

Regime Splits Are Not Interaction Tests: Spurious Amplification in Mixed-Frequency Volatility Models

Fuli Yang

GARCH Institute, Harvard Square, Cambridge, MA, USA

Email: dr@garch.ac · ORCID: 0009-0005-4055-7561

This version: August 2026

Abstract

Applied volatility research routinely infers state dependence by estimating a model separately in subsamples defined by a conditioning variable and comparing slopes. This paper shows that in mixed-frequency volatility models such a comparison is not an interaction test, for two independent reasons. First, the conditioning variable typically enters the long-run variance component directly; omitting it makes each group-specific slope a pseudo-true quasi-maximum-likelihood parameter rather than a structural marginal effect. The resulting distortion does not diminish with sample size. Second, selecting non-contiguous daily observations by a monthly partition breaks the GJR-GARCH recursion at every omitted block, so the split-sample objective is not the likelihood of a valid GARCH process.

We apply the argument to daily commodity futures returns, the New York Fed Global Supply Chain Pressure Index, and the Caldara-Iacoviello geopolitical-risk index. The consequences are not second-order. For copper, the conventional split reports a supply-chain slope 0.069 higher in high-geopolitical-risk months, but that positive difference exists only under the concatenation implementation used in applied work: re-initialising the recursion at each block gives -0.165 and filtering on the full daily sample gives -0.134 . The reported amplification is a sign flip away from what coherent implementations deliver. In a unified specification with both direct effects and their monthly product under a shared Beta weight, the interaction is significantly negative for copper ($\hat{\theta}_3 = -0.1002$, $p = 0.005$) and cocoa (-0.4635 , $p < 0.001$). A first-order decomposition attributes the split difference to the omitted direct effect: across copper, coffee, and cocoa the observed difference and its predicted omitted-variable component correlate at 0.997, and the largest, most threshold-stable apparent amplification occurs in soybeans, the one market where supply-chain pressure has no detectable effect on long-run volatility at all. In 1,000-replication placebo simulations with the true interaction fixed at zero throughout, the naive split-based Wald test rejects at 14 to 25 percent against a nominal 5 percent, and its rejection rate is nearly flat in the true interaction, whereas the unified likelihood-ratio test reaches 86 percent power at the estimated parameters. The same argument applies to splits on uncertainty, financial stress, liquidity, and policy states.

Keywords: GARCH-MIDAS; sample splitting; interaction effects; pseudo-true parameters; commodity volatility; geopolitical risk

JEL classification: C22; C52; C58; G13; Q02

1. Introduction

A familiar empirical sequence appears across finance, commodity economics, and macroeconomics. A researcher estimates a relationship in a low-risk subsample and then estimates the same relationship in a high-risk subsample. The high-risk coefficient is larger. The difference is described as an amplification effect: policy uncertainty amplifies the market impact of monetary surprises; financial stress amplifies the pass-through of macroeconomic news; geopolitical risk amplifies the volatility consequences of supply disruptions.

The economic intuition may be entirely reasonable. The inferential step is not.

It is important to be precise about why, because a weaker version of this point is well known. In a linear regression, splitting the sample on a binary indicator is numerically identical to estimating a single model in which that indicator is interacted with every regressor and with the intercept. Sample splitting is, in that setting, a reparameterisation rather than a distinct estimator. The equivalence fails in the setting studied here for three reasons, each specific to the mixed-frequency volatility problem.

First, the conditioning variable is continuous and is itself a determinant of the outcome. Once the sample is truncated on it, its direct effect cannot be identified within either group, so the saturated-interaction representation that delivers the linear equivalence is unavailable. Second, the object being modelled is a latent multiplicative variance component estimated by quasi-maximum likelihood on daily data. The split is therefore not a reparameterisation of a common likelihood, and the group coefficients are pseudo-true parameters in the sense of White (1982) rather than least-squares projection coefficients. Third, the discretisation is applied at monthly frequency to a daily recursive process, which destroys the recursion itself.

This paper studies the problem in the GARCH-MIDAS setting, which is designed for an attractive and increasingly common task: explaining daily volatility with low-frequency macroeconomic variables. It is natural to use the Global Supply Chain Pressure Index (GSCPI) as a driver of long-run commodity volatility, and equally natural to ask whether that relationship changes when geopolitical risk is elevated. The standard approach divides daily returns into high- and low-GPR observations using the upper tercile of the monthly geopolitical-risk index, estimates separate GJR-GARCH-MIDAS models in each group, and reads the difference in GSCPI slopes as amplification.

That conclusion does not follow from the design, and in these data it does not survive contact with it. The paper makes four contributions.

A pseudo-true characterisation of split-sample GARCH-MIDAS coefficients. The result is not the textbook omitted-variable identity. Because the long-run component is latent and estimation proceeds by quasi-maximum likelihood on daily returns, the pseudo-true parameter solves an exponentially weighted orthogonality condition; the familiar linear expression emerges only as a local approximation. Two corollaries follow, the second of which shows that the high-minus-low difference fails to recover the interaction even when the conditioning variable has no direct effect, because the interaction enters scaled by a within-group projection difference that is not equal to one.

A distinct and, to our knowledge, previously unstated defect. A monthly partition applied to a daily recursive variance process breaks the GJR-GARCH recursion at every omitted block. In this application the high-GPR subsample consists of 83 months distributed across 24 disjoint blocks, ten of them a single month long. Section 5.1 shows that this is not a technicality: the sign of the headline statistic depends on how the resulting gaps are handled.

A quantitative account of a specific empirical result. We re-examine the application using a common long-run component containing both direct effects and their interaction, and decompose the observed split difference into its constituent channels using the estimated parameters and a placebo simulation calibrated to the observed regressors.

A constructive diagnostic. Any researcher holding a split table can compute two within-group projection coefficients from auxiliary regressions and obtain a first-order estimate of the mechanical component of the observed slope difference before drawing any conclusion about state dependence. In these data that calculation, which requires no volatility modelling, accounts for essentially the entire reported amplification effect.

The purpose is not to argue that state dependence is unimportant. Section 7 reports a state-dependent leverage effect that survives every test we apply to it, estimated inside a single likelihood. The purpose is to establish that state dependence must be estimated as state dependence.

Section 2 places the argument in the literature. Section 3 develops the theory. Section 4 describes the data and estimation protocol. Section 5 contrasts split-sample and unified results and presents the decomposition. Section 6 presents the placebo, size, and power simulations. Section 7 reports the leverage result. Section 8 provides robustness checks. Section 9 concludes.

2. Related Literature

2.1 *Mixed-frequency volatility models*

The GARCH-MIDAS framework decomposes conditional variance into a transitory component governed by daily return innovations and a long-run component driven by weighted lags of monthly or quarterly variables. Engle, Ghysels, and Sohn (2013) formalised the idea that macroeconomic variables can explain the low-frequency component of market volatility without forcing those variables into a daily-frequency model, and Conrad and Loch (2015) show that forward-looking macroeconomic expectations carry substantial information about the long-run component. Wang and Ghysels (2015) establish the asymptotic properties of quasi-maximum-likelihood estimation in volatility component models, and Conrad and Kleen (2020) analyse the multiplicative-component specification used here. These results supply the regularity framework within which the pseudo-true argument of Section 3 operates.

Subsequent work has extended the framework to alternative volatility processes, asset classes, and state-dependent environments. Regime-switching GARCH-MIDAS models (Pan, Wang, Wu, and Yin, 2017) show that it is technically possible to let long-run dynamics differ across latent or observed states. This matters here because the question is not whether a unified state-dependent specification is conceptually available. It is. The practical obstacle is estimation: interaction terms, nonlinear MIDAS weights, asymmetric GARCH dynamics, and short samples can generate flat likelihood surfaces or local optima. These difficulties are the usual explanation for reverting to subsample estimation. They do not change what the subsample estimates identify, and Section 6.4 shows that in this application the unified test has ample power.

2.2 *Interactions, omitted variables, and sample splitting*

An interaction model asks whether the marginal effect of one covariate changes with another. In a linear setting that question is answered by including both main effects and their product, and a binary split is equivalent to a saturated version of that model. Ai and Norton (2003) show that the equivalence and the interpretation of the product term both break down in nonlinear models, where the cross-partial derivative of the conditional mean is not the coefficient on the product term.

Feigenberg, Ost, and Qureshi (2025) show that estimated interaction coefficients are biased when an interacted confounder is omitted from a model that already contains the interaction of interest. The present paper addresses a different failure. Our concern is not bias in an estimated interaction coefficient but the gap between

the split-sample estimand and the interaction estimand in a nonlinear latent-component model, where no interaction is estimated at all. The results are complementary: theirs concerns what an included interaction measures, ours concerns whether a slope comparison measures an interaction.

Hansen (1996, 2000) develops inference for threshold models in which the threshold is estimated jointly with the slope parameters; this is the defensible route when a researcher genuinely believes the state is discrete, and it differs sharply from imposing a tercile cut by convention. Amado and Teräsvirta (2013) develop smooth-transition conditional variance models, the natural continuous-state alternative. Leeb and Pötscher (2005) analyse the consequences of conditioning inference on a data-dependent model-selection step, relevant whenever the threshold is chosen after inspecting results.

White (1982) provides the framework used in Section 3: under misspecification, quasi-maximum-likelihood estimators converge to the parameter minimising the Kullback-Leibler discrepancy between the true and the assumed conditional density. Everything that follows applies that idea to a partitioned mixed-frequency volatility likelihood.

2.3 Commodity risk and supply chains

The GSCPI (Benigno, Di Giovanni, Groen, and Noble, 2022) aggregates delivery-time indicators and transport-cost measures into a broad gauge of global supply-chain pressure. The geopolitical-risk index of Caldara and Iacoviello (2022) captures newspaper-based variation in concerns about wars, terrorism, military threats, and geopolitical actions. Both can plausibly affect commodity volatility through separate channels: physical scarcity and transport disruption on one side; uncertainty, sanctions risk, policy response, and changes in hedging demand on the other.

Those mechanisms give a reason to expect direct effects from both variables, and a reason to expect correlation between them during global shocks. That combination is exactly why a GPR-defined split is hazardous. The more economically credible it is that geopolitical risk affects long-run volatility directly, the less credible it is to omit it from each subgroup equation and read the residual slope difference as an interaction.

3. Theory

3.1 Unified mixed-frequency specification

Let r_t be the daily return on day t and $m(t)$ its calendar month. Conditional variance is decomposed as

$$r_t = \mu + \sqrt{\tau_{m(t)} h_t} \eta_t, \eta_t \sim i.i.d.(0, 1), \varepsilon_t \equiv r_t - \mu,$$

with the short-run component following the GJR-GARCH recursion

$$h_t = \omega + \left[\alpha + \gamma 1(\varepsilon_{t-1} < 0) \right] \frac{\varepsilon_{t-1}^2}{\tau_{m(t-1)}} + \beta h_{t-1}.$$

Because τ_m and h_t enter multiplicatively, the intercepts ω and θ_0 below are not separately identified. We impose

$$\omega = 1 - \alpha - \beta - 1/2\gamma,$$

so that $E[h_t] = 1$ and the level of conditional variance is carried entirely by τ_m . Covariance stationarity requires $\alpha + \beta + 1/2\gamma < 1$; non-negativity of the news impact requires $\alpha \geq 0$ and $\alpha + \gamma \geq 0$.

Let g_m be the monthly GSCPI and p_m the standardised monthly GPR index. With $K = 12$ monthly lags and one-parameter Beta weights

$$\varphi_k(\omega_2) = \frac{[(K+1-k)/(K+1)]^{\omega_2-1}}{\sum_{j=1}^K [(K+1-j)/(K+1)]^{\omega_2-1}}, \omega_2 > 1,$$

which impose monotonically declining weights, the long-run component is

$$\log \tau_m = \theta_0 + \theta_g \sum_{k=1}^K \varphi_k(\omega_2) g_{m-k} + \theta_p \sum_{k=1}^K \varphi_k(\omega_2) p_{m-k} + \theta_3 \sum_{k=1}^K \varphi_k(\omega_2) (g_{m-k} p_{m-k}).$$

Three features require comment.

The interaction regressor is the MIDAS-weighted **monthly product** $g_{m-k} p_{m-k}$, not the product of separately aggregated GSCPI and GPR measures. Weighting the product preserves the monthly timing of the joint realisation, which is the economically meaningful event; multiplying two aggregates would allow a high-GSCPI month to interact with a high-GPR month twelve months apart.

The three low-frequency terms share the same weight parameter ω_2 . This ensures that the two direct effects and the interaction are evaluated over a common effective lag distribution, preventing the interaction from acquiring spurious flexibility through a separate lag shape. Section 8.2 reports the unrestricted alternative.

GPR is standardised; GSCPI is retained in published units. This asymmetry is deliberate. The GSCPI is constructed in standard-deviation units relative to its own historical average, so re-standardising would be redundant, and $g_m = 0$ has the interpretable meaning of normal supply-chain conditions. Writing G_m, P_m, Q_m for the three MIDAS aggregates, the marginal effects are

$$\frac{\partial \log \tau_m}{\partial g_{m-k}} = \varphi_k(\omega_2) (\theta_g + \theta_3 p_{m-k}), \quad \frac{\partial \log \tau_m}{\partial p_{m-k}} = \varphi_k(\omega_2) (\theta_p + \theta_3 g_{m-k}).$$

Thus θ_g is the marginal effect of supply-chain pressure at average geopolitical risk, since p_m is mean zero, and θ_p is the effect of geopolitical risk at normal supply-chain conditions. At the sample mean $\hat{g} = 0.1583$ the latter is $\theta_p + 0.1583 \theta_3$.

The parameter of interest is θ_3 , the cross-partial up to the weight φ_k . Because the specification is additive in logarithms, the benchmark $\theta_3 = 0$ is already multiplicative in variance levels; a negative θ_3 indicates a sub-multiplicative joint effect, not a dampening one.

The first expression also identifies what a split design is attempting to approximate. Under the correct model the average marginal effect within group S is approximately $\theta_g + \theta_3 \dot{P}_S$, so the difference a valid split would recover is $\theta_3 (\dot{P}_H - \dot{P}_L)$. Section 5.3 compares that benchmark with what the split design delivers.

3.2 Split-sample estimation targets a pseudo-true parameter

The standard design partitions daily observations by whether their calendar month is in the upper GPR tercile and estimates the restricted long-run equation separately in each group, omitting both the direct geopolitical-risk term and the interaction from both subgroup likelihoods.

Because τ_m is latent and estimation is by quasi-maximum likelihood on daily observations, the split coefficient cannot be derived from a least-squares projection of an observed $\log \tau_m$. No such regression exists: given the low-frequency regressors, the long-run equation is deterministic. The relevant object is the White (1982) pseudo-true parameter. We state the result under a simplification that isolates the long-run block.

Assumption 1. The short-run component is treated as known and normalised, $h_t \equiv 1$, so the quasi-score for the long-run parameters excludes the recursive term $\partial h_t / \partial \theta$.

Assumption 1 is not innocuous and we do not conceal it. In the estimated model the recursion is fed by $\varepsilon_{t-1}^2 / \tau_{m(t-1)}$, so a misspecified long-run component propagates into α , γ , and β . Relaxing the assumption adds terms to the condition below; it does not restore identification of θ_3^0 , and Section 6 quantifies the additional contamination by simulation.

Proposition 1 (Pseudo-true split coefficient). Consider a true process with long-run component $\log \tau_m^0 = \theta_0^0 + \theta_g^0 G_m + \theta_p^0 P_m + \theta_3^0 Q_m$. Let a researcher select a GPR-defined subsample S and estimate $\log \tau_m^* = a_S + \theta_{g,S}^* G_m$. Under Assumption 1 and the regularity conditions of White (1982) and Wang and Ghysels (2015), the estimator converges in probability to $(a_S, \theta_{g,S}^*)$ solving

$$E_S[e^{\Delta_m} - 1] = 0, E_S[(e^{\Delta_m} - 1) G_m] = 0,$$

where $\Delta_m = \log \tau_m^0 - \log \tau_m^* = c_S + (\theta_g^0 - \theta_{g,S}^*) G_m + \theta_p^0 P_m + \theta_3^0 Q_m$ and $c_S = \theta_0^0 - a_S$. Consequently $\theta_{g,H}^* - \theta_{g,L}^*$ is not, in general, an estimator of θ_3^0 .

The two conditions jointly determine the two unknowns. The weighting is exponential, not linear: observations with large Δ_m receive disproportionate weight, so the pseudo-true value is not a projection coefficient of any moment of the data.

Corollary 1 (Local approximation). When Δ_m is small in the relevant range, $e^{\Delta_m} - 1 \approx \Delta_m$ and

$$\theta_{g,S}^* \approx \theta_g^0 + \theta_p^0 \lambda_S^P + \theta_3^0 \lambda_S^Q, \lambda_S^j = \frac{\text{Cov}_S(G_m, j_m)}{\text{Var}_S(G_m)}, j \in \{P, Q\},$$

so that

$$\theta_{g,H}^* - \theta_{g,L}^* \approx \theta_p^0 (\lambda_H^P - \lambda_L^P) + \theta_3^0 (\lambda_H^Q - \lambda_L^Q).$$

The first term is the omitted direct effect; the second is the true interaction, rescaled by a projection difference that need not equal one. The familiar omitted-variable expression is therefore only a first-order approximation to the pseudo-true target, not the exact estimand in a GARCH-MIDAS likelihood. Section 6.3 measures how far the approximation degrades as $|\theta_p^0|$ grows.

Corollary 2 (The split difference does not recover θ_3^0 even without a direct effect). If $\theta_p^0 = 0$, then $\theta_{g,H}^* - \theta_{g,L}^* \approx \theta_3^0 (\lambda_H^Q - \lambda_L^Q)$, and that projection difference equals one only by coincidence. The split difference

recovers the interaction only up to an unknown scale determined by the selection mechanism — a property of the joint distribution of the MIDAS regressors in the selected groups, not a structural parameter.

Corollary 2 removes the last available defence. Even a researcher willing to assume away the direct effect cannot interpret the magnitude, and in general not the sign, of a split difference as an interaction. In these data $\lambda_H^Q - \lambda_L^Q$ ranges from 0.80 to 2.25 across empirically reasonable weighting and threshold choices (Section 5.3).

3.3 Conditions under which a split interpretation would be valid

The split statistic equals the interaction only if

$$\theta_p^0(\lambda_H^P - \lambda_L^P) = 0, \lambda_H^Q - \lambda_L^Q = 1, \Delta_m \approx 0,$$

and, in addition, only if the split likelihood is a valid likelihood, which Section 3.5 shows it is not. The first condition requires either that geopolitical risk has no direct effect or that its within-group projection on supply-chain pressure is identical across groups selected on geopolitical risk. Neither is plausible. The second is a knife-edge restriction on a selection-dependent quantity. The third fails precisely when the direct effect is large enough to matter.

3.4 Why conventional diagnostics do not repair the design

Researchers may check whether the coefficients differ, whether confidence intervals overlap, or whether a Wald statistic rejects equality. None addresses the estimand problem. They can establish that the two pseudo-true coefficients differ; they cannot establish that the difference is an interaction. Section 6.4 shows that in this design the Wald test does not even do the former reliably, rejecting at up to five times its nominal rate.

The problem persists in arbitrarily large samples. Additional data estimate the group-specific pseudo-true coefficients more precisely without removing the omitted terms. Precision is not identification.

The appropriate diagnostic is structural: estimate the unified model, include the direct effect, and test $\theta_3 = 0$. Where a unified model cannot be identified or stably optimised, the correct conclusion is that the available evidence does not identify the state-dependent effect.

3.5 Regime selection breaks the GJR-GARCH recursion

The omitted-variable problem is sufficient to invalidate the amplification interpretation. Section 3.2, however, grants the split design its most favourable reading by assuming that the subgroup objective has a well-defined pseudo-true limit at all. That concession is generous.

The GJR-GARCH recursion is defined over consecutive daily observations. A monthly tercile rule retains daily returns from selected calendar-month blocks and discards every observation from the excluded blocks, so the retained observations are not a contiguous series. At each transition between retained blocks, h_{t-1} and ε_{t-1} are unobserved within the selected sample, and $\tau_{m(t-1)}$ is undefined under the subgroup model because the omitted month has no fitted long-run value.

The practice admits three implementations, which are not equally defensible.

1. **Concatenate** and treat retained observations as adjacent. This splices unrelated volatility levels across gaps and injects an artificial shock at every boundary. It is the implementation ordinarily used.

2. **Re-initialise** the recursion at each block. Internally coherent, but it estimates persistence from short blocks, and the initialisation error is not negligible when blocks are short relative to the memory of the process.
3. **Filter on the full daily sample** and evaluate the likelihood only on retained days, extrapolating the fitted long-run equation to omitted months. This is a legitimate conditioning operation and the correct way to implement a partial-sample evaluation. It does not solve the estimand problem of Section 3.2, but it is not itself invalid.

Only the third is coherent, and it is not the one ordinarily used. Under variants 1 and 2 the subgroup objective is not merely a misspecified long-run equation; it is not a GARCH likelihood at all. This point is distinct from omitted-variable bias: the latter concerns what the split coefficient estimates, the recursion problem concerns whether the estimator is based on a coherent conditional-variance process.

The magnitude is easily quantified. Over the 246 months surviving the twelve-lag requirement, the upper-GPR tercile comprises **83 months distributed across 24 disjoint blocks**, with a mean run of 3.46 months, a median of 2, **ten single-month blocks**, and a longest run of 15. At roughly 21 trading days per month the high-GPR group holds about 1,740 daily observations interrupted by 23 recursion breaks. With persistence near $\beta = 0.94$ the effective memory of the short-run component is approximately $1/(1 - \beta) \approx 16$ trading days, so the breaks contaminate on the order of 370 observations, more than one fifth of the subsample. In the ten single-month blocks a sixteen-day burn-in consumes three quarters of the block. Under a median split the fragmentation is worse in absolute terms: 122 months in 31 blocks with 15 singletons.

Section 5.1 shows that this is not a theoretical concern. The sign of the headline copper statistic depends on which of the three implementations is used.

3.6 A diagnostic for applied work

Corollary 1 yields a screen that requires no additional estimation of the volatility model.

1. Estimate the unified specification once and retain $\hat{\theta}_p$ and $\hat{\theta}_3$.
2. Construct G_m , P_m , Q_m and compute $\hat{\lambda}_S^P$, $\hat{\lambda}_S^Q$ by ordinary least squares within each selected group.
3. Form $\hat{\theta}_p(\hat{\lambda}_H^P - \hat{\lambda}_L^P)$ and $\hat{\theta}_3(\hat{\lambda}_H^Q - \hat{\lambda}_L^Q)$ and compare with the observed split difference.
4. Report $\hat{\lambda}_H^Q - \hat{\lambda}_L^Q$ explicitly. A value far from one shows that the split statistic cannot be read as an interaction even in magnitude.

Where step 1 is infeasible the screen can be run with the direct effect alone, which is the dominant channel when the conditioning variable is a broad macro-financial risk measure. Section 5.3 applies the full procedure and Section 5.5 shows that it predicts the cross-section of split differences almost exactly.

4. Data and Design

4.1 Asset series

The analysis uses daily closing prices for four commodity futures: copper, coffee, cocoa, and soybeans. Copper is the primary application because the industrial-metals case motivates the supply-chain mechanism most directly. Daily log returns are $r_t = 100[\log P_t - \log P_{t-1}]$.

The commodity extension is a falsification design, not a robustness afterthought. Copper is closely linked to industrial production, transport infrastructure, inventories, and global manufacturing. Coffee and cocoa are agricultural commodities exposed to weather, shipping, and concentrated origin-country disruption. Soybeans combine agricultural supply conditions with trade-policy sensitivity and freight exposure. If the split difference reflected an economic amplification mechanism, its cross-sectional pattern should track these physical supply structures. If it is mechanical, it should track $\hat{\theta}_p(\lambda_H^P - \lambda_L^P)$, whose projection component is common across assets because it depends only on the low-frequency regressors. Section 5.5 evaluates the two predictions and they are cleanly separated.

The copper series is COMEX High Grade Copper in U.S. dollars per pound. Contract, exchange, units, and vintage are recorded in the replication package. The analysis makes no claims that depend on London Metal Exchange trading conditions.

4.2 Supply-chain pressure

The Global Supply Chain Pressure Index is obtained from the Federal Reserve Bank of New York and combines delivery-time measures, transportation costs, and other global logistics indicators, expressed in standard deviations from its historical average.

The low-frequency sample runs from January 2005 through June 2026, giving 258 monthly observations. After twelve monthly MIDAS lags the estimation window begins in January 2006 and comprises 246 usable months. In the vintage used here the sample mean is 0.1583 and the maximum is 4.4379 in December 2021. Because the GSCPI is revised retrospectively, the exact vintage is archived in the replication package.

4.3 Geopolitical risk

Geopolitical risk is measured by the monthly global GPR index of Caldara and Iacoviello, with the GPRA and GPRT subindices used in robustness exercises. Over the estimation sample the mean level is 103.6 and the maximum is 330.9 in March 2026. GPR is standardised before entering the unified specification.

For split comparisons, high risk is the upper GPR tercile, reproducing the standard design; median and upper-quartile partitions are reported as sensitivity checks. The paper's argument does not depend on the threshold, and Section 5.4 shows that the dependence of the results on the threshold is itself part of the evidence.

4.4 Preliminary relationships

The two low-frequency variables are related but not collinear. Their full-sample correlation is 0.126 — large enough that a split on one materially changes the conditional distribution of the other, small enough that multicollinearity alone does not explain the results.

The selection mechanism is visible directly. Under the upper-tercile rule, mean GSCPI is 0.269 in high-GPR months against 0.158 in the rest of the sample; under the upper quartile the gap widens to 0.399 against 0.127; under a median split the ordering reverses, 0.118 against 0.273. The composition of the selected samples therefore does not vary monotonically with the threshold, the first indication that split statistics computed at different cut-offs are not measuring a common structural quantity.

Daily returns are matched to monthly aggregates by calendar month; each daily observation inherits the MIDAS aggregates of its month. The estimation samples contain 5,149 daily returns for copper, 5,150 for coffee, 5,150 for cocoa, and 5,154 for soybeans.

4.5 Estimation protocol

All nonlinear estimates use a multi-start protocol. The GARCH-MIDAS likelihood can contain local optima, particularly when the weight parameter, three long-run slopes, and the asymmetry coefficient are all free. A single starting value is insufficient evidence of a global optimum.

The protocol evaluates six starting values for unrestricted models and four for restricted models — the first at a fixed centre, the remainder drawn from a fixed-seed Gaussian perturbation of it — and retains the feasible solution with the highest likelihood. Optimisation uses the Nelder-Mead simplex with inequality constraints imposed by a penalty barrier. The gap between the best and second-best log-likelihood is recorded for every specification in the convergence log supplied with the replication package.

Standard errors are computed from the numerically differentiated Hessian of the quasi-log-likelihood. Because the penalty barrier makes that Hessian unreliable near a constraint boundary, we validate it in two ways in Section 5.2, and we report a case in which a widely used bootstrap alternative fails for a reason the paper itself explains.

This protocol is not a technical footnote. A paper arguing that numerical fragility can masquerade as economic state dependence must hold its own estimates to that standard.

5. Empirical Results

5.1 The split-sample finding is not implementation-invariant

Partitioning by the upper tercile of geopolitical risk produces a larger estimated GSCPI slope in the high-risk group for copper, cocoa, and soybeans. Read descriptively, this says the estimated coefficient differs across GPR-selected samples. It does not establish why — and Table 1 shows it does not even establish that the difference is positive.

Table 1. Sensitivity of the split statistic to the recursion implementation. Upper-GPR-tercile split; $\hat{\theta}_g^H - \hat{\theta}_g^L$; six starting values per fit.

Asset	Sample start	Concatenate	Re-initialise	Full-sample filter
Copper	2006-01	+0.0687	−0.1651	−0.1342
Copper	2010-01	−0.0492	−0.2616	−0.1984
Cocoa	2006-01	+0.2370	−0.0830	−0.0065
Cocoa	2010-01	+0.3008	−0.0416	+0.3455

The result reported in applied work is the first cell of each block. For copper, the positive difference exists only under concatenation. Both coherent implementations return a negative difference of roughly twice the magnitude, and the coherent full-sample filter, -0.1342 , is close to the benchmark $\theta_3(\dot{P}_H - \dot{P}_L) = -0.1183$ implied by the unified estimates at the estimated weight (Section 5.3). For cocoa the concatenated statistic $+0.2370$ falls to -0.0830 and -0.0065 .

Three points follow. First, the recursion problem of Section 3.5 is not second-order; it is the proximate source of the reported sign. Second, the statistic is not robust to a start-date change that alters nothing about the

design: copper's concatenated difference is +0.0687 from 2006 and -0.0492 from 2010, and cocoa's filtered difference moves from -0.0065 to $+0.3455$. Third, an estimator whose headline statistic changes sign with the handling of discarded observations is not estimating a well-defined feature of the data-generating process.

A further feature of the design deserves emphasis. The two subsample fits re-estimate the weight parameter separately, and they do not agree: for copper at the tercile threshold $\hat{\omega}_2$ is 1.37 in the high group and 2.54 in the low group; for soybeans, 1.05 and 5.75; for cocoa at the upper quartile, 290 and 1.70. The group coefficients are therefore slopes on differently filtered regressors, and their difference is not a comparison of marginal effects with respect to a common variable.

We nevertheless designate the concatenated 2006-01 estimate as the headline observed statistic throughout, because it is the object of the critique. For copper that value is $+0.0687$ and for cocoa $+0.2370$.

5.2 The unified model reverses the split-sample conclusion

The central empirical result is not that the unified interaction coefficient is negative. It is that the unified model rejects the no-interaction null in the direction opposite to the split-sample narrative.

Table 2. Unified GJR-GARCH-MIDAS estimates. Hessian standard errors in parentheses. p_{MIDAS} tests $\theta_g = 0$ against the no-MIDAS benchmark; p_{direct} tests $\theta_p = 0$; p_{int} tests $\theta_3 = 0$. All likelihood-ratio tests have one degree of freedom.

	Copper	Coffee	Cocoa	Soybean
Daily observations	5,149	5,150	5,150	5,154
$\hat{\theta}_g$ (GSCPI only)	0.1649 (0.0849)	0.0836 (0.0499)	0.2058 (0.0925)	0.0700 (0.0958)
p_{MIDAS}	0.070	0.122	0.029	0.477
$\hat{\theta}_g$ (unified)	0.0420 (0.0579)	0.0347 (0.0531)	0.1706 (0.0826)	0.0353 (0.0828)
$\hat{\theta}_p$ (unified)	0.3200 (0.0707)	0.1172 (0.0872)	0.8996 (0.2215)	-0.3226 (0.1400)
$\hat{\theta}_3$ (unified)	-0.1002 (0.0344)	0.0830 (0.0659)	-0.4635 (0.1021)	0.2013 (0.0882)
p_{direct}	<0.001	0.011	0.023	0.102
p_{int}	0.005	0.179	<0.001	0.071
$\hat{\omega}_2$	25.79	1.53	1.08	2.05
$\hat{\alpha}, \hat{\gamma}, \hat{\beta}$	0.047, 0.008, 0.938	0.050, -0.039, 0.949	0.032, -0.015, 0.973	0.078, -0.015, 0.917

For copper, the constant-slope GSCPI coefficient is 0.1649. Once geopolitical risk enters the long-run component directly, it falls by three quarters to 0.0420 while the direct GPR coefficient is 0.3200. That collapse is the omitted-variable channel operating in plain view: in the restricted specification the GSCPI slope was absorbing the direct effect of a correlated omitted regressor. In the full specification $\hat{\theta}_3 = -0.1002$,

significant at the one percent level and carrying the opposite sign to the conclusion implied by the regime split. Cocoa gives the same qualitative result more sharply, with $\hat{\theta}_p = 0.8996$ and $\hat{\theta}_3 = -0.4635$ ($p < 0.001$).

These are therefore not cases in which an apparently positive effect becomes statistically uncertain after proper specification. They are cases in which the split procedure reports positive amplification while the correctly specified model supports a significant negative interaction. This is a sign reversal, not an attenuation.

Coffee and soybeans do not support a robust interaction of either sign. We report them in full rather than setting them aside, because the cross-sectional pattern is itself evidence (Section 5.5). Note in particular that for soybeans the GSCPI has no detectable long-run effect at all ($p_{\text{MIDAS}} = 0.477$) and the estimated direct GPR effect is negative.

Validating the standard errors. Because the Hessian is computed under a penalty barrier, we check it two ways. A parametric bootstrap that simulates 400 samples from each fitted model on the observed trading-day calendar and re-estimates gives a sampling standard deviation of $\hat{\theta}_3$ of 0.0411 for copper — against a Hessian value of 0.0344 — with a mean estimate of -0.1059 against a true -0.1002 , so the estimator is essentially unbiased and the Hessian understates dispersion by about a fifth. For cocoa the parametric standard deviation is 0.1790 against a Hessian 0.1021, and the estimator shrinks toward zero (mean -0.3143 against -0.4635). Using the parametric standard deviations, the interaction remains significant for both: $z = -2.44$ for copper and $z = -2.59$ for cocoa. All inference reported as significant survives this correction, and Section 6.4 shows the likelihood-ratio test attains 86 to 87 percent power at these parameter values.

A moving-block bootstrap over daily observations, by contrast, returns standard errors two to six times larger (0.1995 for copper), under which nothing in Table 2 is significant. We report this because the reason it fails is the subject of the paper. Resampling blocks of daily observations imposes a recursion break at every block boundary — roughly twenty per replicate at a 250-day block length — and simultaneously destroys the monthly variation from which the long-run parameters are identified. It is, in miniature, the same error as the split design of Section 3.5, and it is not a valid variance estimator for this model. We use the parametric bootstrap as the reference and flag the block bootstrap as a diagnostic that should not be applied to mixed-frequency component models.

On the identification of $\hat{\theta}_g$. In the unified specification the direct GSCPI coefficient is not individually significant for copper ($z = 0.73$). This matters for interpretation: the point at which the total marginal effect $\theta_g + \theta_3 p$ changes sign is not estimable with any useful precision, and we make no claim about it. What the data identify is the direct geopolitical-risk effect and the interaction, not the decomposition of the supply-chain effect at zero geopolitical risk.

5.3 Decomposing the split difference

Section 3.6 supplies a screen that converts the theoretical identity into a quantitative account. Because λ_S^P and λ_S^Q depend only on the low-frequency regressors, they are common across commodities and can be computed exactly.

Table 3. Within-group projection coefficients. Upper-tercile GPR split, 246 usable months (83 high, 163 low), $K = 12$.

ω_2	λ_H^P	λ_L^P	$\lambda_H^P - \lambda_L^P$	λ_H^Q	λ_L^Q	$\lambda_H^Q - \lambda_L^Q$
1.05	0.0789	-0.1258	0.2047	0.5068	-0.4963	1.0031

ω_2	λ_H^P	λ_L^P	$\lambda_H^P - \lambda_L^P$	λ_H^Q	λ_L^Q	$\lambda_H^Q - \lambda_L^Q$
1.50	0.1058	-0.1370	0.2428	0.5839	-0.5473	1.1312
3.00	0.1748	-0.1395	0.3142	0.7763	-0.6079	1.3842
10.00	0.2846	-0.1169	0.4015	1.0575	-0.6378	1.6953
25.79	0.3045	-0.1029	0.4074	1.1491	-0.6495	1.7986

Across all three thresholds and the full weighting grid, $\lambda_H^P - \lambda_L^P$ ranges from 0.089 to 0.463 and $\lambda_H^Q - \lambda_L^Q$ from 0.795 to 2.253.

Three results follow.

The projection difference has the sign required to manufacture apparent amplification. The within-group projection of geopolitical risk on supply-chain pressure is positive in the high-GPR group and negative in the low-GPR group at every weighting. Selection on GPR does not merely shift the level of GPR across groups; it reverses the sign of its within-group relationship with GSCPI. This is the mechanism of Section 3.2 made visible.

The omitted direct effect alone accounts for the observed statistic. At $\omega_2 = 1.5$ and the copper estimate $\hat{\theta}_p = 0.3200$, the omitted-direct-effect channel predicts $0.3200 \times 0.2428 = +0.0777$, against an observed $+0.0687$. For cocoa, $0.8996 \times 0.2428 = +0.2185$ against an observed $+0.2370$. A researcher applying only step 3 of the diagnostic in Section 3.6, before estimating any interaction, would have found that the entire observed amplification statistic is accounted for by a channel with nothing to do with state dependence.

The interaction enters rescaled, and the scale factor is not one. The quantity $\lambda_H^Q - \lambda_L^Q$ is 1.13 at the baseline weighting and varies by nearly a factor of three across defensible specification choices. Corollary 2 is therefore not a theoretical curiosity: even if the direct effect were zero, the split difference would recover θ_3^0 multiplied by a factor a researcher cannot know without estimating the unified model first.

Adding both channels at $\omega_2 = 1.5$ gives a predicted difference of $0.0777 - 0.1134 = -0.0357$ for copper and $0.2185 - 0.5243 = -0.3058$ for cocoa, against observed values of $+0.0687$ and $+0.2370$. The first-order approximation reproduces the magnitude of the omitted-direct-effect channel but does not reconcile the total, and we do not present it as if it did. Corollary 1 holds only locally; with $\hat{\theta}_p$ of 0.32 and 0.90 the deviation Δ_m is not small; the two groups re-estimate different weight parameters; and Section 3.5 establishes that the observed statistic is not the maximiser of a valid likelihood in the first place. Section 6.3 measures the approximation error directly by simulation and finds that the analytical prediction overstates the true mechanical bias by roughly 85 percent at $\theta_p = 0.32$, so the simulated bias, the analytical prediction, and the observed statistic are all of the same order and the same sign. What the decomposition establishes is sufficient: both contaminating channels are of the same order as, or larger than, the entire observed statistic, and neither is an interaction.

For comparison, a valid split design would recover $\theta_3(\dot{P}_H - \dot{P}_L)$, which at the estimated weights is -0.1183 for copper and -0.3678 for cocoa. The split design delivers a positive statistic where the correctly specified model implies a negative one of comparable magnitude.

5.4 Threshold sensitivity

Table 4. Split differences across thresholds. $\hat{\theta}_g^H - \hat{\theta}_g^L$, concatenation implementation; Wald z from the two subsample Hessian standard errors where both are available.

Asset	Median	Upper tercile	Upper quartile
Copper	+0.1228	+0.0687	+0.2161
Coffee	+0.0498	-0.0155	+0.0034
Cocoa	+0.4234 ($z = 3.65$)	+0.2370 ($z = 2.17$)	+0.0580 ($z = 0.46$)
Soybean	+0.3116 ($z = 2.77$)	+0.3001 ($z = 2.81$)	+0.3163 ($z = 2.90$)

Copper's difference is non-monotone in the threshold, falling from 0.123 at the median to 0.069 at the tercile and rising to 0.216 at the quartile. Cocoa's declines monotonically from 0.423 to 0.058 and loses significance entirely at the most extreme cut-off — the opposite of what an amplification mechanism predicts, since the high-risk group becomes more extreme as the threshold rises.

This is what the decomposition anticipates. The split difference depends on $\lambda_H - \lambda_L$, not on the mean level of GPR in each group, and within-group covariances need not vary monotonically with the cut-off. The composition evidence reinforces the point: moving from a tercile to a median split reverses the ordering of mean GSCPI across the two groups. A statistic whose underlying sample composition reverses with the threshold cannot be measuring a structural marginal effect.

5.5 Cross-commodity evidence

Section 4.1 posed a falsification test. If the split difference reflects an economic mechanism operating through physical supply chains, its cross-sectional pattern should track market structure, storability, and exposure to trade disruption. If it is mechanical, it should track $\hat{\theta}_p$ scaled by the common projection difference.

Table 5. Observed split difference against its predicted omitted-variable component. Upper tercile, $\omega_2 = 1.5$, predicted value $\hat{\theta}_p \times 0.2428$.

Asset	$\hat{\theta}_p$	Observed	Predicted	Residual
Copper	0.3200	+0.0687	+0.0777	-0.0090
Coffee	0.1172	-0.0155	+0.0285	-0.0440
Cocoa	0.8996	+0.2370	+0.2185	+0.0186
Soybean	-0.3226	+0.3001	-0.0783	+0.3784

For copper, coffee, and cocoa the correspondence is close to exact: the correlation between the observed split difference and its predicted omitted-variable component is 0.997, and the residuals are 0.009, 0.044, and 0.019. Three markets with entirely different physical supply structures line up on a single line whose slope is a property of the low-frequency regressors, not of any commodity. That is the signature of a mechanical artefact, and it is not what an economic amplification mechanism would produce.

Soybeans break the prediction, and the way they break it is more damaging to the amplification interpretation than agreement would have been. Soybeans display the largest apparent amplification in the cross-section, +0.30, and it is the only one stable across all three thresholds and significant at every one of them ($z = 2.77, 2.81, 2.90$). It is therefore the single most impressive split-sample result in the study. It occurs in the one market where the unified model finds no supply-chain effect on long-run volatility to amplify ($p_{\text{MIDAS}} = 0.477$), where the estimated direct GPR effect is negative, and where the first-order decomposition predicts a negative split difference. No economic account of supply-chain amplification can accommodate a market in which supply-chain pressure does not affect long-run volatility in the first place; the split design nonetheless delivers its most significant amplification result there.

The pattern across the four markets is therefore consistent throughout: where the unified model is well behaved, the split difference is predicted by the omitted direct effect; where it is not, the split design produces its strongest and most stable “finding” of all.

5.6 Interpretation

A negative interaction coefficient invites the objection that it is economically implausible. The objection rests on a misreading of the functional form. The specification is additive in $\log \tau_m$, so the benchmark $\theta_3 = 0$ already implies that the effects of supply-chain pressure and geopolitical risk **multiply** in variance levels. A negative θ_3 means the joint effect is sub-multiplicative. It does not mean geopolitical risk dampens the volatility consequences of supply-chain stress. As noted in Section 5.2, the direct GSCPI coefficient is not separately identified with precision in the unified model, so we make no claim about where the total marginal effect changes sign.

The supported reading is narrower than the original claim but better founded:

- Geopolitical risk affects the long-run component of commodity volatility directly and substantially, in three of four markets.
 - Supply-chain pressure affects it in copper and cocoa, and not detectably in coffee or soybeans.
 - Separate GPR-regime estimates generate apparent positive amplification through the omitted direct effect, in a magnitude the decomposition predicts and in a sign that depends on the recursion implementation.
 - Where the interaction is precisely estimated it is negative, not positive.
-

6. The Null Distribution of the Split-Sample Difference

6.1 Design

The analytical results show a split difference can be nonzero, and of either sign, when the true interaction is zero. The simulation quantifies the problem in a design calibrated to the application and converts it into a statement about inference.

This is a placebo bias map rather than a generic sampling exercise. The true interaction is fixed at zero, $\theta_3^0 = 0$, at every grid point and in every replication. The design retains the observed monthly GSCPI and standardised-GPR paths, so their persistence, their empirical association, and the selected-sample covariance structure of Table 3 are preserved exactly. The long-run component is

$$\log \tau_m = 0.5 + 0.10 G_m + \theta_p^0 P_m,$$

with Beta shape 1.5 and twelve lags. The direct effect θ_p^0 is varied over a grid extending to 0.90, covering the copper estimate (0.32) and reaching the cocoa estimate (0.90) without extrapolation. Daily returns are generated from the full GJR-GARCH recursion with $\mu=0.02$, $\alpha=0.05$, $\gamma=0.02$, $\beta=0.93$, Gaussian innovations, and the exact observed trading-day calendar over the 246 usable months. Each grid point uses **1,000 replications**. Each replication applies the conventional upper-tercile split under the concatenation implementation, estimates the restricted model in each group, and records the difference and a Wald statistic from the two Hessian standard errors.

6.2 Bias-map results

Table 6. Placebo bias map, 1,000 replications per grid point, $\theta_3^0=0$ throughout. “Analytical” is $\theta_p^0 \times 0.2428$ from Corollary 1. $P(\Delta \geq \text{obs})$ uses the observed copper statistic + 0.0687. Monte Carlo standard errors on the mean range from 0.0056 to 0.0077 across the grid.

θ_p^0	Mean Δ	s.d.	5th pct	95th pct	$P(\Delta > 0)$	$P(\Delta \geq \text{obs})$	Analytical
0.00	+0.0013	0.178	-0.277	+0.290	0.501	0.337	0.000
0.10	+0.0115	0.179	-0.280	+0.292	0.533	0.364	0.024
0.20	+0.0355	0.185	-0.249	+0.331	0.572	0.433	0.049
0.32	+0.0418	0.184	-0.264	+0.343	0.594	0.432	0.078
0.50	+0.0564	0.204	-0.284	+0.365	0.607	0.481	0.121
0.70	+0.0653	0.212	-0.306	+0.394	0.630	0.505	0.170
0.90	+0.0696	0.242	-0.337	+0.440	0.635	0.513	0.219

Four features matter.

The placebo is clean at the origin. With no direct effect the mean split difference is +0.0013 with a Monte Carlo standard error of 0.0056, indistinguishable from zero, and exactly half the replications are positive. The design introduces no bias of its own when the conditioning variable does not enter the long-run equation. This is a stronger placebo than we anticipated and it isolates the mechanism cleanly. A preliminary version of this exercise at sixty replications suggested a small negative mean at the origin; at 1,000 replications that is resolved as simulation noise, and we report the point because the earlier figure was inside its own Monte Carlo error.

The bias map is monotone in the omitted direct effect. The mean rises from +0.0013 to +0.0696 across the grid. The apparent amplification statistic is generated by the direct geopolitical-risk effect that the split equations omit, and by nothing else in the design.

Corollary 1 is directionally right and quantitatively optimistic. The analytical prediction matches the simulated mean at small θ_p^0 (0.024 against 0.012 at 0.10; 0.049 against 0.036 at 0.20) and increasingly overstates it thereafter (0.219 against 0.070 at 0.90). This is exactly the failure mode Corollary 1 announces: it is a local approximation. The screen in Section 3.6 should therefore be read as establishing the order of magnitude and sign of the mechanical channel, not its exact size.

The observed statistic is unremarkable under the null. At copper’s estimated direct effect, 43 percent of placebo replications produce a split difference at least as large as the observed $+0.0687$. Even at $\theta_p^0 = 0$, with no omitted direct effect at all, 34 percent do. For cocoa the corresponding exceedance probability at $\theta_p^0 = 0.90$ is 24 percent. The dispersion of the split statistic under a zero-interaction null is simply too large — a standard deviation of roughly 0.18 to 0.24 — for the observed values to carry evidential weight.

6.3 Size of the naive test

The inferentially relevant object is the behaviour of the test a researcher would actually run. The last column of Table 7 reports the empirical rejection rate of a nominal five percent Wald test of $H_0: \theta_g^H = \theta_g^L$ built from the two subsample standard errors, under $\theta_3^0 = 0$.

Table 7. Empirical size of the split-based Wald test, nominal 5 percent.

θ_p^0	0.00	0.10	0.20	0.32	0.50	0.70	0.90
Rejection rate	0.145	0.152	0.157	0.177	0.216	0.205	0.247

The test rejects nearly three times too often even when the conditioning variable has no direct effect and no interaction exists, and nearly five times too often at the cocoa calibration. The distortion at the origin cannot be attributed to omitted-variable bias, since the mean split difference there is zero; it comes from the recursion breaks of Section 3.5 and from the non-standard behaviour of pseudo-true estimators under selection. The two defects therefore contribute separately: one displaces the estimator, the other invalidates its standard error.

6.4 Power, and the disposal of the feasibility defence

The standard defence of sample splitting is that the unified model cannot be estimated, so the split is all that is available. A second arm sets $\theta_p^0 = 0.32$ and $\theta_3^0 \in \{0, 0.05, 0.10, 0.20\}$ with 500 replications each, and compares the likelihood-ratio test of $\theta_3 = 0$ in the unified model with the split-based Wald test on the same simulated data.

Table 8. Rejection rates at nominal 5 percent, $\theta_p^0 = 0.32$, 500 replications.

True θ_3	Unified LR	Split Wald	Mean $\hat{\theta}_3$	s.d. $\hat{\theta}_3$
0.00	0.104	0.165	0.006	0.111
0.05	0.130	0.204	0.047	0.116
0.10	0.188	0.227	0.096	0.109
0.20	0.500	0.239	0.193	0.111

Two conclusions follow, and the second is the more important.

The unified estimator is essentially unbiased at every value of the true interaction, and the likelihood-ratio test has a rejection rate that rises steeply with it, from 10 percent under the null to 50 percent at $\theta_3 = 0.20$. Its size is roughly twice nominal, which should be acknowledged, but its power is real. Under a design calibrated to the estimated parameters rather than to this generic grid, power is far higher still: a parametric bootstrap at the

copper and cocoa point estimates rejects $\theta_3 = 0$ in 86 and 87 percent of replications respectively. The feasibility defence fails on its own terms.

The split-based Wald test has almost no power to detect the interaction. Its rejection rate moves from 0.165 to 0.239 as the true interaction goes from zero to 0.20, an increase of seven percentage points, most of which is size distortion rather than power. Over the same range the unified test moves by 40 points. A test whose rejection rate is nearly flat in the parameter it is claimed to test is not a test of that parameter. This is the sharpest available statement of the paper's thesis, and it is an empirical statement rather than an algebraic one.

7. State-Dependent Leverage: What Survives

The split estimates also indicate that the GJR asymmetry parameter differs across geopolitical-risk states. For copper at the upper tercile the estimated asymmetry is -0.0581 in the high-GPR group against $+0.0299$ in the low-GPR group, a difference of -0.0880 , implying that negative return shocks raise short-run volatility in the familiar way during tranquil periods but not during geopolitically stressed ones.

Consistency requires that this finding be held to the same standard as the one the paper rejects. It passes both tests.

It survives the placebo. Table 6's design records the split difference in $\hat{y}$ alongside the difference in $\hat{\theta}_g$.

Under $\theta_3^0 = 0$ the mean difference in $\hat{y}$ is within 0.002 of zero at every grid point, with a standard deviation of 0.024 to 0.026, and the observed -0.0880 falls below the 0.2nd percentile of the placebo distribution at every value of θ_p^0 . Unlike the supply-chain slope difference, the leverage difference is not manufactured by the design.

It survives unified re-estimation. A defensible model lets the asymmetry depend on geopolitical risk inside a single likelihood,

$$\gamma_m = \gamma_0 \exp(\gamma_1 P_m),$$

where the exponential form preserves $\alpha + \gamma_m \geq 0$. For copper, $\hat{\gamma}_0 = 0.0169$ and $\hat{\gamma}_1 = -0.4167$ (s.e. 0.1777), significant with a likelihood-ratio p of 0.003 against the constant-asymmetry model. The implied asymmetry falls from 0.026 at one standard deviation below mean geopolitical risk to 0.011 one standard deviation above. The interaction coefficient is essentially unchanged at -0.0955 , so the two results are not competing for the same variation.

The effect does not generalise. For cocoa $\hat{\gamma}_1 = +0.2754$ with $p = 0.16$; for coffee it is indistinguishable from zero; for soybeans the optimiser drives γ_1 to its constraint boundary and the fit is not interpretable. We report the copper result as a single-market finding, estimated properly, and not as a general regularity.

The contrast with Section 5 is the point. Two split-sample differences were computed from the same partition of the same data. One is fully accounted for by an omitted direct effect and reverses sign under coherent recursion handling; the other is far outside its placebo distribution and is confirmed by unified estimation. The methodological critique is not that split-sample differences are always artefacts. It is that a split-sample difference carries no information about which of these two situations obtains, and that distinguishing them requires the unified model.

8. Robustness

8.1 *Alternative geopolitical-risk measures*

The headline GPR index combines geopolitical acts and threats, which have different economic content. The unified model is re-estimated with GPRA and GPRT in place of the headline index, with Table 3 recomputed for each subindex. A difference in results is informative but must be read through the joint model: a larger split difference under GPRT would not show that threats amplify supply-chain pressure, since it may reflect a stronger direct threat effect or a different within-group projection.

8.2 *The shared weight restriction*

The baseline constrains the three low-frequency terms to share ω_2 . Estimating three separate weight parameters and comparing by likelihood ratio with two degrees of freedom, the restriction is not rejected for copper ($L R = 3.97$, $p = 0.137$), and the interaction coefficient is materially unchanged at -0.1086 . For the remaining three commodities the unrestricted search returned a likelihood below the restricted optimum — decisively so for cocoa — which is an optimisation failure rather than evidence, and which illustrates the numerical fragility discussed in Section 4.5. We therefore retain the restricted specification and report the unrestricted fits in the replication package with their convergence diagnostics rather than as estimates.

8.3 *Alternative lag windows*

Re-estimation with shorter and longer lag windows tests whether the association depends on the assumed transmission horizon. These checks matter because the weight parameter can absorb part of the effect of the lag choice. Note that the estimated $\hat{\omega}_2$ for copper is 25.8, which concentrates almost all weight on the first monthly lag, while for cocoa it is 1.08, which is nearly uniform across twelve months — the two markets in which the interaction is significant sit at opposite ends of the lag-shape space, so the result is not an artefact of a particular horizon assumption.

8.4 *Alternative thresholds and estimated thresholds*

Median, tercile, and quartile cut-offs are reported in Table 4 as descriptive checks and are not substitutes for interaction estimation. Where a researcher genuinely believes the state is discrete rather than continuous, the appropriate estimator is a threshold model with the threshold estimated jointly (Hansen, 2000) or a smooth-transition variance specification (Amado and Teräsvirta, 2013), not a cut-off imposed by convention.

8.5 *Distributional assumptions and numerical robustness*

Gaussian quasi-maximum likelihood is conventional in this literature and delivers consistent estimates under correct specification of the conditional variance even when the innovation density is misspecified (Bollerslev and Wooldridge, 1992), but it may be inefficient for commodity returns with pronounced excess kurtosis; Student- t innovations are reported as an extension. Heavy tails alter standard errors and convergence behaviour but do not repair the estimand problem in separate GPR-selected regressions. Because the parametric bootstrap of Section 5.2 also assumes Gaussian innovations, it validates the Hessian standard errors against sampling variability but not against distributional misspecification; this is the main remaining qualification on the reported inference.

9. Conclusion

This paper reconsiders a common empirical claim: that a larger coefficient in a high-risk subsample than in a low-risk subsample demonstrates that risk amplifies the underlying relationship.

In the GARCH-MIDAS setting that claim does not follow, for two independent reasons. The conditioning variable enters long-run volatility directly, so each group-specific slope is a pseudo-true quasi-maximum-likelihood parameter whose probability limit contains the omitted direct effect and a rescaled version of the true interaction; and the monthly partition of a daily recursive process destroys the GARCH recursion, so under the usual implementation the split objective is not a valid likelihood at all.

The application shows that neither defect is second-order. The positive copper split difference of +0.0687 becomes -0.1651 when the recursion is re-initialised at each block and -0.1342 when the filter is run on the full daily sample; the reported finding is an artefact of splicing 24 disjoint blocks into a single pseudo-series. Where the split difference is positive, the omitted direct effect predicts it: across copper, coffee, and cocoa the observed difference and its predicted omitted-variable component correlate at 0.997, and the largest and most threshold-stable apparent amplification in the study occurs in soybeans, where supply-chain pressure has no detectable effect on long-run volatility to amplify. In placebo simulations with the true interaction fixed at zero, the split-based Wald test rejects at 14 to 25 percent against a nominal 5 percent, and its rejection rate rises by only seven points as the true interaction moves from zero to 0.20 — while the unified likelihood-ratio test rises by forty, and reaches 86 percent power at the estimated parameters. The unified model, estimated in a common likelihood, rejects no interaction in the direction opposite to the split-sample narrative for both markets in which it is precisely estimated.

The correct test is straightforward in principle: estimate a unified long-run variance equation containing supply-chain pressure, geopolitical risk, and their monthly product under a common lag structure, and test the coefficient on the product term. Separate-group estimates are descriptive diagnostics, not substitutes. Where the unified specification cannot be identified, the appropriate conclusion is that the data do not identify the interaction, not that sample splitting has established it.

Section 3.6 offers a practical screen. Before interpreting any split table, compute the within-group projection coefficients of the omitted conditioning variable on the retained regressor and multiply by an estimate of the direct effect. In this application that calculation, requiring two auxiliary regressions and no volatility modelling, would have accounted for the entire reported amplification before any interaction was estimated.

The critique is not that state dependence is absent or that split-sample differences are always artefacts. Section 7 documents a state-dependent leverage effect in copper that lies outside its own placebo distribution and is confirmed in a unified likelihood with $p = 0.003$. The point is that a split-sample difference carries no information about which case obtains. Distinguishing a real interaction from a manufactured one requires the model that contains both.

The logic generalises. The same algebra applies whenever researchers divide samples by uncertainty, policy regimes, financial stress, climate states, liquidity conditions, or political risk and compare coefficients across the resulting groups, and the recursion argument applies whenever a low-frequency partition is imposed on a high-frequency recursive process.

References

Ai, C., and Norton, E. C. (2003). Interaction Terms in Logit and Probit Models. *Economics Letters*, 80(1), 123–129. [https://doi.org/10.1016/S0165-1765\(03\)00032-6](https://doi.org/10.1016/S0165-1765(03)00032-6)

- Amado, C., and Teräsvirta, T. (2013). Modelling Volatility by Variance Decomposition. *Journal of Econometrics*, 175(2), 142–153. <https://doi.org/10.1016/j.jeconom.2013.03.006>
- Benigno, G., Di Giovanni, J., Groen, J. J. J., and Noble, A. I. (2022). The GSCPI: A New Barometer of Global Supply Chain Pressures. *Federal Reserve Bank of New York Staff Reports*, No. 1017. <https://doi.org/10.2139/ssrn.4114973>
- Bollerslev, T., and Wooldridge, J. M. (1992). Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances. *Econometric Reviews*, 11(2), 143–172. <https://doi.org/10.1080/07474939208800229>
- Caldara, D., and Iacoviello, M. (2022). Measuring Geopolitical Risk. *American Economic Review*, 112(4), 1194–1225. <https://doi.org/10.1257/aer.20191823>
- Conrad, C., and Kleen, O. (2020). Two Are Better Than One: Volatility Forecasting Using Multiplicative Component GARCH-MIDAS Models. *Journal of Applied Econometrics*, 35(1), 19–45. <https://doi.org/10.1002/jae.2742>
- Conrad, C., and Loch, K. (2015). Anticipating Long-Term Stock Market Volatility. *Journal of Applied Econometrics*, 30(7), 1090–1114. <https://doi.org/10.1002/jae.2404>
- Engle, R. F., Ghysels, E., and Sohn, B. (2013). Stock Market Volatility and Macroeconomic Fundamentals. *Review of Economics and Statistics*, 95(3), 776–797. https://doi.org/10.1162/REST_a_00300
- Feigenberg, B., Ost, B., and Qureshi, J. A. (2025). Omitted Variable Bias in Interacted Models: A Cautionary Tale. *Review of Economics and Statistics*, 107(5), 1260–1274. https://doi.org/10.1162/rest_a_01361
- Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models. CIRANO Working Paper 2004s-20.
- Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance*, 48(5), 1779–1801. <https://doi.org/10.1111/j.1540-6261.1993.tb05128.x>
- Hansen, B. E. (1996). Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis. *Econometrica*, 64(2), 413–430. <https://doi.org/10.2307/2171789>
- Hansen, B. E. (2000). Sample Splitting and Threshold Estimation. *Econometrica*, 68(3), 575–603. <https://doi.org/10.1111/1468-0262.00124>
- Leeb, H., and Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, 21(1), 21–59. <https://doi.org/10.1017/S0266466605050036>
- Pan, Z., Wang, Y., Wu, C., and Yin, L. (2017). Oil Price Volatility and Macroeconomic Fundamentals: A Regime Switching GARCH-MIDAS Model. *Journal of Empirical Finance*, 43, 130–142. <https://doi.org/10.1016/j.jempfin.2017.06.005>
- Wang, F., and Ghysels, E. (2015). Econometric Analysis of Volatility Component Models. *Econometric Theory*, 31(2), 362–393. <https://doi.org/10.1017/S0266466614000334>
- White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1), 1–25. <https://doi.org/10.2307/1912526>
-

Data and Code Availability

The analysis uses daily COMEX High Grade Copper, cocoa, coffee, and soybean futures prices; the New York Fed Global Supply Chain Pressure Index, vintage archived in the replication package; and the monthly Caldara-Iacoviello geopolitical-risk index with its GPRA and GPRT subindices. The replication package contains all source series with vintage identifiers, the multi-start estimation program, the three recursion implementations of Section 3.5, the projection-decomposition script, the placebo, size, power, and parametric-bootstrap simulation code, optimiser logs including convergence codes and best-versus-second-best likelihood gaps, and the full simulation output.