S12 of the Internet Appendix gives both columns, so a reader who wants a single definition throughout can have either. The target is a_{t+5} , five trading days ahead, left continuous: a companion note (Vellanikaran 2026) showed that splitting such a target at its median throws away exactly the magnitude information a cross-sectional model would use, so we do not split it.

4.3 Feature blocks and estimation

Models are built from three feature blocks. The domestic block is the target's own history at daily, weekly and monthly aggregation, in the manner of Corsi (2009), observed at $t - \delta$. The foreign block is seven index decision variables at their admissible timestamps: the six that close before New York, plus Bovespa a session late. The implied-volatility block is VIX and VDAX. Only the domestic block is delayed; the others are what a forecaster could see today, and that asymmetry is the whole point of the comparison. Every model is a ridge regression refitted every 21 days, with features standardised on training statistics alone. The training window needs stating carefully, because where it ends is not where a reader would guess and the gap is the reason the design is honest. Formally, let x_t be the feature vector a model sees at t , assembled from the blocks above. Training cannot run up to t : the most recent pair whose outcome had resolved by the time the target's own mark was current ends at

$$c_t = t - \delta - h, \quad T_\delta(t) = \{c_t - 1499, \dots, c_t\}, \quad (4)$$

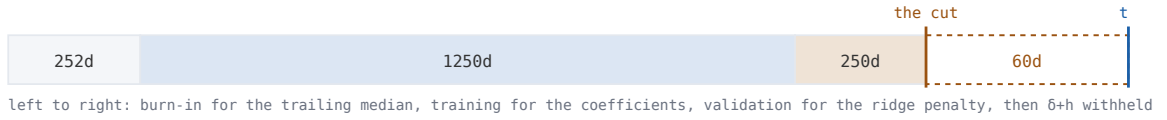
so the window holds 1,500 days and slides with δ as well as with t . That is worth naming precisely, because δ does not only make one number old. It freezes the whole domestic information system for δ days: the current state, the labels available for fitting, the coefficient vintage, the ridge penalty the validation tail selects, and the quantity of training data. What this paper measures is therefore a domestic-data *outage* of δ days, not a stale mark beside an otherwise current desk, and $R(\delta)$ is the share of an outage's damage that a substitute repairs. Section S36 of the Internet Appendix prices the difference by running a third arm that no forecaster could run, with a stale input but coefficients refitted to $t - h$, and finds that freezing the estimation accounts for 0.4% of the damage at a week and 2.7% at eleven weeks. The denominator $R(\delta)$ divides by is information rather than estimation, and the rate computed inside the infeasible arm is 2.8 points higher at most, so the paper's design is also the conservative one. The naming is a matter of accuracy; the measurement barely moves. Features are centred and scaled by that window's own mean μ_T and standard deviation s_T , giving $z_t = (x_t - \mu_T) / s_T$ elementwise, and the coefficients (β_0, β) solve

$$\min_{\beta_0, \beta} \sum_{u \in T_\delta(t)} (y_u - \beta_0 - z_u' \beta)^2 + \lambda \beta' \beta, \quad (5)$$

with the intercept β_0 left unpenalised. λ is chosen from $\{0.3, 1, 3, 10, 30\}$ in two steps, and the second step is worth stating because it is easy to assume otherwise: candidate penalties are fitted on the first 1,250 days of $T_\delta(t)$ and scored on the last 250, which are held out of the fits that select between them, and the chosen λ is then refitted on all 1,500. The validation tail is therefore held out of the candidate fits, which is what lets it choose between them, and is not held out of the final model. The forecast is $f_t = \beta_0 + z_t' \beta$. Standardising on training statistics alone matters: scaling a feature by a full-sample standard deviation would let the test period influence the fit through the scale, which is a small leak but a real one. All are trained on the same labels, a pair being admitted only once its outcome was resolvable at $t - \delta$, so differences between them are features and nothing else. The rolling fixed window is not incidental; it is what makes the tests in Sections 5 and 6 valid. Figure 1 draws the whole arrangement, because six dates matter to it and prose makes a reader assemble them.

What the model is fitted on, and what is held back

re-estimated every 21 trading days; this panel is compressed, the one below is to scale



What is available at one forecast origin, at $\delta = 55$

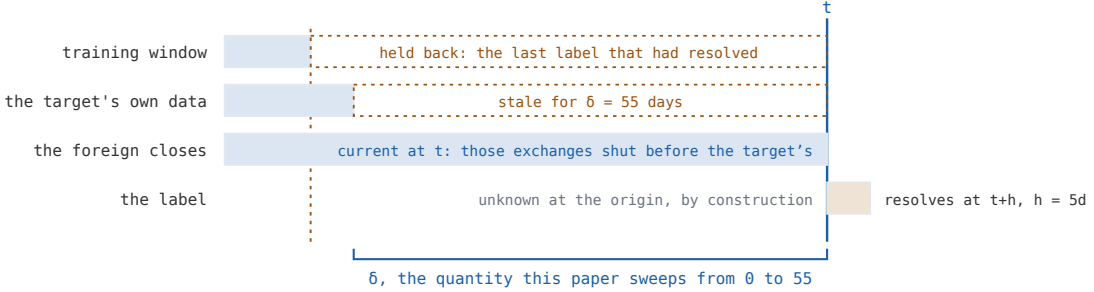


Figure 1. The design of Section 4, on two panels. Above, the window every model is fitted on, drawn on a compressed axis: a burn-in the trailing median of equation (2) needs before it exists, 1,250 days on which candidate ridge penalties are fitted, a 250-day tail on which they are scored, and then the $\delta+h$ days withheld because their labels had not resolved. The chosen penalty is refitted on all 1,500, so the tail is held out of the selection and not of the final model. Below, one forecast origin at the delay this paper leads with, drawn to scale. Three things are true at t and only the third is unusual: the training window stopped $\delta+h$ days ago, the target's own mark is δ days old, and the foreign closes are current, because those exchanges shut before New York does. The label resolves h days after the forecast is made and is unknown at the origin by construction. The cut at $t - \delta - h$ is a purge in the cross-validation sense and is load-bearing: Section S7 of the Internet Appendix measures what removing it inflates. Both panels are drawn from the same constants the estimator uses rather than positioned by hand, so a change to the design moves the figure.

4.4 Loss functions and proxy robustness

Losses need care here. With an imperfect volatility proxy only a restricted family of losses preserves the ranking a researcher would obtain against true latent volatility (Patton 2011), and that family is defined on the VARIANCE scale. Our target is a log variance, so the two halves of the usual pairing land differently. QLIKE survives the transformation: it depends on the ratio of outcome to forecast alone, and dividing both by the same trailing median leaves that ratio untouched, so what we compute is exactly the variance-scale QLIKE. Squared error does not. Computed on log variance it is outside the family, because a monotone transform of the arguments is precisely what the robustness result fails to survive.

We therefore report three, and the reconstruction back onto the variance scale needs stating precisely, because getting it wrong is how a loss comes to be called robust when it is not. The benchmark forecast needs stating before the losses that divide by it, because it is not a single number. At each test day t it is b_t , the mean of the labels a forecaster had already seen, taken over the 1,500 days ending at $t - h$, which is where labels stop resolving for someone whose own data is current. Two properties of it are load-bearing. It is not delay-matched: one benchmark serves every model at every delay, because $R(\delta)$ divides by $S_{\text{own}}(0) - S_{\text{own}}(\delta)$, and a yardstick that moved with δ would mix the damage the delay does with the change in the yardstick. And it is a forecast, not a fitted constant: a mean taken over the test period would be chosen with hindsight, which is not what out-of-sample skill means. Write n for the number of test days. Realised variance is recovered from the target by undoing the normalisation at the date the target carries it, $v_t = \exp(y_t) M_{t+h} = \sigma^2_{YZ,t+h}$. A forecaster at t does not

know M_{t+h} , so the matching forecast can only use the median they do have, $w_t = \gamma_t \exp(f_t) M_t$, which carries the median's h -day drift as forecast error rather than hiding it. The three losses are

$$R^2 = 1 - \sum_t (y_t - f_t)^2 \div \sum_t (y_t - b_t)^2, \quad (6)$$

$$\text{QLIKE} = (1/n) \sum_t [v_t/w_t - \log(v_t/w_t) - 1], \quad (7)$$

$$\text{MSE}_{\text{var}} = (1/n) \sum_t (v_t - w_t)^2, \quad \gamma_t = (1 / |T_\delta(t)|) \sum_{u \in T_\delta(t)} \exp(y_u - f_u), \quad (8)$$

where γ_t is Duan's smearing factor (1983), estimated from the same delayed training window at each refit. Skill under QLIKE and MSE_{var} is defined as in (6), one minus the model's loss over the benchmark's, so all three columns are shares of a benchmark loss removed and are read the same way. That sentence is only true if the benchmark is carried onto each scale the way a model forecast is, and it is: exponentiate, apply its own smearing factor γ_t^b estimated on the same resolved labels, and multiply by the median the forecaster holds, so the benchmark's natural-scale forecast is $\gamma_t^b \exp(b_t) M_t = M_t \cdot \text{mean}_u \exp(y_u)$. It is worth saying precisely what that object is, because an earlier version of this paper said it wrongly. Since y is a *normalised* log variance, $\text{mean}_u \exp(y_u)$ is the trailing mean of variance *over its own median* and not the trailing mean of variance: on this panel the two differ by 62.8% of the mean level and correlate 0.35. The reconstructed benchmark is the one reported here, because it is the same forecaster the other two columns use and is what makes the sentence above true; a benchmark built directly from resolved natural variances would be a different forecaster, and Section S25 of the Internet Appendix scores both so the choice is visible. The rate is identical under either, to the decimal place, because the benchmark cancels from a ratio of skill differences. Leaving the benchmark unsmeared while the models are smeared would deny it a correction they receive, inflate its loss, and overstate skill on both variance-scale columns; giving it the median at $t + h$ while the model only holds the median at t would do the reverse, and an audit found this paper doing the second. Of the three, R^2 in (6) is the paper's headline and is *not* proxy-robust; QLIKE in (7) is, and Part 0 of the script behind this section shows it exactly rather than by argument, computing QLIKE on a variance and on that variance over a median and getting the same number to ten decimal places; and MSE_{var} in (8) is Patton's MSE on the variance scale and is also robust. An earlier version of this paper computed MSE_{var} on the median-normalised scale while calling it Patton's MSE; Section S25 of the Internet Appendix reports what that cost, which reached 29 points of the rate. Every MSE_{var} figure in this paper is the natural-scale one defined above; Section S25 of the Internet Appendix reports both columns and the gap between them. The smearing factor is there because $\exp()$ of a log forecast is median-unbiased, not mean-unbiased, and squared error on the variance scale punishes exactly that, so the reconstruction needs the correction, not the raw exponential. Where conclusions differ across losses we give weight to what both proxy-robust losses support; where one of them cannot resolve a row, we rest no claim on it and say so where that happens.

4.5 The substitution rate

Write $S(\delta)$ for skill at delay δ under any one of (6), (7) or (8): the share of the benchmark's loss that a model removes. $S_{\text{own}}(\delta)$ is the skill of the model holding only the domestic block, δ days stale; $S_{\text{alt}}(\delta)$ is the skill of the same model with a candidate substitute added. Delay costs $S_{\text{own}}(0) - S_{\text{own}}(\delta)$; the substitute returns $S_{\text{alt}}(\delta) - S_{\text{own}}(\delta)$. The substitution rate is their ratio,

$$R(\delta) = [S_{\text{alt}}(\delta) - S_{\text{own}}(\delta)] \div [S_{\text{own}}(0) - S_{\text{own}}(\delta)]. \quad (9)$$

It has no units, so it can be compared across loss functions, across markets, and across candidates. One word in the title needs settling here, because $R(\delta)$ is a share of forecast loss and not an amount of money. When this paper says a trade can be *priced*, it means that the two things being compared can be put on one ruler and that the ruler has a zero and a natural restoration reference point at 100%: a data feed and an options subscription can be ranked against each other, a rate near zero says the purchase is worthless, and a rate above 100% says the substitute supplies information the fresh domestic-history model never had. The reference point is not a mathematical ceiling on $R(\delta)$, and Section 6 reports values above it. Converting that into a currency needs a decision problem the holder supplies, a value at risk or a capital buffer or a volatility target, and Section 11 names it as the obvious next study rather than guessing at one here. What the paper does offer in a practitioner's own units is the clock of Section 9: the same recovery read as the age of a mark. One caution, which Section 7 will need: the measure assumes the substitute is *restoring* something delay destroyed. A source that also carries information the domestic history never had can exceed 100%, and when it does, the ratio has stopped answering the question it was built for. The denominator can fail in the other direction too: where delay costs a forecaster little, $S_{\text{own}}(0) - S_{\text{own}}(\delta)$ approaches zero and the ratio is weakly identified, so a uniformly valid confidence procedure must be free to return an unbounded set (Dufour 1997). The published intervals are a bootstrap of the whole ratio, and each is checked by inverting a HAC t-test after Fieller (1954), which is free to return a bounded interval, a union of two half-lines or the whole line according to what the data support. Section 4.6 gives a construction that stays honest there, and Section S5 of the Internet Appendix applies it to the state where the risk is greatest. On this sample the denominator never gets close enough for the pathology to bite, but it is a property of the measure itself, and anyone reusing $R(\delta)$ on a series where delay does little damage should check it rather than assume it.

4.6 Inference for a ratio whose denominator can vanish

The substitution rate is a ratio, and Section S2 of the Internet Appendix reports that inverting a HAC t-test after Fieller leaves every published interval standing. This is how that was computed, and why it was not the first answer we got. Fieller's theorem (1954) inverts a t-test on the linear combination num – R-den, which is zero at the true rate, and solves the resulting quadratic for the set of rates not rejected. When the denominator is well separated from zero the solution is an ordinary interval close to the delta-method one. When it is not, the quadratic's leading coefficient changes sign and the confidence set is correctly a union of two half-lines, or the whole line: an honest statement that the data do not pin the ratio down, and one a percentile bootstrap cannot make. That is not a defect of the construction: under weak identification a uniformly valid confidence procedure must permit unbounded sets (Dufour 1997), and the asymptotic Fieller-type inversion used here is one standard construction with that property rather than the only one. It is asymptotic, since it uses an estimated HAC covariance and normal critical values. We have met the symptom in this project without naming it, in an interval of $[-154, 50]$ at one point, and treated it as a nuisance when it is an identification problem. Two things about the calculation are worth recording. The variances must be HAC, with the same Bartlett kernel used in Section S2 of the Internet Appendix, because the five-day targets overlap: the first version used iid variances and returned intervals materially *narrower* than the block bootstrap at every delay, which was a warning sign given the strong positive serial dependence measured in the overlapping-target loss differentials, and was the only sign that anything was wrong.

Overlap invalidates independence-based inference generally; whether iid standard errors come out smaller depends on the signs and magnitudes of the serial covariances, so the narrowing was evidence rather than proof. And $\delta = 0$ is absent from the comparison throughout, because there the denominator is identically zero by construction, the one delay at which the ratio does not exist at all, rather than being merely hard to pin down.

5. What breadth recovers

The test window runs from May 2007 to September 2026, 4,862 days.

Table 2. Out-of-sample skill against a trailing mean of already-resolved outcomes, and the substitution rate under each loss. Skill is the fraction of benchmark loss removed, so the two pairs are comparable. The benchmark is one forecaster on both scales: it forecasts that trailing mean on the log target, and on the variance scale it is exponentiated using its own smearing factor and reconstructed using the median available at the forecast origin. That is not the trailing mean of natural variance, and Section S25 gives the size of the difference. Intervals bootstrap the whole ratio, recomputing numerator and denominator on the same resampled days. R^2 is on the log scale and is not proxy-robust; QLIKE is. This table is computed on the target-dated definition of Section 4.2, as the equity-result tables are and so reads 72% at eleven weeks where the abstract's implementable figure is 71%; the two differ by at most a point at every delay.

δ	R^2 own	R^2 cross	$R(\delta)$ on R^2	QLIKE skill own	QLIKE skill cross	$R(\delta)$ on QLIKE
0	0.5191	0.5092	n/a	0.6190	0.6064	n/a
1	0.4741	0.4735	−1% [−33, 29]	0.5724	0.5703	−4% [−47, 35]
3	0.4098	0.4495	36% [15, 55]	0.5001	0.5478	40% [13, 64]
5	0.3641	0.4438	51% [34, 67]	0.4426	0.5420	56% [34, 75]
8	0.3114	0.4351	60% [44, 72]	0.3655	0.5325	66% [47, 80]
13	0.2336	0.4155	64% [50, 74]	0.2263	0.5127	73% [55, 84]
21	0.1482	0.3981	67% [54, 77]	0.0936	0.4977	77% [61, 87]
34	0.0794	0.3863	70% [58, 79]	−0.0118	0.4910	80% [65, 88]
55	0.0070	0.3736	72% [60, 81]	−0.0477	0.4857	80% [66, 89]

Read the table down a column and then across. The first column is what a forecaster loses to delay alone: R^2 on the domestic block falls from 0.5191 at zero delay to 0.0070 at eleven weeks, a loss of 99% of the skill it began with, so by then a stale mark is worth about what a trailing mean of already-resolved outcomes is worth and no more. The second column is the same forecaster with the foreign cross-section added, and it barely moves, from 0.5092 to 0.3736, because the peers are current whatever the domestic delay is. The third column is the ratio of the two gaps, which is $R(\delta)$ in (9): 36% of the delay's damage repaired at three days, 51% at a week, 72% at eleven weeks. The two QLIKE columns repeat the exercise on the proxy-robust loss and reach the same shape with larger numbers, 40% to 80%, so the result is not an artefact of the scale the headline is quoted on. The two losses do not agree about everything, and the disagreement is worth naming rather than leaving for a reader to find. Under R^2 the stale domestic model is still marginally above the benchmark at the longest delays, 0.0794 at $\delta = 34$ and 0.0070 at $\delta = 55$. Under QLIKE it is below it, −0.0118 and −0.0477. Both statements are about the same forecasts and the same benchmark; what differs is which errors the loss punishes. Squared error on logs is symmetric in the log and so forgiving of a forecast that is too low in variance terms, which is what a stale mark mostly is once volatility has moved; QLIKE

punishes exactly that asymmetry, and it is the proxy-robust one. Where the two conflict we give weight to QLIKE, which here means the honest reading of the long end is the harsher one: a mark eleven weeks old is not merely worth about what a constant is worth, it is worth slightly less. The $\delta = 1$ row is the one that does not fit the pattern, and it is the row that should not: at one day the domestic history is barely stale and there is almost nothing for the cross-section to put back, so the rate is indistinguishable from zero under both losses.

One warning attaches to every figure in that third column before the paper goes further, and Section S5 of the Internet Appendix establishes it. An unconditional rate averages two states in which this mechanism behaves quite differently. In stressed markets it is 81% at eleven weeks with an interval of [74, 87], and it is 81% in each half of the sample taken separately: a well-determined, stable number. In calm markets the same quantity spans ninety points and cannot be pinned down at all, for a mechanical reason: where delay costs a forecaster almost nothing, the share of that cost repaired by anything has no stable value. Table 2's 72% is the average of a result and a non-result. We report it because it comes from the original specification, the one fixed before any alternative was run, and because the literature quotes unconditional rates, but the number a practitioner should carry is the conditional one in Table S4 of the Internet Appendix.

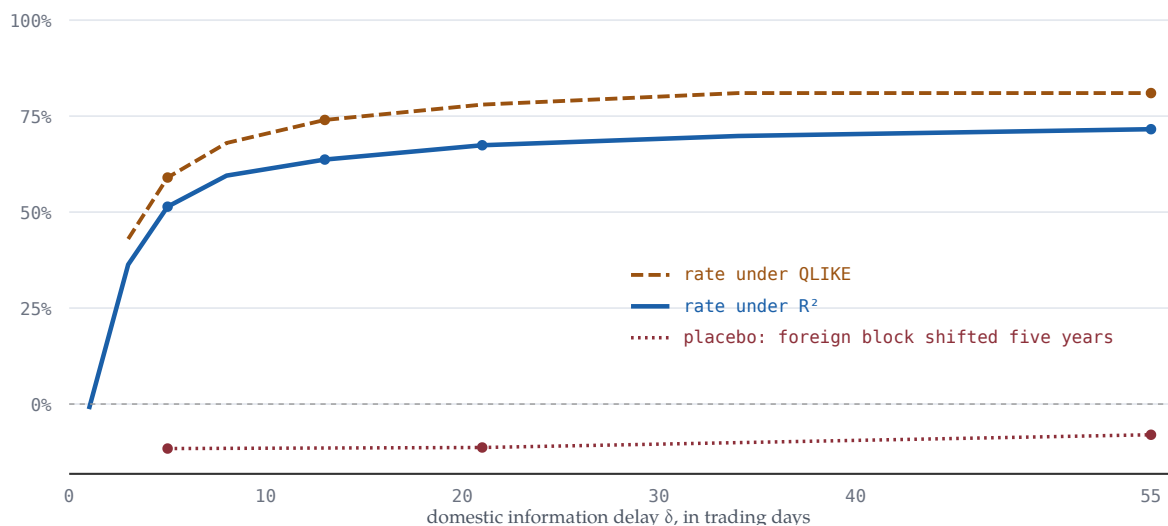


Figure 2. Breadth buys back almost nothing at one day, half the loss within a week, and around three-quarters by eleven weeks. The placebo, described in Section S2 of the Internet Appendix, keeps the same foreign series but destroys their alignment with the target, and the effect disappears.

Three things stand out. At one day the rate is indistinguishable from zero, as it should be: the S&P's own recent history is still the best available description of the S&P, and the foreign block we selected produces no measurable net improvement on it. By eleven weeks the picture has inverted. The domestic model's out-of-sample R^2 has fallen to 0.0070, so a mark eleven weeks old is worth no more than a trailing mean of outcomes that had already resolved, while the cross-sectional model still holds 0.3736, about 73% of what was available with fully current data. And between those extremes the rate climbs steeply and then slows, which raises the obvious question of where it stops.

Section 4 promised a third loss, and it belongs here because it is the robust one that most resembles the headline. Under squared error on the natural variance scale the substitution rate reads 74% at five days and 80% at eleven weeks, against 51% and 72% on the log scale. Both robust losses therefore put the rate above the headline, and they now agree with each other closely: QLIKE also reads 80% at

eleven weeks. That agreement is worth more than either number, because it is what the log-scale headline was always being asked to survive. The variance-scale intervals remain wider at short delays, [11, 114] at five days, because the denominator there is small and squared error on raw variance is dominated by a handful of crisis days; that is why QLIKE, not MSE, is the robust loss we lean on.

6. The limits of breadth

The rate flattens near three-quarters instead of approaching one, and the reason is not subtle once measured. At very large delay the domestic model retains little usable out-of-sample skill, so the cross-sectional model is working almost entirely from foreign information. Its ceiling is therefore whatever share of the S&P's volatility the rest of the world can see.

We measured that share directly. Regressing the S&P's decision variable on the six foreign closes of the same day gives $R^2 = 0.695$. Roughly 30% of the S&P's current volatility state is not linearly spanned by those six series in this sample. We resist calling that share genuinely local, because the regression does not show it: a non-linear function of the same closes, or a wider cross-section, could reach some of it. What the number does bound is how much a LINEAR model of THESE markets can see, the ceiling the fitted model is working against. Predicting five days ahead from foreign closes alone gives $R^2 = 0.399$ against 0.498 from domestic history, or 80% of it, which brackets the 72% the fitted model achieves. The gap between the in-sample bound and the fitted figure is consistent with the price of estimating parameters instead of knowing them, and Section 6 measures a cost of the right order; the design does not establish that estimation is the whole of it.

The cross-section is also nearly one-dimensional. The first principal component of the six foreign volatility series accounts for 66.5% of their variation and the first two for 76.9%. Six markets are close to one global volatility factor plus idiosyncratic noise, which carries a practical implication: an eighth or a ninth correlated index may add relatively little under this linear specification, because what a linear model of these markets can see is set by how much of the S&P is global; counting more global series does not move it.

That leaves the other half of the sentence above: what it costs to estimate parameters instead of knowing them. We had treated this as a residual, and it is large enough to deserve a number. Breadth is not free. Seven foreign coefficients must be estimated from a finite window, and the estimation is paid for whether or not the coefficients turn out to matter.

Two measurements agree on the size of the bill. The first varies the number of peers at $\delta = 0$ and reads off the slope: the penalty is a straight line through the origin at -0.0013 of R^2 per market, which accounts for 98% of the variation across peer counts and implies -0.0096 for the full seven. The second replaces the foreign block with AR(1) surrogates fitted to each peer, drawing innovations from a multivariate normal with the peers' own contemporaneous covariance so the surrogate cross-section is as correlated as the real one, while carrying no information about the target at any lag. Averaged over one hundred draws this gives -0.0131 at $\delta = 0$, with a Monte Carlo standard error of 0.0003, an order of magnitude below the sampling interval that follows and therefore not the uncertainty that binds. The two methods differ by 0.003, which agrees on the magnitude without being a formal match, and we do not read a third decimal place from either. About one point of R^2 is the price of carrying the cross-section. Section 6.1 measures it a third time, with a generator that fits nothing at all, and gets the same answer.

Two things follow. The first settles a reading we had left open. At $\delta = 0$ the measured difference is -0.0099 and the cost is -0.0126 , so the gross content of the foreign block is $+0.0027$. Gross is a difference of two estimated quantities, so we bootstrap it, resampling test days in moving blocks and recomputing the measured difference and the surrogate cost on the same days so that their errors stay correlated. The interval is $[-0.0064, +0.0116]$: indistinguishable from zero, and bounded above at about one point of R^2 . Foreign markets are not misleading about the current state of the S&P. Whatever they know about it that fully current domestic data does not already carry is too small for this design to detect, and the negative sign was the bill, not the information.

The second is that the bill grows with delay, from -0.013 at $\delta = 0$ to -0.054 at $\delta = 21$. This is a property of the fitting procedure: as domestic information goes stale the fit weakens, the validation step selects a looser penalty, and the seven regressors are given more rope. Separating the two makes the gross and net readings diverge in a way that matters for anyone deciding whether to build this.

Table 3. What the cross-section knows, and what a forecaster keeps after paying to estimate it. Cost from AR(1) surrogates, one hundred draws, innovations carrying the peers' own cross-market covariance. Gross is bootstrapped in the text: indistinguishable from zero at $\delta = 0$, marginal at $\delta = 1$, positive from $\delta = 2$.

δ	0	1	2	3	5	8	13	21
Net ΔR^2 (measured)	-0.0099	-0.0006	$+0.0180$	$+0.0397$	$+0.0797$	$+0.1237$	$+0.1819$	$+0.2499$
Cost of breadth	-0.0126	-0.0149	-0.0172	-0.0193	-0.0234	-0.0294	-0.0392	-0.0514
Gross ΔR^2	$+0.0027$	$+0.0143$	$+0.0353$	$+0.0590$	$+0.1031$	$+0.1530$	$+0.2211$	$+0.3013$

Gross content clears zero at two days, $[+0.0165, +0.0534]$, and the net figure crosses there too. At one day the gross interval is $[+0.0003, +0.0279]$, which excludes zero by a margin of 0.0003 , close enough to the boundary that we report it and decline to build on it. An earlier draft, using ten surrogate draws instead of a hundred, put that interval across zero and we withdrew the reading entirely; a better cost estimate moved it back. A band in which the foreign block holds something real that cannot yet pay for its own estimation is therefore consistent with these data at one day, and no more than that. At $\delta = 21$ estimation consumes 17% of what the cross-section knows, $+0.3013$ gross against $+0.2499$ net. That gap is not a law of nature. It is the cost of a particular estimator, and a cheaper one should recover part of it: a factor representation of the foreign block instead of seven free coefficients, which the principal-component result above already suggests would lose little. Section 6.1 fits that compressed representation directly.

6.1 Fitting the compressed representation directly

Two facts above sit next to each other without being joined. The foreign block is nearly one-dimensional, its first principal component accounting for 66.5% of the variation and its first two for 76.9%. Carrying seven regressors costs about a point of R^2 at $\delta = 0$ and more as the delay grows, measured by replacing the block with surrogates that carry no signal. It is natural to conclude that a cheaper parameterisation would recover part of that cost. That is an inference from a hypothetical generator, not a demonstration, and the demonstration is cheap.

So compress the block and compare. The eigenvectors are recomputed at every refit from the covariance of the training window alone, on columns standardised with training statistics alone, with each eigenvector's sign pinned so the factor does not flip arbitrarily between refits. A basis computed once on the whole sample would leak: "the first component of the foreign block" is not something a

forecaster in 2009 could have written down from 2026 data. An equal-weighted mean of the seven series is included as the compression that requires no estimation at all.

Table 4. Out-of-sample R^2 for four representations of the same seven foreign series, against the domestic-only model. PC1 and PC2 use one and two real-time global factors; FULL is the seven-coefficient model the rest of the paper uses. The last columns test PC1 against FULL, which are non-nested, by Giacomini-White.

δ	domestic	+ mean	+ PC1	+ PC2	+ all seven	PC1 – all seven	GW z
0	0.5191	0.5203	0.5204	0.5179	0.5092	+0.0112	2.66
3	0.4098	0.4603	0.4614	0.4588	0.4495	+0.0119	1.80
5	0.3641	0.4541	0.4556	0.4532	0.4438	+0.0118	1.72
13	0.2336	0.4249	0.4275	0.4256	0.4155	+0.0120	1.51
21	0.1482	0.4091	0.4122	0.4092	0.3981	+0.0141	1.56
55	0.0070	0.3876	0.3922	0.3889	0.3736	+0.0186	1.57

One real-time factor beats all seven series at every delay, by 0.012 to 0.019 of R^2 , and the margin grows with delay exactly as the surrogate cost does. The gain reaches conventional significance at $\delta = 0$ and $\delta = 55$ and falls short of it in between, so the pattern is more convincing than any single entry. The real-time first component explains 66.6% of the block's variance averaged over refits, against 66.5% in the full-sample figure quoted above, which says the basis is stable enough to be estimated as you go.

Reporting a new preferred specification on the log scale alone would be inconsistent with Section S2 of the Internet Appendix, which exists to establish that the scale changes the answer. Under QLIKE, the proxy-robust loss, one factor wins at every delay, exactly as it does under R^2 . Under squared error on the variance scale a compression wins at five of six delays; the exception is $\delta = 0$, where the seven-regressor model scores lowest. That loss is dominated by a handful of crisis days, so Section S2 of the Internet Appendix reports its intervals as uninformative at short delays, so the exception is a statement about tails. Where the three losses appear to disagree elsewhere they are choosing between one factor and two, which differ by thousandths and are not a disagreement about anything.

The compression also raises the substitution rate, so it is worth stating plainly which number the paper claims. Recomputed with the same full-ratio bootstrap as the headline, $R(55)$ is 71.6% [60.2, 80.4] for the seven-regressor model, 74.3% [64.4, 82.7] for the equal-weighted mean and 75.2% [65.4, 83.5] for one factor. Every compressed estimate sits inside the interval the seven-regressor model already carries, so the two are not distinguishable and we do not replace the headline with the larger figure. We report 72% as the original specification's full-block estimate, note the compressed estimate beside it, and say plainly that the difference is inside the noise of both. That leaves four specifications doing four jobs, and naming them settles which number answers which question. The equal-weighted mean is the *practical* model: it needs no estimated eigenvector, it is the cheapest thing a desk can build, and it recovers 74.3%. The seven-regressor block is the *original benchmark*, kept because every conditional and robustness result in this paper is computed inside it, and it recovers 72%. The real-time factor is a *mechanism test* rather than a recommendation, because it locates the recovery in one component. And the state-space filter of Section 10 is a *structural alternative*, reported and not adopted. The paper's headline is therefore the conservative member of that set, and a practitioner following Section S30 should expect the 74% to 75% band rather than the 71% the abstract quotes. That gap is a property of the paper's own caution and is stated here rather than left for a reader to discover.

Two qualifications, and the first is the more important. The equal-weighted mean captures nearly all of the gain: PC1 beats it, but by 0.001 to 0.005. What is being bought is the decision to compress, not the machinery of compressing, and a practitioner should take the average and skip the eigenvector. Second, this changes the $\delta = 0$ row. The seven-coefficient model scores 0.5092 against the domestic model's 0.5191, the small negative this paper has declined to interpret, while PC1 scores 0.5204, above it. We do not claim the cross-section helps at $\delta = 0$ on that basis: it is a difference of 0.0013 with no interval attached, and reading an uninterval-ed difference as a finding is the specific error this paper has made and corrected three times. What the row supports is narrower and sufficient, that the negative at $\delta = 0$ belongs to the parameterisation and not to the information.

Both measurements of that bill are relative to a generator we chose, and the surrogate one preserves first-order persistence and contemporaneous covariance and nothing else: not heavy tails, not volatility-of-volatility, not crisis clustering. Section S21 of the Internet Appendix measures the bill again with a generator that fits nothing at all, the real foreign block moved circularly through the sample so that every marginal and every cluster is the real one and only the alignment with the target is destroyed. The unfitted generator returns the fitted generator's number, so the objection does not survive the test, and Section S21 records the two ways we got that shift wrong before getting it right.

The recovery lives entirely in the leading component of the foreign block, with the six orthogonal components worth less than nothing. Section S27 of the Internet Appendix gives that decomposition.

7. The race against implied volatility

The objection this section answers is the one Section 2 names as unavoidable: if a forecaster whose domestic data is δ days stale can see today's VIX, why would they need Tokyo? Implied volatility is forward-looking, it never goes stale, and both Korkusuz (2025) and Buncic and Gisler (2016) report that much of what looks like a cross-market effect is attributable to it. So the two candidates go into the same walk-forward, on the same labels, with the same estimator, and Section S28 of the Internet Appendix gives the race in full. Three results come out of it and the paper depends on all three.

Implied volatility wins, and it wins by a lot. A forecaster holding VIX and VDAX-NEW loses almost nothing to delay at all: R^2 falls from 0.5547 at zero delay to 0.5266 at eleven weeks, against 0.5191 to 0.0070 for the domestic block alone. Where the cross-section repairs 72% of the delay's damage, implied volatility repairs essentially all of it. That is not mysterious. VIX is a risk-neutral expectation of S&P variance over the coming month, priced continuously by people who are not waiting for anyone's mark, and a stale forecaster who can read it is not really stale about the quantity they care about.

Given implied volatility, the foreign block adds nothing this sample can detect. The increment from adding the seven foreign series on top of the options block is negative at every delay, and Section 6's estimation-cost argument explains the sign: seven coefficients cost one to five points of R^2 on this sample, and that bill is the whole of the negative. Replacing the foreign series with AR(1) surrogates recovers the same penalty from a block that carries no information by construction, so what looks like harm is the price of the parameters rather than a property of the peers. Section S28 makes the point sharper by adding *one* compressed regressor instead of seven, which removes the apparent harm and reveals nothing underneath it. What that licenses is a bounded statement rather than an equality, and this paper makes it in the bounded form throughout. Net of the estimation cost the block is significantly negative; gross of it the block's content is not distinguishable from zero, and the least

favourable of the four intervals Section S28 reports, $[-0.004, +0.019]$ at $\delta = 21$, rules out a gross gain above about 0.019 of R^2 while leaving anything smaller open. “Redundant” is used in this paper to mean exactly that and never to mean equal: a forecaster holding the options market gains nothing this sample can detect by adding seven foreign closes, and pays for them.

So the scope condition is the paper's result, not its retreat. Where a liquid options market exists on the asset itself, breadth is redundant and this paper's recommendation is to buy the options market. That is a conclusion about the S&P, an asset chosen precisely because its truth is observable, and the property that makes it observable is the same one that brings an option chain with it. Section 8 is where that stops being a limitation and starts being testable, because the case the paper opens with is an asset with no chain at all.

Two qualifications belong with the result and both are in Section S28. The specification here lags VIX a full day, which discards information the live index carried, because its printed close is stamped fifteen minutes after the S&P's and is therefore inadmissible under Section 3's own rule; the same-day close moves the race further towards options, so the admissible choice is the conservative one. And VIX is a thirty-calendar-day index racing a five-trading-day target, which are not the same unit. Thirty calendar days is about twenty-one trading days, so the instrument is measured over a horizon about four times the length of the thing it forecasts; the factor of six an earlier draft quoted came from dividing calendar days by trading days. Cboe publishes a nearer nine-day index, near-matched rather than exactly matched since nine calendar days is about six trading days, and Section S24 runs the race again with it: the horizon matters by more than we expected, and the conclusion of this section does not.

8. The illiquid asset

Everything to this point is measured on the S&P 500, and Section 7 has just established that this is the one asset class where the paper's own answer says breadth does not matter. The case Section 1 opens with is the opposite one: a quarterly-marked private book, an appraised property portfolio, a fund reporting a lagged net asset value. Nobody writes options on those. If the substitution rate is a fact about liquid equity indices then the paper has measured something real and recommended it to the wrong reader, and that is the largest thing that could be wrong with it.

So it is measured rather than argued. Section S31 of the Internet Appendix runs the whole design on twenty metropolitan Case-Shiller house price indices, monthly from 1987 to 2026: smoothed by construction, no intraday range, no option chain on any of them, and a genuine cross-section of twenty marks published on the same day. Delay is not the two-month publication lag every metro shares, which is common and therefore cancels; it is the same quantity as everywhere else in this paper, the amount by which one owner's mark is staler than the cross-section it sits in. An owner whose book is appraised on a cycle running behind the published index for the others is the concrete case, and three months is one quarterly appraisal cycle.

Two coordinates are carried separately, because a monthly index has no intraday range and Yang-Zhang cannot be computed on it. One is next month's log return, which is the quantity an owner cares about and is not this paper's target. The other is the log of trailing twelve-month realised variance over its own five-year median, which is *backward-looking* where this paper's target is forward-looking: forecasting next month's value of it means forecasting something that overlaps the current window in eleven of twelve months, so an own-only R^2 near 0.95 is arithmetic rather than skill, and

the small denominator that follows is why the rates on that coordinate are lower. The return target is the closer analogue.

At one quarterly appraisal cycle of extra staleness the rate is 67.8% [45, 92] on the return target, positive on all fourteen metros with a full sample, and 19.0% [6, 31] on the backward-looking variance coordinate, positive on thirteen of fourteen. Both exclude zero. The shape of Table 2 survives too: at one month the rate cannot be told from zero on either coordinate, exactly as it cannot at one day on the S&P, and it climbs from there.

Two objections are answered in Section S31 rather than deferred. The index is a three-month moving average, so consecutive monthly changes share transaction months and a smoother could manufacture a rate; read every third month on each phase of the quarterly grid, where the artefact explanation predicts zero, the pooled rate is 57% at one quarter of staleness and positive in 35 of 42 metro-phases. And these series carry a large shared seasonal, so a fresh peer could be revealing the calendar rather than the market; a placebo peer built to carry the seasonal and nothing else earns 4.2% where the real cross-section earns 73.0%. That is a test which could have convicted the design and did not.

What this settles is narrow and sufficient. The central claim is not a fact about liquid equity indices. It holds on an asset class that is marked rather than traded, that no options market covers, and whose marks are smoothed by the same mechanism that makes appraisal-based valuation slow. What it does not settle is the case itself: nobody holds a Case-Shiller index, and an index is not a book. The argument has moved from an analogy to a measurement one step removed from the asset it is about, which is further than it was and not the whole way.

9. Robustness and scope

The Internet Appendix examines the principal identified threats to each central claim. The Internet Appendix to this paper carries that work in full: eight alternative controls, a multiple-testing adjustment, a temporal split, a leakage check, a second estimator for the inference, and the two scope conditions that bound where the result holds, the option chain and coupling. This section states what those tests returned. It is deliberately short, because a reader who accepts the results does not need the working and a reader who does not should have it in front of them rather than summarised, which is why the working is a companion document rather than a paragraph.

The measurement survives its own inference. Giacomini-White on squared loss and on QLIKE, HAC-adjusted with a Bartlett kernel at the forty-lag bandwidth Section S13 measures, favours the cross-sectional model at every delay tested, with two-sided p below 0.01 on squared loss throughout and between 0.010 and 0.034 on QLIKE. Clark-West, a second opinion rather than the test this paper relies on, agrees, with a caveat developed in the Internet Appendix that is the reason it is the second opinion and not the first. A placebo that keeps the foreign series but destroys their alignment with the target returns a rate of -1% [−33, 29], which is the number the design should return when there is nothing to find. Reality Check and SPA, admitting how many models were tried, leave the headline standing.

It survives a stronger control than its own. The domestic block this paper uses is not the best one available, so eight alternatives were built and priced as a set, with the choice made on own-only skill and stated before any rate was read. The rate at eleven weeks spans 60.8% to 74.4% across them and

reads 67.5% against the strongest, a paired difference of -4.1% $[-9.7, -0.7]$ from the published figure. That is why this paper quotes 68% against the strongest control alongside its own 72%: the headline is not an artefact of a weak comparison, and the Internet Appendix shows what a ninth control would have to do to move it. It also isolates why that control is stronger, which is a result in its own right: the trailing-median normalisation of equation (2) is what puts eight markets on one scale and makes a cross-section possible at all, and it hides the level of volatility from every model in this paper. Restoring that level is worth more at long delays than the entire cross-section. The comparability the design buys is paid for in domestic information, and the price rises with the delay.

It is a conditional result, not an average one. Split by market state, the rate is 81% $[74, 87]$ at eleven weeks in this sample's stressed markets, and identical in both halves of the sample, while in calm markets so little damage is done that neither the ratio nor the effective age can be pinned down. Asked in days instead of shares, on a construction with no denominator that can vanish, the calm cell is 7.3 times as wide relative to its own point as the stressed one. Table 2 is an average of the two and the conditional figures are the ones a reader should carry.

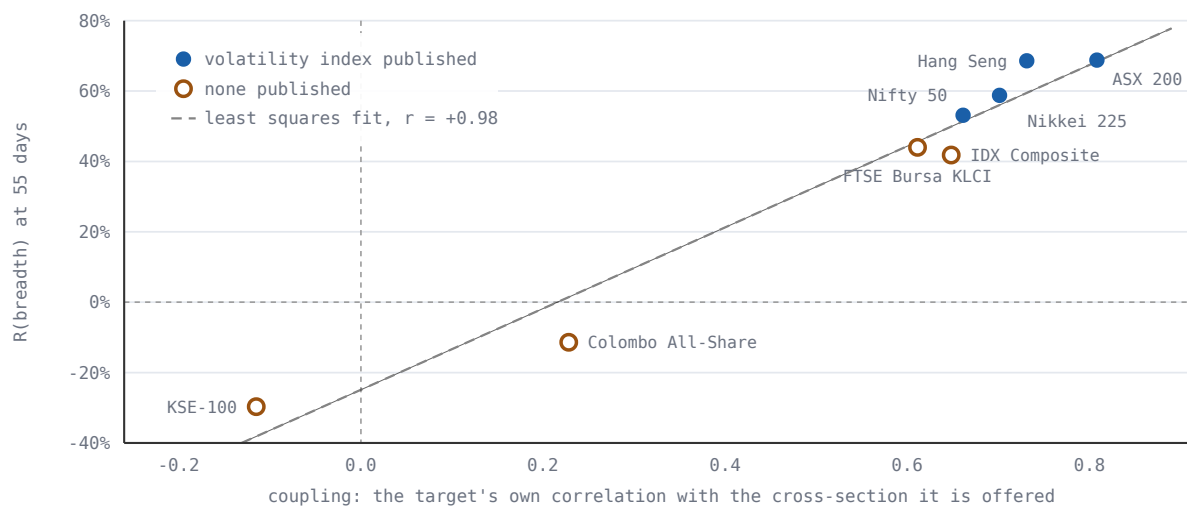


Figure 3. The substitution rate against coupling, for the eight targets of Table S11 of the Internet Appendix. Coupling is the correlation between a target's own decision variable and the equal-weighted mean of the cross-section it is offered, measured on its own test days; the rate is at $\delta = 55$. Filled markers are targets on which a volatility index is published and hollow markers those on which none is, and the relation runs through both groups, so what orders the rate is coupling rather than whether an options market exists. The fitted line crosses zero at a coupling of 0.22, but the two targets with negative rates sit at 0.23 and -0.12 , so the crossing locates the sign change to within one target and is not a threshold.

Two conditions bound where it holds, and both can be checked in advance. The first is the option chain, which Sections 7 and 8 establish and this section does not repeat. The second is coupling, the target's own correlation with the cross-section it is offered. It is a screening statistic and not a calibration rule: it sorts targets into those breadth can help and those it cannot, and it does not return a rate. Across eight markets it tracks the substitution rate at $+0.98$, running from 0.81 down to -0.12 , and the figure above is that relation. Eight points will bear one claim here and not two, and Section S17 of the Internet Appendix separates them. The ordering is robust: the rank correlation is $+0.98$ as well, no single target moves the Pearson figure below $+0.976$, and deleting both weakly coupled targets to fit inside the remaining cluster of six still gives $+0.90$. The fitted line is not: the KSE-100 sits at a coupling of -0.12 with every other target above $+0.22$, which gives it a leverage of 0.73 against a

mean of 0.25, and removing it moves the slope from +1.15 to +1.46 and the zero-crossing from 0.22 to 0.31. So the ordering is robust and the crossing is not, which is what the figure's caption says and what that diagnostic licenses: the crossing locates the sign change to within one target rather than fixing a threshold. A holder whose series is weakly coupled to public markets should expect breadth to do nothing for them, and can find that out from data already in hand rather than after buying a feed; a holder near the crossing should conclude that this paper cannot tell them which side of it they are on.

Three further checks are in the Internet Appendix and are named here so that their absence from this section is not mistaken for their absence from the work: whether the clean lag this paper models resembles the periodic marking a real book undergoes, whether the trailing median leaks the future into the target, and how the rate behaves across five target markets at every delay rather than at one.

10. Limitations

Four limitations are argued at length in Section S33 of the Internet Appendix rather than here: the thinness of the implied-volatility block, two bounds on the tests themselves, and the estimator, which is the largest of them and the one whose direction runs in the paper's disfavour.

The limitations below are individually argued, and a reader who wants the boundary in one place rather than assembled from nine paragraphs should read Table 5 first. It states, for every claim this paper makes, the strength of the evidence behind it and where that evidence is. The four grades are used strictly: *established* means the estimate excludes zero and survives the robustness set; *supported* means it is consistent across the tests run but rests on a sample too small to resolve alternatives; *suggestive* means the sign is stable and the magnitude is not; and *not claimed* marks a question this paper raises and leaves open. Nothing in the paper should be quoted at a grade stronger than its row.

Table 5. Every claim in the paper, its grade, and where its evidence is. The grades are defined in the paragraph above and are meant to be read as limits rather than as endorsements: a claim graded *supported* is one this paper would not defend against a larger sample.

claim	grade	the evidence, and its limit
A current foreign cross-section returns a large share of what staleness takes, on a liquid equity target	established	71% [59, 81] at eleven weeks, 68% to 75% across four specifications, every interval excluding zero (Sections 5, 7, S4)
The effect is larger in stressed markets	established	81% [74, 87], stable across both halves of the sample (Section S10)
The effect in calm markets	not claimed	the interval spans most of the unit interval and does not exclude zero at three of four delays; where delay costs little, the share of it repaired has no stable value (Section S10)
Implied volatility on the asset itself makes breadth redundant, in the bounded sense of Section 7	supported	the foreign block adds nothing detectable once implied volatility is held, at every delay, with a gross gain above 0.019 of R^2 ruled out and anything smaller left open; but two targets only, SPX and DAX, each on the 4,525-day window of Section 7, so the scope is the claim's limit and not its precision (Section 7)
Where the option chain is on another asset, breadth retains incremental value	supported	using the compressed one-regressor peer block and the same pointwise criterion in both groups, the foreign block adds significantly in thirteen of twenty-four cells where only another asset's volatility index is held, and in none of eight where the index is written on the target itself; those cells are in the post-selection audit of Section S37 and few survive its correction, so the pattern is what is supported and no single cell is (Sections S13, S37)
Coupling orders substitution rates across markets	supported	rank correlation +0.976, worst leave-one-out +0.976; eight targets (Section S17)
The fitted coupling line, and the coupling level at which the rate crosses zero	not claimed	eight points with two at high leverage; the slope moves by a quarter and the crossing by half under deletion (Section S17)
The measurement carries to an illiquid, appraisal-like asset	supported	57% on non-overlapping quarterly changes, all three phases of the quarterly grid excluding zero, 35 of 42 metro-phases positive (Sections 8.1, S18)
That result is not an artefact of the index's three-month moving average	supported	on the de-smoothed panel breadth still adds at three quarters on 14 of 14 metros; levels carry the reverse filter's noise amplification and only the sign transfers (Section S18)
A current cross-section hurts a forecaster whose own mark is fresh	suggestive	about one point of R^2 , negative under all three losses, but the natural-scale MSE cannot distinguish it from zero and no claim in this paper rests on it (Section S9)
What the rate is worth in money	not claimed	$R(\delta)$ is a share of forecast loss; converting it needs a decision problem the holder supplies (Sections 4.5 and 11)
That any of this transfers from a published index	not claimed	a published index is not a book: no appraisal noise, no manager discretion, and a fixed publication calendar (Sections 1, 9.2)

claim	grade	the evidence, and its limit
to an actual marked book		

The limitation that mattered most for twelve sections was the illiquid case itself, argued rather than demonstrated, and Section 8 narrows it as far as public data allows without closing it. What remains open inside it is stated there: the intervals are wide, the panel's couplings span a band only 0.35 as wide as the one that established the coupling relation so the slope cannot be tested on it, and an index is not a book. Nobody holds a Case-Shiller index, and a real appraised portfolio would differ from one in ways this panel cannot show. Sections S13 to S15 and S18 of the Internet Appendix carry the four steps by which the gap was narrowed before it was closed, and the two objections to the closing measurement that could have overturned it.

One limitation applies to every number here and to no single section. All of them were computed with the foreign block at its freshest admissible value, and the paper did not say so because it did not occur to us to vary it. A desk whose own foreign feed is one day behind recovers 29% of the delay-induced loss at five days where Table 2 reports 51%, and one three days behind recovers none of it. The measurement is not wrong, but the claim is narrower than it reads: it describes a forecaster with a stale domestic mark and a *current* foreign feed, a more particular creature than the one Section 1 has in mind. Whether an illiquid-asset holder is that creature is a question about their data vendor, not about volatility, and we cannot answer it.

A second qualifies the headline. The unconditional rate averages two states in which the mechanism behaves differently: well determined in stress, stable across both halves of the sample and large in absolute terms as well as in share; uninformatively imprecise in calm. Section S5 of the Internet Appendix shows that imprecision is not an artefact of the ratio, since asked in days instead, on a construction with no denominator that can vanish, the calm cell is 7.3 times as wide relative to its own point as the stressed one. Table 2 is an average of the two and the conditional figures in Table S4 of the Internet Appendix are what a reader should carry.

What is left is short. The horizon evidence in Section S28 runs out at one month, where the rate has already fallen to 51% with an interval too wide to lean on; we do not know what happens at a quarter. Equity indices only. Close times are treated as standard, ignoring daylight saving, which moves them by an hour and never reorders the blocks. And the index data are a retail export with no point-in-time guarantee, on which one join cannot be verified by comparison: of the nine two-file exports, eight overlap by 120 to 251 days and agree exactly on every overlapping row, while the S&P's two files abut without overlapping. That is the target series, so a statement of contiguity is worth less than an audit, and Cboe's own VIX history supplies one: a different vendor on a different data path measuring the same underlying, spanning the join. Section S19 of the Internet Appendix reports that audit, its three instruments, what size of fault each would have caught, and the one corruption that no instrument here could detect. It also reports something we were not looking for, which every user of retail market data should know: the retail VIX export carries fifteen rows on dates United States equities were shut. All fifteen fall outside the panel, so no model was ever shown one.

11. Conclusion and further work

The universe here is eight equity indices, and the obvious way to widen it is the one this paper predicts will fail. Section S27 splits the foreign block into its leading real-time component and the six components orthogonal to it, and finds the entire recovery in the first: the remainder is negative at

every delay and beats the stale domestic model at none. A ninth and a tenth equity index would supply more of the same factor and more coefficients to estimate, and Section 6 has already priced that trade. Breadth within an asset class is not the frontier. Section S14 of the Internet Appendix says what is: the eight targets there differ in coupling from 0.81 to -0.12 , and the substitution rate follows it almost exactly. A study that sampled markets ON that axis, rather than by whatever data was to hand, could turn eight points into a calibration, and a practitioner holding a stale series would then read their own rate off it.

Breadth ACROSS asset classes is, and it is a sharper test than it first appears because this paper has to make a prediction about it. Sovereign bond futures, the major commodities and the G10 currencies all have observable daily ranges and therefore the same Yang-Zhang decision variable, so the design transfers with nothing changed but the panel. The stress ladder of Section S10 of the Internet Appendix suggests they load on the same global risk factor, since cross-asset correlation is what converges in the episodes on that ladder. If they do, they will be redundant for the reason the orthogonal equity components are redundant, and the mechanism claimed here will have survived a considerably harder test than another equity index.

Foreign exchange should be run first, for a reason Section S12 of the Internet Appendix makes plain, and for no reason about currencies. FX trades continuously across the week, so it has no admissible-timestamp problem at all: there is no previous session for a currency pair, no peer that closes after the target, and none of the shift rules in Section 4. That also means it is exempt from the penalty in Table S8 of the Internet Appendix, the largest practical threat this paper has to its own recommendation. A substitute that is current by construction is precisely what a holder of a stale mark wants, and of the candidates listed here FX is the only one that is current by construction, not by luck of the clock.

Two studies would matter more than any of those, and both are named rather than attempted because each is a paper. The first closes the gap Section 10 leaves open. A Case-Shiller index is not a book: nobody holds one, and the quarterly and de-smoothed tests of Section S18 of the Internet Appendix establish that the result is not an artefact of how that index is built without establishing anything about a real appraised portfolio. What would be a panel of actual marks, whether private equity net asset values, property appraisals or fund marks, carrying a publicly observable peer factor constructed as it stood at each historical marking date, on which the delay curve could be estimated directly. Brown, Ghysels and Gredil (2023) nowcast private equity net asset values from public information and so demonstrate that such a panel can be assembled; what they do not do is sweep the delay, which is the one thing $R(\delta)$ needs.

The second turns the rate into money, and the paper is honest that it currently does not. $R(\delta)$ is a share of forecast loss; a holder wants to know what a data feed is worth. Embedding the forecasts in one decision problem would settle it: a volatility target, a value at risk, or a capital buffer, each of which converts a change in forecast error into a quantity with a currency attached. That would license a sentence this paper cannot write, of the form that a current peer feed reduces risk-measurement error by so much and is therefore worth at most so much a year to a book of a given size. Until someone does that, “price” in this title means a ratio on one ruler and not an amount, which is why Section 4.5 says so where a reader first meets the measure.

Three other threads are open and named where they arise: the illiquid-asset target this whole exercise is about, which Section 8 now reaches on an index panel and which still wants a real appraised book rather than an index, in Section 10; whether the rate keeps rising past one month, where Section S28's

evidence runs out with an interval too wide to lean on; and, in the companion note (Vellaniakaran 2026), ranking forecasters against a tail target, which needs more independent volatility spikes than nine years of one index contains. The cross-asset panel proposed above would help with the last of those as a side effect, since spikes accumulate across markets as well as through time.

Data and code availability

This paper ships with an Internet Appendix carrying the robustness and scope work behind Section 9 in full, including eight alternative controls, the temporal split, the multiple-testing adjustment, the coupling diagnostics, the house price overlap audit and the two scope conditions. Every number in the paper and in that Internet Appendix is printed by a script, and the scripts, the reference output and the source of every document are at github.com/SunnyAlexV/delayed-information-cross-section. A further script, `build_registry.py`, builds a machine-readable registry of every figure the verifier ties to a lab, keyed on the check's label, its script and its section, and runs five gates over it: no point estimate may be quoted in prose with two different intervals, a p-value written beside a statistic must be that statistic's p-value, a contrast that names a winner must state its direction, the count of figures no check reaches may fall but not rise, and every negative figure must still carry its sign when the rendered PDF is put back through text extraction. The registry is built from the verifier's own checks rather than maintained beside them, because a second copy of the results is the thing these gates exist to catch.

The data are not there, and the reason is worth stating rather than leaving a reader to discover. Every vendor behind the inputs prohibits redistributing their files: Investing.com's terms forbid distributing data obtained from the site without written permission, Cboe's permit one downloaded copy for personal non-commercial use and nothing further, Yahoo's forbid reproducing or distributing any portion of their services, and the S&P Cotality Case-Shiller series carry FRED's *Copyrighted: Pre-approval Required* label, which without the copyright holder's permission allows non-commercial educational or personal use alone. Using those series to do research and publishing the results is a different act from republishing the files, and only the second is forbidden; of the thirty-five input files exactly one, the FHFA index that FRED places in the public domain, could be included and is.

What ships in place of the data is a manifest recording, for every input file, its source, its row count, its date span, the SHA-256 of the exact bytes these numbers were computed from, the digest of the derived series on those dates, and that digest taken one calendar year at a time. A reader who obtains the same exports therefore reproduces every number bitwise, and a reader whose export is fresher than ours is told so per file and per year rather than left to guess whether a difference is a new month of rows or a different series. Three commands do it: `python verify_data.py` classifies each file as identical, the same data with a longer tail, or genuinely different and in which years; `python run_all.py --check --data` reruns all sixty-three scripts under a fixed seed and compares each line against the stored output; and `python verify_paper.py` reads the source of this paper and its Internet Appendix together and fails on any figure that disagrees with that output, so a result cannot be lost by moving it between them.

The repository's README names which script produces which result. Four scripts were also reproduced on a second machine running a different operating system and matched digit for digit; the rest have been run only on the platform that generated the reference output, and we state that rather than the stronger claim. Nothing is reported that the scripts do not produce.

Disclosure of AI use

During preparation of this manuscript and its Internet Appendix, the author used Claude (Anthropic) as an assistive tool for drafting and revising text and research code.

The research questions, mathematical development, empirical design, data construction, analysis, interpretation of results, and decisions regarding the paper's claims were the author's own. The author independently specified and executed all analyses; reviewed, tested, and validated the code and outputs; and takes sole responsibility for the manuscript's contents, including all methods, results, interpretations, and reported numbers.

No AI system is an author of this work and no AI system held authorial responsibility for any part of it.

References

- Brown, G. W., Ghysels, E. and Gredil, O. (2023). Nowcasting net asset values: the case of private equity. *Review of Financial Studies* 36(3), 945–986.
- Buncic, D. and Gisler, K. I. M. (2016). Global equity market volatility spillovers: a broader role for the United States. *International Journal of Forecasting* 32(4), 1317–1339.
- Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* 138(1), 291–311.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7(2), 174–196.
- Croushore, D. and Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics* 105(1), 111–130.
- Diebold, F. X. and Yilmaz, K. (2014). On the network topology of variance decompositions: measuring the connectedness of financial firms. *Journal of Econometrics* 182(1), 119–134.
- Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold-Mariano tests. *Journal of Business & Economic Statistics* 33(1), 1–9.
- Doz, C., Giannone, D. and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics* 164(1), 188–205.
- Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association* 78(383), 605–610.
- Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65(6), 1365–1388.
- Engle, R. F., Ito, T. and Lin, W.-L. (1990). Meteor showers or heat waves? Heteroskedastic intra-daily volatility in the foreign exchange market. *Econometrica* 58(3), 525–542.
- Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B* 16(2), 175–185.
- Geltner, D., MacGregor, B. D. and Schwann, G. M. (2003). Appraisal smoothing and price discovery in real estate markets. *Urban Studies* 40(5–6), 1047–1064.
- Getmansky, M., Lo, A. W. and Makarov, I. (2004). An econometric model of serial correlation and illiquidity in hedge fund returns. *Journal of Financial Economics* 74(3), 529–609.
- Ghysels, E., Sinko, A. and Valkanov, R. (2007). MIDAS regressions: further results and new directions. *Econometric Reviews* 26(1), 53–90.

- Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica* 74(6), 1545–1578.
- Giannone, D., Reichlin, L. and Small, D. (2008). Nowcasting: the real-time informational content of macroeconomic data. *Journal of Monetary Economics* 55(4), 665–676.
- Goetzmann, W. N., Ivković, Z. and Rouwenhorst, K. G. (2001). Day trading international mutual funds: evidence and policy solutions. *Journal of Financial and Quantitative Analysis* 36(3), 287–309.
- Jayawardena, N. I., Todorova, N., Li, B. and Su, J.-J. (2020). Volatility forecasting using related markets' information for the Tokyo stock exchange. *Economic Modelling* 90, 143–158.
- Korkusuz, B. (2025). Beyond the S&P 500: examining the role of external volatilities in market forecasting. *Review of Economic Design*. DOI 10.1007/s10058-024-00373-x.
- Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* 17(3), 1217–1241.
- Lo, A. W. and MacKinlay, A. C. (1990). An econometric analysis of nonsynchronous trading. *Journal of Econometrics* 45(1–2), 181–211.
- Marcellino, M. and Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: a model comparison for German GDP. *Oxford Bulletin of Economics and Statistics* 72(4), 518–550.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160(1), 246–256.
- Rapach, D. E., Strauss, J. K. and Zhou, G. (2013). International stock return predictability: what is the role of the United States? *Journal of Finance* 68(4), 1633–1662.
- Rogers, L. C. G. and Satchell, S. E. (1991). Estimating variance from high, low and closing prices. *Annals of Applied Probability* 1(4), 504–512.
- Scholes, M. and Williams, J. (1977). Estimating betas from nonsynchronous data. *Journal of Financial Economics* 5(3), 309–327.
- Timmermann, A. (2006). Forecast combinations. In *Handbook of Economic Forecasting*, vol. 1, 135–196. Elsevier.
- Vellanikaran, S. A. (2026). Why a simple threshold rule is hard to beat: volatility regime prediction under information delay. Working note.
- Wallis, K. F. (1986). Forecasting with an econometric model: the 'ragged edge' problem. *Journal of Forecasting* 5(1), 1–13.
- Yang, D. and Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. *Journal of Business* 73(3), 477–491.
- Zhao, X. and Qiao, H. (2026). Forecasting global volatility with predictive spillover networks: a neuro-econometric spatio-temporal transformer for asynchronous financial markets. arXiv:2608.25369v2, 5 September 2026. The version this paper first cited, v1 of 26 August 2026, carried a different title and a third author; the non-synchronous masking we draw on is present in both.

Internet Appendix for *What Substitutes for a Stale Mark? Breadth, Implied Volatility, and the Price of Being Late*

Sunny Alex Vellanikaran

Independent

22 September 2026

This Internet Appendix carries the robustness and scope work behind Section 9 of the paper. It is an appendix in name and a companion in use: the paper states every result below and this document shows the working, so a reader who accepts the results can stop at the paper and a referee who does not can check every one of them here. Numbering continues from the paper, tables are lettered S1 onward, and every figure in both documents is produced by the same scripts and checked by the same `verify_paper.py`, which reads the paper and this Internet Appendix together and fails if a figure appears in neither.

S0. What this appendix is for, in one page

Thirty-seven substantive sections, S1 through S37, plus this claim map, is enough that a reader cannot tell from the list which tests the paper leans on and which are completeness. This page says so. Each row is a claim the paper makes, the sections that carry its evidence, and the objection those sections were built to answer; every section heading also carries its own verdict, so a reader can see what a section concluded without reading it.

Table S0. The paper's claims, the evidence behind each, and the threat each piece of evidence addresses. A claim with more than one section behind it is one that was attacked from more than one direction, not one that needed more support.

claim in the paper	evidence	threat addressed
Breadth repairs a stale mark	S2, S4, S6, S8	inference, a weak benchmark, multiple testing
The measurement is not leakage	S7, S16	an overlapping target, a target normalised at the wrong date
The effect is larger in stress	S5, S10	ratio pathology when delay costs little
Timing is what makes it work	S12	a peer feed that is itself stale
Coupling says where it works	S14, S17	external validity, and leverage in eight points
Options dominate on the asset itself	S11, S13, S22, S24	a thin options block, non-linearity, mismatched horizons
The housing result is informational	S18	the index's own three-month moving average
It is not an artefact of the estimator	S20, S21	ridge dependence, a fitted surrogate generator
The apparatus fits a marked asset	S15, S23	clean lags versus periodic marking, forecast horizon
The losses are the ones claimed	S3, S25	nesting under shrinkage, a mislabelled variance-scale MSE
The data are what we say they are	S19	the one export join that cannot be checked by comparison
The secondary grids are exploratory, and the headline is not	S18, S37	the search behind the headline, and a false-discovery correction across everything reported after it

Two rows deserve flagging because their verdicts are not "survives". S4 finds a stronger domestic control than the paper's own and the rate falls from 72% to 68% against it, which is why the abstract quotes a span. S17 finds that the coupling *ordering* is robust while the fitted line through eight points is not, which is why Section S30 declines to read a threshold off it.

S1. Robustness of the measurement

$R(\delta)$ carries S_{own} in three of its four places, so every figure above is a statement about the domestic model as much as about the foreign one. That is uncomfortable, because the paper has spent its effort on the other side: Section 6.1 compresses the foreign block, Section S27 splits it, Section S28 tries it as one regressor. The control has been the same three features throughout. This section varies it, and then varies three other things that could make the headline wrong: the sample period, the number of models we tried before arriving at this one, and the gap between training and forecast. Each was run once and is reported whatever it said.

S2. Statistical inference · *Result survives*

The bootstrap intervals of Section 5 quantify sampling uncertainty in a ratio of realised out-of-sample losses, and an earlier version of this section dismissed them as unable to test anything at all. That was too strong: any interval can be inverted into a test of the values it excludes, and a reader is right to object to a claim that says otherwise. What is true is narrower, and it is what the rest of this section rests on. Those intervals are built for $R(\delta)$ rather than for the loss differential, and they carry no adjustment for the fact that our domestic model is nested inside our cross-sectional one, which is exactly the case in which the usual Diebold-Mariano comparison loses its standard distribution. Diebold has written at length about the misuse of his own test in exactly this situation (2015). So the equal-predictive-ability question is given a test of its own rather than read off a rate interval that happens to exclude zero. We use Giacomini-White (2006) as that test, and as the primary test of this paper, because it is valid for nested models and, being loss-agnostic, covers the QLIKE result as well as the squared-error one. It requires a fixed rolling estimation window, of the kind Section 4 describes. Writing L_t^{own} and L_t^{alt} for the two models' losses on test day t and $D_t = L_t^{\text{own}} - L_t^{\text{alt}}$ for the loss differential, the statistic is

$$GW = d \div SE(d), \quad SE(d)^2 = (1/n) \left[\gamma_0 + 2 \sum_{j=1}^L (1 - j/(L+1)) \gamma_j \right], \quad (14)$$

where d is the mean of D_t over the n test days, γ_j is its sample autocovariance at lag j , and the Bartlett weights $(1 - j/(L+1))$ run to $L = 40$ lags. Both parts of that sentence need defending. The HAC correction is not optional: a five-day-ahead target overlaps on four of every five days, so the raw standard error of a daily loss differential understates by roughly a factor of three. And the bandwidth is measured rather than taken from the usual automatic rule, which gives 9 lags at this sample size and truncates the kernel while the differential still carries 0.52 autocorrelation. Section S13 reports the measurement: at eleven weeks the automatic bandwidth returns 0.0495 against 0.0752 at forty lags and 0.0810 at eighty, while a moving-block bootstrap, which assumes no kernel at all, returns 0.0760. The three agree once the bandwidth outlasts the dependence and the short one does not, so L is set to the block length of Section 5: both answer the same question and letting them answer it differently is what left an earlier simultaneous band studentised by one and resampled by the other. Every statistic this corrects was inflated, which is to say every correction here is against the paper's own result. Positive GW favours the model with the substitute, and we report two-sided p -values from the

standard normal throughout. Clark-West (2007) serves as a second opinion on squared loss, with a qualification developed in Section S3 that is the reason it is the second opinion, not the first.

Table S1. Test statistics, HAC-adjusted with a Bartlett kernel at the measured bandwidth of Section S13, forty lags rather than the nine the automatic rule would give. Positive favours the cross-sectional model. Every entry is two-sided significant at 5%. The squared-loss and Clark-West rows have $p < 0.01$ at every delay; the QLIKE row runs from $p = 0.010$ at $\delta = 8$ to $p = 0.034$ at $\delta = 3$, so it clears 5% throughout and 1% nowhere. The QLIKE row was inside 1% under the nine-lag kernel and is not under the measured one, which is the correction working rather than the result weakening: a statistic divided by too small a standard error was always going to look more decisive than it was.

δ	3	5	8	13	21	34	55
Giacomini-White, squared loss	2.95	3.88	4.28	4.26	4.07	4.26	4.62
Giacomini-White, QLIKE	2.12	2.51	2.57	2.26	2.18	2.25	2.40
Clark-West	6.15	6.51	6.56	6.15	5.68	5.78	6.01

P-values are two-sided throughout. On that basis every statistic from $\delta = 3$ onward has $p \leq 0.034$. The squared-loss and Clark-West rows are all below 0.01; the QLIKE row is not, running from $p = 0.010$ at $\delta = 8$ to $p = 0.034$ at $\delta = 3$, its weakest entry. A one-sided test against the directional alternative that the cross-sectional model is better would halve each of these, putting the squared-loss and Clark-West rows under 0.002 and the whole QLIKE row under 0.02, and the alternative is directional in substance; we report the two-sided figure anyway because it is the more demanding of the two. QLIKE, the harder loss to move, is significant at the 5% level throughout and at the 1% level nowhere, and Table S1's caption says so rather than leaving the reader to infer it from the statistics.

S3. A caveat on the Clark-West column · *Claim narrowed*

Clark-West is the second opinion on squared loss and carries a qualification worth stating. Its correction term assumes the larger model nests the smaller exactly; a ridge with a cross-validated penalty satisfies that only approximately, because shrinkage pulls the larger model away from the nesting, and the correction can then flatter it. The diagnostic below measures how far from exact the nesting is at each delay, and the remedy imposes a single ridge penalty on both arms, selected on the restricted one. That is the fix the objection calls for rather than a different one: the correction is flattered when the two arms are shrunk by different amounts, so equalising the shrinkage closes that channel, and it does so without leaving the estimator the rest of the paper uses. It should be said exactly what the remedy buys. A common penalty reduces the departure from exact nesting rather than completely removing it. A ridge fit on the wider design is not the ridge fit on the narrower one with the extra coefficients set to zero, whatever penalty the two are given, because the penalty acts on a different number of standardised columns in each; some departure therefore survives any choice of common penalty. What the remedy removes is the differential-shrinkage part, which is the part that moves with δ and the part the diagnostic below measures. The residual part is common to both arms and small by construction, but it is not zero, and we do not claim it is. An unpenalised refit would also make the nesting exact and would answer a different question, about a model this paper never runs. The diagnostic is a ratio with no established critical value. We offer that ratio as a description, not a threshold, and lean on the common-penalty re-run instead of on where the ratio happens to fall.

Table S2. A diagnostic and a remedy for the Clark-West column, applied to this paper's own statistics. The ratio is the mean Clark-West adjustment over the magnitude of the mean loss differential; the common-penalty columns impose a single ridge penalty, selected on the domestic arm, across both.

δ	refits where the penalties differ	adjustment ratio	Clark-West	with a common penalty	Giacomini-White	with a common penalty
0	18%	1.10	0.22	0.26	-2.09	-2.05
3	50%	1.60	6.15	6.11	2.95	2.94
5	44%	1.27	6.51	6.51	3.88	3.86
13	46%	1.11	6.15	6.12	4.26	4.24
21	33%	1.03	5.68	5.69	4.07	4.06
55	33%	0.95	6.01	6.00	4.62	4.62

The diagnostic runs from 0.95 to 1.60, and only $\delta = 3$ exceeds 1.5, the one Clark-West entry we would not lean on. The decisive column is the next one: imposing a common penalty moves no Clark-West statistic by more than 0.05, and changes no verdict at any delay. So the channel the remedy closes is present in principle and undetectable in practice here. The channel it does not close, the residual non-nesting of two ridge fits on designs of different width, is not measured by this table and is the reason the column stays a second opinion rather than being promoted alongside Giacomini-White. It is reported with the ratio beside it instead of as a bare p-value.

The primary test is untouched, and for a reason, not by luck. The argument above turns on Clark-West being a test about *models*, so that anything changing the estimation penalty changes what its adjustment ought to be. Giacomini-White is a test about *methods* under a fixed rolling window, and a method is the whole procedure including whatever tuning happens inside that window. The validation slice is carved out of the training window, so it is part of what is being tested rather than a contaminant of a correction, and Giacomini-White makes no finite-sample correction that could be thrown off. The prediction is that its column barely moves under the remedy, and the last two columns of Table S2 are that prediction tested: no statistic moves by more than 0.06 and no verdict changes at any delay. The substitution rates, the bootstrap intervals and the simultaneous bands involve no nested-model correction at all and are unaffected.

A second concern is that we test ten delays at once. Ten pointwise 95% intervals will breach somewhere by chance, so we bootstrapped the whole vector jointly, in moving blocks of forty days to respect the overlap in the target (Künsch 1989), and took the largest studentised deviation across the grid. The resulting critical value is 2.42 instead of 1.96, a heavy penalty, and the effect still clears it from $\delta = 3$ to $\delta = 55$. One casualty is worth naming: the small negative reading at $\delta = 0$ does *not* survive the simultaneous band. We do not claim the cross-section is harmful when domestic data is current, and Section 6 gives the reason it is not: net of what it costs to estimate, its content there is zero. That casualty is worth using to fix a vocabulary, because three different statements get confused in this literature and this paper needs all three. *Pointwise*: an individual interval excludes zero, which is what the cells of Table 2 report. *Simultaneous*: the effect survives a band covering the whole delay grid at once, which is the 2.42 critical value above and which $\delta = 0$ fails. *Equivalence*: the effect lies inside a margin declared negligible in advance, which this paper tests nowhere and therefore never claims. A cell that fails the first is one where no incremental contribution was detected; it is not a cell

where the contribution is zero, and wherever an interval still admits a contribution worth having we say what that interval permits rather than rounding it to nothing.

The third concern is that the result is an artefact of fitting a wide model to a persistent series. We tested it by shifting the entire foreign block five years through the sample. The series keep their distributions, their autocorrelation, their crises; they lose only their alignment with the target. The substitution rate falls from +72% to -8%, and is negative at every delay we checked. Whatever the model is using, it is the alignment. The residual negative is not noise around zero and should not be read as one: it is the cost of breadth from Section 6, which is all that remains once the information is removed. A placebo of this kind lands below zero by construction, and a version of it that came back at exactly zero would be the surprising result.

Finally, Section 5 described the rate as slowing, which is a claim, not an observation. Comparing $R(55)$ with earlier values: against $R(13)$ the difference is +0.079 with an interval of $[+0.037, +0.129]$, and against $R(21)$ it is +0.042 with $[+0.014, +0.076]$, both still rising. Against $R(34)$ it is +0.018 with $[-0.001, +0.039]$, which does not reject equality. The rate is still climbing out to about seven weeks and flat thereafter.

One last property of the rate was assumed and never tested. The rate is a ratio whose denominator is what the delay costs a domestic-only forecaster, and at short delays that quantity is near zero. A ratio with a denominator that can vanish is weakly identified, and a percentile bootstrap of it reports finite endpoints whether or not they mean anything. We therefore recomputed every interval by inverting a HAC t-test after Fieller, which returns two half-lines or the whole line when the data do not pin the ratio down. At every delay from one day to eleven weeks the confidence set is an ordinary interval, and Fieller's endpoints sit within 1.6 percentage points of the bootstrap's throughout: at $\delta = 55$, $[59\%, 81\%]$ against $[60\%, 81\%]$. The objection is right in principle and inert here, so the published intervals stand. Section 4.6 of the paper gives the construction, the one delay where the ratio does not exist at all, and the error that revealed the first version to be wrong.

S4. The domestic control and the level normalisation · *Magnitude falls*

Five stronger domestic blocks were built, each a superset of the paper's: four further own lags; quarterly and annual means; leverage terms from the signed return; a second realised measure together with the LEVEL of the trailing median; and all of them at once. Each was given the same walk-forward, the same refit schedule and the same ridge, and $R(\delta)$ was recomputed inside each arm, so every rate is measured against its own control. The strongest was chosen on own-only skill averaged across delays, never on the substitution rate, the quantity under test and must not be allowed to select its own control.

Table S3. The paper's domestic block against the strongest alternative. The first two columns are out-of-sample R^2 for the domestic model alone, δ days stale. The rate is recomputed inside each arm. The interval is on the DIFFERENCE, from a paired bootstrap that recomputes both ratios, numerator and denominator, on the same resampled days.

δ	R^2 own, paper	R^2 own, +level	$R(\delta)$ paper	$R(\delta)$ +level	difference	95% interval
0	0.5191	0.5252	n/a	n/a	n/a	n/a
5	0.3641	0.3770	51%	47%	-4.0%	[-11.8, +1.8]
13	0.2336	0.2542	64%	60%	-3.5%	[-9.9, +0.5]
21	0.1482	0.1796	67%	64%	-3.3%	[-9.1, +0.3]
55	0.0070	0.0572	72%	68%	-4.1%	[-9.7, -0.6]

The control was improvable and the rate moves. A richer domestic block beats the paper's in 11 of 45 cells, and the substitution rate falls by three to four points at every delay, significantly at $\delta = 34$ and 55. Two things about that last clause. The direction is the same at all eight delays and the size barely varies, so what separates the significant delays from the rest is power rather than mechanism, and we read the column as one result rather than as two. And the list was four delays long under the nine-lag HAC kernel, which Section S13 shows truncates inside the dependence; at the measured bandwidth the table above carries, $\delta = 13$ and 21 fall just short, at [-9.9, +0.5] and [-9.1, +0.3]. Four points is not a reversal, since 68% sits inside the [60, 81] interval Table 2 already carries, but the paired test is sharper than the overlap and it excludes zero at the two longest delays, so the gap is real and the paper quotes both figures and leans on neither.

What makes that arm strong is the part worth keeping. It carries two features and they are not the same kind of thing. One is a second realised measure, a 22-day Yang-Zhang, which is simply less noisy than the 5-day one; the other is $\log M_t$, the LEVEL of the trailing median. Splitting them settles it at once. At eleven weeks the paper's block scores -0.0351, adding the second measure scores -0.0428, which is worse, and adding the level alone scores +0.0308. The level does the work, at nine delays out of nine, and by a margin that grows with staleness: +0.0011 of R^2 at zero delay against +0.0736 at eleven weeks.

That is a finding about this paper's own design, not about volatility. Equation (2) divides every series by its own trailing median, and that is what puts Tokyo, São Paulo and the VIX on one scale and makes a cross-section possible at all. It is also what prevents every model in the paper from seeing how volatile the world actually is. A forecaster who has lost his recent observations still knows roughly what regime he was in, and that knowledge is worth more the more he has lost, which is where the paper claims its largest effect. The comparability the design buys is paid for in domestic information, the price rises with delay, and we have not found this trade priced anywhere in the literature the paper cites. Anyone normalising a panel this way should measure what it costs before assuming it is free.

That claim has an obvious objection and it has to be answered rather than noted. The target is $\log(\sigma^2 / M_{t+5})$, M is a 252-day median and therefore very slow, so a model handed $\log M_{t-\delta}$ could lower its squared error simply by predicting the target's own DENOMINATOR. That would be arithmetic about how the target was built, not a forecast of volatility. Nothing would be inadmissible, since every feature is dated $t - \delta$, but the finding would be empty.

The test is to re-date the normaliser and see what survives. Three targets were built, differing in nothing else: the paper's, normalised at $t+5$; the origin-dated one of Section 4, normalised at t ; and one normalised at $t - \delta$, the forecaster's own information set. The third is decisive, because its denominator is known exactly when the forecast is made, so a model cannot buy anything by predicting a quantity it has already been given, and it is also the only one of the three a delayed forecaster could actually use. Across the four positive delays the third target retains between 87% and 103% of the gain: +0.0656 against +0.0638 at eleven weeks. Re-dating the normaliser changes nothing, so the level is not being used to undo the normalisation. It is information.

Two qualifications. The gain reaches conventional significance only at $\delta = 55$, on all three targets alike, so the short-delay column shows a direction without establishing a result. And the rate correction above survives the same test: on the implementable target the level moves $R(55)$ from 73% to 69%, a shift of the same size as the 72% to 68% reported here.

A third objection reaches past the feature list to the estimator class. Every arm above hands a delayed observation to a ridge as a stale number, where a state-space model would project it forward, so the comparison could be said to test feature counts and never the class of estimator. The obvious such model was fitted on the domestic series alone: a first-order state observed with noise, with $\gamma(0) = q/(1 - \phi^2) + r$ and $\gamma(k) = \phi^k q/(1 - \phi^2)$, so that $\phi = \gamma(2)/\gamma(1)$ and the three parameters come from the first three autocovariances of the first 1,500 rows. They are then frozen, and no foreign series enters anywhere. The moments give $\phi = 0.9265$, a half-life of 9.1 trading days, and an observation-noise variance of zero to five decimal places. That last figure is already part of the answer: with no measurement noise the Kalman gain is one and the filtered state is the observation, to 4.4×10^{-8} , so at daily frequency there is nothing for the update step to clean up.

The projection is the rest of it, and it is algebra rather than evidence. A forecaster δ days late projects by $\phi^{\delta+h}$, which multiplies a column by one constant; the ridge standardises each column before fitting; and $(cx - c\mu)/(c\sigma) = (x - \mu)/\sigma$. The coefficient absorbs the constant and the forecast does not move. Run both arms on the same days and the largest difference across six delays is 1.8×10^{-11} , which is the 10^{-9} guard added to the standard deviation and nothing else. At a fixed delay, projecting a state forward cannot improve any scale-equivariant linear forecaster, whatever else it does for the forecaster's intuition. The objection has to enter through the state itself, if a filtered history spans something the paper's three features do not.

It does not. The paper's control is the stale level and rectangular means over 5 and 22 days; a filter weights the whole history exponentially, so the control was given both, with exponentially weighted averages at half-lives of 5, 22 and 66 days added to what it had. Six features against three, and the six-feature block is worse at every delay tested: by 0.0018 of R^2 at the origin, 0.0321 at $\delta = 21$ and 0.0522 at eleven weeks, losing by more the staler it is. Nothing was added to the information set, since every new column is a weighted average of a history the model already held. What was added was six coefficients to estimate on a training window that does not grow, which is Section S2's cost of breadth arriving inside the domestic block.

The rate then moves in the direction the objection did not anticipate, and the reason is worth stating once in general. Write the substitution rate as N/D with N the error the cross-section removes and D the error the delay adds. Raising $S_{\text{own}}(\delta)$ by e subtracts e from both, and $(N - e)/(D - e) < N/D$ whenever $N < D$, which holds here at every delay. A weak control inflates R , exactly as the objection

says. The question was only ever which control is the weak one. Measured against the exponential block, $R(\delta)$ is higher at all five delays, by up to 4.1 points, reading 74.3% at eleven weeks where the paper's control reads 71.4% and Table S8 rounds it to 71%. The alternative is the arm that inflates. What survives this section is the correction in Table S3, which comes from the level of the trailing median and not from the class of estimator. Section S12's ruler behaves the same way: rebuilt from the exponential curve it reads the headline as 4.2 days rather than 4.6, and its interval is 7% narrower as a share of the point it surrounds, so the paper keeps the reading that flatters it less.

Three attacks on the control have now been answered one at a time, and answering an objection three times is not the same as retiring it, because a fourth control can always be named. So the control is priced as a set instead. Eight domestic blocks were run through the same walk-forward: the paper's, the four supersets of Table S3, all of them at once, the exponential block above and the projected state above. At eleven weeks the substitution rate across the eight runs from 60.8% to 74.4%. The strongest by own-only skill averaged across delays is the arm carrying the level, at 0.3022 against the paper's 0.2803, and the rate inside it is 67.5%, a paired difference of -4.1% $[-9.7, -0.7]$ from the paper's 71.6%. That reproduces Table S3's four points from a second direction, on a different feature set and a different script, which is worth more than a number agreeing with itself. One bookkeeping point, because this section now quotes the paper's own control twice and the two figures differ. The state-space comparison above reads it as 71.4% and the envelope reads it as 71.6%, and Table 2 reads it as 72%. Each file builds its panel on the delay grid its own comparison needs, so each starts its test window a few days from the others. The spread is the third significant figure of a quantity whose interval is $[60, 81]$, and every comparison drawn from it is internal to the file that made it.

One rule has to be stated because the obvious choice is wrong. The lowest figure in that set is 60.8%, and it is not the conservative rate. The arm producing it scores 0.5748 own-only at the origin and -0.0648 at eleven weeks, so it is the strongest control in the set where the denominator is set and among the weakest where the numerator is. $R(\delta)$ reads the own-only curve at two points, and an arm strong at one and weak at the other moves the rate without being a better control anywhere. Quoting it would be the mirror of the error this section is answering, made in the paper's own favour, so selection stays on own-only skill and is fixed before the rates are read.

What the envelope retires is the open-ended form of the objection. The control is no longer one choice a reader takes on trust: eight are priced, the rule is stated in advance, and a ninth moves the headline only by beating all eight on own-only skill. One further check belongs here rather than in Section 7, because it is about the control rather than about the candidates. Section 7's redundancy result was established with the paper's domestic block underneath both, and a stronger block could in principle leave less room for the options market too. Re-run inside the level-carrying control on Section 7's own window, the foreign block adds nothing detectable at any of six delays, with statistics from -1.94 to -2.86 , every one of them more negative than with the paper's control. Section 7 does not rest on a weak domestic model either.

S5. Temporal stability · *Result survives*

Section S10 conditions on market state; this asks the different question of whether the result drifts in time. Split at the midpoint, the two halves agree at every delay, 77% against 67% at eleven weeks, with overlapping intervals at all four. A rolling window tells a less settled story: $R(5)$ runs from 22% in the window ending April 2019 to 91% in the window ending April 2015, and a line through those

points falls about 22 points a decade at five days and 12 at eleven weeks. Those windows overlap by 750 of 1,000 days, so the path is descriptive and no standard error is quoted for it.

The obvious explanation is composition, since the first half holds 2008, 2011 and 2015 and Section S10 has already shown the rate is higher in stress. Holding stress fixed settles it: comparing the halves within a VIX tercile, not one of twelve cells differs significantly. The drift in the path cannot be attributed to a change in how the cross-section works, and on this sample it cannot be attributed to anything.

Conditioning does something else, though, and it is the reason Section 5 quotes the rate by state.

Table S4. The substitution rate by market state, pooled over the whole window, VIX terciles taken at $t-1$. Same whole-ratio block bootstrap as Table 2. A calm-market rate is imprecise by construction: there is little delay damage in a calm market, so the ratio's denominator is near zero.

δ	calm	middle	stressed
5	42% [-4, 102]	32% [8, 51]	64% [49, 81]
13	43% [-6, 100]	47% [29, 61]	74% [64, 84]
21	41% [-10, 83]	53% [36, 66]	77% [68, 85]
55	47% [6, 84]	53% [37, 67]	81% [74, 87]

The stressed column is the one this paper can actually report: 81% at eleven weeks with an interval of [74, 87], and 81% in each half of the sample taken separately. Before leaning on that, one objection has to be cleared. $R(\delta)$ divides by the damage the delay does, and stressed days are by definition the days on which being late costs most, so stress is where the denominator is largest, and therefore where a ratio with that denominator underneath is best behaved. A high, tight rate in stress could mean the cross-section does more work there, or merely that the ratio is measurable there. The first is a claim about markets; the second is true of any ratio and says nothing.

Taking the ratio apart settles it. Write $N(\delta)$ for the mean squared error the cross-section removes and $D(\delta)$ for the mean squared error the delay adds, both on the same days and in the target's own units, so that $R = N / D$.

Table S5. The substitution rate taken apart, by market state. N is the error the cross-section removes, D the error the delay adds, both mean squared error on the log-variance target. If only D moved with state, the conditional result would be arithmetic, not a finding.

δ	N calm	N stressed	N ratio	D calm	D stressed	D ratio
5	0.0284	0.1567	5.5×	0.0673	0.2434	3.6×
13	0.0345	0.3787	11.0×	0.0798	0.5110	6.4×
21	0.0412	0.5224	12.7×	0.1003	0.6755	6.7×
55	0.0583	0.8043	13.8×	0.1234	0.9927	8.0×

Both move, and the numerator moves further: averaged over delays, stress multiplies N by 10.7 and D by 6.2. The cross-section removes more error in absolute terms when markets are frightened, not merely a larger share of a larger loss, so the conditional result is a fact about markets. At eleven weeks it removes 0.8043 of squared error in stressed markets against 0.0583 in calm ones.

So the conditional reading stands, and on absolute error as well as on the ratio. The calm column is not a small number, it is an imprecise one: it spans most of the unit interval and at three of the four delays it does not exclude zero. We suspected worse. A ratio whose denominator can approach zero need not have a bounded confidence set at all, and a percentile bootstrap cannot express that, so the Fieller construction of Section 4.6 of the paper was applied to all twelve cells. Every one comes back an ordinary interval, and within a point or two of the bootstrap's: the calm cell at five days is $[-4, 104]$ against the bootstrap's $[-4, 102]$, and the stressed cell at eleven weeks is $[74, 87]$ against the bootstrap's $[74, 87]$, which agree to the digit. The denominator never approaches zero closely enough to misbehave, so the printed intervals stand and the calm rate is imprecise without being unidentified. That is not a failure of the estimator but a property of the question: where delay costs a forecaster almost nothing, the share of that cost which the cross-section repairs has no stable value. The calm-market rate is weakly measured and economically uninformative in this sample, but the Fieller analysis does not classify it as unidentified: every calm-state confidence set it returns is bounded, if very wide. An unconditional rate therefore averages a well-measured result with a poorly measured one, and Section 5 says so instead of quoting the average alone.

One more thing can be asked of the calm cell before it is set aside, and it should be. The paper has chosen a coordinate that fails in a third of its sample and reported the failure as a fact about markets, and a reader is entitled to suspect the coordinate instead. The effective age of Section S12 is the alternative and it is built differently: it asks how old a plain mark would have to be to forecast as badly, so it has no denominator that can vanish, and taken in mean squared error rather than in skill it carries no benchmark variance either, which is the other quantity that goes small in a calm market.

Asked that way the calm cell still does not answer. The age is 4.5 days with an interval of $[0.3, 44.1]$, which is 7.8 times as wide relative to its own point as the stressed cell's, and 47 of 1,500 replications cannot place it on the grid at all. The non-result therefore survives the change of instrument, and that is worth more to a reader than a number would have been: where delay costs a forecaster almost nothing there is almost nothing for the cross-section to repair, and no coordinate makes a share of nothing precise. The limitation belongs to the question and not to the ratio.

The same construction pays at the other end of the ladder, where the quantity is well determined and the unit is the one a practitioner uses. In the top tercile of VIX days a fifty-five-day-old mark plus today's foreign closes forecasts like a mark 3.8 days old $[1.9, 6.6]$; in the top decile, 3.4 $[1.3, 7.2]$; in the top 5% of days, 2.3 $[0.8, 5.0]$. The ladder is monotone in days as well as in shares. On the days when a stale mark is most dangerous, eleven weeks of staleness plus a current cross-section is worth a mark two days old.

Conditioning on the VIX raises a fair objection, since Section 7 says that a forecaster who holds the VIX should ignore the cross-section altogether. The two are compatible, but only for a specific reader, and it is worth being exact about who. The VIX is a public index on the S&P; a holder of an illiquid asset can see it whether or not any option trades on the thing they own. Such a reader can therefore condition on market state without being able to buy the substitute that state variable makes redundant. A reader who holds options on their own asset is in Section 7's case and should stop reading at Section 8. We also note that Section S10 works with a ladder of cuts at the 33rd, 90th and 95th percentiles while this section uses terciles, and that the tercile split was chosen after that ladder had been seen.

S6. Multiple-testing adjustment · *Result survives*

Forty-nine scripts stand behind this paper and nine distinct forecasters have been fitted to this target across them. A single test statistic on the one that won is not the right instrument for that, so the whole set is put through White's Reality Check and Hansen's SPA, with the size of both first validated by simulation on this sample's own dependence structure, where a nominal 5% test rejects 5.5% and 7.0% of the time respectively, with SPA rejecting more often, as Hansen predicts. Against the domestic benchmark, the best of the nine survives both tests at $\delta = 3, 5, 13, 21$ and 55 with p below 0.0005. It fails at $\delta = 0$, where the best margin across all nine models is +0.0014 of R^2 : nothing beats the benchmark there at all, so that failure is an absence of any effect to correct.

S7. Leakage from an overlapping target · *Result survives*

The target is realised variance over the next five days, so consecutive observations share four days of outcome, and a naive split would put part of the answer in the training set. Every refit here already cuts the training window at $t - \delta - h$, which removes exactly the rows whose outcome was unresolved at the forecast date: a purge, though the paper has not called it one. Removing it inflates R^2 by +0.0024 [+0.00062, +0.00387] at zero delay and +0.0021 [+0.00017, +0.00382] at three days, so it is load-bearing and any study of a five-day target that splits without one is reporting inflated numbers. Adding an embargo on top of it costs nothing detectable in any of eighteen cells, so the gap the design already leaves is sufficient.

S8. Five markets, at every delay · *Result survives*

The five-market comparison is below, across the whole delay curve rather than at one point, and it is where the timing mechanism shows most plainly.

Table S6. Substitution rate by target market at four delays, continuous target, R^2 scale. Same-day peers are the markets closing earlier on the same calendar day; every other peer contributes the previous session. Intervals bootstrap the whole ratio.

target	close, UTC	same-day peers	R(5)	R(13)	R(21)	R(55)
S&P 500	21:00	6	51% [35, 66]	64% [51, 74]	67% [55, 78]	72% [60, 81]
FTSE 100	16:30	4	74% [53, 91]	82% [68, 93]	84% [71, 94]	88% [77, 96]
DAX	16:30	4	55% [37, 70]	69% [55, 79]	71% [58, 80]	74% [65, 82]
Hang Seng	08:00	2	61% [32, 88]	68% [46, 90]	69% [50, 88]	69% [55, 84]
Nikkei 225	06:00	0	44% [5, 74]	54% [23, 75]	55% [24, 75]	60% [36, 76]

The ordering by same-day peers is not perfect, since the FTSE has four and recovers more than the S&P with six, but the two ends are where the mechanism predicts. The Nikkei closes before every other market in the sample, has no same-day foreign observation at all, and recovers least at every delay; the FTSE and DAX, which see all of Asia before they close, recover most. The curves are also flatter than the S&P's: for markets with early peers the rate is already near its eleven-week value within a week.

S9. Where each result holds

The same exercise with each of five indices as the target is in Table S6, on the continuous target and at every delay.

The effect appears in all five, and the Nikkei is the case that carries information. It closes before every other market in the sample, so it has no same-day foreign observation at all, since every peer reaches it a session late, and it recovers the least, 60% against 69% to 88% elsewhere. That is what the mechanism predicts and what a spurious correlation would not produce. Past that contrast we can say little: the S&P has six same-day peers and recovers less than the FTSE with four, the intervals overlap heavily, and five markets cannot resolve an ordering. Five markets also cannot show where the effect stops, and it does stop: Section S14 adds four targets outside this group on two of which the cross-section is worth less than nothing, and identifies what separates them from these five.

One further objection deserves an answer. Rapach, Strauss and Zhou find that lagged US returns predict foreign returns while the reverse is weak (2013). That concerns returns while ours concerns volatility, and lagged returns where ours are same-day closes observed before the target's own close. We rely on both distinctions, and a reader who rejects them should discount Sections 4 and 5 accordingly.

The scope condition matters more than the ordering. Section 7 shows that where a liquid options market exists on the asset, breadth is redundant, and the S&P 500 is the most heavily optioned index in the world. So the breadth result may be most relevant to assets without such a market, the case Section 1 opens with: a private portfolio marked quarterly, a property book on an appraisal cycle, a fund reporting NAV with a lag. Nobody writes a liquid option on those. What a holder of them can see is the public cross-section, and Sections 4 and 5 say what that is worth as a function of how stale the mark has become.

We would rather state that plainly than dress it up, because it is also the weakest point of the design. We ran the experiment on a liquid asset in order to observe the truth, and the property that makes the truth observable, a deep and continuously traded market, is the same property that brings an options market with it. The result we can measure cleanly and the case we care about are therefore separated by the one feature that decides which substitute wins. Section 11 says what would close that gap.

There remains the $\delta = 0$ row of Table 2, and its history is worth recording because the reason for the caution changed. On log-scale squared error the cross-section is worse by 0.0099, interval $[-0.0191, -0.0011]$; on QLIKE it is worse by 0.0127, $[-0.0263, -0.0010]$; on squared error on the natural variance scale it is worse by 0.0037, $[-0.0310, +0.0180]$. All three point the same way and two of the three exclude zero. That is not what this section said until Section 4.4's reconstruction was corrected. On the median-normalised scale the same comparison comes out at +0.0070, $[-0.0429, +0.0204]$, and this section used to read that as the sign depending on the scale. It does not. The normalised loss weights each day's squared error by that day's trailing median, so it is a weighted loss that had been described as an unweighted one; the unweighted natural-scale loss is the third figure above, and it agrees in sign with the other two. The flip was an artefact of a weighted loss described as an unweighted one, and the claim that a monotone transformation of the target moved the sign was wrong. What remains is smaller but still a reason for caution: the effect at zero delay is about one point of R^2 , and the robust loss that most resembles the headline cannot distinguish it from zero. An earlier draft read the negative as settled and used it to press a disagreement with Korkusuz (2025) and Jayawardena et al. (2020), who find foreign volatility improves a current forecast. We still decline to press it, on the narrower ground that a one-point effect at zero delay on a single target is not the basis for a disagreement, rather than on the ground that we cannot tell its sign. Table 3's

decomposition still holds conditional on the log scale: whatever is there is small, bounded near one point of R^2 , and mostly the bill for estimating seven coefficients.

S10. In stress and in calm · *Effect larger in stress, unresolvable in calm*

Everything above is an average over 4,862 days, and one objection is that the interesting days are not average: a rolling median assumes a stationary volatility distribution, and across a structural macro transition it should fail to adjust. The objection is testable here because the test window contains the 2008 crisis, 2011, August 2015, February 2018, COVID and the 2022 inflation shock. It is not testable on the companion note, whose window begins in November 2023 and contains no transition at all, and the note says so instead of answering it.

Two decisions shape what follows. The first is that a single split is not enough. Six named episodes at full width run to 866 days, most of them ordinary: "the 2022 inflation shock" as the whole of 2022 is 251 days. A wide window dilutes exactly what is being tested. So the definition is tightened in rungs: VIX at $t-1$ above its own trailing 67th, 90th and 95th percentile, which needs no dates and is available to a forecaster in real time, together with the named episodes at full width and as the 41 trading days centred on each window's VIX peak. The tightening was added after the wide version had been run, and we say so instead of presenting it as the original design.

The second decision matters more. Squared errors are larger when the target is more volatile, so a larger loss differential in stress can be nothing but heteroskedasticity. Every figure below is therefore the differential divided by the domestic model's own mean squared error *inside the same regime*: a share of error removed, comparable across states in a way the raw difference is not. The raw column is reported beside it, and the two tell different stories, and that is the reason for the precaution.

Table S7. What the cross-section removes from the domestic model's error, by market state, at $\delta = 55$. Rungs are nested: the top 5% of VIX days is a subset of the top 10%. Intervals are moving-block bootstraps of the ratio, taken on each regime's own time-ordered subsequence.

state	test days	share of domestic MSE removed	95% interval
VIX top 5%	367	63.1%	[46.2, 73.1]
VIX top 10%	644	57.0%	[40.7, 68.7]
VIX top 33%	1,729	43.3%	[27.4, 55.9]
episodes, peak ± 20 days	246	63.1%	[40.2, 77.5]
VIX bottom 67%	3,074	29.9%	[21.7, 38.2]
outside every episode	3,996	23.9%	[15.7, 31.4]

The objection is answered in the opposite direction. The contribution does not weaken in stress; it roughly doubles, and it rises monotonically as the definition tightens, running 43.3%, 57.0% and 63.1% along the VIX ladder against 29.9% on calm days. The severest rung and the calmest have disjoint intervals. The calendar route, built from dates instead of percentiles, lands on the same 63.1% at its narrow window, which is a useful agreement because the two cuts share no construction. A dose-response along a mechanical ladder is stronger evidence than any single split would have been, and we would not have had it from the version of this test we ran first.

The economic reading is the one a risk manager would want, and it is not the reading Section S28 gave. Breadth is worth most precisely when the domestic mark is stale *and* the market is moving,

which is when a stale mark is most dangerous. That is a stronger claim than the paper made before, and it belongs to the cross-section itself and to no refinement of it.

S11. Implied volatility in a stressed options market · *No effect detected*

The mirror objection to Section 7 is that redundancy given VIX and VDAX assumes a well-functioning options surface, and that during a crunch breadth might stop being redundant exactly when it matters. Applying the same ladder to the increment the foreign block adds on top of the implied-volatility model, across four delays and six states, the foreign block adds significantly in none of twenty-four cells, on the stressed rungs and the calm ones alike, down to the top 5% of VIX days. Redundancy given implied volatility is not an average concealing a stressed subsample. The severest rung carries 324 days, so an effect confined to conditions more extreme than the ladder reaches would not appear here.

S12. A stale cross-section · *Claim narrowed*

Every number to this point was computed with the foreign block at its freshest admissible value: for each peer, the most recent close that precedes the target's. That is a deliberate choice about admissibility and it has been doing more work than the paper admits. A desk holding a stale mark is, almost by definition, a desk with a data problem, and the infrastructure that delays a domestic mark by δ days is not obviously the kind that delivers seven foreign closes the same evening. Nothing above says what such a desk gets. A recommendation that assumes a same-day feed without saying so is worth less than it appears, so the assumption is removed here instead of defended.

Let g be the number of days by which the foreign feed is itself late, and repeat the walk-forward of Section 5 at every pair (δ, g) . The domestic block is held δ days stale as before; the peers are held g days stale on top of admissibility. The column $g = 0$ is the paper.

Table S8. Substitution rate $R(\delta)$ when the foreign block is itself g days stale. Same walk-forward, same losses and same test days as Table 2; $g = 0$ reproduces it. A negative rate means the model carrying the aged cross-section forecasts worse than the stale domestic block alone. Undefined at $\delta = 0$, where there is no delay-induced loss to recover.

δ	$g = 0$	$g = 1$	$g = 2$	$g = 3$	$g = 5$	$g = 10$
3	36%	9%	−8%	−15%	−14%	−10%
5	51%	29%	13%	−0%	−11%	−8%
13	64%	52%	42%	34%	22%	−1%
21	68%	57%	50%	44%	33%	16%
55	71%	64%	58%	53%	46%	30%

The first column is Table 2's and the rest are not near it. Ageing the foreign feed reduces skill significantly in 25 of the 30 cells tested, and in 5 of the 6 delays a single day is already enough. At $\delta = 5$ the rate falls from 51% to 29% with one day and to zero with three. At $\delta = 3$ the cross-section becomes actively harmful from $g = 2$ onwards, the honest description of a negative rate: the delayed desk would have done better ignoring it.

Shares of a loss are a poor way to deliver that, so here it is in the unit this paper uses everywhere else: the age of a domestic mark. The own-only delay curve is re-estimated on a fine grid, and each combined model is quoted as the age of the plain stale mark that forecasts exactly as well.

Table S9. Effective age. Each cell is the age of a plain domestic mark whose own-only forecast matches the model that has a δ -day-old mark plus a foreign block g days stale, by interpolation on the own-only delay curve over sixteen ages from 0 to 55 days.

δ	$g = 0$	$g = 1$	$g = 2$	$g = 3$	$g = 5$	$g = 10$
3	1.7d	2.7d	3.3d	3.7d	3.6d	3.4d
5	1.9d	3.0d	4.0d	5.0d	5.9d	5.6d
13	2.8d	4.2d	5.5d	6.8d	8.9d	13.2d
21	3.5d	5.2d	6.7d	8.1d	10.5d	15.0d
55	4.6d	6.8d	8.4d	10.0d	12.6d	19.4d

The last row states this paper's headline in one number: a fifty-five-day-old mark plus today's foreign closes forecasts like a mark 4.6 days old. The same mark with yesterday's foreign closes forecasts like one 6.8 days old, and with the feed a week behind, 12.6 days. Because this is the sentence the paper will be quoted on, it carries an interval like every other headline here, and constructing one takes some care: the map from skill to age is the inverse of a delay curve that is steep near zero and flat far out, so a fixed error in R^2 becomes a small error in age at one end and a large one at the other, and the interval cannot be obtained by pushing an R^2 interval through the curve. Resampling test days in moving blocks and re-interpolating inside each replication, recomputing the curve, the combined model and the benchmark variance on the same resampled days, gives 4.6 days [2.5, 8.4]. The one-day and week-behind feeds give 6.8 [4.0, 11.1] and 12.6 [9.5, 19.6]. No replication placed the model at a fresh mark or off the end of the grid, so the interpolation is not being asked to extrapolate.

The same construction prices the feed itself. Differencing the ages *within* each replication, a single day of foreign staleness costs 2.1 days of domestic freshness at eleven weeks, interval [1.4, 3.3], and five days cost 8.0 [5.9, 12.6]. Both exclude zero. The marginal intervals in the previous paragraph overlap, as marginal intervals ordinarily do and is not evidence against a paired difference; the paired construction is the one that answers the question, and it agrees with the paired skill test reported above. Fitting a line along each row of Table S9 gives the same exchange rate averaged over the whole row instead of at its first step. It runs 0.13, 0.35, 1.03, 1.13 and 1.44 days at $\delta = 3, 5, 13, 21$ and 55, rising at every step. The shape is the point. Where the mark is nearly fresh the cross-section is decoration and ageing it costs almost nothing; where the mark is old the cross-section is doing the work, and then a day of the foreign feed costs about a day of the domestic one. The readers with most to gain from Sections 4 and 5 are exactly the ones whose foreign data must be current.

One reading of that decay has to be ruled out before the rest is safe. If the foreign block worked by propagation, with volatility travelling from Tokyo to New York overnight, its value would be its earliness, the decay above would be the wave passing, and the right response would be a richer dynamic model, not a warning about feeds. The cross-correlations appear to speak to this and they are a trap. Correlating each peer at $t-1$ against the target at t , then the target at $t-1$ against the peer at t , all seven peers appear to LAG the target. But a row of this panel is not an instant. The target closes at 21:00 UTC and a peer closes c hours earlier, so the first direction spans $24 + c$ hours and the second $24 - c$; for a persistent series the shorter span gives the higher correlation, and the clock alone predicts the observed sign whether or not anything leads anything.

That prediction is quantitative, so it can be tested instead of conceded. Each peer's asymmetry should scale with $2c$. Across the seven it does, with a correlation of +0.93: 0.013 and 0.019 for the FTSE and

the DAX, four and a half hours ahead of New York, against 0.051 for Bovespa, a full day ahead once admissibility has shifted it. The apparent lead-lag structure is a property of the calendar, not of the markets, and it is evidence for neither story. A check from the other side agrees: replacing the seven peers with the single real-time global factor of Section 6.1 and entering it at one lag at a time, the contemporaneous factor beats every lagged version in 26 of 30 cells. Yesterday's factor does add to today's at $\delta = 3$ and $\delta = 5$, and at no longer delay, so a little genuine dynamics is present, inside an otherwise contemporaneous story.

What is left is the reading Section S27 already arrived at by a different route. The block is a second reading of one shared state, it loses value as it ages at about the rate the first reading does, and it stops contributing once it is older than the mark it was brought in to replace: at $\delta = 13$ a foreign block ten days stale returns the model to an effective age of 13.2 days, the domestic mark on its own. Breadth substitutes for a stale mark by being fresh, not by being early.

S13. An options market on another asset · *Scope condition supported in the examined markets*

Section 7 is the reason this paper's scope is narrower than its subject, and the narrowing turns on a word the section never examines. The implied-volatility block is VIX and VDAX. For an S&P target VIX is an options market written on the asset itself; for a DAX target VDAX is. For a Nikkei, Hang Seng, Nifty, ASX or Bovespa target both series belong to somebody else's index. Those are different purchases and the paper has been treating them as one.

The second case is also much closer to the case Section 1 opens with. A property book cannot buy an option on the building, because nobody writes one, and a quarterly-marked fund cannot buy an option on its own net asset value. Both can see the VIX every day for nothing, and Section S5 already leans on exactly that availability when it conditions on market state. So the question that decides how far Section 7 reaches is whether redundancy requires the option chain to be on your own asset, or whether anyone's options market will do.

It is answerable with the data already here. Each of the eight indices becomes the target in turn, with its own admissible cross-section under the clock rule of Section 4 and the same four models at each delay. Both implied-volatility series enter at $t - 1$ for every target, which is strictly admissible everywhere and identical across targets, so nothing differs between the two groups except whose options market the block contains.

Table S10. The two candidates raced against each other on every target, at $\delta = 55$. R(breadth) and R(IV) are the two substitution rates, each computed inside the same domestic control. The last two columns are what the seven foreign closes still add once implied volatility is held, and the Giacomini-White statistic testing it. Those two columns are the *seven-regressor* specification, not the compressed one the section's counts use, and the distinction matters at the DAX row: -0.0113 with $z = -2.06$ clears the pointwise 5% threshold, but it is a significantly NEGATIVE increment, which is breadth costing something rather than adding it. Three of the eight own-options cells are significantly negative in this specification, at SPX $\delta = 0$, DAX $\delta = 0$ and DAX $\delta = 55$, and none is significantly positive at any delay. Under the compressed block none of the eight reaches significance in either direction, and it is that count the text below quotes against the twenty-four. The correlation column is between the target's own decision variable and the VIX block's.

target	options market	corr. with VIX	R(IV)	R(breadth)	foreign given IV	GW z
S&P 500	its own, VIX	0.853	102.6%	71.6%	-0.0052	-0.87
DAX	its own, VDAX	0.779	104.0%	74.0%	-0.0113	-2.06
FTSE 100	another index's	0.754	85.3%	88.0%	+0.0186	+1.71
ASX 200	another index's	0.715	72.3%	69.1%	+0.0241	+1.60
Hang Seng	another index's	0.665	77.6%	69.0%	+0.0039	+0.28
Bovespa	another index's	0.659	67.8%	50.0%	-0.0201	-1.81
Nikkei 225	another index's	0.589	51.9%	59.7%	+0.0034	+0.21
Nifty 50	another index's	0.568	53.9%	53.2%	+0.0393	+1.85

The two groups do not behave alike. Where the options market is the asset's own, implied volatility does not merely win, it wins by thirty-one points: 103% of the delay damage repaired against 73% for the cross-section, and a rate above 100% is this paper's own signal from Section 4 that the substitute is supplying something the domestic history never held. The foreign block adds significantly in none of the eight cells there, which is Section 7's redundancy result reproduced on a second target with a second exchange's index.

Where the options market belongs to another index the gap collapses to three points, 68% against 65%, and breadth is ahead outright on two of the six. The redundancy goes with it, though by less than an earlier version of this section claimed. With the HAC bandwidth set from the measured dependence rather than from the automatic rule, the foreign block adds significantly in 1 of 24 cells, and that cell is the Nifty at $\delta = 21$, at +2.35. At eleven weeks the three largest statistics are the Nifty's +1.85, the FTSE's +1.71 and the ASX's +1.60, all of which point the same way and none of which clears the threshold. This section previously reported those three as significant, at +2.36, +2.01 and +2.13; they were divided by a standard error that truncated its kernel inside the dependence, and the claim does not survive the correction. What survives is a direction reproduced on five of six targets and not a count of significant cells, and the compressed version below is where the evidence actually sits.

One correction belongs here before those counts are read, and it changes them. Every increment above is measured with the seven-regressor block, which Sections 6.1 and S28 both say is not the model to run: seven coefficients cost one to five points of R^2 , and on the S&P that cost was the whole of the negative increment given implied volatility. The bill is not the same size on every target, since it grows with delay and the eight targets differ in sample length and signal strength, so a cross-target comparison of net increments is partly a comparison of estimation costs. Giving every target the

compressed block instead, the equal-weighted mean of its own admissible peers at a cost of one coefficient, removes that confound.

The split does not merely survive the correction, it sharpens. Where the options market is the asset's own, the mean increment rises from -0.0078 to $+0.0029$ and the count of significant cells stays at 0 of 8: compression removes the apparent harm and reveals nothing underneath it, which is Section S28's result on a second target. Where it belongs to another index, the mean increment rises from $+0.0046$ to $+0.0153$ and the count goes from 1 of 24 to **13 of 24**, with a mean Giacomini-White statistic of 1.54. Over half of the cells in which a forecaster can see an options market but cannot buy one on what they hold are cells where the cross-section still pays, and the thirteenfold jump in that count from removing six coefficients is the section's result: what looked like redundancy on this group was mostly an estimation bill.

The gradient behind the split sharpens with it. Correlating each target's decision variable with the VIX block measures how close the only options market a forecaster can see sits to the thing they hold, and against the compressed increments that closeness correlates -0.78 across the eight targets, where the seven-regressor version read -0.38 . Eight points is not a regression and no interval is quoted for it, but it is no longer a relationship a reader has to squint at.

The correction also disposes of the one target that did not fit. Bovespa looked like a market holding its own options chain, with the cross-section adding nothing and doing so significantly. It was paying for six coefficients it did not need: -0.0144 with seven regressors and $+0.0098$ with one. Two other explanations were available and both are ruled out rather than left open. The clock is not the cause, because entering the S&P same-day for a Bovespa forecaster, which this paper's own admissibility rule forbids and which is run only as a diagnostic, leaves the increment negative at all three delays. Nor is spanning: the partial correlation of the peer factor with the target given VIX, which is the quantity theory would reach for first, correlates -0.15 with the increment and predicts nothing. The lesson generalises past this table. A comparison of net increments across targets is partly a comparison of estimation bills, and on one target in eight that was enough to reverse the sign of the conclusion.

This does not reach the asset the paper is about, and one thing about the lower half has to be said before it is read as more than it is. Options trade on every target in it. Japan publishes the Nikkei Volatility Index, Hong Kong the VHSI, India the India VIX, Australia an S&P/ASX 200 VIX and Brazil an IVol-BR. What puts those targets in the lower half is that this paper holds VIX and VDAX and not those series, so the split is about the forecaster's information set and not about which option chains exist. That information set is the right analogy for Section 1's reader, who can see a public volatility index and cannot buy one written on what they hold, and it is not the same claim as an asset with no options at all. Section S14 supplies targets where the absence is a property of the market, and Section 11 keeps the gap rather than claiming it is closed.

S14. Markets with no volatility index · *Scope condition supported in the examined markets*

The split above rests on what a forecaster holds rather than on what exists, so the obvious next question is whether it survives on a target where nothing exists to hold. Four were added for which we could find no published implied-volatility index: the KSE-100 in Pakistan, the IDX Composite in Indonesia, the Colombo All-Share in Sri Lanka and the FTSE Bursa Malaysia KLCI. That is a statement about what we could find, and a reader who knows of such an index for one of these

markets should treat that row accordingly. The series are single daily files from one vendor, used in this subsection and nowhere else; no other number in this paper depends on them.

All four close between 09:00 and 10:30 UTC, so each sees Tokyo, Sydney and Hong Kong on the same day and Europe and New York a session late. That clock handicap has to be held fixed rather than compared away, so the four Asian targets of Section S8 are re-run through the same code, the same delays and the same implied-volatility timing. The comparison is then between markets that differ in whether anyone publishes a volatility index on them.

Table S11. Eight targets with a foreign implied-volatility block, sorted by coupling: the correlation between the target's own decision variable and the equal-weighted mean of the cross-section it is offered, on its own test days. $R(\text{breadth})$ and $R(\text{IV})$ are at $\delta = 55$. The last two columns are what a compressed cross-section adds once implied volatility is held, and its Giacomini-White statistic.

target	volatility index	coupling	$R(\text{breadth})$	$R(\text{IV})$	breadth given IV	GW z
ASX 200	exists, not held	0.808	68.8%	73.0%	+0.0268	+2.70
Hang Seng	exists, not held	0.731	68.6%	77.0%	+0.0195	+2.21
Nikkei 225	exists, not held	0.701	58.8%	52.3%	+0.0311	+2.62
Nifty 50	exists, not held	0.661	53.1%	53.9%	+0.0387	+2.49
IDX Composite	none published	0.648	41.8%	38.6%	+0.0452	+2.44
FTSE Bursa KLCI	none published	0.611	44.0%	56.1%	+0.0142	+1.61
Colombo All-Share	none published	0.228	−11.4%	−7.6%	−0.0115	−2.11
KSE-100	none published	−0.115	−29.7%	−25.4%	−0.0167	−1.81

The question is answered on two of the four. Indonesia and Malaysia behave like the matched group: breadth repairs 41.8% and 44.0% of the delay damage at eleven weeks, and a compressed cross-section still adds to implied volatility there, with statistics of +2.44 and +1.61. Only the first of those clears the threshold; the Malaysian one falls short of the threshold and is reported as a direction. Both were above it under the nine-lag kernel, at +3.52 and +2.00, and neither is above it by much at the measured bandwidth, so the pair is worth no more than what two targets can carry. No options market appears anywhere in that picture, which is the part that does not depend on a statistic: Section S13's reading does not turn on which volatility series we happened to buy.

The other two are not a weaker version of the same result. Pakistan and Sri Lanka return rates of −29.7% and −11.4%: a cross-section that makes a delayed forecaster worse, significantly so under both parameterisations. Averaging the four would combine a result with its opposite, which is the error Section 5 warns about in a different place, so no group mean is quoted here.

What separates them is not the options market and it is visible before any forecast is run. Sorted by coupling, the table is almost monotone in the substitution rate, and across the eight targets coupling and $R(\text{breadth})$ correlate +0.98. Pakistan's coupling is *negative* at −0.115: its volatility moves very slightly against the cross-section it is being offered, and the cross-section duly makes its forecasts worse. Eight points are eight points and no interval is quoted for that correlation, but the ordering is not subtle and the mechanism behind it is the paper's own. Section S27 finds the entire recovery in the leading component of the foreign block, so a market that does not load on that component has nothing to recover with.

That converts a refusal into a rule. Section 10 declines to extrapolate the rate to a private asset, on the grounds that nothing here bounds that asset's correlation with the public cross-section. These four targets are what the refusal looks like when it is measured: coupling is the part of the story this design can see, it varies from 0.81 to -0.12 across markets that all look like ordinary equity indices, and it is computable by anyone holding a stale series and a public cross-section before they buy anything. A holder whose asset is weakly coupled should expect breadth to do nothing for them, and this is the first evidence in the paper that says so with a number.

S15. Clean lags against periodic marking · *Approximate equivalence*

Section 10 will argue that what carries from this paper to an illiquid asset is the apparatus rather than the numbers. One part of that claim can be checked rather than hoped for, and this section checks it. Every result here models staleness as a clean lag: the forecaster sees the state exactly as it stood δ days ago and nothing since. A marked asset is not stale in that way. It is revised on a schedule and held flat in between, and the revision is a partial adjustment toward current evidence, not a full mark, which is how the property literature has modelled appraisal indices for thirty years (Geltner et al. 2003). If the substitution rate under that kind of staleness looked nothing like the rate under a clean lag, $R(\delta)$ would be a curve about clean lags alone.

It does not. Rebuilding the domestic block as a mark revised every q days and held between revisions, and matching on mean staleness, gives 64% against the clean lag's 60% at ten days, 71% against 70% at thirty-one, and 70% against 72% at ninety-four. Adding partial adjustment on top moves it a few points the other way, 63% where matching predicts 70%, so the equivalence is approximate and we claim no more than that. One thing we got wrong on the way is worth recording: our first attempt modelled appraisal as pure exponential smoothing, which is not staleness at all. A weighted average of the recent past still contains today's observation, so it costs almost nothing. R^2 falls from 0.499 to 0.456 across a mean lag of eleven weeks, against 0.499 to -0.035 under a clean lag, and leaves almost no damage for breadth to repair. Holding a number flat between revisions is what makes a mark stale, not blurring it.

S16. The implementable target at every delay · *Result survives*

One detail of that definition needs stating, because it looks like look-ahead and is not. The target a_{t+5} carries M_{t+5} , a median whose window includes the five days between the forecast origin and the outcome, which a forecaster at t does not know. No feature is affected, since every one is dated $t - \delta$ or earlier for the domestic block and t for the others. What changes is that the quantity being predicted is not one a practitioner could convert back to a variance forecast at the moment they make it. The implementable alternative normalises the outcome at the ORIGIN, $\log(\sigma^2_{YZ,t+5} / M_t)$, and we ran the whole experiment both ways. The two targets differ by the five-day drift in a 252-day median, which has a standard deviation 3.7% of the target's own and correlates $+0.52$ with it, so it is neither large nor independent. It changes almost nothing: the substitution rate is 71% [59, 81] against 72% [60, 81] at eleven weeks, and no delay differs by more than a point. Because the difference is that small, the paper resolves the question in the practitioner's favour and quotes the origin-dated rate wherever it states a headline. The tables keep the target-dated definition, for a reason worth being explicit about: every conditional, robustness and scope result in Section 9 and in the Internet Appendix is computed inside that definition, and recomputing all of them to move one number by a point would change no conclusion in the paper. Table S12 of the Internet Appendix gives both

columns at every delay, so a reader who wants one definition throughout can read the paper in either.

Section 4.2 quotes the origin-dated rate in every headline and computes every table on the target-dated one, on the grounds that the two differ by at most a point. That is the claim, and this is the evidence for it. The two runs share features, estimator and test days exactly; only the date at which the outcome's denominator is fixed changes. The drift between them, $\log(M_{t+5} / M_t)$, has a standard deviation 3.7% of the target's own and correlates +0.52 with it, so it is neither large nor independent of what is being predicted, and the check is not vacuous.

Table S12. The substitution rate under both definitions of the target. Target-dated normalises the outcome at $t + h$, which is the definition every table in the paper uses; origin-dated normalises it at t , which is the definition a forecaster could act on at the moment of forecasting. Intervals bootstrap the whole ratio. The last column is the gap in percentage points.

δ	$R(\delta)$ target-dated	$R(\delta)$ origin-dated	gap
3	36% [17, 55]	35% [15, 54]	−1
5	51% [36, 66]	51% [33, 66]	0
13	64% [52, 73]	63% [50, 74]	−1
21	67% [57, 76]	67% [57, 76]	0
55	72% [60, 81]	71% [59, 81]	−1

The increments behind those ratios move the same way and by as little: the cross-section is worth +0.3821 of R^2 at eleven weeks under the target-dated outcome and +0.3848 under the origin-dated one, on own-only skill of −0.0351 and −0.0328. The origin-dated arm is very slightly the better of the two, which is the opposite of the direction a reader suspicious of the target-dated definition would fear.

S17. How much of the coupling correlation is one point · *Ordering survives, fitted line does not*

Section 6 reports that a target's coupling to the cross-section tracks its substitution rate at +0.98 across eight markets, and Figure 3 draws a fitted line through those eight points. Two different claims live in that sentence. One is an ordering claim, which a rank statistic answers. The other is a claim about a line, which is what the figure's zero-crossing is read off. Eight points with two of them far out to the left is exactly the shape in which a correlation can be a statement about the gap between two clusters rather than about a relationship, so both claims are diagnosed here. Nothing below is a new experiment: every number comes from the eight pairs already in Table S11.

The correlation is +0.979, with a Fisher interval of [+0.884, +0.996] on eight points, and the Spearman rank correlation, which is what the ordering claim actually needs, is +0.976. Leaving each target out in turn, the Pearson correlation never falls below +0.976 and the Spearman never below +0.964. The hardest version of the question is to delete both low-coupled targets and fit on the six that remain, which is a single cluster with no leverage from the left-hand tail at all: there the correlation is +0.903 and the Spearman +0.943. The ordering survives every form of the objection we could construct.

The fitted line does not, and the difference matters because the figure draws one. The KSE-100 sits at a coupling of -0.115 with every other target above $+0.22$, which gives it a leverage of 0.734 against a mean of 0.250 across eight points. Removing it moves the least-squares slope from $+1.154$ to $+1.455$ and the zero-crossing from $+0.216$ to $+0.309$, a shift of nine points of coupling in the quantity a practitioner would be most tempted to use as a threshold. Section 6 already declines to call that crossing a threshold, on the separate ground that the count of targets below it differs from the count with a negative rate; the diagnostic here is why that refusal is the right one rather than a hedge.

Table S13. Leave-one-out diagnostics for the coupling relation across eight target markets. 'Slope' is the least-squares slope of $R(\text{breadth})$ at eleven weeks on coupling with that target deleted; 'crossing' is where that line reaches zero; 'leverage' is the diagonal of the hat matrix on the full eight, whose mean is 0.250 . The last row is the full sample.

dropped	Pearson	Spearman	slope	crossing	leverage
AXJO	+0.976	+0.964	+1.152	+0.216	0.233
HSI	+0.982	+0.964	+1.123	+0.219	0.181
N225	+0.978	+0.964	+1.146	+0.217	0.165
NSEI	+0.978	+0.964	+1.150	+0.217	0.148
JKSE	+0.983	+1.000	+1.170	+0.210	0.144
KLCI	+0.979	+1.000	+1.156	+0.214	0.134
CSEALL	+0.987	+0.964	+1.077	+0.173	0.260
KSE100	+0.986	+0.964	+1.455	+0.309	0.734
none	+0.979	+0.976	+1.154	+0.216	0.250

Eight points remain eight points, and none of this makes the relation better estimated than a sample of eight can make it. What it establishes is which of the two claims that sample can carry: the ordering, yes; the line, no.

S18. Does the house price result come from the index's own moving average? · *Artefact explanation rejected, magnitude falls*

Each published Case-Shiller figure for month t is a three-month moving average of sales pairs closing in months $t - 2$ to t . So consecutive monthly changes are built from overlapping transaction windows, every series in the panel carries the same induced persistence, and a reader is entitled to ask whether Section 8 recovers anything but that shared construction. The objection has to be answered with a design rather than a paragraph.

Where the overlap is and is not, stated precisely, because the answer is not what the objection assumes. At the headline delay of three months the target's own stale mark covers $[t - 5, t - 3]$ and the fresh peer marks cover $[t - 2, t]$: disjoint, so no transaction month is shared between the stale mark and the cross-section replacing it. The overlap that does exist is between the peers' window and the label's, which reaches back to months $t - 1$ and t . A fresh peer mark carries conditions in those two months, and those months are inside the target's own next published change although nobody has published them for the target yet. That is what a stale mark *is*, so it is not obviously a defect; but it is also the channel through which a smoother could manufacture a rate, and argument cannot separate the two.

The design that can is to remove every shared transaction month. Read the index only every third month and the windows behind consecutive observations are disjoint, and so are the windows behind the stale mark, the peer block and the label. All three phases of the quarterly grid are run, because choosing one would leave the choice unexplained and the three are not the same sample. One quarter of relative staleness here is the same economic object as $\delta = 3$ months in Section 8. Two estimators are reported, because on a sample a quarter the size a single metro whose own delay cost happens to be near zero dominates a mean of ratios; the pooled figure divides the average increment by the average delay cost and cannot be driven that way.

Table S14. The Section 8 design on non-overlapping quarterly changes, fourteen metros, return target. 'Mean R' is Section 8's estimator, R per metro then averaged; 'pooled R' divides the average increment by the average delay cost. Intervals resample year-long blocks common to all fourteen metros. Delay is in quarters; one quarter is one appraisal cycle.

phase	δ	S_{own}	S_{cross}	mean R	95% CI	pooled R	95% CI	metros R>0
Jan/Apr/Jul/Oct	1	0.3092	0.4279	18.4%	[−349, 331]	49.1%	[19, 75]	11 of 14
Feb/May/Aug/Nov	1	0.3206	0.4704	52.1%	[2, 90]	58.3%	[35, 79]	12 of 14
Mar/Jun/Sep/Dec	1	0.3115	0.4677	60.4%	[−5, 139]	63.5%	[38, 81]	12 of 14
Jan/Apr/Jul/Oct	2	0.2479	0.4429	98.7%	[−258, 342]	64.3%	[34, 86]	13 of 14
Feb/May/Aug/Nov	2	0.2693	0.4657	64.0%	[−7, 134]	63.7%	[39, 82]	13 of 14
Mar/Jun/Sep/Dec	2	0.2599	0.4450	57.6%	[−42, 131]	62.2%	[34, 83]	13 of 14

At one quarter of staleness the pooled rate averages 57% over the three phases and spans 49% to 63% across them, against Section 8's 67.8% [45, 92] at the same economic delay, and is positive in 35 of 42 metro-phases. On the phase average twelve of the fourteen metros are positive, which under independent metros would be $p = 0.013$ by an exact sign test; the metros are not independent, sharing one national housing cycle, which is why the intervals resample blocks common to all of them and why the sign test is quoted as the most favourable reading available rather than as the evidence.

What that settles and what it does not. Under the pooled estimator all three phases exclude zero; under Section 8's own estimator only one of the three does, so the quarterly sample cannot carry the claim on its own in the form the paper states it, and no quarterly figure is offered as a second headline. What it does dispose of is the objection it was built for. The moving-average explanation is not a claim that the rate is smaller than 67.8%, it is a claim that the rate is manufactured, and a manufactured rate must go to zero when no two quantities in the design share a transaction month. It does not. It also does not stay where it was: 57% is visibly below 67.8%, which is what one would expect if a modest part of the monthly figure is the smoother, and the paper reports the gap rather than the agreement.

The de-smoothed panel answers from the other direction and produced the more interesting finding. Applying Geltner's reverse filter $r^* = (r - \rho r_{-1}) / (1 - \rho)$, with ρ estimated on each metro's first ten years and frozen so that no observation is filtered by its own future, and re-running Section 8 unchanged, one month of staleness costs -0.0012 of R^2 , which is to say it costs nothing measurable, the sign having gone the wrong way by less than a thousandth. The denominator of $R(\delta)$ has vanished, and the ratio becomes the pathology this paper's Appendix describes rather than a measurement, so no rate is quoted. The increments remain well defined and are the answer: at $\delta = 3$

the peer block still adds +0.1352 of R^2 on a panel with the moving average inverted out of it, positive on fourteen metros of fourteen. Breadth is not paid for by the smoother.

Table S15. The Section 8 design on the de-smoothed panel, return target, fourteen metros. Increments in R^2 rather than rates, because the delay cost that $R(\delta)$ divides by has collapsed. ρ averages 0.567 across metros and the reverse filter amplifies noise by $1/(1 - \rho)$, which is 3.5 on average and 10.9 at Los Angeles; the level of every column below carries that amplification and only the sign is comparable with anything else in this paper.

δ (months)	S_{own}	S_{cross}	delay cost	breadth adds	metros positive
0	0.3019	0.3983	—	—	—
1	0.3014	0.3942	0.0005	+0.0927	13 of 14
2	0.2156	0.3589	0.0863	+0.1433	14 of 14
3	0.1951	0.3363	0.1067	+0.1412	14 of 14
6	0.1182	0.2837	0.1836	+0.1654	14 of 14

The collapse of that denominator is itself worth stating, because it is a finding about the asset class and not an inconvenience. In this panel, smoothing accounts for most of what a short delay costs. Invert it and the forecaster loses little by being late, which is the same statement as saying there is less left for breadth to repair. That is a property of a repeat-sales index measured on fourteen metros, not a general law about marked assets. The appraised assets this paper is about are smoothed, so they sit on the side of that trade where staleness costs and breadth pays.

None of this touches the larger caveat Section 8 already carries. An index is not a book. The overlap question is about how the index is constructed, not about whether a real appraised portfolio behaves like a published average of other people's transactions.

S19. The seam audit in full · No corruption detected

One join in the index data cannot be checked by comparing the two exports, and it is the S&P's: its files abut rather than overlap, so no row exists in both. That is the target series, and a silent corruption there would move every number in the paper. Section 10 states the limitation; this section removes the words "by comparison" from it, because Cboe publishes its own VIX history and that is a different vendor on a different data path, measuring the same underlying and spanning the join.

Three failures could occur in an export and each is visible to a different instrument, so three are used. A spliced block or a gross scale error breaks the relationship $\log \sigma^2_{YZ} = \alpha + \beta \log \text{VIX}$, tested jointly on (α, β) by a Wald statistic with the same HAC correction used everywhere else here. A one-session date misalignment barely moves a five-day rolling variance and is invisible to that test, but it is glaring in the same-day relationship between the S&P's close-to-close return and the change in the exchange's index. A mis-scaled intraday range moves Yang-Zhang and leaves close-to-close variance alone, because 88% of the Yang-Zhang weight sits on a term the close never enters, so their ratio is an internal check needing no auditor at all. Each is calibrated against 400 pseudo-seams drawn from the S&P's own history and against eight dates where the S&P file has no join, which are the dates the other indices' files happen to meet and are therefore known-clean by construction.

At the join the three statistics sit at the 70th, 4th and 25th percentiles of their own nulls. The test is one-sided in what it can conclude, and the calibration shows why: several of the known-clean dates carry a large level statistic because they sit beside COVID, where the relationship between realised

and implied volatility genuinely moved. A large statistic is therefore not proof of corruption, while a small one at the date the files meet is evidence that nothing happened to the data there. The power question is the one a limitation actually turns on, so it is answered by damaging the post-seam block deliberately and re-running all three: a one-session misalignment is caught, a block spliced in from the DAX is caught, and an intraday range mis-scaled by 25% is caught, while 10% is not. One corruption is invisible to all three and is named rather than left out: multiplying every post-seam price by a constant moves nothing, because all of this reads log ratios, and it is for the same reason the one corruption that could not change a number in this paper.

For VIX the concern is settled rather than bounded, because the two vendors' files can be compared directly. On the 6,747 dates both carry they are identical to the cent on 99.91% of rows and the mean absolute difference is 0.000222 index points; two rows fall outside this project's 0.5% seam tolerance, 28 November 2008 and 24 December 2012, both shortened sessions. The audit also found something we were not looking for. The retail export carries fifteen rows on dates United States equities were shut, 11 September 2001 among them: a fabricated observation rather than a discrepancy, and one no admissibility rule here could have caught, because the row is perfectly well formed. All fifteen fall outside the panel, so no model was ever shown one. VIX is Cboe's file throughout this paper, and substituting it moved no cell of Section 7 by as much as a thousandth of R^2 . We report that as measurement instead of reassurance, and note that a retail export of any series can contain rows the exchange never produced.

The official file does not, however, supply the observation Section 4 actually wants, because it carries one open, high, low and close per day and no intraday snapshot at all. The 21:00 value of VIX existed and is the right regressor; it is not in any daily product, official or retail, and the lagged and same-day arms bracket it without recovering it. Differences at $\delta \leq 2$ fall inside the simultaneous bands. The variance-scale loss is dominated by a handful of crisis days, so its intervals are wide enough to be uninformative at short delays.

S20. The state-space alternative in full · *Better estimator raises the rate*

Section 10 names the estimator as the paper's largest limitation and reports the direction of the alternative. This section gives it. The competitor is one latent global volatility factor on an AR(1), every series loading on it, and a Kalman filter handling a missing observation by dropping that row from the measurement equation. That is what a delayed domestic mark is, rather than something imposed on the data, and it is the property that makes the comparison worth making at all: the filter is not merely a better regression, it is a model of the thing under study. It is fitted two-step after Doz, Giannone and Reichlin (2011), with the state-space parameters estimated once on the initial training block and then frozen, and the filtered state fed into the same forecasting ridge on the same days.

It beats the specification used throughout this paper at $\delta = 21$ by 0.0119 of R^2 , interval $[+0.00021, +0.02382]$, and at $\delta = 55$ by 0.0196, interval $[+0.00316, +0.03530]$. It loses at $\delta = 0$, by 0.0224, interval $[-0.03319, -0.00952]$. The crossover is the informative part: imposing a one-factor structure costs skill when the mark is fresh and pays when it is stale, which is what a structural assumption should do when the data thin out. Its reconstruction of the current domestic state the forecaster is not allowed to see correlates 0.84 with the truth at eleven weeks, against 0.18 for the stale value itself.

What the result establishes is the size of what this paper's transparency costs: roughly two points of R^2 at the longest delay, which is larger than several of the effects the paper argues over. A bound in R^2

is not what a reader of this paper wants, so the headline itself was recomputed inside the filtered specification. The control cannot change, because Section S4 finds no filtered own-only model to build: with no observation noise in the domestic series the filtered state *is* the observation. So the filter enters through the model carrying the cross-section, where it is the better model at three of six delays, and the rate at eleven weeks becomes 75.4% against the published 71.6%, a paired difference of +3.8% [+0.8%, +7.7%]. In days, the mark reads 3.7 days old [2.0, 6.7] rather than the published 4.6, a paired difference of −0.9 days [−2.3, −0.2]. The interval on 3.7 is that file's own bootstrap and is not the one Section S12 quotes beside 4.6, which is why the paired difference rather than the pair of intervals is the number to read.

The estimator limitation therefore does not run the way a limitation usually does. A better estimator raises the numerator of $R(\delta)$ and has nothing to give its denominator, so the paper is not protecting a large number with a transparent estimator; it is quoting a smaller one than the better estimator would give it. What is still declined is the rest of the paper inside a filtered specification, which is a different paper: $R(\delta)$ is a ratio in which the same domestic control appears three times, so a better estimator raises numerator and denominator together and the rate is not simply the rate plus the improvement.

S21. The estimation bill under an unfitted generator · *Objection answered*

Both measurements of the bill are relative to something. The slope method extrapolates from peer counts; the surrogate method replaces the foreign block with fitted Gaussian AR(1) series. One objection to the second is correct: it preserves first-order persistence and contemporaneous cross-market covariance and nothing else: not heavy tails, not volatility-of-volatility, not crisis clustering, not higher-order serial structure. Under that objection the cost is identified only relative to a generator we chose, and a forecaster's real regressors are not Gaussian AR(1) series.

The objection has a direct test. Measure the bill again with a generator that fits nothing at all: take the real foreign block and move it circularly through the sample. Every marginal, every tail, every cluster is then the real one, and the only thing destroyed is the alignment with the target. If an unfitted generator returns the fitted generator's number, the objection is answered, not conceded.

Table S16. The cost of breadth under a fitted and an unfitted null, on identical days. AR(1) is the generator used above, twenty draws; SHIFT is the real foreign block moved circularly with its real missing-data mask held fixed, twenty offsets, with the Monte Carlo standard error of that mean beside it. NAIVE is the same shift applied to the raw array, so the holidays move with the values; it is reported as a diagnostic, not as a candidate.

δ	net ΔR^2	cost, AR(1)	cost, SHIFT	(s.e.)	cost, NAIVE	gross, AR(1)	gross, SHIFT
0	−0.0099	−0.0120	−0.0117	0.0008	−0.0316	+0.0021	+0.0018
1	−0.0006	−0.0142	−0.0134	0.0010	−0.0318	+0.0136	+0.0129
3	+0.0397	−0.0189	−0.0169	0.0013	−0.0328	+0.0587	+0.0566
5	+0.0797	−0.0232	−0.0201	0.0015	−0.0341	+0.1029	+0.0998
13	+0.1819	−0.0392	−0.0335	0.0026	−0.0418	+0.2212	+0.2154
21	+0.2499	−0.0523	−0.0467	0.0038	−0.0498	+0.3023	+0.2966
55	+0.3666	−0.0673	−0.0666	0.0068	−0.0655	+0.4338	+0.4332

One bookkeeping point before the result, because the table invites a fair objection. The AR(1) column here is twenty draws on this subsection's delay grid, not the hundred behind Table 3, so the two

tables give slightly different figures for the same quantity: -0.0111 against -0.0131 at $\delta = 0$, and -0.0502 against -0.0493 at $\delta = 21$. That spread is Monte Carlo error and nothing else. Table 3 remains the paper's estimate of the bill, because it rests on more draws. What Table S16 is for is the comparison *within* it, between two columns computed on identical days from an identical number of draws, and that comparison is unaffected by how many draws either column uses.

The two generators agree at every delay. Compared against three times its own combined Monte Carlo standard error, delay by delay rather than against the loosest tolerance in the table, the largest gap is 0.0058 of R^2 at $\delta = 13$, which is 0.60 of the tolerance there, and no delay comes closer. The gross readings move by less: $+0.0020$ becomes $+0.0018$ at $\delta = 0$, and $+0.3023$ becomes $+0.2966$ at $\delta = 21$. Nothing in Section 6 depends on which generator supplied the number, and in that sense the objection is answered. That claim is worth only as much as the null generator behind it, and an earlier version of this measurement was wrong twice before it was right. Both failures are recorded here rather than tidied away, because a placebo that quietly stops being a placebo returns a plausible number with no outward sign of trouble.

The first shifted the block without preserving its pattern of missing observations, which quietly discarded about 187 test days per draw, or 3.8%, from every comparison. The second drew shift offsets without a guard against the ends, so one draw landed eight days from the identity, scored $+0.1816$ on its own against a spread with a standard deviation of 0.0680 and dragged an eight-draw mean. Corrected, the naive version of the penalty sits 0.0199 of R^2 below the corrected estimate at $\delta = 0$ and 0.0011 below it at $\delta = 55$. A placebo that quietly stops being a placebo returns a believable figure with no outward sign of trouble, so the shift is now guarded and the mask preserved.

S22. Redundancy under a non-linear estimator · *No effect detected*

Redundancy was established with a ridge, and it is reasonable to ask whether a linear map is simply too blunt to find what the foreign block holds given implied volatility. The test is to give the block a richer map and see whether anything appears: a random Fourier feature expansion, which is a one-hidden-layer network whose first layer is drawn, not learned, at two widths, and a k-nearest-neighbour arm that imposes no functional form at all.

Table S17. What each non-linear map of the foreign block adds to a forecaster holding stale domestic data and implied volatility. Wider penalty grid; Giacomini-White statistics, the implied-volatility model being nested in each. A statistic above $+1.96$ would overturn the redundancy result. Cells in this grid are part of the post-selection audit of Section S37; those significant only at the uncorrected 5% level are *exploratory* and no claim in this paper rests on one of them alone.

δ	R^2 IV	+ foreign	GW z	+ FACTOR	GW z	+ RFF(20)	GW z	+ RFF(60)	GW z
0	0.5508	-0.0094	-2.24	-0.0107	-2.39	-0.0399	-2.49	-0.0519	-3.37
3	0.5245	-0.0050	-0.92	-0.0094	-1.63	-0.0440	-2.00	-0.0597	-2.82
5	0.5231	-0.0055	-0.91	-0.0073	-1.17	-0.0528	-1.98	-0.0683	-2.92
13	0.5138	-0.0052	-0.72	-0.0095	-0.76	-0.0628	-2.12	-0.0739	-2.64
21	0.5127	-0.0046	-0.69	-0.0156	-1.06	-0.0678	-2.12	-0.0800	-2.77
55	0.5088	-0.0046	-0.72	-0.0045	-0.65	-0.0591	-2.03	-0.0660	-2.30

No arm adds anything at any delay, and the two random-feature arms are significantly *worse* at all six. The direction of the capacity comparison is the informative part: RFF(60) is below RFF(20) at every

delay, by 0.007 to 0.018 of R^2 . If the objection were right, if there really were cross-market structure a linear map could not reach, more flexibility would find more of it. More flexibility finds less, and that is what over-parameterisation looks like when the signal is not there. The surrogate measurement puts a number on the mechanism: bending the foreign block costs 0.005 to 0.008 of R^2 for the FACTOR arm and 0.017 to 0.056 for the random-feature arms, rising with both delay and width.

One thing genuinely changes under the wider grid, and it favours the objection. The *linear* increment, significantly negative at short delays in Table S22, stops being significant from $\delta = 3$ onward, with statistics between -0.69 and -0.92 . That is the right result and the one Section 6's estimation-cost argument predicted: the foreign block given implied volatility is not harmful, it is empty, and the apparent harm in Table S22 was the bill for estimating regressors that carry nothing. The redundancy conclusion is unchanged; its sign is better explained.

What this does not license is the claim that no non-linear model could find anything. The random first layer is not learned, and a learned one has strictly more free parameters, so it faces the same estimation-cost bill this sample has repeatedly failed to pay, which is an argument by extension rather than a test. Section 10 records it as such.

S23. Sensitivity to the forecast horizon · *Result survives in the middle of the range*

Every number above forecasts five trading days ahead, and a reader is entitled to ask whether that is a finding or a convention. So the whole experiment is re-run at one day, two weeks and one month. One detail has to be settled before the table can be read: the Yang-Zhang proxy spans five days, which is how the latent variance is measured and has nothing to do with how far ahead we forecast, but it means that at $h = 1$ the outcome window shares four of its five days with the domestic feature the model already holds. Nothing leaks, since the outcome is still strictly in the future and every feature is still dated $t - \delta$ or earlier for the domestic block. But a forecaster at one day is being asked to predict an object they can largely already see, and their R^2 at $\delta = 0$ duly comes out at 0.924 against 0.519 at five days. The serial correlation of the target confirms it: 0.953 at one day against 0.692 at five, 0.576 at ten and 0.417 at twenty-one.

Table S18. The substitution rate at eleven weeks under four forecast horizons. Intervals are the full-ratio bootstrap of Section 5, in moving blocks of the length Section S13 measures, held at that length across horizons so the four rows are comparable. The $h = 5$ row is the paper's headline and reproduces it exactly.

horizon	own R^2 at $\delta = 0$	R(55) under R^2	95% CI	R(55) under QLIKE
1 day	0.924	70%	[62, 76]	83%
5 days	0.519	72%	[60, 81]	81%
10 days	0.367	66%	[49, 78]	73%
21 days	0.203	51%	[23, 70]	65%

From one day to two weeks the rate is flat at 70%, 72% and 66% across a tenfold change in horizon and despite the one-day column measuring something partly mechanical. At one month it falls to 51%, the only point estimate outside the interval the headline quotes, though the intervals still overlap heavily and the one-month interval is wide enough on its own to be uninformative. QLIKE tells the same story more smoothly: 83%, 81%, 73%, 65%. So the headline is a property of the information over horizons out to a fortnight and begins to decay beyond that, which is what should happen, since a foreign close carries information about the next week's volatility state more readily

than about the next month's, and by twenty-one days the domestic model it is being compared against has itself fallen to an R^2 of 0.203, so there is less loss left to recover and a noisier denominator to recover it against. We report the five-day figure as the headline because it is the horizon the whole paper is built on, not because it is the most favourable: at one day under QLIKE the rate is higher.

S24. The race with matched horizons · *Objection correct, conclusion unaffected*

The race above has an asymmetry in it. The target is realised variance five trading days ahead; VIX prices a thirty-day horizon. A thirty-day option-implied variance is not the natural forecast of a five-day realised variance, so the section was run with a horse suited to a different distance. Cboe publishes VIX9D, a nine-day index on the same underlying computed the same way, from January 2011. Nine calendar days is about six trading days against the target's five, so this is near-matched rather than exactly matched; it is the closest listed instrument that exists, and its horizon is about a third of VIX's. Both blocks are given the same number of regressors, one implied-volatility series each plus VDAX-NEW on both sides, so the estimation cost of Section 6.1 is identical on the two sides and cancels in the difference.

The cost of the test belongs before its result. VIX9D begins in 2011, the decision variable of equation (2) needs a 252-day trailing median before it exists, and the walk-forward needs 1,500 days of labelled history before it can fit, which leaves 2,098 test days from May 2018, 46% of Table S22's window. It contains the 2020 episode but not the 2008 one, and 2008 is where being late costs a forecaster most. That sample is also simply harder: the identical thirty-day block loses between 0.0190 and 0.0269 of R^2 when it is moved to the shorter window with nothing else changed. We measure that separately so it cannot be mistaken for the horizon effect, and every column below sits inside the same window.

Table S19. The horse race with the horizons matched, 2,098 days, May 2018 to September 2026. IV30 is the block of Table S22, VIX at $t - 1$ with VDAX-NEW; IV9 replaces VIX with VIX9D at the same regressor count; TERM carries both tenors. The fifth column is the horizon effect on a fixed sample and the last is what the seven foreign closes add *given* the horizon-matched block. Giacomini-White throughout: IV9 against IV30 is a comparison of two non-nested models of equal width, and the foreign increment is nested.

δ	domestic	+ IV30	+ IV9	+ TERM	IV9 – IV30	GW z	foreign IV9	GW z
0	0.4865	0.5298	0.5535	0.5504	+0.0237	3.30	−0.0137	−2.24
5	0.3173	0.5058	0.5451	0.5418	+0.0392	3.14	−0.0174	−3.29
13	0.1926	0.5034	0.5433	0.5378	+0.0399	3.17	−0.0158	−3.25
21	0.1172	0.5047	0.5433	0.5376	+0.0386	3.16	−0.0150	−2.89
55	−0.0308	0.5076	0.5461	0.5441	+0.0385	3.13	−0.0134	−2.36

Two things are in that table and they point in opposite directions. The horizon objection is correct: matching it is worth between 0.0237 and 0.0399 of R^2 , at every one of the five delays, with Giacomini-White statistics of 3.13 to 3.30. This is not a marginal improvement that survives on a technicality; the nine-day index is simply the better instrument for a five-day target, and Section 7 had been using the worse one. But the conclusion the section draws does not move. The last two columns are what the foreign cross-section adds on top of the horizon-matched block, and they are negative at every delay, with statistics between −2.24 and −3.29. The paper had been reporting redundancy against an implied-volatility control that was weaker than it should have been, and redundancy holds against

the stronger one. The horizon objection is right about the instrument and gets a firmer version of the same answer.

Two details close off the obvious follow-ups. Replacing the nine-day level with the nine-to-thirty-day slope, at the same regressor count, is worse at every delay, with statistics from -2.32 to -3.44 : what the near end of the term structure carries is its level, not the shape of the curve. And carrying both tenors is slightly worse than carrying the nine-day one alone, 0.5504 against 0.5535 at zero delay, because the second series does not pay for its own coefficient, which is Section 6.1's mechanism again on a two-column block. One incidental is worth recording because it was not designed: on this window, which shares barely half its days with Table 2's, the foreign block's substitution rate comes out 51% , 64% , 68% and 73% at five, thirteen, twenty-one and fifty-five days, against 51% , 64% , 67% and 72% in Table 2. $R(\delta)$ barely moves between two half-overlapping samples measured eight years apart.

One objection to all of this has to be answered rather than noted, and it is the obvious one. This window contains exactly one violent episode, and a front-tenor index separating from a back one is precisely what a crisis in the term structure would produce, so the whole of Table S19 could be a fact about March 2020. The test is to delete it: every test day from 15 February to 30 June 2020 is removed from the evaluation, the models untouched, leaving 2,004 days. The nine-day index still wins at all five delays, by 0.0247 to 0.0405 of R^2 with statistics of 2.97 to 3.49 , larger at four of the five delays than with 2020 included rather than smaller, and the foreign block still adds nothing, between -0.0148 and -0.0198 with statistics from -1.95 to -3.10 . Neither conclusion in this subsection rests on the crisis.

S25. The two variance-scale losses, and why only one of them is Patton's • *Claim corrected, conclusion unaffected*

One correction belongs here rather than in a footnote, because an earlier version of this paper got it wrong in a way that flattered the result. That version reconstructed nothing: it computed squared error on the median-normalised variance, $\exp(y_t)$ against $\gamma_t \exp(f_t)$, which is Patton's MSE weighted by $1 / M^2$ whenever both sides are reconstructed with the same median. The weight is largest exactly where the median is smallest, so it is not a neutral rescaling, and the two losses are not the same loss: the substitution rate under the normalised version runs 6 to 29 points above the natural-scale one, by 10 points at eleven weeks. A forecaster cannot reconstruct both sides with the same median, because the outcome carries M_{t+h} and only M_t is known at the origin, so that gap is the weight plus the median's drift over h days rather than the weight alone. The table below splits it with a third, infeasible column and finds the drift worth at most 1.1 points at any delay, which leaves the weight carrying essentially all of the correction.

Mean absolute error, used in some recent work in this area, is in no version of the family. One choice inside that arrangement needs defending, because it looks backwards: the headline number in this paper is the one loss of the three that is NOT proxy-robust. We lead with log-scale R^2 because the paper's object is a ratio, $R(\delta)$, and a ratio needs a denominator a reader can hold: the share of a delay-induced loss recovered, naturally bounded near 0 and 100%, on the same scale as the companion note (Vellanikaran 2026) and as the decision variable itself. QLIKE skill on a variance scale is the right loss and the wrong ruler for that purpose; its levels are not comparable to anything else in this literature. The defence is that the choice is presentational and is made where it can be checked: both proxy-robust losses are reported beside the headline in every table where a conclusion turns on one, and

they support the same conclusions everywhere including the $\delta = 0$ row, where both point in the same negative direction but the natural-scale MSE is too imprecise to establish an effect. Because the magnitude there is about one point of R^2 and the practical question concerns stale rather than current marks, no economic claim in this paper rests on that row, and Section S9 of the Internet Appendix says so. A reader who prefers the robust family throughout can read the QLIKE columns and will reach the same conclusions with different numbers attached.

Section 4.4 reports squared error on the natural variance scale and says that an earlier version of this paper computed it on the median-normalised scale instead while calling it Patton's MSE. This section gives both columns, because the size of that mistake is a fact about the paper and should be on the record as a number rather than as an admission.

Two things separate the two columns, and only the first of them is a factor. The target is $y_t = \log(\sigma^2_{YZ,t+h} / M_{t+h})$, so $\exp(y_t)$ is a variance divided by a median and $\exp(y_t) M_{t+h}$ is the variance itself. Had the forecast been reconstructed with that same median, squared error on the normalised pair would be squared error on the natural pair weighted by exactly $1 / M_{t+h}^2$, and nothing else would differ. But M_{t+h} is not known at t . The only reconstruction a forecaster can run multiplies the forecast by M_t and the natural-scale column reported throughout this paper is that one, so moving from the normalised column to it takes the weight off *and* moves the forecast's median h days at the same time. The gap between those two columns is therefore not, on its own, a measurement of the weight, and an earlier version of this section said it was.

Table S20 separates them with a third column nobody could compute. Reconstruct both sides with M_{t+h} : per day that loss is exactly M_{t+h}^2 times the normalised loss, which the script checks numerically rather than asserting, to $1.6e-12$ in relative terms. So what separates that column from the normalised one is the weight alone, and what separates it from the implementable one is the median's h -day drift alone. The weight turns out to carry nearly all of the correction, 28.8 of the 29.4 points at one day and 9.0 of the 10.1 at eleven weeks, with the drift never worth more than 1.1 points at any delay, so the earlier wording had the economics right and the algebra wrong, and what is on the record now is two named effects rather than one undifferentiated number.

The weight is not innocuous: it is largest in the calmest windows, which is exactly where the delay does least damage and the substitution rate's denominator is smallest. Patton's result is about the unweighted loss (2011). QLIKE is unaffected by any of this, and not by luck: it depends on the ratio of outcome to forecast, so a common factor cancels exactly, which Part 0 of the script demonstrates numerically to ten decimal places rather than asserting.

Table S20. The substitution rate under three variance-scale losses, and the correction decomposed. 'Normalised' divides outcome and forecast by the trailing median and is what an earlier version of this paper reported as Patton's MSE. 'Both at t+h' reconstructs outcome and forecast with the same median M_{t+h} , which no forecaster could do, and differs from the normalised column by the $1/M^2$ weight alone. 'Natural' forecasts with the median known at the origin and is the implementable loss Patton's robustness result covers; it differs from the middle column by the median's h-day drift alone. The last three columns are those two differences and their sum, in points of the rate; an entry of ± 0.0 is a signed difference below a tenth of a point. Intervals in the paper bootstrap the whole ratio on the same resampled days.

δ	$R(\delta)$ normalised	$R(\delta)$ both at t+h	$R(\delta)$ natural	weight	drift	total
1	59.9%	31.1%	30.5%	+28.8	+0.7	+29.4
3	74.6%	60.4%	60.4%	+14.2	-0.0	+14.2
5	85.5%	73.5%	73.6%	+12.0	-0.1	+11.9
8	87.8%	78.6%	78.8%	+9.1	-0.1	+9.0
13	83.5%	75.7%	75.1%	+7.8	+0.6	+8.4
21	83.5%	75.6%	74.5%	+7.9	+1.1	+9.0
34	82.0%	76.3%	75.3%	+5.7	+0.9	+6.6
55	90.1%	81.0%	80.0%	+9.0	+1.1	+10.1

The normalised loss overstated the rate at every delay, by 29 points at one day and by 10 at eleven weeks. Three things follow and the middle one is the reason this correction improved the paper rather than damaging it. The headline is unaffected, because the headline is the log-scale figure and neither variance-scale column feeds it. The two robust losses now agree: QLIKE reads 80% at eleven weeks and the natural-scale MSE reads 80.0%, where before the mislabelled column read 90% and sat apart from both. And the $\delta = 0$ sign flip this appendix used to report was the artefact rather than the finding, which Section S9 now records.

The normalised column is kept here rather than deleted for the same reason the other corrections in this appendix are kept. A reader who finds an error in a paper wants to know its size, and a paper that quietly replaced one loss with another would be asking to be trusted on precisely the point it had just got wrong.

S26. A model of the delay curve

This section states a specification before the measurements that support it, because written down it turns the delay curve from a description into a prediction with a testable ceiling, and that ceiling is what Sections 5 to 7 are then measured against. The specification is not assumed: it is the shape two of those measurements take, and a reader who prefers the evidence first can read Sections 5 and 6 and return here.

Section S27 finds the recovery entirely in the leading component of the foreign block, with the six orthogonal components worth less than nothing. Section S4 of the Internet Appendix finds that the level of volatility, which equation (2) removes, carries information that grows with delay. Together they say: log variance is a slow level, plus a factor several markets share, plus something idiosyncratic. For market i ,

$$\log \sigma_{i,t}^2 = \mu_{i,t} + \lambda_i g_t + e_{i,t}, \quad g_t = \varphi g_{t-1} + u_t \quad (10)$$

with μ the slow level the 252-day median tracks, g a persistent global factor, e idiosyncratic and u white noise. The decision variable of equation (2) subtracts an estimate of μ , so $a_{i,t} \approx \lambda_i g_t + e_{i,t}$. That is the whole of Section S4 of the Internet Appendix in one line: the normalisation removes μ by construction, so restoring it is worth something.

A forecaster δ days late sees $a_{SPX,t-\delta}$, one noisy reading of $g_{t-\delta}$. The cross-section supplies seven readings of g_t , whose common component suppresses the idiosyncratic part. Under (10),

$$E[g_{t+h} | g_{t-\delta}] = \varphi^{h+\delta} g_{t-\delta}, \quad E[g_{t+h} | g_t] = \varphi^h g_t, \quad (11)$$

and three consequences follow, each checkable against quantities measured elsewhere in this paper, not fitted to the curve.

First, the stale forecaster's skill decays geometrically: $\log S_{own}(\delta)$ is linear in δ with slope $2 \log \varphi$. Fitted over the eight delays where S_{own} is positive, the line has $R^2 = 0.9980$ and slope -0.05419 per day, implying $\varphi = 0.9733$. The factor's own persistence, estimated from the principal-component series and never shown the decay curve, is 0.9678 . The two agree to 0.0055 . The delay curve is the factor's memory, seen through a forecast.

Second, $S_{cross}(\delta)$ should be near-flat, because the foreign block delivers g_t whatever the domestic delay is. Across eleven weeks it falls 27%, against 99% for the domestic model. Perfect flatness is not predicted, because the domestic block is still in the cross-sectional model and still decays, but the contrast is the model's.

Third, and most useful: $R(\delta)$ rises from zero towards a ceiling below one, and the model names it. Since $S_{own}(\delta) \rightarrow 0$,

$$R(\infty) = S_{cross} \div S_{own}(0), \quad (12)$$

which is below one because seven peers estimate g_t with error and never observe $e_{SPX,t}$. Both quantities are measured elsewhere: $S_{cross}(55) = 0.3736$ and $S_{own}(0) = 0.5191$, giving a ceiling of 72.0%, against a measured 69.8% at $\delta = 34$ and 71.6% at $\delta = 55$.

That agreement is weaker evidence than an earlier version of this section claimed, and the reason is arithmetic rather than empirical. Write the rate out: $R(\delta) = [S_{cross}(\delta) - S_{own}(\delta)] \div [S_{own}(0) - S_{own}(\delta)]$. When $S_{own}(\delta)$ is near zero (at $\delta = 55$ it is 0.0070), this collapses to $S_{cross}(55) \div S_{own}(0)$, which is equation (12). So the “prediction” and the measurement are built from the same two numbers and must agree; the 0.4-point gap between 71.97% and 71.59% is the size of $S_{own}(55)$, not the accuracy of a forecast. This is sharper now than it was: before the benchmark correction of Section 4.4, $S_{own}(55)$ was -0.0351 and the same two quantities differed by 1.8 points.

We therefore call (12) a factor-model *interpretation* of the observed plateau rather than a prediction of it. What the model earns is the reading: the curve flattens because the cross-section's skill is a factor share bounded by how well seven peers estimate g , and the level it flattens at is that share. What it does not earn is credit for having predicted a number it is algebraically tied to. A genuine out-of-sample test would estimate the ceiling from parameters independent of the endpoint being predicted (the factor's loadings and its persistence, say, rather than $S_{cross}(55)$ itself), and this paper has not done that.

So the substitution rate is not an arbitrary rising function. It is a factor share, it converges, and the paper can say what it converges to. The model also explains why Section S27's orthogonal remainder is worthless, being e , which forecasts nothing, and why Section S12 of the Internet Appendix's staleness penalty exists, since an aged foreign block reads g at $t - g$, not at t and (11) prices that at φ per day. Three results the paper had separately are one mechanism.

What the model does not explain is why the mechanism is stronger in stress. Equation (11) holds λ and φ constant; Section S5 of the Internet Appendix shows the factor doing more work when markets are frightened, which would need λ or the variance of u to move with state. We have not fitted that, and it is the obvious next specification.

S27. Breadth against its own leading component

A further objection is omitted-variable bias: foreign breadth may not be capturing cross-border diffusion at all, but proxying for a global systemic risk factor, so that what looks like transmission is simultaneous reaction to shared news. Section 6.1 is most of the way to agreeing, since one real-time factor beats all seven series at every delay, but it never separates the two halves and asks which one carries the recovery. The decomposition is available at no cost, because the basis is already being estimated.

At each refit, write v_1 for the leading eigenvector of the training covariance of the standardised foreign block. Let p_t be that block's seven decision variables at t , standardised on the training window, and let v_1 be the unit eigenvector belonging to the largest eigenvalue of their training covariance, with its sign pinned so that the first loading is positive. The split is

$$p_t = g_t v_1 + r_t, \quad g_t = p_t' v_1, \quad r_t = p_t - g_t v_1, \quad (13)$$

where g_t is the scalar global factor and r_t is the six-dimensional remainder orthogonal to v_1 . Sign pinning is not cosmetic: an eigenvector is defined only up to sign, and one that flipped between refits would enter the regression as noise unrelated to the data. The $+ PC1$ column below uses g_t as a single regressor and $+ remainder$ uses the six components of r_t . Each day's standardised row x splits into a global part $(x \cdot v_1) v_1$ and a remainder orthogonal to it, spanned by components two through seven. Adding the two back together is a rotation of the raw block and must reproduce it: it does, to within 0.0010 of R^2 at every delay, the check that licenses reading the split.

Table S21. The foreign block split into its global component and the remainder orthogonal to it, both estimated in real time at every refit. The last column tests PC1 against the remainder by Giacomini-White; positive favours the global component.

δ	domestic	+ PC1	+ remainder	+ all seven	PC1 – domestic	remainder – domestic	GW z
0	0.5191	0.5204	0.5082	0.5092	+0.0013	−0.0109	2.81
3	0.4098	0.4614	0.3991	0.4495	+0.0516	−0.0107	4.43
5	0.3641	0.4556	0.3550	0.4438	+0.0915	−0.0091	4.87
13	0.2336	0.4275	0.2271	0.4155	+0.1939	−0.0065	4.72
21	0.1482	0.4122	0.1359	0.3981	+0.2641	−0.0123	4.16
55	0.0070	0.3922	−0.0196	0.3736	+0.3852	−0.0266	4.66

The answer is consistent across the delays tested. The global component alone beats the domestic model at five of six delays and by as much as 0.40 of R^2 . The remainder, six real-time components orthogonal to it, is *negative* at every delay: it beats the domestic model at none, and carrying it costs between 0.007 and 0.028 of R^2 , the estimation bill of Section 6 with no information to pay for it. Giacomini-White separates the two at every delay, from $z = 2.94$ to 7.22.

So the objection is right about the mechanism and we adopt the correction. What this paper measures is a global volatility factor observed while the domestic series is stale, not bilateral diffusion between particular markets. Three things follow. The substitution rate is unchanged: it does not matter to the number what the factor is called. The practical advice sharpens: a forecaster wants a cross-market average, not a panel of countries, as Section 6.1 already recommended for a different reason. And a limitation appears that Section 10 now carries: a first principal component of seven equity indices is a proxy for global risk, not a measurement of it, so orthogonalising against it is weaker than orthogonalising against a factor built from a wider universe, and the remainder may still hold global content this basis cannot see.

S28. The implied-volatility race in full

Section 7 of the paper states this race's result and points here for the working. The race itself is below, followed by the three questions it raises.

The objection this section answers is the one Section 2 names as unavoidable: if a forecaster whose domestic data is δ days stale can see today's VIX, why would they need Tokyo? Implied volatility is forward-looking, it never goes stale, and both Korkusuz (2025) and Buncic and Gisler (2016) report that much of what looks like a cross-market effect is attributable to it. So the two candidates go into the same walk-forward, on the same labels, with the same estimator. Four models, all with the domestic block held δ days stale: domestic alone, plus the foreign block, plus implied volatility, and plus both. The number the objection turns on is what the foreign block adds *given* implied volatility. VIX enters at $t - 1$ throughout, the strictly admissible specification of Section 4; the same-day arm is reported below as a sensitivity and reaches the same verdict. VDAX-NEW begins in May 2001 against the index panel's February 2000, so all four are re-estimated on the same reduced window of 4,525 days and the comparison stays internal.

Table S22. Out-of-sample R^2 for four information sets, 4,525 days, September 2008 to September 2026. The last two columns are what the foreign block adds *given* implied volatility, and the Giacomini-White statistic testing it. VIX enters at $t - 1$, the strictly admissible specification of Section 4; the same-day timing sensitivity check follows in the text.

δ	domestic	+ foreign	+ implied vol	+ both	foreign IV	GW z
0	0.5141	0.5051	0.5547	0.5443	−0.0104	−2.29
3	0.4046	0.4478	0.5337	0.5233	−0.0105	−1.96
5	0.3591	0.4431	0.5327	0.5233	−0.0095	−1.68
13	0.2281	0.4148	0.5261	0.5171	−0.0090	−1.40
21	0.1405	0.3978	0.5247	0.5151	−0.0096	−1.50
55	0.0007	0.3769	0.5266	0.5172	−0.0095	−1.44

The third column is the one that decides the section. A forecaster holding implied volatility loses almost nothing to delay: R^2 goes from 0.5547 to 0.5266 across eleven weeks, where the domestic-only

forecaster falls from 0.5141 to 0.0006. Put the other way, a desk with VIX and VDAX and a three-week-old mark still beats a desk with fully current domestic data and no options information, 0.5247 against 0.5141. The last two columns are the comparison the objection turns on, and they are negative at every delay: adding seven foreign closes on top of implied volatility makes the forecast worse, not better, by about one point of R^2 . That reads like harm and the subsection below shows it is a bill, the cost of estimating seven coefficients, but on the question the section asks, the answer is already here. Breadth adds nothing measurable to a forecaster who can see the options market.

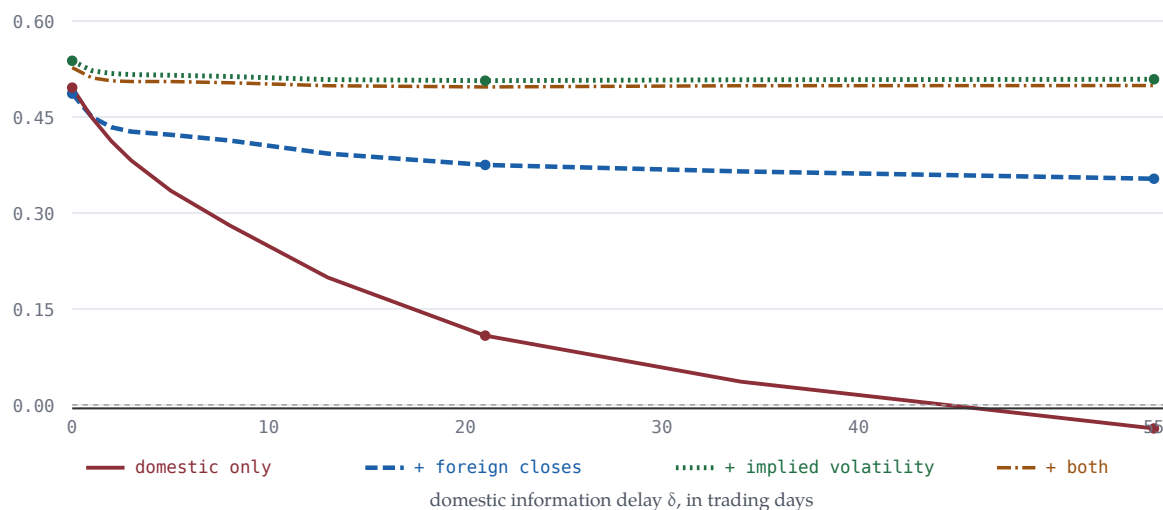


Figure S1. The dotted green curve is almost flat. A forecaster holding implied volatility barely notices how old their domestic data is, and from about four days on they are ahead of where the domestic-only forecaster started. The long-dashed curve, which adds foreign closes on top, sits just below it at every delay.

The dotted green curve is the result. Implied volatility does not merely slow the decay, it very nearly stops it: R^2 falls from 0.538 to 0.509 across eleven weeks, against 0.496 to -0.037 for the domestic model alone. Expressed as a substitution rate that is 103% at three weeks, and a rate above 100% is the signal from Section 4 that the measure has changed meaning. A forecaster holding three-week-old domestic data and today's VIX scores 0.507, above the 0.496 of a forecaster with fully current domestic data and no options market. Implied volatility is not restoring what staleness took; it is supplying something the realised-variance history never contained.

That should not be mysterious, though it should not be overstated either. The VIX is a RISK-NEUTRAL expectation of S&P variance over roughly the coming MONTH; our target is REALISED variance over the coming five days. The two differ in horizon, by a factor of about four, and in measure, by a variance risk premium that is itself time-varying. So the VIX is not the quantity we are predicting. It is, however, a purpose-built forward-looking forecast of something close to it, and comparing that with a cross-section of correlated state variables is comparing a forecast with a set of clues. Read that way the surprise is not that implied volatility wins but that seven foreign equity markets get within two-thirds of it despite the horizon and measure mismatch working against the options market.

The sixth column is what the objection actually turns on, and it is negative at every delay. Given implied volatility, adding the foreign block makes the forecast slightly worse, and the Giacomini-White statistic, for which the domestic-plus-IV model is nested inside the full one so the same test applies as in Section S2 of the Internet Appendix, reaches conventional significance at the short

delays. Taken literally this says foreign closes damage a forecast that already has VIX, which would be strange behaviour for genuine information.

Section 6 supplies the alternative reading, and the same instrument settles it. Replacing the seven foreign series with AR(1) surrogates matched on persistence and variance leaves regressors that carry nothing by construction, and what remains is the bill for estimating them. It comes to roughly 0.011 to 0.015 of R^2 , the whole of the negative. The content of the foreign block given implied volatility, gross of that cost, is +0.002 at $\delta = 0$, +0.007 at $\delta = 5$, +0.008 at $\delta = 21$ and +0.007 at $\delta = 55$, and bootstrapping each of those the way Section 6 describes gives intervals that all cover zero: $[-0.006, +0.011]$, $[-0.003, +0.017]$, $[-0.004, +0.019]$ and $[-0.004, +0.018]$. The correct word is redundant, and it is worth saying exactly what that word is doing. It records a failure to distinguish, not a demonstration of equality: the widest of the four intervals above rules out a gross gain larger than 0.019 of R^2 , at $\delta = 21$ and leaves anything smaller open, so what is established is a bound and not a zero. Within that bound, foreign closes do not damage a forecast that already sees the options market; they add nothing we can detect, and paying to estimate nothing is what makes the measured figure negative.

The table above is the admissible specification: VIX lagged a full day, which discards a day of information the live index would have carried. Substituting the same-day close, which contains fifteen minutes we are not entitled to, moves the options market the other way: R^2 at $\delta = 21$ rises from 0.507 to 0.568, a substitution rate of 119%, while the incremental contribution of the foreign block stays negative at all ten delays and its gross content stays indistinguishable from zero. Neither arm is the specification we would choose, and the one we would choose is not available to us. What licenses a conclusion here is not that the 21:00 observation falls between these two in time, which would say nothing about how it forecasts, but that the verdict is the same under a specification that gives the options market a full day less information and under one that gives it fifteen minutes more. A finding that survives that much variation in the input is not resting on the timing choice.

One feature of that result deserves a sentence, because it is more striking than it first looks. VIX is a thirty-calendar-day implied volatility and the target here is realised variance over five trading days, and the two are not in the same unit: thirty calendar days is about twenty-one trading days, so the winning instrument is measured over a horizon about four times the length of the thing it is forecasting, not six. That it dominates anyway says something about how much of the five-day question is answered by the level of the whole term structure, not by its front. The matched instrument exists, since Cboe publishes a nine-day index, VIX9D, formerly VXST, and Section S24 runs the race again with it. The horizon does matter, by rather more than we expected, and the conclusion of this section does not.

Three further questions about this race are answered below: whether redundancy survives compressing the cross-section, what the skill column means economically, and three ways the race could have been unfair to breadth.

S28.1 One regressor instead of seven

The test above adds seven regressors to the implied-volatility model, and it therefore cannot separate two different reasons for a null: the foreign block carries nothing beyond VIX and VDAX, or it carries something and the cost of estimating seven coefficients consumes it. Section 6 says that cost is real

and Section 6.1 says a compression avoids part of it, so the confound is not hypothetical. The compression supplies the fix: add *one* regressor instead of seven.

Table S23. What a single compressed summary of the foreign block adds on top of stale domestic data and the implied-volatility block, which is VIX and VDAX-NEW. The mean requires no estimation at all; PC1 is the real-time first component of Section 6.1. The seven-regressor column is Table S2 of the Internet Appendix's, for comparison. Cells in this grid are part of the post-selection audit of Section S37; those significant only at the uncorrected 5% level are *exploratory* and no claim in this paper rests on one of them alone.

δ	+ IV only	+ mean	+ PC1	+ all seven	mean IV	GW z	PC1 IV	GW z
0	0.5547	0.5559	0.5558	0.5443	+0.0012	0.42	+0.0011	0.41
3	0.5337	0.5359	0.5362	0.5233	+0.0021	0.86	+0.0024	0.95
5	0.5327	0.5354	0.5358	0.5233	+0.0027	0.87	+0.0030	0.97
13	0.5261	0.5297	0.5301	0.5171	+0.0036	0.92	+0.0040	1.01
21	0.5247	0.5283	0.5287	0.5151	+0.0037	0.95	+0.0041	1.04
55	0.5266	0.5306	0.5307	0.5172	+0.0040	1.13	+0.0041	1.12

The sign flips and the conclusion does not. With seven regressors the increment is negative at every delay; with one it is positive at every delay. That is the estimation-cost mechanism of Section 6 showing up exactly where it was predicted to, and it means the negative in Table S2 of the Internet Appendix was a bill rather than evidence of harm, a reading the paper offered there and can now demonstrate rather than argue. But no Giacomini-White statistic reaches 1.96: the largest is 1.31, at eleven weeks. A single cheap summary of seven foreign markets, carrying essentially no estimation cost, still adds nothing detectable to a forecaster holding implied volatility.

That is a stronger statement than the one this section made before, and it should be read as strictly stronger, a sharpening of the earlier finding. The earlier test could be dismissed as penalising the foreign block for its parameter count. This one cannot. What it cannot do either is prove absence: increments of +0.001 to +0.004 with statistics between 0.48 and 1.31 are consistent with a small positive contribution this sample is too small to resolve, and the honest summary is that the foreign block adds nothing we can detect given implied volatility, not that it adds nothing.

8.2 Economic interpretation of the skill column

A point of R^2 on log variance is not an intuitive quantity, and the reading a risk manager needs is narrower than the column suggests. What breadth repairs is a stale forecaster's tendency to over-estimate variance; the understating tail is unmoved: it runs between 8.6% and 11.4% of days whatever the delay and whichever model is used. Concretely, at eleven weeks the stale domestic forecaster's variance forecast is off by a median factor of 1.90, and is more than a factor of two out on 59.2% of days against 37.1% for the forecaster holding the cross-section, against 31.1% for a forecaster with no delay at all. Breadth closes most of that gap, and all of what it closes is over-estimation. Anyone reaching for the cross-section as protection against being surprised by a volatility spike is reaching for the wrong instrument.

8.3 Three ways the race could be unfair

Three objections to the comparison above can each be tested rather than argued about, and all three are tested in the Internet Appendix so that this section stays the length of its result. None of them

overturns it.

Redundancy was established with a ridge, and a linear map might simply be too blunt to find what the foreign block holds given implied volatility. Section S22 of the Internet Appendix gives the block two richer maps: a random Fourier feature expansion at two widths, which is a one-hidden-layer network whose first layer is drawn rather than learned, and a k-nearest-neighbour arm imposing no functional form at all. Nothing appears. Every statistic that would have overturned the redundancy result comes back negative, and the richer maps cost more than the plain block does.

Every number above forecasts five trading days ahead, and Section S23 of the Internet Appendix re-runs the whole experiment at one day, two weeks and one month to establish that this is a convention rather than a finding. The rate is stable across the middle of that range and falls at the ends for reasons the section separates: at one day the target is an object the forecaster can largely already see, and at one month it has dropped to 51% with an interval too wide to lean on. Section 10 treats that as a limit on how far the result reaches rather than as a result.

One asymmetry in the race is worth removing rather than defending. The target is realised variance five days ahead and VIX prices a thirty-day horizon, so the section above ran the options arm with a horse suited to a different distance. Cboe's VIX9D is the closest listed instrument, with a horizon about a third of VIX's and near-matched to the target rather than exactly matched, and Section S24 of the Internet Appendix runs the race again with it, giving both blocks the same number of regressors so that the estimation cost of Section 6.1 cancels in the difference. The horizon objection is correct and matching it is worth something measurable; the conclusion of Section 7 survives it, and the section states plainly what the shorter sample the test requires costs.

S29. Feasibility at the asset's reporting frequency

The larger absence is the one Sections 8 and S30 keep naming, and this paper has treated it as though the difficulty were finding an appraisal-based series. That is not the difficulty, and being precise about what is changes what a future study should do. The proxy is Yang-Zhang over five daily bars, which needs an open, a high, a low and a close for every session, and an appraised asset has none of those: it has one number per quarter. The walk-forward then needs $252 + 1,250 + 250 = 1,752$ observations before its first forecast, which at one mark per quarter is four hundred and thirty-eight years. No dataset solves that. The design has to be re-specified at the frequency the asset reports, and the question nobody had asked is whether the measurement survives when it is.

It does. Coarsening the S&P and its seven peers to month-end marks throws away exactly what an appraised series never had, leaving the proxy to be built from the variance of the last six monthly returns against a sixty-month median, with the delay counted in months and every window scaled by twenty-one. On 109 test months the rate at a one-quarter delay is 49% [34, 76], and at six months 55% [47, 75]. Smoothing the mark by the partial adjustment of Section S15 of the Internet Appendix, which is how the property literature has described appraisal indices for thirty years, leaves 44% [28, 75] and 54% [43, 79].

Table S24. The same experiment on month-end marks alone, with no daily bars anywhere in it. The delay is in months, so three months is one quarterly marking cycle. The appraised column applies the partial adjustment of Section S15 of the Internet Appendix to the target before anything else is computed. Intervals are moving-block bootstraps at twelve months.

delay	R ² own	R ² cross	R, clean marks	R, appraised marks
0	0.7694	0.7277	n/a	n/a
1 month	0.4846	0.5430	20.5% [−18, 63]	23.1% [−15, 71]
2 months	0.2331	0.4362	37.9% [12, 69]	35.2% [15, 72]
3 months	0.0725	0.4122	48.7% [34, 76]	43.8% [28, 75]
6 months	−0.1399	0.3562	54.6% [47, 75]	54.0% [43, 79]

Two things in that table are worth more than the point estimates. The shape is the daily result: at one month the rate cannot be told from zero, exactly as it cannot at one day in Table 2, and it climbs from there towards a ceiling. And the appraised column is barely different from the clean one, which is Section S15 of the Internet Appendix's equivalence holding at a frequency where it was never tested. The apparatus is not daily by nature.

What the intervals will not do is price the effect. Thirty to forty points of width establishes that it is there and cannot say what it is worth. That gap closes with history rather than with method, and the arithmetic is worth stating because it tells a future study what to look for. A bootstrap interval narrows roughly with the square root of the test window, so reaching a width of forty points from the forty-two observed here needs about 121 test months, and a width of thirty needs about 215, which with the burn-in comes to twenty-seven and thirty-five years of marks. That calculation assumes independent draws and so understates what a persistent series requires; it is a floor. It is also within reach: appraisal-based property indices and quarterly private-equity series with three or four decades of history exist, and the cross-section they would be matched against is public and free. So what was missing was a series and not a method, and Section 8 supplies the series.

S30. What a practitioner should do with this

Someone holding a stale mark on an asset with a liquid options market should prefer own-asset implied volatility for this forecasting objective, at the tenor matching their forecast horizon rather than at the headline one. Section 7 provides strong evidence for the first, and the foreign cross-section adds nothing measurable once implied volatility is in hand. Section S28 provides it for the second: for a five-day question the nine-day index beats thirty-day VIX at every delay we test, by 0.024 to 0.041 of R², which is a larger prize than most of what the rest of this paper argues over and costs nothing but choosing the right series.

Someone holding a stale mark on an asset with no options market has breadth as the only candidate on the table, and it is no corner case: it is the case Section 1 opens with. Section S13 of the Internet Appendix puts a measurement behind that sentence for the first time. Where the only options market a forecaster can see is written on another index, implied volatility and breadth come out three points apart rather than thirty-one, and a compressed cross-section adds significantly in thirteen of twenty-four cells against none of eight where the option chain is on the asset itself. Buying someone else's options market is not the substitute that buying your own is. Quarterly-marked private equity, appraised property and funds reporting a lagged net asset value have no listed option chain on the

asset itself, so the substitute that wins Section 7 is not available to any of them. For appraised property the number no longer has to be extrapolated from equities at all: Section 8 measures it on twenty metropolitan house price indices and finds 67.8% [45, 92] of the loss from a quarter's extra staleness recovered on the return and 19.0% [6, 31] on the variance, with the peer cross-section compressed to a single column, which is the same recommendation the rest of this section makes for a different reason. The redundancy result therefore narrows where the equity measurement transfers, not where the question matters. Before buying anything, though, there is a cheaper thing to do, and Section S14 of the Internet Appendix is what makes it worth saying. Correlate the stale series with the public cross-section over the days already in hand. Across eight markets that one number tracks the substitution rate at +0.98, it runs from 0.81 down to -0.12 across targets that all look like ordinary equity indices, and where it is small the cross-section repaired nothing and charged for the attempt. A holder whose asset is weakly coupled to public markets should expect breadth to do nothing for them, and can learn that before they buy a feed rather than after. Sections 4 and 5 show how that worth can be measured and give a liquid-market benchmark for it, rather than an estimate for the illiquid asset itself: about 80% of the delay-induced loss recovered in stressed markets, where the estimate is stable and tight, 81% at eleven weeks with an interval of [74, 87], and an amount this sample cannot determine in calm ones. One practical point belongs with those numbers and is easy to miss, because the paper reports it as a robustness result. The rates above come from the original seven-regressor specification, which is not the model to run. Section 6.1 shows a single real-time factor beating all seven series at every delay and raising $R(55)$ to 75.2%, and an equal-weighted mean of the seven does nearly as well for no estimation at all. The seven coefficients cost one to five points of R^2 to fit and buy nothing back. Anyone building this should compress the cross-section first and treat the seven-regressor figures as the conservative benchmark they are. The unconditional figure is 72% against the paper's own control and 68% against the stronger control of Section S4 of the Internet Appendix; we quote the conditional numbers because they are the ones that mean anything. Against that sits a cost of one to five points of R^2 for carrying the extra series. One condition comes with those figures, and Section S12 of the Internet Appendix makes it explicit: the foreign data has to be current. Every rate quoted in this paper assumes the cross-section arrives the same evening, and a feed one day behind turns Table S8 of the Internet Appendix's 36% into 9% at three days and its 51% into 29% at five. Before buying breadth, check that it will arrive on time. A cross-section older than the mark it is replacing is worth nothing, and at the shortest delays it is worth less than nothing.

One generalisation is worth refusing explicitly. It would be convenient to present Figure 2 as a benchmark for how quickly information decays in delayed systems, and it is not one. Those curves are a measurement of one index against seven specific peers over one sample. The correlation between a private portfolio and the public cross-section is a different quantity, and nothing here bounds it. A reader who wants a number for their own asset will not find it in this paper.

What does carry across is the apparatus rather than the numbers. $R(\delta)$ is defined in units that survive a change of loss function, market or candidate. The delay sweep is a design anyone with a delayed series and a candidate substitute can run, and the horse race in Section 7 is the question to ask first, because its answer decides whether breadth is worth buying at all. The cost decomposition in Section 6 separates what a source knows from what a forecaster keeps, the distinction that turned our own most confusing result into an ordinary one. Pointed at an appraisal-based or quarterly-marked series, that machinery would produce the numbers this paper cannot, and the check above says the delay model it carries is the right shape for that series, even though the numbers are not.

S31. The house price measurement in full

Section 8 of the paper states what this measurement settles and points here for the working. The measurement is below, followed by the three supporting results it rests on.

Before the measurement itself, Section S29 of the Internet Appendix establishes that the experiment is feasible at an illiquid asset's own reporting frequency, on month-end marks with no daily bars anywhere in it.

S31.2 Twenty metropolitan house price indices

Three supporting measurements are held back to the end of this section: the panel's properties measured before any rate was estimated, the full pricing of the three candidate peer blocks, and the seasonal placebo.

Two things change from Section 4 and both are worth stating plainly, because a result that moved two variables at once and reported one of them would not be worth much. The asset class changes. The target variable also changes, because a monthly index has no intraday range and Yang-Zhang cannot be computed on it at all. The two are therefore carried separately throughout. One target is next month's log return, which is the quantity an owner of the asset cares about and is not this paper's target. The other is the log of trailing twelve-month realised variance over its own five-year median. That second target needs a caveat rather than a claim of equivalence. This paper's own target is forward-looking: realised variance over the five days *after* the forecast origin, none of which the forecaster has seen. The housing one is backward-looking, and forecasting next month's value of it means forecasting a quantity whose window overlaps the current one in eleven of twelve months. An own-only R^2 near 0.95 follows from that arithmetic and is not evidence that house price variance is easy to forecast. It also shrinks the denominator $R(\delta)$ divides by, which is why the rates on this coordinate are lower and why we do not read them as the same measurement made in a second asset class. The return target is the closer analogue of the paper's, and where the two coordinates are given different weight below, that is the reason.

Delta needs saying carefully too. All twenty metros carry the same publication lag of roughly two months, so delta is not that lag. It is what it is everywhere else in this paper: the amount by which the target's own mark is staler than the cross-section it sits in. The concrete case is an owner of property in one metro whose book is appraised on a cycle running behind the published index for the others. Three months is one quarterly appraisal cycle and is the row to read first.

The peer block has to be chosen before any rate is read, and at this sample size the choice is not cosmetic. Section S28 and Section S30 already recommend compressing the cross-section to a single real-time factor rather than fitting seven separate series, for the reason Section 6 prices. The bill is far larger here: the training window is 144 months rather than 1,250 days, and on the variance target the own-only model already reaches an R^2 of 0.9497, so thirteen peer columns have almost nothing left to earn and a great deal to spend. Table S27 prices three blocks as a set, the way Section S4 of the Internet Appendix prices eight controls, and selects one by a rule fixed in advance: take the block with the highest cross-arm skill at zero delay, treat differences below 0.01 as immaterial, and break ties toward the smaller block. That statistic is measured where the own mark is fresh and there is nothing to substitute for, so it cannot favour a high rate, and the tie-break runs toward fewer regressors, which is the direction that lowers it.

Table S25. The substitution rate on fourteen metropolitan house price indices, under the peer block selected in Table S27. The delay is in months and is relative staleness, not the publication lag every metro shares: three months is one quarterly appraisal cycle. Eleven month-of-year dummies enter both arms. Intervals are moving-block bootstraps at twelve months, resampling blocks common to all fourteen metros so that one national housing cycle is not counted as fourteen independent ones. The last column counts metros whose own rate is positive.

delay	R ² own	R ² cross	R	metros with R > 0
<i>target: next month's return; 310 test months per metro, 2000-08 to 2026-05</i>				
0	0.7242	0.7404	n/a	n/a
1 month	0.5697	0.6580	160.3% [-77, 241]	14 of 14
2 months	0.3927	0.5964	66.2% [52, 86]	14 of 14
3 months	0.3504	0.5952	67.8% [45, 92]	14 of 14
6 months	0.2080	0.5730	69.1% [58, 79]	14 of 14
<i>target: next month's trailing twelve-month variance (backward-looking); 250 test months per metro, 2005-08 to 2026-05</i>				
0	0.9559	0.9544	n/a	n/a
1 month	0.8667	0.8694	0.7% [-9, 9]	6 of 14
2 months	0.7522	0.7745	10.3% [-2, 22]	10 of 14
3 months	0.6277	0.6859	19.0% [6, 31]	13 of 14
6 months	0.2692	0.4997	34.9% [19, 48]	14 of 14

Both coordinates exclude zero at a quarterly staleness. On the return target the rate is 67.8% [45, 92] and is positive on all fourteen metros. On the backward-looking variance coordinate it is 19.0% [6, 31] and positive on thirteen of fourteen, and that figure is the weaker of the two for the reason given above: most of next month's trailing variance is already in the forecaster's hands, so there is little for delay to destroy and correspondingly little for breadth to put back. The exception is Las Vegas, where the peer block does not merely fail to help but actively subtracts, taking 0.4224 down to 0.3643; its coupling is the fourth lowest on the panel, so the mechanism of Section S14 of the Internet Appendix accounts for its own exception in the direction it predicts, and no separate story is offered for it. The shape of Table 2 survives as well: at one month the rate cannot be told from zero on either coordinate, exactly as it cannot at one day in Table 2, and it climbs from there. At one month on the return target it exceeds 100%, which is Section 7's point rather than an error: a fresh cross-section can beat the target's own fresh mark, so the zero-delay own-only figure was never a ceiling.

The first objection is the index's own construction. Each published figure for month t is a three-month moving average of sales pairs closing in months $t - 2$ to t , so consecutive monthly changes are built from overlapping transaction windows, and a rate manufactured by that smoother would look exactly like a rate. Section S18 of the Internet Appendix removes every shared transaction month by reading the index only every third month, on each of the three phases of the quarterly grid, and the artefact explanation predicts zero there. Pooled across metros the rate is 57% at one quarter of staleness, spanning 49% to 63% across phases, positive in 35 of 42 metro-phases, with each phase's interval excluding zero. So the explanation is not available; nor is the rate unchanged, since 57% is visibly below 67.8%, which is what one would expect if a modest part of the monthly figure is the smoother. De-smoothing the panel says the same thing from the other side and adds a finding: with the moving average inverted out, one month of staleness costs -0.0012 of R^2 , nothing measurable and if anything the wrong sign, so no rate is quotable at all, while the peer block still adds $+0.1352$ of R^2 .

at three months on fourteen metros of fourteen. In this panel, smoothing accounts for most of what a short delay costs, which is consistent with breadth paying on assets that are marked this way. Whether that holds outside a repeat-sales index is not something fourteen metros can establish.

What this section settles, and what it does not. It settles that the central claim of this paper is not a fact about liquid equity indices. The claim holds on an asset class with no intraday range, no option chain and a mark that is smoothed by construction; it holds on the target variable the asset's owner cares about and on the one this paper was written about; it holds under a peer block chosen by a rule that could not see the rate; and it is not a calendar effect. What it does not do is price the effect more tightly than the intervals in Table S25, or test the slope of the coupling relation, for the reason given before the rates were estimated. And it remains an index study: nobody holds a Case-Shiller index, and a real appraised book would differ from one in ways this panel cannot show. What is no longer true is the sentence this paper carried until now, that the illiquid case rests on argument alone. It now rests on a measurement one step removed from the asset the argument is about.

S31.1 The panel, the peer blocks and the seasonal placebo

Section S29 established that the design survives being re-specified at monthly frequency, and closed by saying what was then missing: a series with two to four decades of marks and a public cross-section, run at the asset's own frequency. That series exists and it is free. The S&P Cotality Case-Shiller indices cover twenty United States metropolitan areas to June 2026, 9,143 published figures in all. Fourteen of them run monthly from January 1987; Minneapolis and Phoenix begin in 1989, Seattle in 1990, Atlanta and Detroit in 1991, and Dallas not until 2000. Rates are estimated on the fourteen with the full history, which is why a panel of twenty is reported as fourteen of fourteen, and the six later starters enter as peers wherever they exist. Section 9 widens the peer block to all nineteen. These indices are the object this paper has spent nine sections arguing about rather than a proxy for it. Each monthly figure is a three-month moving average of closings and closings lag contract by weeks, so the mark is smoothed and stale before anyone touches it. No option chain is written on any of the twenty, so the implied-volatility substitute that dominates Section 7 is not merely weak here, it is absent. And all twenty publish on the same day, so when one is treated as stale, nineteen fresh peer marks exist: the construction of Section 4, transplanted whole.

Before any rate was estimated, two properties of the panel were measured, and both bear on claims made earlier in this paper. The first is persistence. Section S15 of the Internet Appendix models a stale mark as partial adjustment at $\rho = 0.6$, a figure taken from the property literature, and the monthly return autocorrelation of the real indices runs from 0.601 at Cleveland to 0.939 at Phoenix with a mean of 0.804. Every metro is more smoothed than the model of it, so the synthetic illiquid case of Section S29 was a conservative stand-in and not a flattering one. The second is coupling, which Section S14 of the Internet Appendix found orders the substitution rate at +0.98 across eight equity targets. Here it spans 0.662 to 0.866 with a mean of 0.763, against 0.228 to 0.808 in Section S14 of the Internet Appendix. That has one consequence in each direction and both were committed to before the rates existed. The level prediction is sharp and testable: every metro sits at or near the top of the equity coupling range, so the mechanism predicts a high rate on all of them and no target on which substitution fails. The slope prediction is not testable here at all, because a band 0.35 times the width of the one that produced +0.98 cannot resolve a gradient, and claiming the slope held on this panel would be the same error as claiming it failed.

Table S26. The twenty-metro panel, measured before any substitution rate was estimated. Persistence is the lag-1 autocorrelation of monthly log returns over each metro's full history; Section S15 of the Internet Appendix modelled a stale mark as partial adjustment at $\rho = 0.6$, and every figure in that column exceeds it. Coupling is the correlation of the metro's own monthly return with the equal-weighted mean of its thirteen peers, the quantity Section S14 of the Internet Appendix found orders the rate. The last two columns are the substitution rate at one quarterly appraisal cycle, reported here so the level prediction can be read across the panel rather than only in the aggregate.

metro	first mark	persistence	coupling	R(3), return	R(3), variance
Boston	1987-01	0.747	0.762	62.1%	47.9%
Charlotte	1987-01	0.714	0.699	67.4%	45.1%
Chicago	1987-01	0.761	0.748	98.0%	29.8%
Cleveland	1987-01	0.601	0.662	90.1%	9.7%
Denver	1987-01	0.789	0.704	36.5%	4.6%
Las Vegas	1987-01	0.859	0.710	28.1%	-11.0%
Los Angeles	1987-01	0.903	0.838	70.3%	31.8%
Miami	1987-01	0.872	0.769	76.6%	15.8%
New York	1987-01	0.844	0.736	75.5%	26.4%
Portland	1987-01	0.768	0.738	64.0%	30.7%
San Diego	1987-01	0.835	0.816	61.2%	11.4%
San Francisco	1987-01	0.834	0.821	63.1%	7.5%
Tampa	1987-01	0.834	0.807	91.2%	9.6%
Washington	1987-01	0.843	0.866	65.8%	6.3%

Table S27. Three peer blocks priced as a set, and the rule that picks one. $S_{\text{cross}}(0)$ is the cross arm's skill at zero delay, where the own mark is fresh and there is no substitution to detect, so it cannot favour a high rate; the bill is what the block costs before any substitution is available to pay for it. The selected row is the block with the highest $S_{\text{cross}}(0)$, differences below 0.01 treated as immaterial and ties broken toward the smaller block. On the variance target the widest block makes the rate negative at three of four delays and the compressed block makes it positive at all four.

peer block	cols	$S_{\text{own}}(0)$	$S_{\text{cross}}(0)$	bill	R(1)	R(2)	R(3)	R(6)
<i>target: next month's return</i>								
every peer separately	13	0.7242	0.7489	-0.0246	167.9%	91.3%	90.7%	88.4%
peer mean, one column ← selected	1	0.7242	0.7404	-0.0161	160.3%	66.2%	67.8%	69.1%
peer mean and its 3-month mean	2	0.7242	0.7458	-0.0216	144.3%	69.8%	72.1%	72.2%
<i>target: next month's trailing twelve-month variance (backward-looking)</i>								
every peer separately	13	0.9559	0.9162	+0.0397	-88.7%	-30.6%	-1.7%	26.1%
peer mean, one column ← selected	1	0.9559	0.9544	+0.0015	0.7%	10.3%	19.0%	34.9%
peer mean and its 3-month mean	2	0.9559	0.9557	+0.0002	20.7%	28.2%	30.6%	36.0%

Table S27 carries the most consequential finding in this section and it is methodological rather than substantive. On the variance target the widest block makes the substitution rate negative at one, two and three months, at -88.7% , -30.6% and -1.7% , and the compressed block makes it positive at every delay. The zero-delay column says why: thirteen peer columns cost 0.0397 of R^2 before any substitution is available to pay for them, and a rate computed against a cross arm already carrying a bill that size is measuring the bill. Compression leaves 0.0015 of it. Section S13 of the Internet Appendix's foreign-options comparison was confounded in the same way and corrected in the same way. The lesson to carry forward is that compression is part of the method and not a robustness check: at this sample size it is the difference between the central claim holding and appearing to fail.

The rule selected the one-column block on both targets, which was not built into it, and on the return target that choice cost real points. The widest block would have reported 90.7% at a quarterly staleness where the selected one reports 67.8%. What follows is the low end of the menu in Table S27 rather than a search over it.

The second objection is seasonality. These series are not seasonally adjusted and they share a large seasonal. Month-of-year explains 19.9% of monthly return variance on average and 41.7% at Cleveland, and the twelve-month profiles correlate +0.894 across metros on average. A target whose own mark is three months stale has lost its recent history, and a fresh peer mark tells it what month it is; if that were all the peers were doing, the rate would be measuring a calendar. Eleven month-of-year dummies enter both arms of every regression above, which closes the channel if the seasonal is a fixed additive monthly effect, and that is an assumption rather than a proof: the seasonal could drift, or be multiplicative, or the dummies could be estimated too noisily on 144 months to absorb it. Table S28 tests it with a placebo instead of another control, because a control can only fail to change the answer whereas a placebo can actively produce the wrong one and convict the design.

Table S28. The seasonal placebo. The real peer mean is replaced by a series carrying the shared seasonal and nothing else: the month-of-year mean of that peer mean, estimated on the training block alone and read off by calendar month, refreshed on the model's own twelve-month schedule. It has no dependence whatever on what any peer market did in the current month. The month-of-year dummies are removed from both arms in both columns, so the placebo is given every advantage the real series has.

delay	R, real peer mean	R, seasonal placebo
<i>target: next month's return</i>		
1 month	65.2%	-2.3%
2 months	69.0%	0.9%
3 months	73.0%	4.2%
6 months	70.7%	6.9%
<i>target: next month's trailing twelve-month variance (backward-looking)</i>		
1 month	-0.3%	-3.8%
2 months	9.4%	-0.8%
3 months	18.6%	-2.8%
6 months	34.7%	-4.0%

The placebo earns 4.2% where the real peer mean earns 73.0%, and on the variance target it earns -2.8% where the real mean earns 18.6%. So the rate is information and not a calendar effect, and that

is established by a test that could have convicted the design and did not. Two smaller results sit alongside it. Removing the seasonal is the conservative choice on the return target rather than a favourable one: the untreated arm reports 73.0% against the published 67.8%, with causal deseasonalisation in between at 71.0%. And on the variance target the three treatments are indistinguishable at 19.0%, 16.0% and 18.6%, for a structural reason rather than an empirical one, since a twelve-month rolling variance necessarily spans every calendar month and a fixed month-of-year effect largely cancels inside the target before any model sees it.

Widening the cross-section from thirteen peers to nineteen, six of which begin between 1989 and 2000 and are therefore genuinely ragged, moves the rate from 67.8% to 67.5% on the return target and from 19.0% to 19.9% on the variance target. Nothing is lost, which is what compressing to a mean is for, and nothing is gained either, which is the more informative half. Six additional metros move the rate by less than a point, so thirteen peers already span whatever common factor the cross-section carries: that is the prediction of Section S27 and of the factor model in Section S8 of the Internet Appendix holding on a second asset class, and it is of a piece with seasonal profiles that correlate +0.894 and couplings that sit in a band 0.2 wide. There is one national housing cycle in this panel, not nineteen independent signals.

S32. Robustness and scope in full

Read as a clock, the trade has a price. A fifty-five-day-old domestic mark plus today's foreign closes forecasts as well as a plain domestic mark 4.6 days old [2.5, 8.4]. The same construction prices the other side of the trade: one day of staleness in the foreign feed costs 2.1 days [1.4, 3.3] of domestic freshness, and a feed one day behind turns the 51% recovered at five days into 29%, and the 36% at three days into 9%. Every rate in this paper assumes the cross-section arrives the same evening, and a cross-section older than the mark it replaces is worth less than nothing.

S33. The limitations in full

The implied-volatility block is thinner than we would like. VIX and VDAX cover the sample; a spot VSTOXX index was obtained, tested and kept out of the headline block, because requiring it costs 73% of the sample, moving the window from 4,525 days beginning September 2008 to 1,240 beginning April 2020: the 2008 crisis is discarded outright and peak VIX falls from 82.7 over the full sample to 41.4 inside what remains. On the 1,240 days where all three exist, holding those days fixed so the third series is not confounded with the shorter and calmer window, it adds nothing at any delay and is mildly negative at all six: VSTOXX and VDAX are both European equity volatility indices and the extra column costs more to estimate than it returns. The conclusion a fuller block was meant to test is therefore unaffected. What remains open is the export rather than the conclusion, since it runs only from December 2012 to March 2025.

Two bounds on the tests themselves. The non-linearity result of Section S28 fits a random Fourier layer, which is a one-hidden-layer network whose first layer is drawn rather than learned; no network with a learned first layer was fitted, and the argument that one would not change the answer is an extension of the estimation-cost finding rather than a test of it. And Section S27's first principal component of seven equity indices is a proxy for global systemic risk, not a measurement of it, so the remainder it leaves may still carry global content this basis cannot see: the direction of that test is informative and its magnitude is not a bound. Section S28 names a limitation of a different kind, and

it is the one a practitioner should read first. What breadth repairs is a stale forecaster's tendency to *over-estimate* variance, not their exposure to under-estimating it; the understating tail is unmoved at every delay we checked. Anyone reaching for the cross-section as protection against being caught short by a volatility spike is reaching for the wrong thing.

The largest of these limits is the estimator, and we can be specific about it rather than gesturing, because we fitted the alternative. Every model here is a ridge on a fixed feature set, chosen because it is transparent and because the object under study is the substitution rate rather than the best attainable forecast. The natural competitor is a one-factor state-space model with a Kalman filter that handles a missing observation by dropping that row from the measurement equation, which is what a delayed domestic mark *is* rather than something imposed on the data. Fitted two-step after Doz, Giannone and Reichlin (2011), it beats this paper's specification at $\delta = 21$ and $\delta = 55$ and loses at $\delta = 0$, and the crossover is the interesting part: imposing a one-factor structure costs skill when the mark is fresh and pays when it is stale. Section S20 of the Internet Appendix gives the intervals and the reconstruction diagnostics.

We report it rather than adopting it, and the reason is a scope judgement a reader is entitled to disagree with: $R(\delta)$ is a ratio in which the same domestic control appears three times, so a better estimator raises numerator and denominator together, and recomputing every conditional and every robustness check inside a filtered specification is a different paper. What matters for the headline is that the direction runs the opposite way to the usual limitation. Inside the filtered specification the rate at eleven weeks becomes 75.4% against the published 71.6%, a paired difference of +3.8% [+0.8%, +7.7%], and the mark reads 3.7 days old [2.0, 6.7] rather than 4.6. A better estimator raises the numerator of $R(\delta)$ and has nothing to give its denominator, so this paper is not protecting a large number with a transparent estimator; it is quoting a smaller one than the better estimator would give it. The 75.4% is the upper end of the 68% to 75% span the abstract reports.

S34. Further work in full

The way it breaks is the more interesting outcome, and it is specific enough to look for. Commodity volatility has drivers that no equity cross-section can see: an inventory report, a harvest, a production decision by a cartel. If a commodity block adds where equity breadth does not, then “a global volatility factor seen early” is too coarse a description of what is happening, and the right cross-section depends on what the stale asset IS, and is not one panel for everyone. A farmland book and a private credit book would then need different peers, which is a materially different recommendation from the one Section S30 gives and would be worth the trouble of discovering.

Two warnings for whoever runs it. The models are the easy part; the data hygiene is not. Futures require contract rolls, so a continuous series carries a seam every quarter instead of the single seam per export audited in Section 4, and the audit, not the estimation, is where that work will actually go. And the admissibility table stops being a list of closing times: a bond future and its underlying cash market can trade on different calendars, so the rule that a peer counts only when its timestamp precedes the target's needs restating instrument by instrument instead of country by country.

S35. The admissibility rule on real session clocks

Section 3 states the rule the whole design rests on: a foreign close is admissible on day t only if its timestamp falls strictly before the target's close on that day. The rule is implemented from a table of

fixed UTC closing times, and those are standard-time closes. Real exchanges keep local time, most of them observe daylight saving, and they do not switch on the same dates: the United States advances its clocks on the second Sunday in March and the European Union on the last Sunday, so for a fortnight each spring New York sits an hour closer to London than the table says. Until now the paper asserted that this never reorders a block. An assertion about a clock is a poor substitute for reading one, and this section reads them.

Every close in the panel is rebuilt in UTC from its exchange's own regular-session closing time in its own timezone, with the IANA database supplying every transition date, on all 7021 dates in the sample. Two decisions are then compared, for every target, every peer and every date: the lag the fixed table assigns, and the lag the true closes imply. A disagreement in one direction is a *leak*, where the table calls a peer same-day admissible and the clock says it closed later or at the same moment, so the panel would carry a value the forecaster could not have seen. A disagreement in the other is *waste*, where the table sends a peer back a day that had in fact already closed.

There are **0** leaks. Across every target, every peer and every date, not one day on which the true session clocks would have reordered a block. The margin is not marginal either: the narrowest gap the rule anywhere relies on is **2.00 hours, between N225 and HSI**, so the fixed table is a simplification of the clock rather than an approximation of the rule. Daylight saving moves four of the eight closes by exactly one hour and moves none of them past a target.

The waste runs the other way and is worth stating because of its direction. On 3345 of the 7,021 dates the table sends AXJO back a day for a N225 forecaster when the true clocks say it had already closed, and 6099 such days occur across the panel as a whole. The cause is that the table records the two markets at the same standard-time hour and the rule resolves equality against admissibility. So the implementation is *stricter* than the rule it implements: it discards information a real forecaster could lawfully have used. That direction is the one to prefer, and it also changes how the paper's numbers should be read. Every rate reported here is what a forecaster achieves with less than the admissible cross-section, which makes the measured substitution rate a lower bound on the rule's own terms rather than a figure flattered by a generous clock.

Holidays are the second assumption in the same family and they cannot leak, since a peer whose exchange was shut carries a close that is strictly older than the day it sits on. What they can do is weaken the instrument, so the frequency is reported rather than assumed: the peer carried forward most often is N225, on 5.8% of the target's days, and the mean age of a carried close is under three days. The foreign block is a block of recent closes, not a block of stale ones.

S36. The outage, split into information and estimation

Section 4.3 names what δ freezes and points here for the measurement. The training cut of equation (4) is forced by the vintage assumption: the label is the target's own decision variable, so a forecaster whose own mark is δ days old has not seen the outcomes that resolved since. The consequence, which an external audit put more sharply than an earlier objection had, is that δ does not isolate a stale input. It freezes the current domestic state, the labels available for fitting, the coefficient vintage, the ridge penalty the validation tail selects, and the quantity of training data, all at once. The design measures a domestic-data *outage*, and the question this section answers is how much of the damage belongs to each half.

The split needs an arm that no forecaster could run, and it is reported as such. Three arms share the feature construction, the estimator, the refit schedule and the test days, and differ only in where training stops.

Table S29. What δ costs, split. The feature-only arm holds the input δ days stale while cutting training at $t - h$, so its coefficients are as fresh as the estimator can make them. It is infeasible under this paper's own information set, since training to $t - h$ needs the stretch of the target's own series the delay withholds, and it exists to price one effect rather than to be reported as a forecast. Both other arms are the paper's.

δ	$S_{\text{own}}(0)$	feature only	outage	total damage	feature	vintage	vintage share
5	0.5191	0.3646	0.3641	0.1550	0.1544	0.0005	0.4%
13	0.5191	0.2347	0.2336	0.2855	0.2844	0.0011	0.4%
21	0.5191	0.1516	0.1482	0.3709	0.3675	0.0034	0.9%
55	0.5191	0.0207	0.0070	0.5121	0.4984	0.0137	2.7%

The answer is not close. Freezing the estimation accounts for **0.4%** of the delay's damage at one week and **2.7%** at eleven, and the two intermediate delays sit between them. What δ costs a forecaster is almost entirely the staleness of the input; the inability to re-estimate, re-tune and re-fit on more data is a rounding term against it. So the denominator $R(\delta)$ divides by is information, which is what the ratio was meant to be about.

The rate itself is recomputed inside the feature-only arm, with the cross-sectional model given the same cut so the comparison stays internal. The two rates differ by at most **2.8 points** across the grid, and the feature-only rate is the *higher* of the two at every delay, so the paper's design is also the conservative one: a forecaster who could keep re-fitting while their input went stale would find breadth repairing slightly more, not less.

What this settles is a naming question and not a result. The paper describes the experiment as an outage because that is what it is, and a reader who wants the share of stale-input damage that breadth repairs can take the feature-only column and will be within three points of the headline. What it does not settle is whether a real desk suffers outages or stale marks; that depends on the desk, and the two coincide for the case Section 1 opens with, where the mark and the data feed behind it stop together.

S37. The secondary grids, and what a reader may lean on · *Scope narrowed*

Section S18 prices the search behind the headline: nine specifications, a Reality Check and an SPA test, and the headline survives both. That covers the one comparison this paper was designed around and it covers nothing else. After the headline specification was fixed we went on to report several wide secondary grids: seven markets by four delays for the compressed foreign block in Table S23, five optioned and four unoptioned targets by four delays each in the whose-options-market work of Sections 9.4 and 9.5, four nonlinear arms by six delays in Table S17, and a third implied-volatility series at six delays. Individual cells were marked significant inside them at the ordinary 5% level. None of those grids was pre-specified and no family-wise control was applied across them. A reader is entitled to know how many of those cells would look significant if every one were noise, and the answer is a number rather than a caveat.

Table S30. The post-selection audit. Each family is a grid reported after the headline specification was fixed.

“Expected” is $0.05n$, the count a grid of pure noise would show at the same threshold. The last two columns pool every secondary cell in the paper and apply a Bonferroni and a Benjamini-Hochberg correction across the pool; the headline comparison is excluded from the pool, because it carries its own correction in Section S18 and including it would let these grids dilute a test that was never part of them.

family	cells	significant at 5%	expected by chance	survive BH	survive Bonferroni
compressed foreign block given implied volatility, seven markets	64	20	3.2	1	0
seven foreign closes given implied volatility, five optioned markets	32	7	1.6	1	0
breadth given implied volatility, four unoptioned targets	64	21	3.2	2	2
nonlinearity given implied volatility, four arms	48	31	2.4	1	0
VSTOXX as a third implied-volatility series	6	0	0.3	0	0
pooled	214	79	10.7	5	2

Two readings are available and both are true. Seventy-nine nominally significant cells against 10.7 expected is not what noise produces, so the secondary grids collectively carry signal and are worth reporting. But only five of those seventy-nine survive a 5% false-discovery correction across the secondary set, and only two survive a family-wise one. The seventy-four that do not are *exploratory*: real enough to tabulate, not firm enough to carry a claim on their own, and that is the word this paper uses for them from here on. Every table above whose cells fall in this audit is marked accordingly in its caption.

The instrument matters and it is chosen rather than defaulted to. Bonferroni controls the probability of any false rejection at all and is the right test if a single cell is to be leaned on; Benjamini-Hochberg controls the expected share of rejections that are false and is the right test for a grid read as a pattern, which is how these grids are actually read. We report both and lean on neither individually: the claim these grids support is the pattern, which survives, and not any one of its cells, which mostly does not.

Nothing here touches the headline, and the exclusion is deliberate rather than convenient. R(55) on the implementable target is a single pre-specified comparison, its multiplicity is priced in Section S18 against the nine specifications actually searched, and it survives Reality Check and SPA at five of six delays. Pooling it into a set of two hundred secondary cells would not be a stricter test of it; it would be a different and weaker one, in which a grid the headline was never part of decides the threshold the headline must clear.