S&P 500 Dealer Gamma and the Next-Session Range of the Ibovespa: A Registered Forecast Test That Fell Just Short of Its Mark

Fernando Varani Espoladore Independent researcher This version: 24 November 2026

Abstract

We ask whether a published estimate of S&P 500 options-dealer gamma exposure (GEX), taken from the last US session before the target session, improves forecasts of how wide the next Ibovespa session will be. The forecast is for the size of the range only, not its direction. The data are daily Ibovespa high/low ranges from 2010 to September 2026, the Cboe Brazil ETF volatility index (VXEWZ) and the SqueezeMetrics daily GEX series. The baseline is a HAR model of the log day range plus log VXEWZ. The candidate adds the trailing one-year percentile of GEX and a flag for negative GEX. Both models are re-estimated on an expanding window, so every forecast uses only earlier rows. Both are scored by squared log error on the same 3,411 sessions (January 2013 to September 2026). The test and its six pass marks were written down in the author's own registration record before the outcomes were computed. That record is not public and has no outside timestamp. The registration records two tries: one counting step that compared no forecast errors, and one outcome look. An earlier, redesigned version of the same idea adds a third try for the idea as a whole. The candidate's loss was 0.9917 of the baseline's, a 0.83% reduction. The registered Newey–West t of the per-session loss difference was −2.11 (one-sided normal $p \approx 0.017$). Because the candidate nests the baseline, this p-value is only approximate. Five of the six marks passed, including a placebo test at the smallest p-value its design allows (0.007). The first mark, a loss ratio of 0.99 or lower, failed. **The registered test KILLED the claim.** The data show a small improvement that did not reach the registered size. The statistics are consistent with a non-zero effect but do not establish one, and they do not show a useful forecasting gain.

Keywords: Ibovespa; day range; volatility forecasting; dealer gamma exposure; GEX; VXEWZ; HAR model; nested forecast comparison; pre-registration; negative result

Introduction

When options dealers are net short gamma, their hedging trades go with the market: they buy as prices rise and sell as they fall. This is widely argued to widen intraday ranges in US equity indices. Non-resident participants trade a large share of Brazilian equities, and the Ibovespa moves closely with US stocks on the same day. That raises a plain question. Does the US dealer-gamma state at the last US close before a Brazilian session add anything to a forecast of that session's range, beyond what Brazil's own recent ranges and a market-implied volatility already contain?

Registration. In this paper, "the registration" (called "the card" in the author's records) means a dated text written before the test. It fixes the claim, data, timing rules, models, loss, statistics, control, guard, pass marks and predicted outcomes. It also logs every look at the data with its numbers and the resulting status. **The registration is not public.** It is held only in the author's private records and has no outside timestamp: no OSF or AsPredicted entry, and no public commit. The registration time (24 November 2026, 08:45:25 CET), the two analysis steps that followed within a minute, and this paper all fall on the same day. Nothing outside the author's own record backs up that order. Readers should treat the claim that the marks were fixed before the look as the author's statement, not as something they can check. Its full content is reproduced in the Method section.

History of the idea, and a disclosed redesign. This study is the second registered version of the idea. The first version used Bovespa mini-index futures over 2022–2026. It was registered about four hours earlier on the same day and stopped at its timing guard before any forecast error was computed. The same-day correlation with the S&P 500 was the peak (0.322), but it fell short of a fixed level of 0.35. The present version then made two changes. It switched to the cash index from 2010, which gives about three and a half times the sample. It also replaced the fixed level with a rule on the *shape* of the correlation peak. The second change was made after seeing that the first version failed the fixed level, so it is a researcher degree of freedom informed by an earlier result. Counting that first version, the idea has had **three tries** in total: one for the first registration and two for this one. The redesign, the registration text and the analysis script were drafted by an AI assistant and checked by a second AI review, without step-by-step human review.

The registration predicted only a small gain, because VXEWZ already carries part of the US stress signal and low GEX tends to come with high implied volatility. It predicted a loss ratio of 0.97–0.99 (most likely about 0.985) and put the chance of clearing every pass mark at about one in four to one in five.

Data

- **Ibovespa daily bars**, January 2010 to 30 September 2026, from public daily index data (Yahoo Finance, symbol ^BVSP). After cleaning, 4,151 sessions remain. Where a date appears twice, the last row is kept. Rows that repeat the previous close with zero

volume are treated as holidays and dropped. High > low > 0 and close > 0 are required. The day range is rng = ln(high/low), and the modelled quantity is its logarithm, ln(rng).

- **VXEWZ**, the Cboe implied-volatility index on the iShares MSCI Brazil ETF (FRED series VXEWZCLS, available from 16 March 2011), in logs. It is quoted in US dollars, so it includes exchange-rate volatility. It is a weaker stand-in for an implied volatility on the Ibovespa itself. The Ibovespa-specific alternative the author had was an implied volatility computed from options with strikes near the current price on BOVA11, the iShares Ibovespa ETF listed on B3, using B3's public daily quotes file (COTAHIST). It is a series the author computed, not a published index, and the author's copy starts only in January 2021. Published Brazilian implied-volatility indices were not examined.
- **S&P 500 dealer gamma exposure (GEX)** from the SqueezeMetrics daily DIX/GEX file, which starts on 2 May 2011. GEX is the vendor's own model estimate. Two features are built on the file's own calendar. *gpct* is the share of the last 252 GEX values (all 252 required, the current one included) that are at or below the current value. *gneg* equals 1 when GEX is negative.

Timing. For a target session T, each outside input (VXEWZ, gpct, gneg) is the value with the last date strictly before T. It is used only if it is at most five calendar days old; otherwise the row is dropped. The registration also described B3's cash close as aligned with New York's 16:00 close. That is not true for the whole sample. The gap between the two closes has changed with the US daylight-saving calendar, with Brazil's own daylight saving (in force until 2019) and with B3's choice of trading hours. The rule above makes the exact alignment unimportant for the forecast. Every input is dated on an earlier calendar day than T, so the underlying market data exist before B3 opens on T, whatever the close times were. (A real-time version would also depend on the vendor posting GEX before B3 opens; see the limits below.)

Vintage of GEX. The GEX file was downloaded as a single current file, not collected day by day. Nothing shows that its historical values are point-in-time. The vendor may have recomputed or revised past values since they were first published. The study therefore assumes, without evidence, that the historical values match what was available at each date. If they were backfilled, the results could differ in either direction from what a real-time user would have seen.

From sessions to model rows. The 4,151 cleaned sessions become 3,575 model rows, so 576 rows are dropped:

- The binding constraint at the start is gpct. It needs 252 GEX values, and the GEX file starts in May 2011, so gpct first exists around the start of May 2012. Every Ibovespa session from January 2010 up to that point drops out. By calendar arithmetic this is roughly 575 sessions, and it also covers the 21 sessions the 22-session average needs.
- The last session (30 September 2026) has no next session to forecast and is dropped.

- VXEWZ starts in March 2011, earlier than gpct, so it does not bind at the start. Its coverage on the judged window was 100%.
- The script does not log a step-by-step breakdown. A small number of rows dropped by the five-day staleness rule therefore cannot be ruled out, but the start-up period explains nearly all of the 576.

From model rows to judged targets. The 3,575 rows give 3,411 judged targets, a gap of 164. The first 150 rows (from about May 2012) are warm-up and get no forecast. The next 14 rows do get forecasts, but their targets fall in December 2012, before the judged window opens on 1 January 2013. This is exact: 150 + 14 = 164. The first judged target is 2 January 2013. The pre-2020 half has 1,730 targets and the half from 2020 has 1,681. GEX was negative on 309 model rows, 297 of them judged.

Method

All of the following was fixed in the registration before the outcome look.

Both models are ordinary least squares on logs. Row t is an Ibovespa session; its target T is the next Ibovespa session. Let l1 be the log range of session t, and l5 and l22 the averages of the log range over sessions t−4..t and t−21..t. Let iv be the log of the VXEWZ value with the last date strictly before T (at most five calendar days old). It is *not* a value from session T itself. gpct and gneg are taken in the same way, from the last GEX date strictly before T.

- **Baseline (HAR + implied):** y = const + l1 + l5 + l22 + iv
- **Candidate:** baseline + gpct + gneg

Here y is the log range of session T. Both models are refit on an expanding window: the forecast for row i uses only rows before i, from row 150 on. The loss is the squared error of y. This is one choice among several possible losses, such as QLIKE or squared error on the range itself. All results here apply to this registered loss only, and no other loss was looked at. As a simple reference, the study also scores the log of the mean range over the last 20 sessions (avg20).

Statistics. The headline number is the *loss ratio*: the candidate's summed loss divided by the baseline's, where below 1 means better. The registered significance statistic is a Newey–West t (Bartlett kernel, 10 lags) of the per-session loss difference, candidate minus baseline, where negative means better (t_NW). For context the registration also reports a *naive t*, which uses the plain standard error with no autocorrelation correction. It is not a Diebold–Mariano statistic in the sense of Diebold and Mariano (1995), which uses a long-run variance, and it carries no mark. P-values are one-sided and come from the standard normal.

Nested models. The candidate nests the baseline, and both are re-estimated on an expanding window. In that setting, under the null of no gain, a t statistic on the squared-error difference does not follow the standard normal. The larger model pays for estimating extra parameters, which shifts the loss difference against it. Normal critical values are then typically conservative (Clark and McCracken, 2001). Clark and West (2007) propose

an adjusted statistic for this case. It was not registered and has not been computed for this version. The normal p-values below are therefore approximate. If anything they are likely to overstate the p-value, but they are not exact.

Control (placebo). The gpct and gneg columns are shifted together circularly by k rows over all model rows, for k = 252, 273, ..., up to n − 252 in steps of 21. That gives 147 shifts. The candidate is refit each time against the same baseline losses. The placebo p-value is (1 + the number of shifts with a ratio at or below the real one) / (1 + the number of shifts).

Timing guard (checked before any forecast error). This is the rank correlation of the daily Ibovespa close-to-close log change at date d with the S&P 500 log change at d − 1, d and d + 1, from May 2011. It passes only if the same-day correlation is the largest and beats both neighbours by at least 0.15. If it fails, no forecast loss is computed.

Power rule (from error spreads only). Before the look, the smallest loss ratio the test could detect at t = 2 with 80% power was estimated. The estimate used the spread and autocorrelation of the baseline's expanding errors and 20 randomly shifted copies of gpct (fixed seed). It never used a mean difference between the two models' losses. This counting step did compute the baseline's errors, but only their spread entered the rule.

Count marks (checked before the outcome look). All seven had to hold: at least 3,000 judged targets; at least 1,400 in the smaller half; the first judged target within 10 days of 1 January 2013; at least 100 negative-GEX rows; VXEWZ coverage of at least 0.97 on the judged window; the timing guard passing; and a detectable loss ratio (from the power rule) of at least 0.975.

Pass marks for the outcome look. The registration lists six marks, and all six had to hold:

1. loss ratio ≤ 0.99;
2. t_NW ≤ −2;
3. loss ratio below 1 in both halves (before 2020 and from 2020), counted as one mark;
4. a negative gpct coefficient in the final fit;
5. placebo p ≤ 0.05;
6. candidate loss below the avg20 reference.

Tries. The registration records **2 tries**: the counting step, which ran two seconds after registration, and the single outcome look, about thirty seconds after registration. There were no repeated outcome looks and no amendments to the marks. With the earlier version described in the Introduction, the idea has had three tries. Before registration, the script was run on synthetic data. On pure noise it failed a mark, and with a planted GEX effect it passed them all. A later planned stage, scoring the frozen model on sessions after registration, was never reached.

Results

Count step. All seven count marks passed (Table 1).

Table 1. Count marks, checked before the outcome look.

Count mark	Threshold	Observed	Outcome
Judged targets	≥ 3,000	3,411	pass
Smaller half	≥ 1,400	1,681 (other half 1,730)	pass
Days from 1 Jan 2013 to first judged target	≤ 10	1	pass
Negative-GEX model rows	≥ 100	309 (297 judged)	pass
VXEWZ coverage, judged window	≥ 0.97	1.000	pass
Timing guard: same-day peak, gap to neighbours	peak at d, gap ≥ 0.15	0.436 vs −0.015 (d−1), −0.023 (d+1); gap 0.451	pass
Detectable loss ratio (power rule)	≥ 0.975	0.9893	pass

Outcome look. Table 2 gives the outcome look against the six registered marks.

Table 2. The outcome look against its six registered pass marks (3,411 judged sessions, 2 January 2013 to 30 September 2026).

#	Measure	Value	Registered mark	Outcome
1	Loss ratio, candidate / HAR + VXEWZ	0.9917	≤ 0.99	**fail**
2	Newey–West t (10 lags) of loss difference	−2.11 (one-sided normal p ≈ 0.017, approximate)	≤ −2	pass
3	Loss ratio in each half: 2013–2019 (n = 1,730) / 2020–2026 (n = 1,681)	0.9943 / 0.9891	both < 1	pass
4	gpct coefficient, final fit	−0.112	< 0	pass
5	Placebo p (147 circular shifts)	0.007	≤ 0.05	pass
6	Loss ratio, candidate / avg20 reference	0.8948	< 1	pass
—	Naive t, no autocorrelation correction (reported, no mark)	−2.28 (one-sided normal p ≈ 0.011)	—	—

Five of the six registered marks passed, and the first failed. Under the registered rule the claim is therefore killed. The candidate cut the baseline's squared log error by 0.83%; the registration asked for at least 1%.

Other numbers, not marks.

- *Placebo.* The 147 shifted GEX series had a median ratio of 1.0019, about what two useless extra regressors cost. None did as well as the real series. The placebo p of 0.007 is the smallest value this design can produce, 1/148, so the test cannot say how far into the tail the real series lies. The circular shifts move the GEX features across market periods. They cannot tell whether one period, such as 2020, drives the result.
- *Halves.* Within the halves, the Newey–West t was −0.99 for 2013–2019 and −2.02 for 2020–2026. Most of the statistical evidence comes from the later half.
- *Reference.* The baseline alone had a loss ratio of 0.9023 against the avg20 reference.
- *Coefficients.* The VXEWZ coefficient fell from 0.237 in the baseline to 0.182 once GEX was added. That fits the registered expectation that the two inputs overlap.
- *Sign of gneg.* In the final fit the coefficient on negative GEX was **+0.050**. Its sign was not a pass mark; the registration predicted it would be near zero, and +0.050 is a modest departure from that. Because the dependent variable is the log range, a positive coefficient means a *wider* forecast range on negative-GEX days. Taken alone, this sign therefore goes the same way as the motivating mechanism (negative gamma widening ranges), not against it. A sign check against the mechanism would have needed a negative coefficient on gpct, which is what was observed. The size of the gneg coefficient should still not be read on its own. Negative GEX is rare (309 of 3,575 rows), so negative-GEX days typically sit low in the trailing GEX distribution, where gpct is also low. The two regressors are collinear, and their separate coefficients are not well identified. As an illustration, a negative-GEX day with gpct near zero, against a median day (gpct = 0.5), gets about $-0.112 \times (0 - 0.5) + 0.050 \approx +0.11$ in the forecast log range. This decomposition is an interpretation; the correlation between the two regressors was not computed.

Table 3. Loss ratio of the candidate over the baseline by calendar year of the target session (below 1 = candidate better).

Year	Ratio	Year	Ratio
2013	0.9988	2020	0.9488
2014	0.9802	2021	1.0002
2015	1.0179	2022	1.0176
2016	0.9820	2023	0.9809
2017	0.9665	2024	1.0034
2018	1.0232	2025	0.9908
2019	1.0001	2026 (to Sep.)	0.9834

The candidate did better in 8 of 14 years and worse in 6 of 14: most clearly in 2015, 2018 and 2022, and barely in 2019, 2021 and 2024. The largest single-year gain was in 2020 (0.9488). The recorded numbers give only per-year ratios, not per-year loss totals, so a ratio for the whole sample without 2020 cannot be derived from them. That ratio was not regis-

tered and has not been computed. It would be the most useful single check of how much the result depends on 2020.

Table 4. Registered predictions against outcomes.

Quantity	Predicted at registration	Observed
Loss ratio	0.97–0.99 (likeliest ≈ 0.985)	0.9917
t_NW	−1.5 to −3.5	−2.11
gpct coefficient	−0.05 to −0.15	−0.112
gneg coefficient (no mark)	near 0	+0.050
Halves	both a little below 1, 2020–26 better	0.9943 / 0.9891
Placebo p	0.01–0.2	0.007 (design minimum)
Loss ratio vs avg20	≈ 0.85–0.92	0.8948
Judged sessions	≈ 3,400	3,411
Detectable loss ratio	≈ 0.98	0.9893

What this does and does not show

What it shows. The registered test killed the claim. On 3,411 Ibovespa sessions, adding S&P 500 dealer-gamma features to a HAR model with Brazil-ETF implied volatility reduced squared log forecast error by 0.83%. That is below the 1% minimum the registration set. It is also slightly smaller than 0.9893, the smallest ratio the power rule said this sample could reliably detect. The other numbers mostly point the same way: a Newey–West t near −2.1, both halves below 1, the expected sign on gpct, and no placebo shift doing as well. **The statistics are consistent with a non-zero effect, but they do not establish one.** The normal p-values are only approximate for nested models. The placebo p sits at its floor. The evidence is concentrated from 2020 on, and the 2013–2019 half alone has $t = -0.99$.

What it does not show.

- It does not show that US dealer gamma is a useful input for Ibovespa range forecasts. A gain under 1% in squared log error is small, and the registered rule rejects it.
- The registration is not public and has no outside timestamp, so readers cannot check that the marks came before the look.
- The timing-guard rule was redesigned after an earlier version of the idea failed the fixed level. The p-values here describe this single registered look. They are not adjusted for the three tries the idea has had, or for that informed redesign.
- The test is a pseudo-out-of-sample evaluation on historical data, not a test on data unseen by the author. Every forecast used only earlier rows, but the window ends before the registration date. The author had also examined the Ibovespa's day ranges and returns over this period in earlier, unrelated studies (calendar effects, volatility risk premium). GEX had not previously been used with the Ibovespa.

- The historical GEX values may not be point-in-time (see Data). A real-time version would also require the vendor's evening post to arrive before B3's 10:00 opening.
- The loss is squared error of the log range only. Other losses were not examined, and the 0.83% figure applies to this loss alone.
- VXEWZ is a dollar-based implied volatility. A better Ibovespa implied-volatility measure, such as one from BOVA11 options (available to the author only from 2021), might absorb some or all of the GEX contribution.
- The role of 2020 has not been isolated (no leave-2020-out ratio has been computed), and 6 of 14 calendar years went the wrong way.
- Nothing here concerns the direction of prices, and nothing here is a recommendation to trade any instrument.

Reproducibility

The study is fully specified above: data, feature timing, models, loss, statistics, placebo, timing guard, power rule, count marks and pass marks. The three input series are public: daily Ibovespa bars (Yahoo Finance, ^BVSP), FRED series VXEWZCLS, and the SqueezeMetrics DIX/GEX daily file, each cut at 30 September 2026. The models use ordinary least squares and the Newey–West variance in plain numerical code, with no tuned hyperparameters. The placebo shifts are deterministic, and the power rule uses the fixed seed 20261124.

Determinism check, not a replication. The unchanged script was rerun on the same data snapshot. All 16 recorded outcome numbers matched exactly, and the script had not changed. This shows only that the computation is deterministic. It is not an independent replication: it used no new data, no new code and no other analyst. **The script and the exact data snapshot are not publicly available** at this version. Because the public sources can revise past values, notably the vendor's GEX, a reader who downloads the series today may get slightly different numbers.

Parts of the analysis and the writing were done with an AI assistant. The hypothesis, the redesign, the pass marks and the analysis script were drafted with AI assistance and, by the author's record, fixed before the outcome look. The marks were not amended afterwards.

References

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1), 85–110.

Clark, T. E., & 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.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

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